UNIVERSITY OF ADELAIDE LIBRARY
26/6/86

PROTEIN STRUCTURE AND FUNCTION

Brian W. Matthews, Ph.D. (Adelaide)

Professor and Chairman, Department of Physics;

Member, Institute of Molecular Biology, University of Oregon

Thesis presented for the degree of

Doctor of Science of the University of Adelaide

August, 1985

Awarded 14-3-86

CONTENTS

Summary	i
Signed Statement	ii
Acknowledgements	iii
PREFACE	1
CONTRIBUTIONS OF COLLABORATORS	2
INTRODUCTION	6
INDEX TO PUBLICATIONS	7
Section A. Thermolysin	8
Section B. Bacteriophage Lysozyme	10
Section C. Bacteriochlorophyll Protein	11
Section D. DNA-Protein Interaction	12
Section E. Protein Evolution	14
Section F. Methods for Protein Structure Determination	15
Section G. Reviews on Protein Structure and Function	16
Section H. Serine Proteases	17
Section I. Miscellaneous	18
PUBLICATIONS	
The Publications for each section follow a section title page	

SUMMARY

This thesis presents a series of publications in which X-ray crystallography and related techniques are used to study the structure and function of biological macromolecules. The work includes the determinations of the three-dimensional structures of thermolysin, bacteriophage T4 lysozyme, a chlorophyll-containing protein, goose lysozyme and "Cro" repressor protein, together with mechanistic, functional and evolutionary studies that followed from the knowledge of these different protein structures. In addition there are two review articles that provide an overview of the field as a whole.

This thesis contains no material that has been submitted for a degree in any University by the author or other persons, except where due reference is made in the text.

B.W. Matthews
Institute of Molecular Biology and Department of Physics
University of Oregon
Eugene, Oregon, USA
August 12, 1985

ACKNOWLEDGEMENTS

I regard myself as singularly fortunate to have been trained by Drs. Harry Medlin, Stan Tomlin, David Blow and David Davies. These are my scientific parents. All were generous to a fault.

Much of the work described herein would not have been possible without the help of a series of talented graduate students, postdoctoral associates and visitors who have passed through my laboratory.

Special thanks go to Helen for her understanding and support.

PREFACE

In support of my candidature for the Degree of Doctor of Science I have chosen to submit a restricted subset of my total publications. Subject to the notes below, the articles submitted are either my own work or are results obtained under my supervision in my laboratory at the University of Oregon.

For completeness I also list my other related publications that are not included in this thesis, although in this case no attempt is made to differentiate between the contributions of the respective coauthors.

With the exception of publications G(1) and G(2) the submitted articles are reports of original research. Publications G(1) and G(2) are invited articles in which the author was asked to review the field of protein crystallography. These two articles contain some original material but are principally a synthesis and review of the work of others.

CONTRIBUTIONS OF COLLABORATORS

(a) General Statement

As mentioned in the Preface, the articles submitted are a restricted subset of my total publications. I have not included papers arising from my Ph.D. thesis. Neither have I included articles in collaboration with my postdoctoral mentors, Drs. D.M. Blow and D.R. Davies.

The submitted articles are my own work or are results obtained under my supervision in my laboratory at the University of Oregon.

In cases of multiple authorship it is often difficult to delineate precisely the contributions of each author. Subject to this caveat, and to the notes listed below, the articles included in this thesis are ones for which I was the senior author and made major contributions to the work described.

(b) Collaborators at the University of Oregon

For most of the submitted publications my coauthors are postdoctoral research associates and graduate students who worked directly under my supervision at the University of Oregon. These collaborators, with publications in parentheses, are as follows:

W.F. Anderson. Postdoctoral Research Associate.
[B(2), B(4), D(1), D(2), D(3), E(2)]

P.M. Colman. Postdoctoral Research Associate.
[A(1), A(2), F(4)]

R.E. Fenna. Postdoctoral Research Associate.
[C(1), C(2)]

R.G. Fisher. Postdoctoral Research Associate.
[D(3)]

M.G. Grütter. Postdoctoral Research Associate.
[B(3), B(4), B(5), E(2), E(3), F(6)]

M.A. Holmes. Graduate Student.
[A(7), A(8)]

W.R. Kester. Graduate Student.
[A(4), A(5), A(6)]

D.H. Ohlendorf. Postdoctoral Research Associate.
[D(1), D(2), D(3), F(6)]

S.J. Remington. Graduate Student; Postdoctoral Research Associate.
[B(1), B(2), B(4), E(1), E(2), F(6)]

R.A. Reynolds. Graduate Student.
[F(6)]

M.F. Schmid. Postdoctoral Research Associate.
[C(2), F(6)]

L.F. TenEyck. Postdoctoral Research Associate.
[B(2)]

L.H. Weaver. Graduate Student; Postdoctoral Research Associate.
[A(3), A(5), B(4), E(3), F(6)]

(c) Collaborators From Other Institutions

In a number of instances the submitted publications include coauthors from other institutions. Their contributions can be summarized as follows:

(1) B.P. Schoenborn and D. Dupourque (Publication A(1)).

Drs. Schoenborn and Dupourque developed a method to obtain large crystals of thermolysin.

(2) Dr. J.N. Jansonius (Publications A(1) and A(2)).

Dr. Jansonius spent a one-year sabbatical leave from the University of Groningen in my laboratory and played an important part in the determination of the structure of thermolysin. His principal contributions were to write several of the computer programs that were used in the structure determination and also help develop conditions for removing zinc from the active site of thermolysin.

(3) K. Titani, K.A. Walsh and H. Neurath (Publication A(2)).

The three-dimensional structure of thermolysin was determined in my laboratory by P.M. Colman, J.N. Jansonius and B.W.M. At the same time the amino acid sequence of the protein was determined at the University of Washington by Dr. Titani, Prof. Walsh and Prof. Neurath. The results were published in Nature as a set of three papers. The first paper was from the Seattle group and described the amino acid sequence. The second paper was from my laboratory (Publication A(1)) and described the three-dimensional structure of the protein. The third paper (Publication A(2)) was written by B.W.M. and was a joint publication in which the findings from the two laboratories were combined to delineate the active site and the calcium binding regions of the molecule.

(4) J.Owen, L.F. Grainger and R.B. Hawkes (Publications B(2) and B(3)).

The structure determination of T4 phage lysozyme and studies of mutant lysozymes described in publications B(1) - B(5) were

principally the work of Mr. S.J. Remington, Dr. M.G. Grütter, Dr. W.F. Anderson and B.W.M. The other coauthors had minor roles.

(5) J.M. Olson (Publication C(2)).

Dr. Olson provided the purified bacteriochlorophyll protein used for the structural studies described in publications C(1) - C(3).

(6) Y. Takeda (Publications D(1) - D(3)).

Dr. Takeda provided the purified Cro protein used for the determination of the structure of this DNA-binding protein.

(7) W.F. Anderson (Publications D(1) - D(3)).

Dr. Anderson began working on structural studies of Cro protein while he was a postdoctoral research associate in my group and continued to collaborate on the project after he assumed a faculty position at the University of Alberta. His first major contribution was to grow crystals of Cro protein (Publication D(4), not included). B.W.M. obtained the first heavy-atom derivative used in the Cro structure determination (Publication D(5), not included). Dr. Anderson found several additional derivatives (Publication D(1). He also began the first comparisons of the amino acid sequence of Cro protein with other DNA-binding proteins (Publication D(6), not included). Although included as a coauthor, Dr. Anderson did not participate directly in any of the work described in publications D(2) and D(3).

INTRODUCTION

The publications are grouped under the following headings:

A. Thermolysin

B. Bacteriophage Lysozyme

C. Bacteriochlorophyll Protein

D. DNA-Protein Interaction

E. Protein Evolution

F. Methods for Protein Structure Determination

G. Reviews on Protein Structure and Function

H. Serine Proteases

I. Miscellaneous

Under each heading a list is given of the publications submitted, followed by other articles on the same topic that are not submitted as part of this thesis.

INDEX TO PUBLICATIONS

SECTION A. Thermolysin

Publications submitted

(1) Matthews, B.W., Jansonius, J.N., Colman, P.M., Schoenborn, B.P. and D. Dupourque. Three-dimensional structure of thermolysin. Nature New Biology 238, 37-41 (1972).

(2) Matthews, B.W., Colman, P.M., Jansonius, J.N., Titani, K., Walsh, K.A. and H. Neurath. Structure of thermolysin. Nature New Biology 238, 41-43 (1972).

(3) Matthews, B.W., Weaver, L.H. and W.R. Kester. The conformation of thermolysin. J. Biol. Chem. 249, 8030-8044 (1974).

(4) Kester, W.R. and B.W. Matthews. Crystallographic study of the binding of dipeptide inhibitors to thermolysin: Implications for the mechanism of catalysis. Biochemistry 16, 2506-2516 (1977).

(5) Weaver, L.H., Kester, W.R. and B.W. Matthews. A crystallographic study of the complex of phosphoramidon with thermolysin: A model for the presumed catalytic transition state and the binding of extended substrates. J. Mol. Biol. 114, 119-132 (1977).

(6) Kester, W.R. and B.W. Matthews. Comparison of the structures of carboxypeptidase A and thermolysin. J. Biol. Chem. 252, 7704-7710 (1977).

(7) Holmes, M.A. and B.W. Matthews. The binding of hydroxamic inhibitors to crystalline thermolysin suggests a penta-coordinate intermediate in catalysis. Biochemistry 20, 6912-6920 (1981).

(8) Holmes, M.A. and B.W. Matthews. The structure of thermolysin refined at 1.6Å resolution. J. Mol. Biol. 160, 623-639 (1982).

Publications not submitted

(9) Colman, P.M., Jansonius, J.N. and B.W. Matthews. The structure of thermolysin: An electron density map at 2.3Å resolution. J. Mol. Biol. 70, 701-724 (1972).

(10) Matthews, B.W. and L.H. Weaver. Binding of lanthanide ions to thermolysin. Biochemistry 13, 1719-1725 (1974).

(11) Weaver, L.H., Kester, W.R., Ten Eyck, L.F. and B.W. Matthews. The structure and stability of thermolysin. Symposium on Enzymes and Proteins from Thermophilic Microorganisms, Zurich, 1975. H. Zuber, ed., 31-39, Experientia Supplement 26.

SECTION A. Thermolysin (continued)

(12) Bolognesi, M.C. and B.W. Matthews. Binding of the bi-product analog L-benzylsuccinic acid to thermolysin determined by X-ray crystallography. J. Biol. Chem. 254, 634-639 (1979).

(13) Monzingo, A.F. and B.W. Matthews. Structure of a mercaptan-thermolysin complex illustrates the mode of inhibition of zinc proteases by substrate-analog mercaptans. Biochemistry 21, 3390-3394 (1982).

(14) Holmes, M.A., Tronrud, D.E. and B.W. Matthews. Structural analysis of the inhibition of thermolysin by an active-site-directed irreversible inhibitor. Biochemistry 22, 236-240 (1983).

(15) Monzingo, A.F. and B.W. Matthews. Binding of N-carboxymethyl dipeptide inhibitors to thermolysin determined by X-ray crystallography: A novel class of transition state analogs for the zinc peptidases. Biochemistry 23, 5724-5729 (1984).

(16) Hangauer, D.G., Monzingo, A.F. and B.W. Matthews. An interactive computer graphics study of thermolysin catalyzed peptide cleavage and inhibition by N-carboxymethyl dipeptides. Biochemistry 23, 5730-5741 (1984).

SECTION B. Bacteriophage Lysozyme

Publications Submitted

(1) Matthews, B.W. and S.J. Remington. The three-dimensional structure of the lysozyme from bacteriophage T4. Proc. Natl. Acad. Sci. USA 71, 4178-4182 (1974).

(2) Remington, S.J., Anderson, W.F., Owen, J., Ten Eyck, L.F., Grainger, C.T. and B.W. Matthews. The structure of the lysozyme from bacteriophage T4: An electron density map at 2.4Å resolution. J. Mol. Biol. 118, 81-91 (1978).

(3) Grütter, M.G., Hawkes, R.B. and B.W. Matthews. Molecular basis of thermostability in the lysozyme from bacteriophage T4. Nature 277, 667-669 (1979).

(4) Anderson, W.F., Grütter, M.G., Remington, S.J., Weaver, L.H. and B.W. Matthews. Crystallographic determination of the mode of binding of oligosaccharides to T4 bacteriophage lysozyme: Implications for the mechanism of catalysis. J. Mol. Biol. 147, 523-543 (1981).

(5) Grütter, M.G. and Matthews, B.W. Amino acid substitutions far from the active site of T4 phage lysozyme reduce catalytic activity and suggest that the C-terminal lobe of the enzyme participates in substrate binding. J. Mol. Biol. 154, 525-535 (1982).

Publications Not Submitted

(6) Matthews, B.W., Dahlquist, F.W. and A.Y. Maynard. Crystallographic data for lysozyme from bacteriophage T4. J. Mol. Biol. 78, 575-576 (1973).

(7) Remington, S.J., Ten Eyck, L.F. and B.W. Matthews. Atomic coordinates for T4 phage lysozyme. Biochem. Biophys. Res. Comm. 75, 265-270 (1977).

(8) Grütter, M.G., Weaver, L.H., Gray, T.M. and B.W. Matthews. Structure, function and evolution of the lysozyme from bacteriophage T4. In "Bacteriophage T4", C.K. Mathews, E.M. Kutter, G. Mosig and P.M. Berget, eds., American Society for Microbiology, Washington, D.C., 356-360 (1983).

SECTION C. Bacteriochlorophyll Protein

Publications Submitted

(1) Fenna, R.E. and B.W. Matthews. Chlorophyll arrangement in a bacteriochlorophyll protein from Chlorobium limicola. Nature 258, 573-577 (1975).

(2) Matthews, B.W., Fenna, R.E., Bolognesi, M.C., Schmid, M.F. and J.M. Olson. Structure of a bacteriochlorophyll a-protein from the green photosynthetic bacterium Prosthecochloris aestuarii. J. Mol. Biol. 131, 259-285 (1979).

(3) Matthews, B.W. Lipid-protein interactions in a bacteriochlorophyll-containing protein. In "Lipid-Protein Interactions", O.H. Griffith and P. Jost, eds., John Wiley, 1-23 (1982).

Publications Not Submitted

(4) Fenna, R.E., Matthews, B.W., Olson, J.M. and E.K. Shaw. Structure of a bacteriochlorophyll-protein from the green photosynthetic bacterium Chlorobium limicola: Crystallographic evidence for a trimer. J. Mol. Biol. 84, 231-240 (1974).

(5) Fenna, R.E., Ten Eyck, L.F. and B.W. Matthews. Atomic coordinates for the chlorophyll core of a bacteriochlorophyll a-protein from green photosynthetic bacteria. Biochem. Biophys. Res. Comm. 75, 751-756 (1977).

(6) Fenna, R.E. and B.W. Matthews. Bacteriochlorophyll proteins from green photosynthetic bacteria. In "The Porphyrins", Vol. III, D. Dolphin, ed., Academic Press, 473-494 (1979).

(7) Fenna, R.E. and B.W. Matthews. Structure of a bacteriochlorophyll protein from Prosthecochloris aestuarii. Brookhaven Symposium in Biology 28, 170-182 (1976).

(8) Matthews, B.W., Fenna R.E. and S.J. Remington. An evaluation of electron micrographs of bacteriochlorophyll a-protein crystals in terms of the structure determined by X-ray crystallography. J. Ultrastruct. Research 58, 316-330 (1977).

(9) Matthews, B.W. and R.E. Fenna. Structure of a green bacteriochlorophyll protein. Acct. Chem. Res. 13, 309-317 (1980).

(10) Schmid, M.F., Tronrud, D.E. and B.W. Matthews. Structural studies of a bacteriochlorophyll-protein. Chemica Scripta 21, 69-71 (1983).

SECTION D. DNA-Protein Interaction

Publications submitted

(1) Anderson, W.F., Ohlendorf, D.H., Takeda, Y. and B.W. Matthews. Structure of Cro repressor from bacteriophage λ and its interaction with DNA. Nature 290, 754-758 (1981).

(2) Matthews, B.W., Ohlendorf, D.H. Anderson, W.F. and Y. Takeda. Structure of the DNA-binding region of lac repressor inferred from its homology with cro repressor. Proc. Natl. Acad. Sci. USA 79, 1428-1432 (1982).

(3) Ohlendorf, D.H., Anderson, W.F., Fisher, R.G., Takeda, Y. and B.W. Matthews. The molecular basis of DNA-protein recognition inferred from the structure of cro repressor. Nature 298, 718-723 (1982).

Publications not submitted

(4) Anderson, W.F., Takeda, Y., Echols, H. and B.W. Matthews. The structure of a repressor. Crystallographic data for the cro regulatory protein from bacteriophage λ. J. Mol. Biol. 130, 507-510 (1979).

(5) Anderson, W.F., Takeda, Y., Ohlendorf, D.H. and B.W. Matthews. Structural studies of the Cro regulatory protein from bacteriophage λ. Proceedings of the Seventh Katzir Conference on Structural Aspects of Recognition and Assembly in Biological Macromolecules, M. Balaban, ed., International Science Services, Philadelphia, 911-917 (1981).

(6) Anderson, W.F., Takeda, Y., Ohlendorf, D.H. and Matthews, B.W. Protein-DNA interaction and recognition: Comparison of the bacteriophage λ Cro repressor with other DNA binding proteins. J. Mol. Biol. 159, 745-751 (1982).

(7) Steitz, T.A., Ohlendorf, D.H., McKay, D.B., Anderson, W.F. and B.W. Matthews. Structural similarity in the DNA binding of catabolite gene activator and Cro repressor proteins. Proc. Natl. Acad. Sci. USA 79, 3097-3100 (1982).

(8) Ohlendorf, D.H., Anderson, W.F. and B.W. Matthews. Many gene-regulatory proteins appear to have a similar α-helical fold that binds DNA and evolved from a common evolutionary precursor. J. Mol. Evol. 19, 109-114 (1983).

SECTION D. DNA-Protein Interaction (continued)

(9) Matthews, B.W., Ohlendorf, D.H., Anderson, W.F., Fisher, R.G. and Y. Takeda. Cro repressor and its interaction with DNA. Cold Spring Harbor Symp. Quant. Biol. 47, 427-433 (1983).

(10) Matthews, B.W., Ohlendorf, D.H., Anderson, W.F., Fisher, R.G. and Y. Takeda. How does Cro repressor recognize its DNA target sites? Trends Biochem. Sci. 8, 25-29 (1983).

(11) Ohlendorf, D.H. and B.W. Matthews. Structural studies of protein-nucleic acid interactions. Ann. Rev. Biophys. Bioeng. 12, 259-284 (1983).

(12) Ohlendorf, D.H., Anderson, W.F., Lewis, M., Pabo, C.O. and Matthews, B.W. Comparison of the structures of cro and λ repressor proteins from bacteriophage λ. J. Mol. Biol. 169, 757-769 (1983).

(13) Takeda, Y., Ohlendorf, D.H., Anderson, W.F. and B.W. Matthews. DNA-binding proteins. Science 221, 1020-1026 (1983).

(14) Ohlendorf, D.H., Anderson, W.F., Takeda, Y. and B.W. Matthews. High resolution structural studies of Cro repressor and implications for DNA recognition. J. Biomolec. Struc. Dynamics 1, 553-563 (1983).

(15) Ohlendorf, D.H., Anderson, W.F., Takeda, Y. and B.W. Matthews. Structure of Cro repressor and its implications for protein-DNA interaction. UCLA Symposium on Molecular and Cellular Biology, New Series, "Mechanisms of DNA Replication and Recombination". Alan R. Liss, Inc., 77-89 (1983).

(16) Matthews, B.W. Structural studies of DNA-protein interactions. Pontificiae Academiae Scientiarum Scripta Varia 55. "Specificity in Biological Interactions". C. Chagas and B. Pullman, eds., 83-98 (1984).

SECTION E. Protein Evolution

Publications submitted

(1) Remington, S.J. and B.W. Matthews. A systematic approach to the comparison of protein structures. J. Mol. Biol. 140, 77-99 (1980).

(2) Matthews, B.W., Grütter, M.G., Anderson, W.F. and S.J. Remington. Common precursor of lysozymes from hen egg-white and bacteriophage T4. Nature 290, 334-335 (1981).

(3) Grütter, M.G., Weaver, L.H. and B.W. Matthews. Goose lysozyme structure: An evolutionary link between hen and bacteriophage lysozymes? Nature 303, 828-831 (1983).

Publications not submitted

(4) Remington, S.J. and B.W. Matthews. A general method to assess the similarity of protein structures with applications to T4 bacteriophage lysozyme. Proc. Natl. Acad. Sci. USA 75, 2180-2184 (1978).

(5) Matthews, B.W., Remington, S.J., Grütter, M.G. and W.F. Anderson. Relation between hen egg-white lysozyme and bacteriophage T4 lysozyme: Evolutionary implications. J. Mol. Biol. 147, 545-558 (1981).

(6) Weaver, L.H., Grütter, M.G., Remington, S.J., Gray, T.M., Isaacs, N.W. and B.W. Matthews. Goose-type, chicken type and phage-type lysozymes suggest that extensive changes occur during evolution in both amino acid sequences and three-dimensional structures of proteins. J. Mol. Evol. 21, 97-111 (1985).

SECTION F. Methods for Protein Structure Determination

Publications submitted

(1) Matthews, B.W. The extension of the isomorphous replacement method to include anomalous scattering measurements. Acta Cryst. 20, 82-86 (1966).

(2) Matthews, B.W. The determination of the position of anomalously scattering heavy atom groups in protein crystals. Acta Cryst. 20, 230-239 (1966).

(3) Matthews, B.W. Solvent content of protein crystals. J. Mol. Biol. 33, 491-497 (1968).

(4) Matthews, B.W., Klopfenstein, C.K. and P.M. Colman. A computer controlled film scanner for X-ray crystallography. J. Phys. E. Sci. Instrum. 5, 353-359 (1972).

(5) Matthews, B.W. Determination of molecular weight from protein crystals. J. Mol. Biol. 82, 513-526 (1974).

(6) Schmid, M.F., Weaver, L.H., Holmes, M.A., Grütter, M.G., Ohlendorf, D.H., Reynolds, R.A., Remington, S.J. and B.W. Matthews. An oscillation data collection system for high resolution protein crystallography. Acta Cryst. A37, 701-710 (1981).

Publications not submitted

(7) Matthews, B.W. Determination and refinement of phases for proteins. In "Crystallographic Computing", F.R. Ahmed, ed., Munksgaard, 146-159 (1970).

(8) Colman, P.M., Weaver, L.H. and B.W. Matthews. Rare earths as isomorphous calcium replacements for protein crystallography. Biochem. Biophys. Res. Comm. 46, 1999-2005 (1972).

(9) Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442-451 (1975).

(10) Matthews, B.W. and E.W. Czerwinski. Local scaling: A method to reduce systematic errors in isomorphous replacement and anomalous scattering measurements. Acta. Cryst. A31, 480-487 (1975).

SECTION G. Reviews on Protein Structure and Function

Articles submitted

(1) Matthews, B.W. X-ray crystallographic studies of proteins. Ann. Rev. Phys. Chem. 27, 493-523 (1976).

(2) Matthews, B.W. X-ray structure of proteins. In "The Proteins", 3rd Ed., Vol. III, H. Neurath and R.L. Hill, eds., Academic Press, New York, 403-590 (1977).

Articles not submitted

(3) Matthews, B.W. and S.A. Bernhard. Structure and symmetry of oligomeric enzymes. Ann. Rev. Biophys. Bioeng. 2, 257-317 (1973).

(4) Matthews, B.W. Protein structures. In "CRC Handbook of Biochemistry and Molecular Biology", 3rd Ed., Vol. II, G.D. Fasman, ed., 760-766 (1976).

(5) Matthews, B.W. Relations between protein structures. Comments Mol. Cell. Biophys. 1, 209-222 (1982).

(6) Matthews, B.W., Grütter, M.G., Weaver, L.H., Anderson, W.F., Ohlendorf, D.H. and R.G. Fisher. X-ray crystallographic studies of the structure, function and evolution of proteins. In "Proteins in Biology and Medicine", R.A. Bradshaw et al., eds., Academic Press, New York, 265-279 (1982).

SECTION H. Serine Proteases

None of these publications are submitted

(1) Matthews, B.W., Sigler, P.B., Henderson, R. and D.M. Blow. Three-dimensional structure of tosyl-α-chymotrpysin. Nature 214, 652-656 (1967).

(2) Sigler, P.B., Blow, D.M., Matthews, B.W. and R. Henderson. Structure of crystalline α-chymotrypsin. II. A preliminary report including a hypothesis for the activation mechanism. J. Mol. Biol. 35, 143-164 (1968).

(3) Matthews, B.W. Some crystal forms of bovine chymotrypsinogen B and chymotrypsinogen A. J. Mol. Biol. 33, 499-501 (1968).

(4) Matthews, B.W., Cohen, G.H., Silverton, E.W., Braxton, H. and D.R. Davies. Relation between γ- and α-chymotrypsin. J. Mol. Biol. 36, 179-183 (1968).

(5) Cohen, G.H., Matthews, B.W. and D.R. Davies. The relation between γ-and α-chymotrypsin II. Direct comparison of the electron densities at 5.5Å resolution. Acta Crystallographica B26, 1062-1069 (1970).

(6) Anderson, W.A., Matthews, B.W. and R.G. Woodbury. Crystallographic data for a group specific protease from rat intestine. Biochemistry 17, 819 (1978).

(7) Reynolds, R.A., Remington, S.J., Weaver, L.H., Fisher, R.G., Anderson, W.F., Ammon, H.C. and B.W. Matthews. Structure of a serine protease from rat mast cells determined from twinned crystals by isomorphous and molecular replacement. Acta Crystallographica B41, 139-147 (1985).

SECTION I. Miscellaneous

None of these publications are submitted

(1) Terry, W.D., Matthews, B.W. and D.R. Davies. Crystallographic studies of a human immunoglobulin. Nature 220, 239-241 (1968).

(2) Colman, P.M. and B.W. Matthews. Symmetry, molecular weight and crystallographic data for sweet potato β-amylase. J. Mol. Biol. 60, 163-168 (1971).

(3) Matthews, B.W. The γ turn. Evidence for a new folded conformation in proteins. Macromolecules 5, 818-819 (1972).

(4) Holmes, M.A., Remington, S.J., Schwendimann, B., Christie, G.E. and B.W. Matthews. Crystallographic data for sturgeon holo D-glyceraldehyde-3-phosphate dehydrogenase. J. Mol. Biol. 112, 651-652 (1977).

(5) Grütter, M.G., Rine, K. and B.W. Matthews. Crystallographic data for lysozyme from egg-white of Embden goose. J. Mol. Biol. 135, 1029-1032 (1979).

(6) Gray, T.M. and B.W. Matthews. Intrahelical hydrogen bonding of serines, threonines and cysteines within α-helices and its relevance to membrane-bound proteins. J. Mol. Biol. 175, 75-81 (1984).

(7) Swanson, R., Weaver, L.H., Remington, S.J., Matthews, B.W. and T.O. Baldwin. Crystals of luciferase from Vibrio harveii. J. Biol. Chem. 260, 1287-1289 (1985).

PUBLICATIONS

SECTION A. THERMOLYSIN

(Reprinted from Nature New Biology, Vol. 238, No. 80, pp. 37–41, July 12, 1972)

Three-dimensional Structure of Thermolysin

B. W. MATTHEWS, J. N. JANSONIUS* & P. M. COLMAN
Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403

B. P. SCHOENBORN & D. DUPOURQUE
Department of Biology, Brookhaven National Laboratories, Upton, New York

The electron density map of thermolysin indicates how the polypeptide chain is displaced throughout the overall shape of the molecule.

We describe here a 2.3 Å resolution electron density map of thermolysin which shows the overall folding of the polypeptide chain and has made possible the identification of many of the 316 amino-acid residues[1,2].

Thermolysin (Calbiochem) was crystallized without further purification. (This is the same source as for the enzyme used in the sequence analysis[1]; Calbiochem is the American distributor for Daiwa K.K.) The protein was dissolved in a mixture of 45% dimethyl sulphoxide/55% water to give a solution of approximately 330 mg/ml. protein 1.4 M calcium acetate, and 0.05 M Tris acetate, *p*H 7.3. After this viscous mixture was clarified by centrifugation, 0.15 ml. aliquots were placed in 1 dram vials which were left in a sealed container along with a dish of water. During several weeks water vapour diffused into the crystallizing vials to lower the solubility of the protein and large hexagonal crystals formed. The space group is $P6_122$ with cell dimensions $a=94.2$ Å, $c=131.4$ Å. From density measurements there are twelve molecules per unit cell; that is, one molecule per asymmetric unit. After the crystals were grown they were slowly equilibrated with solutions of lower ionic concentration. A solution of 0.01 M calcium acetate, 0.01 M Tris acetate, and 5% by volume dimethyl sulphoxide, adjusted to *p*H 7.3, was used as a "standard mother liquor".

* On leave from the Laboratorium voor Structuurchemie, The University, Bloemsingel 10, Groningen, The Netherlands.

Collection of Data

X-ray diffraction data were recorded photographically using an Elliott rotating anode generator and Enraf-Nonius precession cameras. Integrated intensities were measured with a computer-controlled rotating drum film scanner. Altogether about 92,500 intensities were measured for each of the four isomorphs. After symmetry averaging and merging data from different films this number reduced to about 13,700 unique reflexions, not counting the Friedel related measurements. This includes over 93% of the diffraction data to Bragg spacings of 2.4 Å, and about one third of the data between 2.4 Å and 2.3 Å. Excluding the very weak reflexions the mean error in reflexion amplitude obtained in scaling together the respective films for

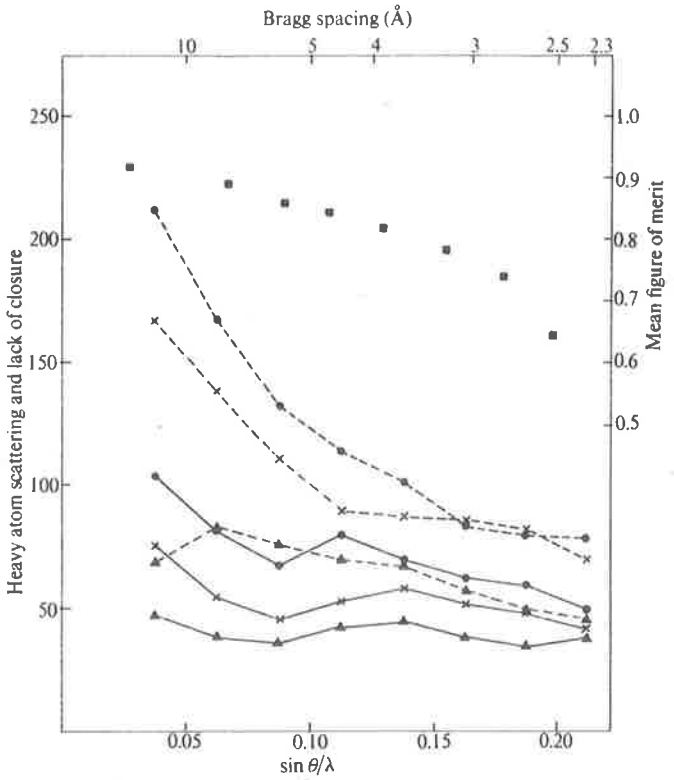


Fig. 1 Angular dependence of refinement statistics for thermolysin isomorphous derivatives; ●, PtI_6; ×, DMA; ▲, Hg–Zn. Root mean square heavy atom scattering denoted by broken lines; lack of closure error by solid lines. ■, Mean figure of merit.

the native and derivative data sets ranged from 2.0% for the native data to 3.7% for the PtI_6 derivative. The space group was shown by two different methods to be $P6_122$ rather than $P6_522$. Additional details of the methods used to determine the Bijvoet differences in this space group and to process and check the data will be given in a subsequent publication.

Heavy Atom Derivatives

Three isomorphous heavy atom derivatives[3] of thermolysin were used. The first was obtained by soaking the crystals for about 2 weeks in a solution of 0.001 M dimercury acetate (DMA) in standard mother liquor, *p*H 7.5. The second isomorph was obtained by soaking the crystals in a saturated solution of K_2PtI_6 in standard mother liquor, *p*H 5.6, for 2–3 weeks. The third derivative "Hg–Zn" was obtained by substituting mercury for the zinc bound at the active site of the enzyme.

After merging all the data for each derivative, a further refinement of the heavy atom parameters was performed by applying Hart's[4] method to the quasi three-dimensional[5] set of centric reflexions which constitute about 18% of the total data. The final parameters are shown in Table 1. Many of the weaker sites of binding of the PtI_6 derivative were found from a 2.5 Å resolution three-dimensional difference map of PtI_6 minus native with phase angles determined from the DMA plus Hg–Zn derivatives.

The variation of some refinement parameters as a function of scattering angle is shown in Fig. 1. As judged by the fact that the r.m.s. heavy atom scattering factor always exceeds the r.m.s. lack of closure error E, the three derivatives all contribute significantly to the overall phase determination out to the resolution limit. We do not regard the current refinement of the heavy atom parameters as complete and therefore prefer to regard the electron density map described here as a preliminary one.

"Best" phases were calculated using the method of Blow and Crick[6], adapted to include the contribution of the anomalous scattering data[7,8]. For the final 13,734 reflexions, the mean figure of merit[6] was 0.79, its variation with the angle being shown in Fig. 1. In the phase angle calculation the overall r.m.s. values of the ratio (lack of closure/E) at the most probable phase was 1.05 (that is, close to the expected value[9] of 1.0), indicating that the error estimates E and E' used for the isomorphous replacement and anomalous scattering measurements were realistic.

Location of Active Site

Latt, Holmquist and Vallee[10] showed that thermolysin contains 1.05 mol of zinc/34,600 g of enzyme and that the zinc is essential for activity We found that by soaking crystals of thermolysin for at least 5 days in a solution of 0.01 M EDTA, 0.01 M 1,10-phenanthroline, 0.005 M calcium acetate, 0.005 M Tris acetate, 0.5% DMSO, adjusted to *p*H 5.5, zinc could be removed without disrupting the crystals. These crystals are less stable than the native crystals, but can be slowly equilibrated with standard mother liquor and photographed. A (0*kl*) difference Fourier projection with amplitudes equal to the difference between the zinc-free and native enzymes and phase angles derived from the DMA derivative is shown in Fig. 2*a*. The interpretation of this map is not obvious, but a detailed study shows that three of the strongest regions of electron density are negative, and are consistent with the removal of an atom with coordinates (0.12, 0.45, 0.06). On readdition of zinc to the crystals, their diffraction pattern becomes indistinguishable from that of the native enzyme, suggesting that zinc has been removed. Crystals of the zinc-free enzyme exposed briefy (15 min) to solutions containing cadmium or mercury show different intensity changes. A (0*kl*) difference Fourier synthesis of the mercury-thermolysin minus native-thermolysin is shown in Fig. 2*b*. This map has three positive peaks at the three positions which were negative in Fig. 2*a*. This result confirms that mercury occupies essentially the same site as zinc in the native enzyme. These experiments not only determined the location of the active site, but also provided a useful heavy atom derivative.

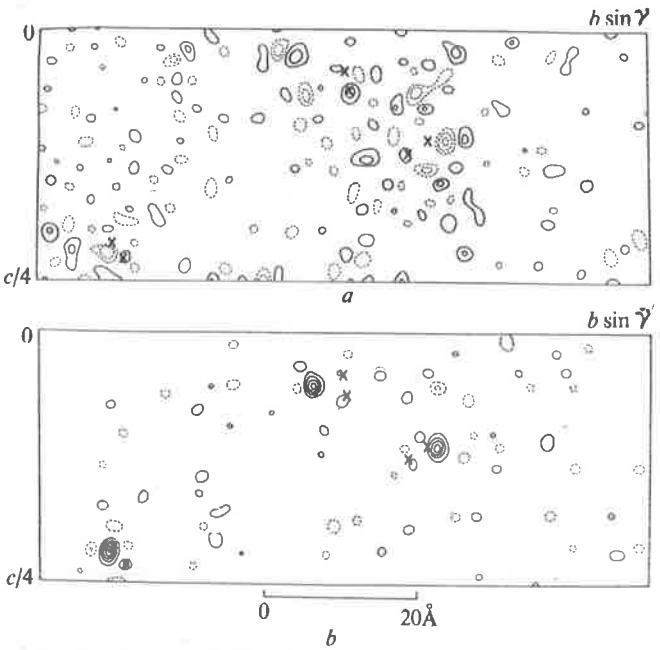


Fig. 2 (*a*) 3 Å resolution (0*kl*) difference Fourier synthesis showing the projected difference in electron density between zinc-free and native thermolysin crystals. Contours are drawn at arbitrary equal intervals, positive contours drawn solid, negative contours broken, zero contour omitted. The sites of the DMA derivative, from which the phase angles were determined, are shown by crosses. (*b*) 2.4 Å resolution (0*kl*) difference Fourier synthesis of mercury–thermolysin minus native thermolysin.

Table 1 Heavy Atom Parameters

Derivative	Z	x	y	z	B	N	$\langle f_H \rangle$	E	R
DMA	32.0	0.123	0.509	0.044	23	2,480	94	52	0.50
	24.6	0.146	0.504	0.060	12				
Hg-Zn	25.0	0.118	0.457	0.053	16	2,554	62	39	0.55
PtI_6	17.4	0.074	0.388	0.481	8	2,490	104	68	0.55
	11.4	0.052	0.401	0.479	15				
	9.6	0.057	0.343	0.503	37				
	12.2	0.072	0.371	0.491	10				
	11.9	0.083	0.399	0.496	11				
	20.8	0.612	0.310	0.600	26				
	10.6	0.075	0.171	0.067	24				
	13.3	0.409	0.397	0.072	14				
	9.3	0.720	0.762	0.024	16				
	11.1	0.379	0.747	0.066	24				

Z, Occupancy of heavy atom site; x, y, z, fractional coordinates; B, isotropic temperature factor $-Å^2$; N, number of centric reflexions included; $\langle f_H \rangle$, root mean square heavy atom scattering; E, root mean square lack of closure error for centrosymmetric reflexions; R, reliability index for centric reflexions. Z, $\langle f_H \rangle$ and E are on an equal arbitrary scale.

Electron Density Map

The electron density distribution was calculated in sections 0.73 Å apart ($c/180$) normal to c, using a programme from G. N. Reeke. The contours were traced onto sheets of transparent 'Dayco' film stretched over aluminium frames, which were then mounted vertically between sets of rails. A view of a portion of the electron density map including the zinc atom is shown in Fig. 3.

The general quality of the map encouraged us to try to follow and identify the sequence of the amino-acids of thermolysin just from the electron density map without reference to the chemically determined amino-acid sequence[1].

From an arbitrary starting point markers were placed in the map at the assumed α-carbon position of each successive residue, and at the same time, on the basis of the appearance of the side chain density, an assignment of the amino-acid residue was made. This preliminary interpretation of the electron density map is summarized in the following article[2] and there were no serious incompatibilities between the tentative X-ray interpretation and the sequence as determined chemically.

A model of thermolysin was constructed on the same scale as the electron density map using the usual Kendrew–Watson components. We used the chemically determined sequence[1], checking that each residue was compatible with the observed electron density. Model building was aided by an optical comparator modified from that of Richards[11], but of modified design. Instead of mounting the mirror at 45°, it is placed directly in front of the electron density map so that both the map sections and the mirror are vertical and parallel to each other. A similar design, proposed independently for small molecule studies, has been described by Rudko and Low[12]. The model is placed in front of the mirror and is illuminated by slab lights above and below, and by having the mirror a foot or so wider than the electron density sections it is possible to work conveniently from either side of the model without the need to look at the mirror through the model. The overall arrangement is straightforward to construct, is very easy to align, and worked very satisfactorily for thermolysin.

Description of the Molecule

The general arrangement of the polypeptide backbone is illustrated in Fig. 4. The molecule is distinctly aspherical, about 64 Å long, 38 Å wide, and up to 37 Å thick. There is a deep groove or cleft across the middle of the molecule with the zinc atom situated in the bottom of this cleft, within about 6 Å of the centre of the gravity of the molecule. Next to the zinc is a hydrophobic pocket suggestive of a substrate specificity site. The cleft is formed between two distinct halves of the molecule, the upper part being made up of the amino terminal half of the molecule, while the carboxyl terminal part lies essentially below the active site cleft. Immediately behind the zinc, running roughly parallel to the cleft, and through the centre of the molecule is a helix (residues 137–150) which contributes two ligands (residues 142 and 146) to the zinc. Apart from its carboxyl terminal residues this helix is totally internal. The only covalent connexion between the upper and lower halves of the molecule is through this central helix. Behind the helix and across the waist of the molecule the upper and lower helices make a number of hydrophobic contacts. Also in the waist region is a double calcium site which is located near the amino terminus of the same central helix[2].

The third zinc ligand (residue 166) occurs in another long internal helix running from residues 160 to 180. This helix lies below the central helix and at about 135° to it. The lower helix runs right through the molecule, passing just below the zinc, with only its terminal residues exposed to the solvent. Other helical regions in addition to 137–150 and 160–180 include residues 67–89, 235–246, 260–268, 281–295, and 301–313. These helices appear to be straight and have the characteristic lack of electron density along the helix axis first noted in myoglobin[13], but are by no means perfectly α-helical[14]. The carbonyl oxygens are almost invariably tilted outward from the helix axis, but a detailed description of the hydrogen bonding must await further analysis. Altogether about 34% of the structure is helical, somewhat higher than the value of 20% estimated

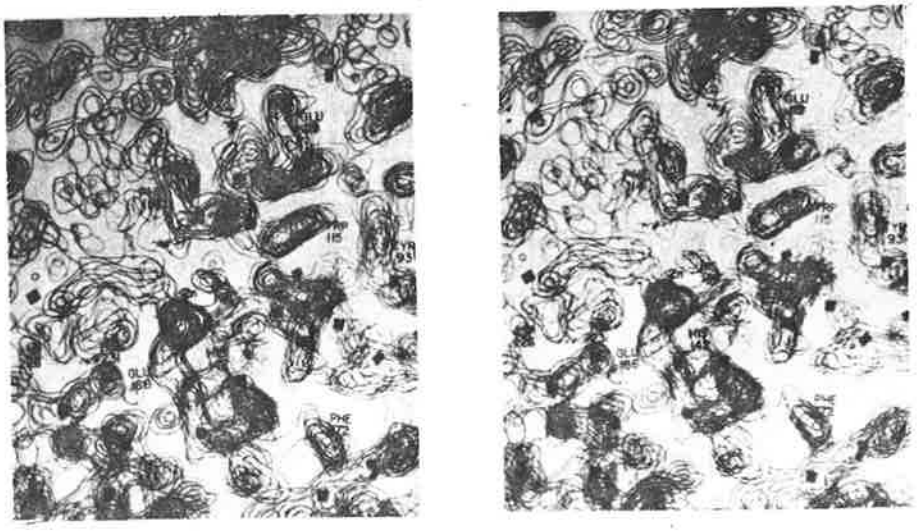


Fig. 3 Stereophotograph of a portion of the thermolysin electron density map, including part of sections $z=-7/180$ through $-15/180$. A number of the amino-acid residues and the zinc atom are labelled. The highest contours at the centre of the zinc atom have been omitted.

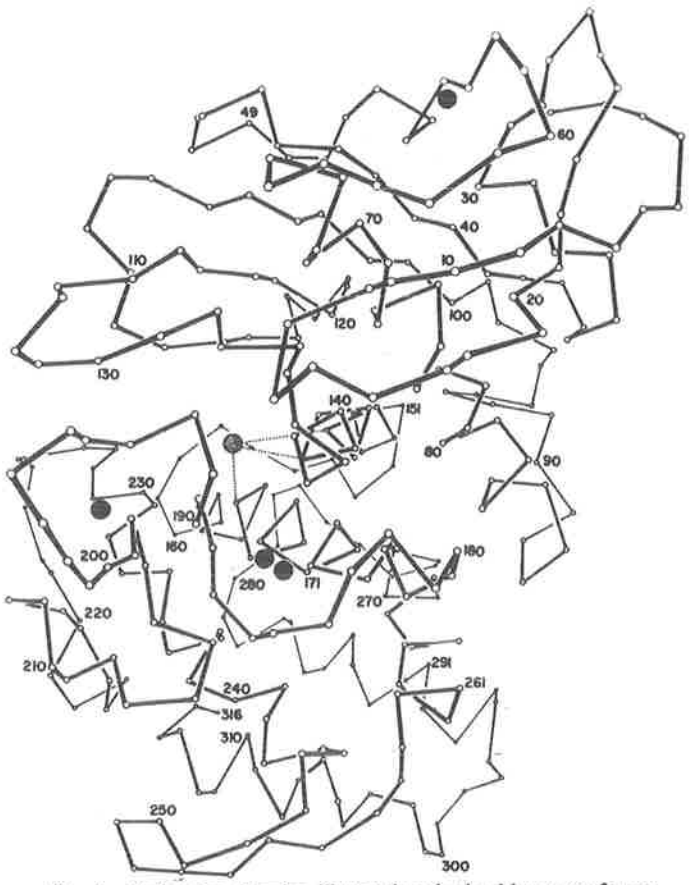


Fig. 4 Projection drawing illustrating the backbone conformation of the thermolysin molecule. The open circles correspond to the α-carbon positions obtained from the electron density map. The zinc atom is drawn stippled with its three protein ligands shown diagrammatically as broken lines. The four calcium atoms discussed in the following article[2] are shown as solid black circles.

by Drucker and Yang[15] from optical rotatory dispersion measurements.

The lower half of the molecule consists of a large hydrophobic core enclosed by the five helices in the C-terminal part of the molecule. Of the totally internal hydrophobic residues in this half of the molecule, most occur within these five helical regions. This hydrophobic region extends up through the waist into the amino terminal part of the molecule where there are several smaller clusters of non-polar residues. Helix 67–89, which is partially internal, contributes to these hydrophobic interactions.

In studying the overall conformation of the polypeptide chain one gains the impression that the molecule could be built up by the polypeptide chain folding back and forth on itself to form locally defined regions, rather than having a final conformation stabilized by interactions between residues far apart in the linear sequence. The most obvious β-structure in the molecule is a rather ill defined region of twisted parallel and antiparallel chains across the top of the molecule. The final polypeptide chains participating in this region of β-structure extend down to form one side of the active site cleft, with the last chain stretched in a fully extended conformation across and above the zinc atom. It remains to be seen whether or not this extended chain participates in the substrate binding. The active site cleft extends right across the molecule, consistent with the prediction of Morihara and Oka[16] that up to six residues may participate in the binding of thermolysin substrates.

Unusual features of the structure which might contribute to its thermostability include the calcium binding sites discussed in the following article[2] and a number of other ionic interactions on the surface of the molecule. In common with several other proteins the trace of polypeptide chain is, for the most part, very tortuous, and where regions of secondary structure do occur they are by no means idealized.

The asymmetric shape of the molecule means that the surface to volume ratio is larger than would result from a compact spherical structure. On the other hand, the molecule can be roughly approximated by two spheres in contact. It has been observed before[17–20] that proteins of molecular weight greater than about 20,000 are often built up not as a single unit but by a combination of two or three large substructures. This finding suggests that a three-dimensional structure based on the principle of a polar exterior surrounding a hydrophobic core can be conveniently achieved with a polypeptide molecular weight of about 10,000–16,000.

We thank Doug Nelson for taking many X-ray photographs and for technical assistance. Others whom we would like to thank include W. Kester, L. Weaver, A. Colman, H. Jansonius and H. Matthews. We are grateful to Drs D. M. Blow and G. N. Reeke for computer programs; Dr M. G. Rossman for a gift of heavy atom compounds, including DMA; and Dr P. B. Sigler for advocating the merits of 'Dayco' film. This work was supported in part by grants from the US National Science Foundation and the National Institutes of Health. J. N. J. thanks the Netherlands Organization for the Advancement of Pure Research (ZWO) for the support of a NATO science fellowship, and B. W. M. is an Alfred P. Sloan research fellow.

Received November 26, 1971; revised March 8, 1972.

1 Titani, K., Hermodson, M. A., Ericsson, L. H., Walsh, K. A., and Neurath, H., *Nature New Biology*, **238**, 35 (1972).
2 Matthews, B. W., Colman, P. M., Jansonius, J. N., Titani, K., Walsh, K. A., and Neurath, H., *Nature New Biology*, **238**, 41 (1972).
3 Green, D. W., Ingram, V. M., and Perutz, M. F., *Proc. Roy. Soc.*, A, **225**, 287 (1954).
4 Hart, R. G., *Acta Cryst.*, **14**, 1188 (1961).
5 Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C., and Sarma, V. R., *Nature*, **206**, 757 (1965).
6 Blow, D. M., and Crick, F. H. C., *Acta Cryst.*, **12**, 794 (1959).
7 North, A. C. T., *Acta Cryst.*, **18**, 212 (1965).
8 Matthews, B. W., *Acta Cryst.*, **20**, 82 (1966).
9 Matthews, B. W., in *Crystallographic Computing* (edit. by Ahmed, F. R.), 146 (1970).
10 Latt, S. A., Holmquist, B., and Vallee, B. L., *Biochem. Biophys. Res. Commun.*, **37**, 333 (1969).
11 Richards, F. M., *J. Mol. Biol.*, **37**, 225 (1968).
12 Rudko, A. D., and Low, B. W., *J. Appl. Cryst.*, **3**, 545 (1970).
13 Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., and Shore, V. C., *Nature*, **185**, 422 (1960).
14 Pauling, L., Corey, R. B., and Branson, H. R., *Proc. US Nat. Acad. Sci.*, **37**, 205 (1951).
15 Drucker, H., and Yang, J. T., *Fed. Proc.*, **28**, Biochem. Abstr., 3306 (1969).
16 Morihara, K., and Oka, T., *Biochem. Biophys. Res. Commun.*, **30**, 625 (1968).
17 Blow, D. M., *Biochem. J.*, **112**, 261 (1969).
18 Drenth, J., Jansonius, J. N., Koekoek, R., Swen, H. M., and Wolthers, B. G., *Nature*, **218**, 929 (1968).
19 Wright, C. S., Alden, R. A., and Kraut, J., *Nature*, **221**, 235 (1969).
20 Shotton, D. M., and Watson, H. C., *Nature*, **225**, 811 (1970).

Printed in Great Britain by Flarepath Printers Ltd., St. Albans, Herts.

(Reprinted from Nature New Biology, Vol. 238, No. 80, pp. 41–43, July 12, 1972)

Structure of Thermolysin

B. W. MATTHEWS, P. M. COLMAN & J. N. JANSONIUS*
Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403

K. TITANI, K. A. WALSH & H. NEURATH
Department of Biochemistry, University of Washington, Seattle, Washington 98195

Results of the chemical and X-ray analyses are combined to locate the active site and the calcium binding sites.

We bring together here the results obtained from the chemical sequence and three-dimensional electron density map of the thermolysin molecule. The terms chemical sequence and X-ray sequence shall denote the amino-acid sequence determined respectively by the chemical analysis[1] or by inspection of the electron density map[2].

Comparison of Sequences

A preliminary comparison showed that the chemical and X-ray sequences were similar and that there were no serious incompatibilities between them. Aligning the X-ray and chemical sequences to give maximum agreement and assuming that the X-ray method cannot distinguish Glu from Gln or Asp from Asn, 53% of the assignments made by X-ray analysis were found to be in agreement with the chemically determined sequence. This may be compared with agreements of 48% and 60% obtained, respectively, for the X-ray identifications of the chemically unknown portions of the sequences of myoglobin[3,4] and carboxypeptidase A[5].

Model building in the optical comparator permitted the comparison of the chemical identification, residue by residue, with the electron density map. This suggests that the chemical sequence[1] is compatible with the electron density map, and most of the differences between the X-ray and chemical identifications could be attributed to the difficulty in distinguishing between amino-acids with rather similar electron density (such as Val, Leu, Ile, and Thr). Other apparently more serious discrepancies were often caused by a lack of electron density both on the surface and for residues which were heavy atom ligands in one of the isomorphous derivatives used to determine the protein phase angles. The residues which were most often correctly identified were those with the smallest side chains (Gly, Ala) and those with the bulkiest (Phe, Tyr, Trp). The tentative X-ray sequence included an additional residue between positions 54–55 and 299–300, and omitted Asp 191, Pro 195, Gly 223, and Gly 264. Model building subsequently showed that in each of these cases the chemical sequence could be accommodated by the electron density. We attribute a number of errors in the X-ray identification to the cursory way in which the X-ray identifications were made within a period of 3 days, and without any attempt to build a model consistent with the electron density map.

The sequence used in building the model differed from the chemical sequence in that residues 248 (Gly) and 249 (Thr) were transposed. The electron density map was incompatible with Thr-Gly and suggested Gly-Thr as the correct sequence. This result was communicated to the authors of ref. 1 who had meanwhile reached the same conclusion by chemical analysis. At positions 301 and 302 the chemical composition indicates (Glu, Gln) but the ordering is uncertain. In the electron density map residue 302 appears to make a salt link on the surface of the molecule to Lys 262. We conclude therefore that the most probable sequence is Gln (301), Glu (302).

In building the atomic model of thermolysin, the knowledge of the chemical sequence was particularly helpful in the active site region where the residues possibly involved in the catalytic activity could be identified with certainty and in the region of the molecule from residues 184 to 200 which had been difficult to interpret from the electron density map alone. These two regions are described below.

The Active Site

The location of the zinc atom was determined by difference Fourier syntheses[2] and was confirmed by prominence of the metal in the electron density distribution of the native protein. The three zinc ligands were shown to be residues 142, 146, and 166 and were tentatively identified as Met, His, and Asp, the first identification being based on the high side chain density of residue 142 more than on its shape. The chemical sequence[1] has His, His, Glu at positions 142, 146, and 166 which model building shows to be compatible with the electron density map. The combination His, His, Glu is, of course, the same as that found for the zinc ligands in bovine carboxypeptidase A[5,6].

The finding that thermolysin and carboxypeptidase, in spite of their different tertiary structures, have common features at the active site is reminiscent of the relation between subtilisin and the mammalian serine proteases[7]. These two types of serine enzymes have been shown to have different three-dimensional structures, yet to be virtually identical in their catalytically important residues[8]. We will therefore describe the active sites of thermolysin in relation to that of carboxypeptidase A[5].

Looking from the front of the cleft in thermolysin, the three zinc ligands are below and behind the zinc, and are disposed approximately tetrahedrally. Above and at the back of the zinc and partially buried is a glutamic acid (residue 143) which might be a counterpart of Glu 270 in carboxypeptidase. Below and to the right of the zinc is a salt link between Asp 170 and Arg 203. The electron density map of thermolysin suggests that Asp 170 is inaccessible to solvent and besides participating in the salt link with Arg 203 is within hydrogen bonding distance of the zinc ligand His 142. It is not obvious what correlation if any might be expected between Arg 203 in thermolysin, an endopeptidase, and Arg 145 in carboxypeptidase which has been observed to interact with the terminal carboxylate group of peptide substrates and presumably contributes to the absolute specificity of carboxypeptidase for cleavage of peptide bonds adjacent to a terminal free carboxyl group[5]. The other residue thought to be catalytically important in carboxypeptidase is Tyr 248 which in the native enzyme is directed away from the active site, but on binding substrate

* On leave from the Laboratorium voor Structuurchemie, The University, Bloemsingel 10, Groningen, The Netherlands.

is thought to undergo a large conformational change so that its phenolic hydroxyl moves to within hydrogen bonding distance of the nitrogen of the scissile peptide bond[5]. There is also a tyrosine (residue 157) in the active site of thermolysin, but its position does not seem to correspond to that of Tyr 248 in carboxypeptidase. In the thermolysin electron density map the side chain of Tyr 157 is not clearly defined, but it appears to lie along the active site cleft, to the left of the zinc, with its hydroxyl directed approximately toward the zinc, and possibly hydrogen bonded to the zinc ligand Glu 166.

In spite of these possible similarities the active site of thermolysin is strikingly different from that of carboxypeptidase A in that the imidazole ring of a third histidine residue, His 231, is in front of the zinc atom of thermolysin and about 5 Å away. Furthermore, this histidine participates in a salt link with Asp 226, an interaction so far characteristic of the serine proteases including chymotrypsin and its homologues[7] and subtilisin[8,9]. Therefore from this comparison it appears that the relation between thermolysin and carboxypeptidase A is not comparable with the relation between subtilisin and the mammalian serine proteases. Thermolysin and carboxypeptidase may have in common some elements of their catalytic mechanism, but (in contrast to subtilisin and chymotrypsin) not all.

Calcium Binding Sites

Calcium is necessary to stabilize the structure of thermolysin[10]. There is no convincing evidence for the direct participation of calcium in the function of the enzyme, although there has been speculation that this might be the case. Latt *et al.*[11] found different preparations of thermolysin to contain, on the average, 3.3 atoms of calcium per 34,600 molecular weight of thermolysin, and Feder has recently shown that calcium is required for the thermostability of the enzyme[12].

In the electron density map of thermolysin there are three features, and possibly a fourth, which may be attributed to the presence of bound calcium ions. The three presumed calcium ions appear in the electron density map as strong roughly spherical regions of density, not accounted for by any part of the amino-acid sequence. Two of the calcium ions are situated close together, with a centre to centre distance of about 3.8 Å. This pair of ions is surrounded by a cluster of ordered water molecules, backbone carbonyl groups, and acid groups including Asp 138, Glu 177, Asp 185, Glu 190, and Asp 191, all of which participate in a network of salt links and other interactions. The resultant complex is partly internal, and it appears that one of the calcium ions and residues Asp 138 and Glu 177 are not accessible to solvent. Lys 182 makes an additional salt link to Asp 191 so that the overall network includes five positive and five negative charges. Inspection of the electron density map suggests that the coordination of both calcium ions may be octahedral. It was this region of the molecule that was most difficult to interpret without knowledge of the chemical sequence.

The third presumed calcium site is situated in an exposed region on the surface of the molecule, interacting with Asp 57 and possibly with Asp 59. The region of the electron density map which we have most difficulty in interpreting in detail includes residues 196–200 which are at the end of an exposed loop on the surface of the molecule.

Initial inspection of the electron density map suggested that another calcium ion might be bound in this part of the molecule, but the interpretation was uncertain. Subsequent crystallographic experiments[14] in which the calcium ions were replaced with strontium, barium, and lanthanide ions have confirmed the presence of three calcium ions and have shown that a fourth ion does in fact bind at Asp 200.

The location of the four calcium ions is consistent with the recent results of Feder[12] revealing four calcium ions bound to the thermolysin molecule, and that these ions are essential for thermostability but not for enzymatic activity. Also the arrangement of the calcium ions confirms the proposal of Drucker *et al.*[13] and of Feder[12] that the calcium sites occur in two groups. On the other hand, the observation that the closest calcium is about 13 Å from the active site zinc does not support the suggestion of Drucker *et al.*[13] that calcium is involved directly in the activity of thermolysin.

We summarize our findings as follows. Combination of sequence and X-ray diffraction analyses has led to a three-dimensional model of thermolysin at 2.3 Å resolution which shows the following features.

The active site is formed by amino-acid residues of the central section of the polypeptide chain. It displays a deep cleft across the middle of the molecule with the zinc atom situated at its bottom. As in the case of carboxypeptidase A, the three zinc ligands are two histidines and glutamic acid (in this instance residues 142, 146, and 166) but otherwise these two metallo-proteases are not homologous. Residues of the active site that may contribute to the catalytic mechanisms include the following: Glu 143, which is partially buried; Asp 170 and Arg 203, which probably form an ion pair; Tyr 157; and in front of the zinc atom, a third histidine residue, His 231, which forms a salt link to Asp 226.

There are two adjacent calcium ions seemingly firmly bound inside the surface of the molecule by chelation to five acidic groups: Asp 138, Glu 177, Asp 185, Glu 190, and Asp 191. Two additional calcium binding sites are at exposed surface regions, one chelated by Asp 57 and possibly also by Asp 59, and the other chelated by Asp 200.

The structure was determined from the crystalline enzyme in the absence of substrate or inhibitor and hence does not necessarily describe the functionally active conformation of the enzyme. For this reason and in the absence of a specific hypothesis of the catalytic mechanisms, it seems impossible at this time to identify with certainty the functional residues of the active site.

This work was supported in part by grants from the National Institutes of Health, the American Cancer Society, and the National Science Foundation. J. N. J. thanks the Netherlands Organization for the Advancement of Pure Research for the support of a NATO science fellowship. B. W. M. is an Alfred P. Sloan research fellow.

Received November 26, 1971; revised March 8, 1972.

1 Titani, K., Hermodson, M. A., Ericsson, L. H., Walsh, K. A., and Neurath, H., *Nature New Biology*, **238**, 35 (1972).
2 Matthews, B. W., Jansonius, J. N., Colman, P. M., Schoenborn, B. P., and Dupourque, D., *Nature New Biology*, **238**, 37 (1972).
3 Kendrew, J. C., Watson, H. C., Strandberg, B. E., Dickerson, R. E., Phillips, D. C., and Shore, V. C., *Nature*, **190**, 666 (1961).
4 Edmundson, A. B., *Nature*, **205**, 883 (1965).
5 Lipscomb, W. N., Hartsuck, J. A., Quiocho, F. A., and Reeke, G. N., *Proc. US Nat. Acad. Sci.*, **64**, 28 (1969).
6 Lipscomb, W. N., Reeke, G. N., Hartsuck, J. A., Quiocho, F. A., and Bethge, P. H., *Phil. Trans. Roy. Soc.*, B, **257**, 177 (1970).
7 Blow, D. M., and Steitz, T. A., *Ann. Rev. Biochem.*, **39**, 63 (1970).
8 Kraut, J., Robertus, J. D., Birktoft, J. J., Alden, R. A., Wilcox, P. E., and Powers, J. C., *Cold Spring Harbor Symp. Quant. Biol.*, **36** (in the press).
9 Wright, C. S., Alden, R. A., and Kraut, J., *Nature*, **221**, 235 (1969).
10 Endo, S., *J. Fermentation Tech.*, **40**, 346 (1962).
11 Latt, S. A., Holmquist, B., and Vallee, B. L., *Biochem. Biophys. Res. Commun.*, **37**, 333 (1969).
12 Feder, J., Garrett, L. R., and Wildi, B. S., *Biochemistry* (in the press).
13 Drucker, H., Borchers, S., and Yang, J. T., *Fed. Proc.*, **29**, *Biochem. Abstr.* 2361 (1970).
14 Colman, P. M., Weaver, L. H., and Matthews, B. W., *Biochem. Biophys. Res. Commun.*, **46**, 1999 (1972).

Printed in Great Britain by Flarepath Printers Ltd., St. Albans, Herts.

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 249, No. 24, Issue of December 25, pp. 8030–8044, 1974
Printed in U.S.A.

The Conformation of Thermolysin*

(Received for publication, May 30, 1974)

BRIAN W. MATTHEWS, LARRY H. WEAVER, AND WILLIAM R. KESTER

From the Institute of Molecular Biology, and the Departments of Physics and Chemistry, University of Oregon, Eugene, Oregon 97403

SUMMARY

An improved electron density map has been obtained for the thermostable protease thermolysin. The map, calculated at a nominal resolution of 2.3 A, was based on the three heavy atom isomorphs described previously and on data for the zinc-free and europium-substituted enzyme. The zinc-free and native enzymes are closely isomorphous, and the difference in electron density between them shows that the only structural changes which occur on the removal of the zinc ion are slight adjustments in residues in the immediate vicinity of the metal ion. The approximate conformation of most of the thermolysin molecule is revealed without ambiguity, and preliminary atomic coordinates are presented. Some aspects of the molecular stereochemistry are discussed, including tertiary folding, helices, turns, and calcium binding sites. The peptide dihedral angles for most of the residues lie within or near the "allowed" regions for an empirical hard sphere model, but the conformations of a few residues lie in "disallowed" regions. There are several indications that the angle τ at the α carbon atom must be significantly greater than the tetrahedral for a number of residues. With the possible exception of the four calcium binding sites, there is nothing particularly unusual in the thermolysin conformation to which one might attribute the thermostability of the molecule. It is suggested that the enhanced stability of thermostable proteins relative to thermolabile ones cannot be attributed to a common determinant such as metal ion or hydrophobic stabilization, but in a given instance may be due to rather subtle differences in hydrophobic character, metal binding, hydrogen bonding, ionic interactions, or a combination of all of these.

The amino acid sequence (1) and three-dimensional structure (2–4) of the thermostable proteolytic enzyme thermolysin have recently been determined. The protein, mol wt 34,600, crystallizes in space group $P6_122$, cell dimensions $a = b = 94.2$ A, $c = 131.4$ A, with 12 molecules per unit cell, 1 per asymmetric unit.

* This work was supported in part by grants from the National Science Foundation (GB30823X) and the National Institutes of Health (GM 20066, GM 15423) and by the award to B.W.M. of an Alfred P. Sloan Research Fellowship and a Public Health Service Career Development Award (GM 70585) from the Institute of General Medical Sciences.

An improved electron density map has been obtained by the further refinement of the three heavy atom derivatives used previously, and by the inclusion of the data for two additional isomorphs. Atomic coordinates, obtained from the new electron density map, are currently being refined, but this process will take some time. In the meantime, in the belief that the present 2.3 A resolution electron density map shows the approximate conformation of most of the thermolysin molecule without ambiguity, we present in this communication the atomic coordinates available at this time, and discuss some aspects of the molecular stereochemistry.

HEAVY ATOM DERIVATIVES

The electron density map of thermolysin described previously (2, 4) was based on the following three heavy atom modifications: reaction with dimercury acetic acid; reaction with potassium platinum iodide; and replacement of the active site zinc with mercury. The resultant isomorphous derivatives have now been analyzed in detail for the sites of minor substitution by calculation of three-dimensional difference Fourier maps in the usual way, and further refined using Hart's (5) procedure for the approximately 2500 centrosymmetrical reflections in the zones (0*kl*), (*hhl*), and (*hk*0).

A fourth derivative was obtained by the substitution of europium for calcium at three of the four calcium binding sites (6, 7) and in addition, the protein phase angles were improved further by the incorporation of data for the zinc-free enzyme. The heavy atom parameters for each isomorphous pair are given in Table I, and some of the refinement statistics are included in Fig. 1 and Table II.

Dimercury Acetic Acid—It reacts exclusively with N^{ϵ} of His 231 in the active site. Difference Fourier syntheses revealed a peak of positive density, clearly above background, approximately 3.4 A from each mercury, and apparently corresponding to the carboxyl group. This feature, labeled COO^- in Table I, is 4.8 A from the active site zinc ion, suggesting that although the carboxyl group is close to the zinc, neither of the oxygens acts as a ligand.

Potassium Platinum Iodide—The major site of reaction of this derivative is at His 250, although there is also extensive partial iodination of tyrosine residues, as indicated in Table I.

High resolution three-dimensional difference Fourier syntheses and heavy atom refinement suggest that after a reaction with His 250, only 3 iodine atoms, I1-I3 in Table I, remain bound to the platinum, at respective calculated distances of 2.7, 3.0, and 2.2 A. The mean of these three values, 2.6 A,

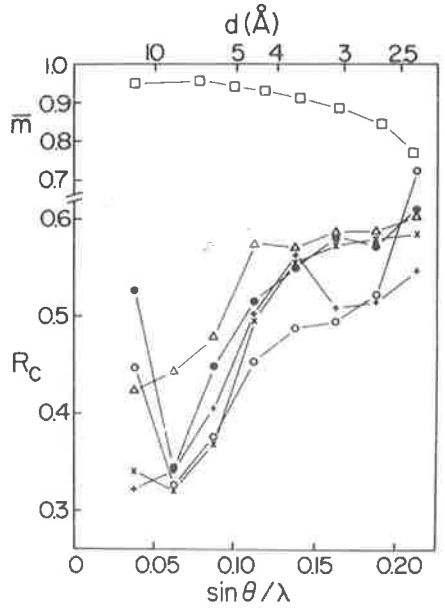


FIG. 1. Graphs showing the variation with angle of the residual R_C for centrosymmetrical reflections, defined in Table II, for the five isomorphous pairs used in calculating the electron density map: ×, Eu^{3+}; +, dimercury acetic acid; △, PtI_6; ●, Hg *versus* native; ○, Hg *versus* Zn-free. Also shown is the mean figure of merit $\bar{m}$ (□).

agrees well with the average Pt-I bond length of 2.68 A observed for the PtI_6 octahedra in the crystals of PtI_4 (8), however the variation in the individual distances indicates that the refined coordinates for the atoms involved should be regarded with caution. The calculated distance from the platinum to N^{δ} of His 250 is 2.2 A. It also seems likely that Lys 239, that has its ε-amino group 2.3 A from the platinum if it does not move on reaction, participates in the platinum coordination. Whether or not water molecules act as additional ligands remains uncertain, as does the geometry of the platinum coordination.

Mercury and Zinc-free Thermolysin—By soaking crystals of thermolysin in solutions of EDTA and 1,10-phenanthroline (2), the active site zinc ion can be removed. On exposing the crystals of the apoenzyme to solutions containing mercuric ion (0.001 M $HgCl_2$, 0.03 M LiCl), essentially complete mercury substitution occurs at the active site within 15 min. Three-dimensional difference Fourier syntheses also revealed that during the brief exposure to mercuric ion, some reaction occurred at His 231 (Site Hg 2, Table I). It also became clear that the major site of mercury substitution (Hg 1) did not coincide exactly with that of the displaced zinc ion, and it was found that in comparing mercury thermolysin and native thermolysin, superior refinement statistics were obtained by considering the mercury atom as a positive scatterer, and the displaced zinc ion as a negative one ((−Zn), Table I). The distance from the zinc to the mercury site is 0.7 A, which may be compared with 1.2 A for carboxypeptidase A (9) and 0.6 A for carbonic anhydrase (10).

A full set of three-dimensional data to a nominal resolution of

TABLE I

Refined heavy atom parameters

Z is the occupancy in electrons; x, y, z, the fractional atomic coordinates, and B the thermal factor in A^2.

Isomorphous pair	Heavy atom	Z	x	y	z	B	Location
DMA[a] *versus* native	Hg	48.0	0.1233	0.5097	0.0437	23.5	
	Hg	37.1	0.1465	0.5042	0.0598	11.4	His 231
	COO^-	15.2	0.1097	0.4956	0.0676	33.9	
Hg *versus* native	Hg 1	52.8	0.1175	0.4547	0.0531	21.0	Zn
	-Zn	−24.2	0.1175	0.4475	0.0549	45.6	Zn
	Hg 2	12.2	0.1340	0.4856	0.0598	74.5	His 231
Hg *versus* Zn-free	Hg 1	49.7	0.1170	0.4542	0.0531	12.4	Zn
	Hg 2	9.6	0.1426	0.4906	0.0651	41.9	His 231
K_2PtI_6 *versus* native	Pt	25.5	0.3137	0.3872	0.0176	13.7	His 250, Lys 239
	I 1	17.9	0.3467	0.4018	0.0209	12.0	
	I 2	21.3	0.2935	0.3623	0.0019	21.7	
	I 3	17.9	0.3149	0.3986	0.0031	12.5	
	I 4	29.7	0.6126	0.3096	0.5996	26.1	Tyr 24
	I 5	15.6	0.0758	0.1710	0.0665	23.9	Tyr 296
	I 6	2.3	0.0990	0.0461	0.0581	20.8	Tyr 296
	I 7	20.3	0.4089	0.3971	0.0718	13.2	Tyr 46
	I 8	11.7	0.7192	0.7623	0.0243	16.5	Tyr 179
	I 9	5.0	0.7651	0.7996	0.0218	21.8	Tyr 179
	I 10	18.2	0.3794	0.7476	0.0656	23.9	Tyr 28
	I 11	7.2	0.6719	0.6889	0.0767	21.6	Tyr 75
	I 12	6.0	0.7629	0.4729	0.0417	19.2	Tyr 211
Eu^{3+} *versus* native	Eu 1	30.3	0.7533	0.6156	0.0385	18.0	Ca 1
	-Ca 2	−22.7	0.7580	0.6259	0.0666	40.7	Ca 2
	Eu 3	36.3	0.4325	0.8765	0.0420	29.9	Ca 3
	Eu 4	33.8	0.6999	0.4903	0.0821	23.2	Ca 4
	W 1	4.5	0.7803	0.6246	0.0796	10.1	Ca 2
	-W 2	−4.1	0.7570	0.6091	0.0798	5.8	Ca 2
	-W 3	−2.9	0.7285	0.6225	0.0674	12.7	Ca 2

[a] DMA represents dimercury acetic acid.

TABLE II

Thermolysin refinement statistics for centrosymmetrical reflections

$|\overline{F_{PH} - F_P}|$ is the mean isomorphous difference, $\langle F_H \rangle$, the rms heavy atom scattering, E, the lack of closure error, all in electrons, and $R_C = \Sigma \, | \, |F_{PH} \pm F_P| - F_H \, | / \Sigma \, |F_{PH} \pm F_P|$.

Isomorphous pair	Reflections	$\overline{\lvert F_{PH} - F_P \rvert}$	$\langle F_H \rangle$	E	R_C
					%
DMA[a] *versus* native	2480	118	146	78	48.2
Hg *versus* native	2554	84	95	60	53.0
Hg *versus* Zn-free	2563	112	135	92	48.3
K_2PtI_6 *versus* native	2490	137	156	109	54.8
Eu^{3+} *versus* native	2539	117	141	82	49.8

[a] DMA represents dimercury acetic acid.

TABLE III

Data processing statistics for zinc-free and europium thermolysin

The respective R values are as defined previously (4, 11). R_{sym} gives the agreement between all symmetry related intensities recorded on the same film and R_{sca}, the agreement between symmetry-averaged intensities on the stronger and weaker film in a film pack. R_{merge} gives the agreement between structure amplitudes measured on different films, omitting approximately the weakest 30% of the data.

Derivative	Native	Zn-free	Eu^{3+}
Intensity Statistics			
Films	24	22	22
Av. R_{sym}	0.054	0.062	0.064
Av. R_{sca}	0.053	0.039	0.061
R_{merge}	0.020	0.025	0.023
Cell dimensions			
a, b (A)	94.2	94.4	94.2
c (A)	131.4	131.4	131.3

2.3 A was collected photographically (4, 11) for the zinc-free enzyme. As for the other isomorphs (2, 4), about 92,000 integrated intensities were reduced to obtain a final set of about 14,000 unique reflections. The data processing statistics are given in Table III. The cell dimensions of the zinc-free crystals (Table III), averaged over the 22 crystals used to collect the high resolution data, are much closer to those of the native crystals than as appeared from the preliminary study (2).

A 2.3-A resolution difference electron density map phased with the three isomorphs used previously (2, 4) revealed that the removal of the zinc ion caused very little perturbation of the protein structure. A portion of the section of this map passing through the zinc site is shown in Fig. 2. The obvious negative feature, of height 50 σ, where σ is the rms background difference density, is due to the loss of the zinc. Adjacent features, less than ⅕ the height, but still extending to ± 10 σ, are certainly significant and indicate small localized adjustments which must occur in the positions of the three zinc ligands and in the side chains of Glu 143 and His 231.

The similarity of the native and zinc-free structures demonstrates that the conformation of the active site region can be maintained in the absence of zinc, and suggests that the zinc ion is probably not an essential structure-forming element in the folding process.

The isomorphism of the zinc-free and native enzyme crystals encouraged us to include the zinc-free data in the calculation of the protein phase angles. For several reasons we chose as an

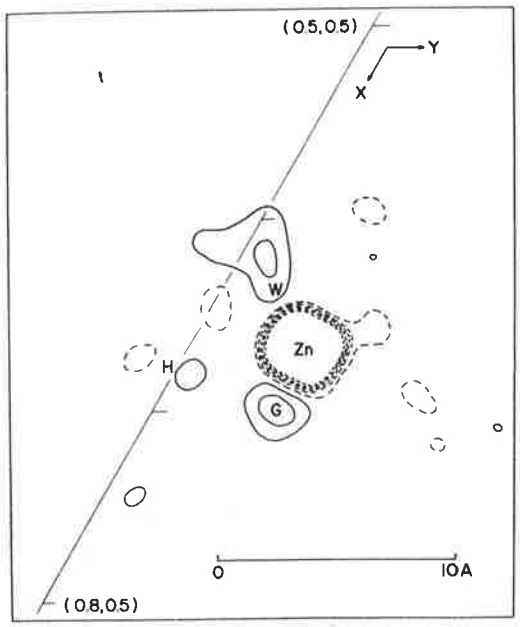


FIG. 2. Difference electron density between zinc-free and native thermolysin. Contour level 3 σ. The illustration includes a portion of the section $z = -10/180$ and the central 11 contours at the zinc site have been omitted. The letters G, H, and W are respectively at the approximate positions of O^{ϵ} of Glu 166, the center of the imidazole of His 231, and the zinc-liganded water molecule.

isomorphous pair zinc-free *versus* mercury thermolysin, rather than zinc-free *versus* native thermolysin. In the first place, the refinement statistics for this pair (Table II) were comparable with the best derivative. Although the resultant phase angles tend to be those of the zinc-free rather than the native enzyme, this essentially results only in a slight diminution of the density at the zinc position, which is immaterial. Also, the use of the zinc-free data as "native" in one contribution to the over-all phase-determining process (12) tends to counteract possible phase errors introduced by the imperfections in the structure amplitudes of the native protein. As is well known, the Blow-Crick treatment of errors in the isomorphous replacement method (12) assumes that the amplitude of the native protein is error-free. This assumption is satisfactory when the number of isomorphous pairs is small, but becomes more questionable as the number of derivatives is increased.

Europium Substitution for Calcium—Europium ions readily replace calcium at three of the four calcium binding sites (6, 7), and provide an excellent isomorphous derivative. Intensity statistics for the europium data set are given in Table III. An estimate of the reproducibility with which the anomalous scattering differences were measured was obtained by evaluating the correlation between these differences measured on one film with their values estimated on other films. The correlation coefficient, defined previously (4), has a value of unity for perfect agreement, and zero for random disagreement. Of the 153 pairwise correlation coefficients which could be evaluated for the 18 films containing noncentrosymmetrical reflections, 141 coefficients exceeded 0.5, 10 were between 0.3 and 0.5, 2 between zero and 0.3, and none of the coefficients were negative. The preponderance of strongly positive correlation coefficients indicates that for the europium derivative the anomalous scattering differences were well above the noise level, and were measured with good reproducibility. Correlation coefficients for the other derivatives (4) were satisfactory, although not as consistently good as those for the europium data.

Refinement statistics for the europium derivative are given in Tables I and II. The minor sites, labeled W1-W3, are presumed water molecules displaced or bound on the displacement of Ca 2 (7).

DETERMINATION OF ATOMIC COORDINATES

Phase angles for 13, 834 reflections to a nominal resolution of 2.3 A were calculated in the usual way using the five isomorphous pairs described above and the anomalous scattering information for the four derivatives containing heavy atoms (12, 13). The mean figure of merit, $\bar{m}$, was 0.88 and the ratio (lack of closure at the most probable phase/E (or E')) averaged over-all phase contributors and all reflections was 1.16. A value close to unity confirms that the lack of closure estimates E and E' are essentially correct (14). At a given Bragg angle, $\bar{m}$ was essentially independent of structure amplitude, except for the very weak reflections, where it decreased markedly.

The new electron density map is similar to that described previously (2, 4), but improved in detail. In the new map, the course of the polypeptide backbone is quite clear, as is the approximate orientation of almost all of the peptide groups. Also, knowing the amino acid sequence (1), the conformation of the majority of the side groups is unambiguous. A representative portion of the electron density map is shown in Fig. 3.

A model of the enzyme was constructed of Kendrew-Watson components with the aid of a modified Richards optical comparator (4, 15). Using the model as a guide, markers were then placed in the electron density map at atomic sites which were

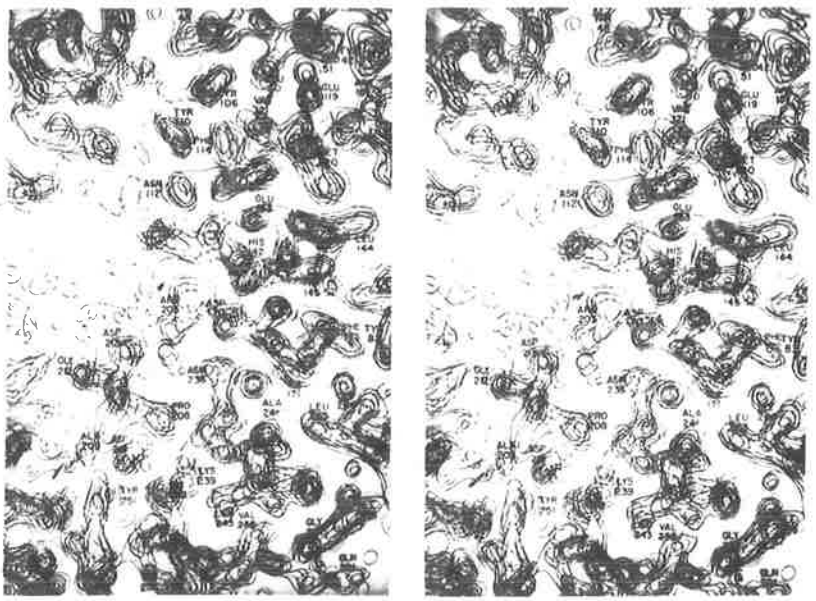


Fig. 3(a)

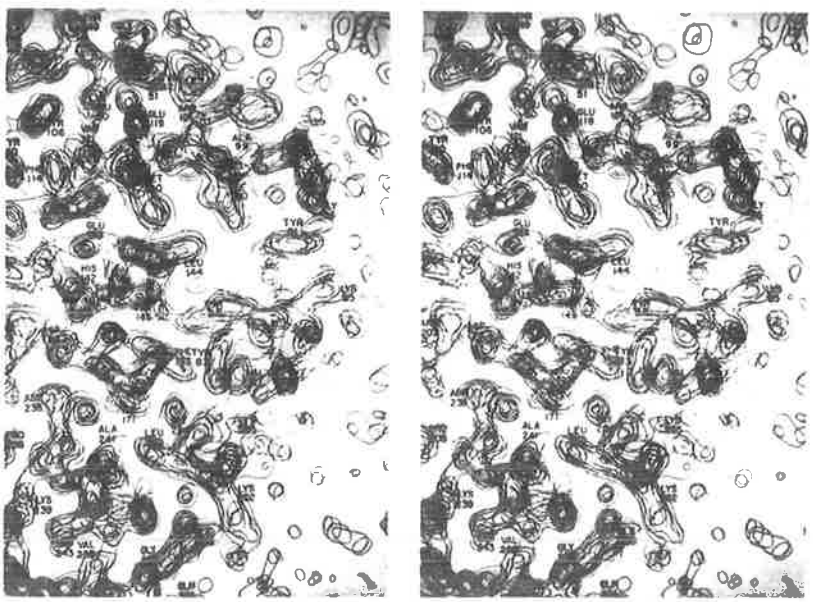


Fig. 3(b)

FIG. 3. Thermolysin 2.3 A resolution electron density. The views include portions of two sets of contiguous sections through the center of the molecule. *a*, *left half*, sections $z = -3/180$ to $-9/180$; *b*, *right half*, sections $z = -3/180$ to $-9/180$; *c*, *left half*, sections $z = -9/180$ to $-15/180$; *d*, *right half*, sections $z = -9/180$ to $-15/180$. In *a* and *c*, the active site cleft extends to the left and above the zinc atom which is labeled in *c*. The labels identifying each of the residues are placed near the side chain rather than the backbone density. A number of characteristic side chains can be identified. Two helical regions including residues 142-148 and 172-167 extend through the sections and are almost normal to them. The contrast between the molecule and the surrounding solvent can be seen most clearly on the right hand side of *b* and *d*.

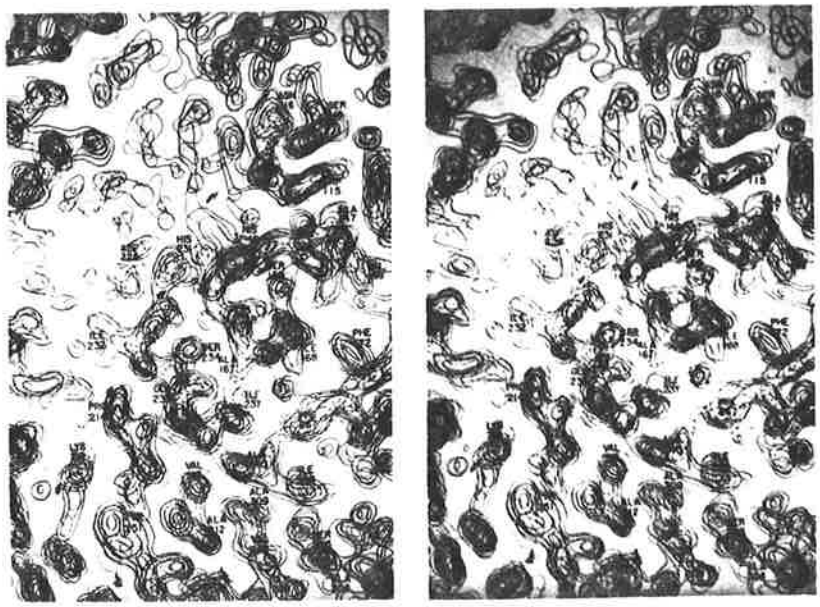


Fig. 3(c)

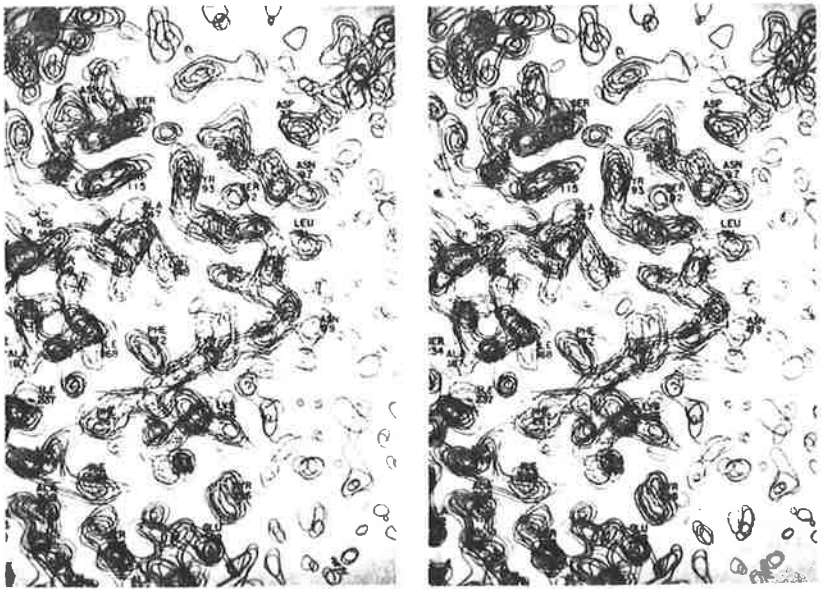


Fig. 3(d)

See legend on preceding page

stereochemically reasonable, as indicated by the model, but which also agreed as closely as possible with the observed electron density. A somewhat similar procedure was used for carboxypeptidase A (16). Although it would have been easier to measure the coordinates directly from the model, the procedure adopted circumvents the problem, which becomes acute for larger proteins, of keeping the model strictly in register with the electron density map in the optical comparator. If one is interested only in preserving the local stereochemistry, slight bodily shifts of the model are not important, but if the model coordinates are intended as a starting point for crystallographic refinement, then such shifts should be avoided at all cost. The procedure of placing atomic markers in the electron density map, using the model only as a guide, also has the advantage that the coordinates need not be constrained by the mechanical limitations of the model components. On the other hand, the finite spacing between the map sections does lead to some inaccuracy in the coordinate normal to the sections. For thermolysin we attempted to minimize this error by estimating z values to one-quarter of a section (0.2 A).

The preliminary coordinates were checked in several ways, *e.g.* by calculating the $C_i^\alpha - C_{i+1}^\alpha$, $C_i^\alpha - C_i^\beta$, and $C_i^\alpha - O_{i-1}$ distances. The rms difference between these distances and their respective means was about 0.16 A, which gives an estimate of accuracy with which the atomic markers were placed, and their positions measured. Of course, this error estimate does not include any possible misinterpretations of the electron density map.

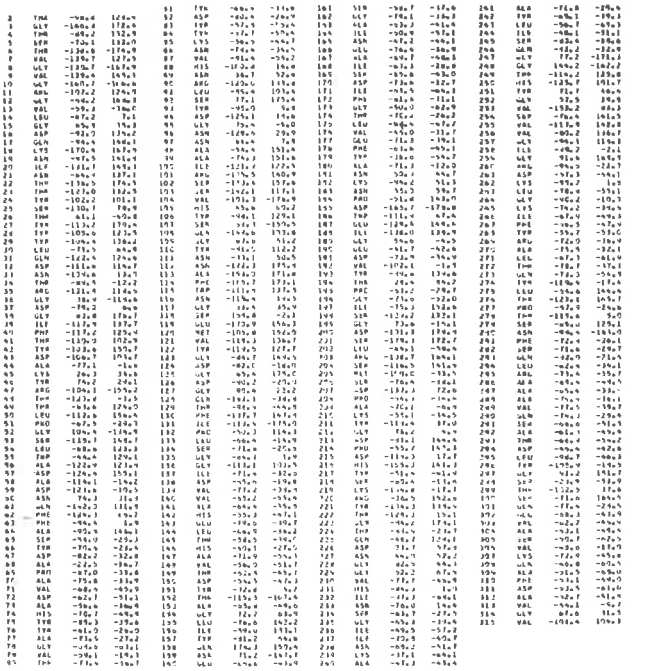
The coordinates were checked further by using Diamond's (17) procedure to construct an "idealized" model, which matched the measured coordinates as closely as possible. Such a model of thermolysin, in which the usual dihedral angles were allowed to vary, but the angle $\tau(N\text{-}C^\alpha\text{-}C)$[1] at the α carbon atom was

[1] Conventions adopted in this paper adhere to the recommendations of the IUPAC-IUB Commission on Biochemical Nomenclature (18).

TABLE IV

Conformational angles for thermolysin

Respective columns give the residue number, residue type, and values of ϕ and ψ which were determined as described in the text.



fixed at 109.65°, differed from the measured atomic coordinates by 0.38 A, on the average. This discrepancy is due in part to errors in the measured coordinates, but also to the restriction on the model that τ be kept constant. Although the Diamond model building procedure provides an extremely useful check on the atomic coordinates, it takes no account of short van der Waals contacts, or other energetic constraints between nonbonded atoms. The consequences of this limitation in the model building procedure may be illustrated by consideration of the peptide dihedral angles (ϕ, ψ) determined before and after the model building process. The angles (ϕ, ψ) are normally defined in terms of planes defined by atoms within consecutive peptide units (18, 19), but can alternatively be derived from planes defined by (C^{α}_{i-1}, $C_i{}^{\alpha}$, $C_i{}^{\beta}$), ($C_i{}^{\beta}$, $C_i{}^{\alpha}$, C^{α}_{i+1}), assuming the geometry at C^{α} to be tetrahedral. For the raw coordinates, the values of ϕ and ψ were determined both ways (except for glycine). In this way, an estimate was obtained of the uncertainty in these angles arising from the errors in measuring the atomic coordinates. The rms difference between the two determinations was 10° for ϕ and 11° for ψ, the values given in Table IV being the average.

A plot of these angles, superimposed on the allowed regions for a hard sphere model (19, 20) is illustrated in Fig. 4*a*. For comparison, the dihedral angles of the standardized model structure are plotted in Fig. 4*b*, in which it can be seen that the conformations of several residues have moved away from the "allowed" into strongly "disallowed" regions. It has been noted previously (16, 21)[2] that in a real protein the angle $\tau(C^{\alpha})$ is, on the average, greater than tetrahedral, and that the model building process can be made more realistic by allowing this angle to vary. Although this has not yet been tested for thermolysin, it is clear that a model in which $\tau(C^{\alpha})$ is fixed at 109° is not completely satisfactory.

In our opinion the "best" thermolysin coordinates available at the present time are those obtained directly from the electron density map and listed in Table V, although if, *e.g.* stereoscopic representations of the structure are desired, then the idealized model coordinates are aesthetically more pleasing. Both sets of coordinates are available from the Protein Data Bank (22).

The coordinates are in Angstroms relative to a right-handed coordinate system, with the origin at the crystallographic origin and X parallel to a^*, Y parallel to b, and Z parallel to c. Fractional crystallographic coordinates (x, y, z) may be derived from the Cartesian ones by the following transformation: $x = 0.01225\ X$, $y = 0.01061\ Y + 0.00612\ X$, and $z = 0.00761\ Z$.

CONFORMATION OF MOLECULE

A detailed description of the conformation of thermolysin must await the refinement of the structure which is in progress at the present time. The scope of this section is therefore deliberately limited, and is intended only to point out some of the more obvious features of the stereochemistry of the molecule.

Tertiary Folding—The over-all conformation of the molecule, which is folded into two lobes with the active site and a double calcium binding site in the equatorial region, has been described previously (2–4). One method of displaying the regions of the polypeptide chain which are in close proximity in the three dimensions is to plot the distances between all of the pairs of α carbon atoms, as shown in Fig. 5 (after Phillips (23)). Helical regions are indicated by clusters of close contacts along the diagonal, parallel β structure by off-diagonal contacts parallel to the diagonal, and anti-parallel β structure by regions of contact at right angles to the diagonal.

As is clear from the diagram, there are very few close approaches between the amino- and carboxyl-halves of the molecule. There is a tendency for regions of the polypeptide chain about 60 residues apart to come together, *e.g.* residues near 10 and 70 form a region of parallel β structure, as do residues near 40 and 100. In the vicinity of residues 80 and 140, two helices cross "behind" the active site region. Residues near 140 and 190 form part of the double calcium binding site and the presumed substrate specificity pocket, and residues including 165 and 230 form part of the active site region. There are a few isolated regions where portions of the polypeptide chain far removed in the linear sequence come close together in three dimensions, but the only such region which is at all extended encompasses residues near 160 and near 280. The latter portion of the polypeptide chain forms an open loop on the surface of the molecule at one end of the active site cleft.

Helices—Residues which form part of a helix are listed in Table VI. The helices are not completely regular, as has been observed for a number of proteins (24). The longer helices tend to be predominantly α helical, in four cases tightening to the 3_{10} or α_{II} conformation at the COOH terminus, and in two cases at the NH_2 terminus of the helix. The dihedral angles for the α, 3_{10} and α_{II} helical conformations are somewhat similar, namely, (−57, −47), (−49, −26), and (−93, −18), respectively (18, 19, 25) and it must be emphasized that the uncertainty in the present thermolysin dihedral angles is such that it is not always possible to distinguish with certainty between these different conformations. Furthermore, when the α_{II} or 3_{10} conformations do occur at the end of an α helix, they are essentially transitional, and never include more than 2 consecutive residues. There is no instance of a sequence of 3 or more

[2] J. Deisenhofer and W. Steigmann, personal communication (submitted for publication).

residues which could be described as forming a turn of α_{II} helix, although in three cases, 3 or 4 consecutive residues do form short 3_{10} helices.

Residues Asp 226-Asn 227-Gly 228-Gly 229 form a single turn of a left-handed α helix, the first observed in a globular protein.

β Structure—Residues of thermolysin which participate in extended β structure are listed in Table VII. The criterion for their inclusion is that the residue be in approximately the fully extended conformation and participate in at least one β-type hydrogen bond. A schematic illustration of the extended β structure has been given previously (4).

Turns—There has been recent interest in "turns" or "hairpin bends" in the polypeptide chain of globular proteins. Such features occur relatively frequently, and can be considered as a category of secondary structure. Venkatachalam (26) studied the possible conformations of 3 linked peptide units in which a hydrogen bond could be formed between the 1st and 3rd unit. He differentiated six general classes of bend, Types I, II, III and Types I′, II′, III′, although in practice, Types I and III are almost indistinguishable, as are Types I′ and III′. Also Type I is somewhat similar to Type II′, as is Type II to Type I′ (24).

The various examples of turns present in thermolysin are

Table V

Preliminary atomic coordinates for thermolysin

The atom names correspond to the standard nomenclature (18) with transliteration of the Greek superscripts, and using EE for eta (η). The distinction between the amino group and the oxygen atom of amide side chains is not indicated by the diffraction results and these atoms are labeled NOD1, NOD2 for asparagine, and NOE1, DOE2 for glutamine. The atom labeled 1 is believed to be oxygen on stereochemical grounds. The coordinates of the bound metal ions, which are not included in the table, are as follows: Ca 1, 61.2, 22.3, 5.0; Ca 2, 61.8, 23.3, 8.5; Ca 3, 35.8, 32.5, 16.3; Ca 4, 57.1, 12.8, 10.5; Zn, 54.7, 20.4, −6.9.

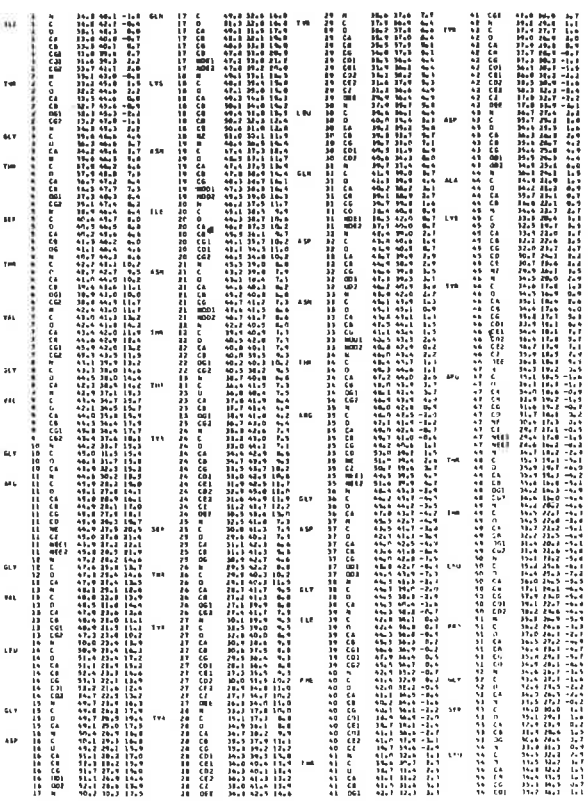
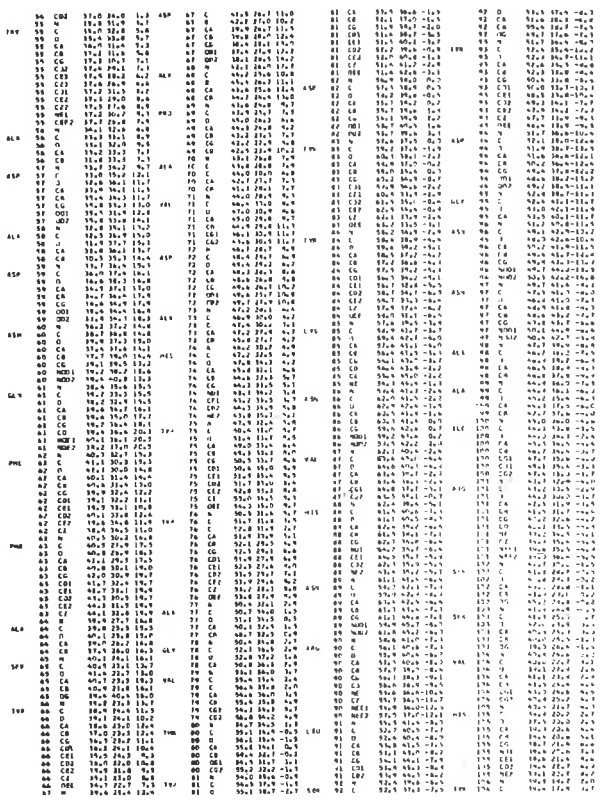


Table V *continued*



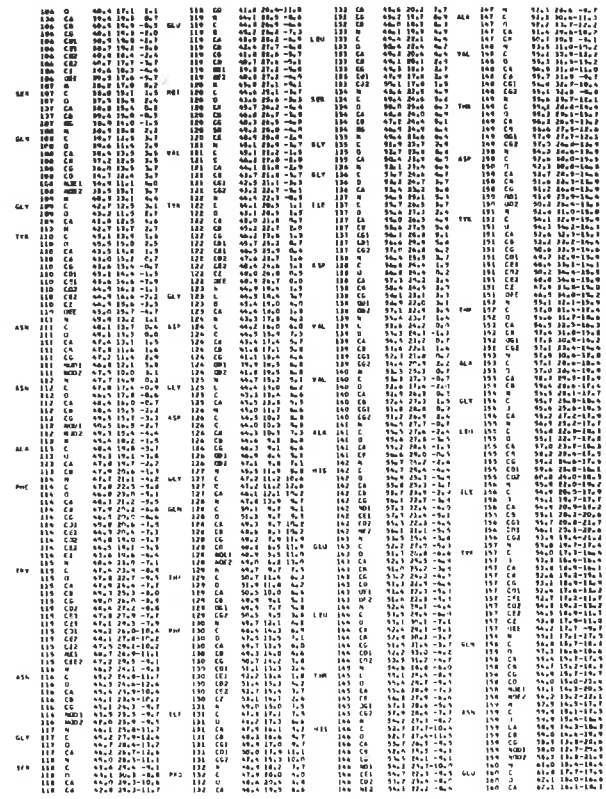


TABLE V *continued*

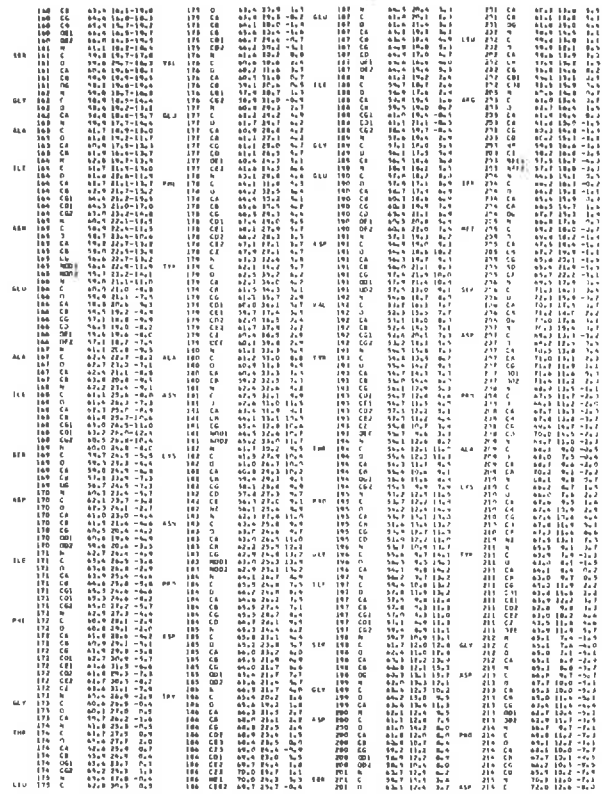
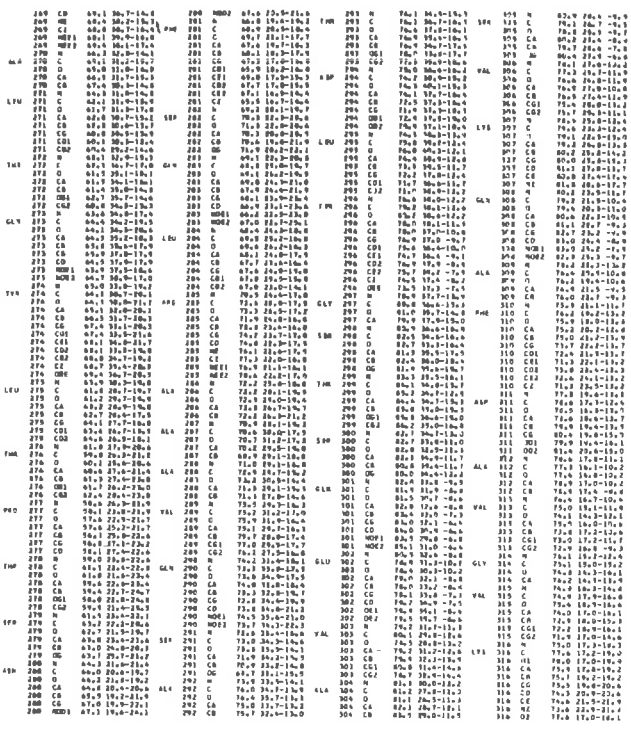
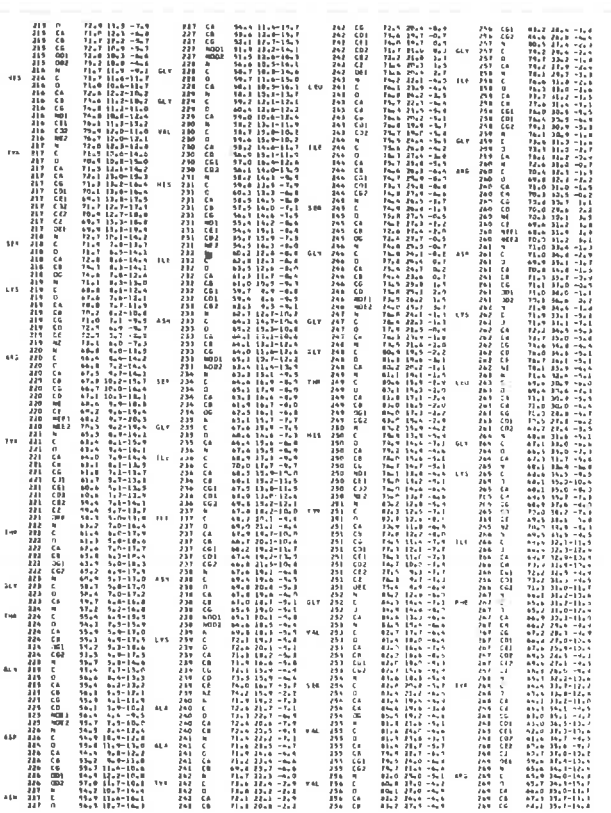


TABLE V *continued*





listed in Table VIII. Details of the so-called γ turn (27) are given elsewhere (28). It should be noted that a number of the examples in the table occur either at the ends of α helices, or in short 3_{10} helices, and these residues also appear in Table VI. The 3_{10} conformation is, of course, identical with that found in a Type III bend, and very similar to that in a Type I bend. The fact that residues with identical conformations can occur either in "bends" or "helices" is a complication which is sometimes ignored in the attempts to differentiate between the different classes of secondary structure on the basis of amino acid sequence.

Using a "hard sphere" model of allowed peptide conformation (26), there is no restriction (except proline) on the residue occupying position 2 or 3 in the Types I and III bends, but for the other four categories, a glycine is expected at one or both of these loci. The Types I and III (*i.e.* approximate 3_{10}) bends occur much more frequently in thermolysin than in any other class. The only exceptions are a Type II′ bend at residues 35 to 38, where glycine occupies the 2nd position, as predicted, and a Type III′ bend at residues 44 to 47. In the latter case, a glycine is expected at both the 2nd and 3rd positions, but the actual sequence is Lys 45-Tyr 46. The electron density map in this region of the molecule is well defined, and we believe that our current interpretation is unlikely to be in error. Presumably the angle τ at the α carbon atom of 1 or both of these residues must be greater than tetrahedral in order to relieve the otherwise too close van der Waals contacts. The conformations of both Lys 45 and Tyr 46 correspond to a left-handed 3_{10} helix, which is "allowed" for $\tau = 115°$ (20).

Further evidence for the existence of residues with $\tau > 109°$ comes from a consideration of the Type I and Type III turns. In an extension of Venkatachalam's study of the conformation of 3 linked peptide units (26), Chandrasekaran *et al.* (29) calculated the low energy conformations for this system, assuming "standard" dimensions for the polypeptide backbone and empirical potential functions (20, 30). The results of these calculations are in poor agreement with the conformations of the

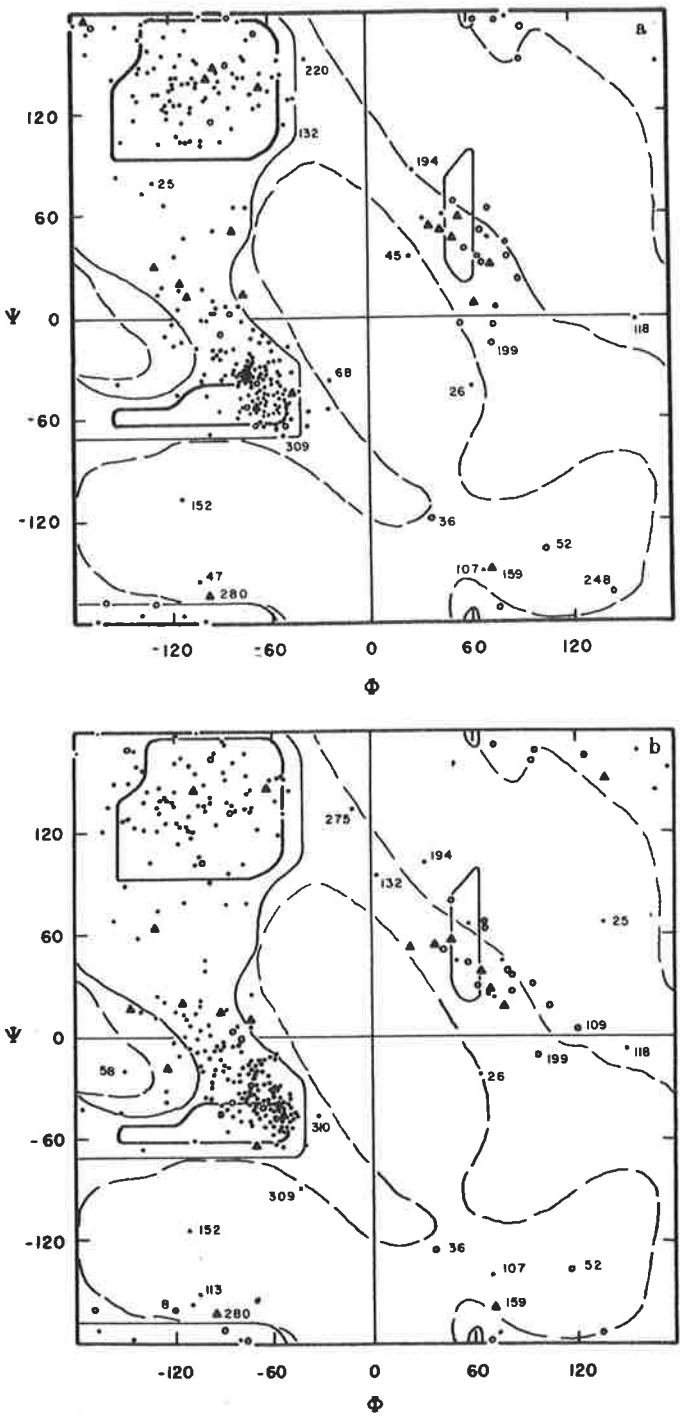


FIG. 4. Conformational angles for thermolysin; glycine residues are indicated by *open circles* and asparagine residues by *triangles*. The allowed regions for a hard-sphere model (19, 20) are indicated by *solid lines*, and those predicted by a quantum mechanical method (33) are indicated by a *broken line*. *a*, coordinates obtained from the electron density map; *b* "idealized" coordinates (see text).

bends observed in thermolysin. For example, the average dihedral angles for the 13 Type I bends in thermolysin (Table VIII) are (−61°, −35°), (−96°, 6°) whereas Chandrasekaran *et al.* (29) predict a maximum in the stabilization energy for bends with conformational angles (−50°, −50°), and (−110°, 40°). The apparent preference for ψ_3 to be near zero (*cf.* Fig. 4*a*) rather than near −40°, as predicted by theory, is particularly noteworthy. Furthermore, although there are eight Type III

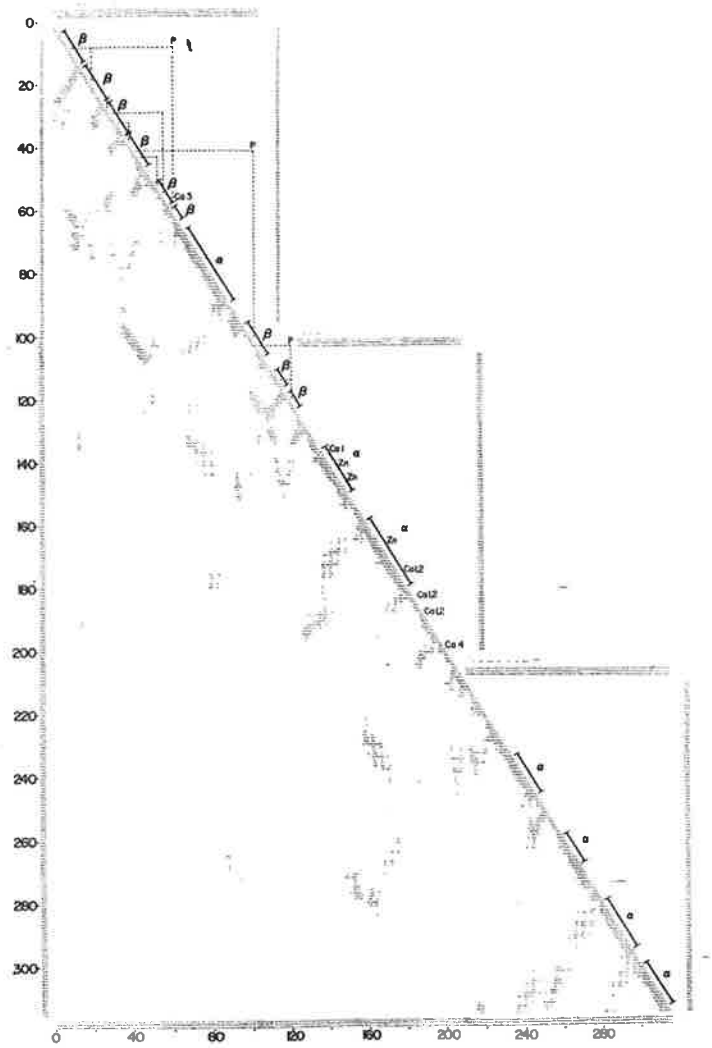


FIG. 5. Plot showing all pairs of α carbon atoms which are within 9 A of each other. Residues participating in extended secondary structure and in metal binding are indicated along the *diagonal*.

TABLE VI

Helices

The ends of the helices are defined by the first and last residues which participate in at least one helical-type backbone hydrogen bond and have essentially a helical conformation. The conformation of residues not in parentheses is approximately α helical, whereas those in parentheses are closer to a 3_{10} or α_{II} conformation. Residues 226–229 are left-handed α helical.

Residues included	Type of helix
(65)–(66)–67–87	3_{10}, α_{II}, α
(133)–(134)–(135)	3_{10}
137–150–(151)	α, α_{II}
160–179–(180)	α, 3_{10}
190–191–(192)	α, 3_{10}, α_{II}
(208)–(209)–(210)–(211)	3_{10}
(217)–(218)–(219)	3_{10}
[226–229]	Left-handed α
(233)–(234)–235–246	3_{10}, α
260–261–(262)–263–274	α, α_{II}, α
281–296	α
301–312–(313)	α, 3_{10}

TABLE VII

β Structure

Residues included in this list are in approximately a fully extended conformation and participate in at least one β type hydrogen bond. Residues in parentheses essentially continue the β sheet, but either do not participate in the hydrogen bonding or depart significantly from an extended conformation.

Residues included	Type of participation[a]
(2)–(3)–4–11	p, a
16–25	a, a
27–29	a
31–32	a
39–43	a, a
52–(54)–(55)–57	a
61–62	p
100–104–(105)–106	p, p
112–115	a
119–122	p, a
(187–188)	a
(203–204)	a
248–249–(250)	a
(253)–(254)–255	a

[a] p, parallel; and a, anti-parallel.

TABLE VIII

Hairpin bends

Residue number	Sequence	Type	Location if in extended secondary structure
12–15	Gly-Val-Leu-Gly	I	β sheet
25–27	Ser-Thy-Tyr	γ	β sheet
35–38	Arg-Gly-Asp-Gly	II′	β sheet
44–47	Ala-Lys-Tyr-Arg	III′	β sheet
64–67	Ala-Ser-Tyr-Asp	III	*N*-end, α helix
92–95	Ser-Tyr-Asp-Gly	I	
132–135	Pro-Leu-Ser-Gly	III	3_{10} helix
133–136	Leu-Ser-Gly-Gly	I	
149–152	Thr-Asp-Tyr-Thr	I	*C* end, α helix
178–181	Phe-Tyr-Ala-Asn	III	*C* end, α-helix
189–192	Gly-Glu-Asp-Val	III	Single helical turn
190–193	Glu-Asp-Val-Tyr	I	
207–210	Asp-Pro-Ala-Lys	I	
208–211	Pro-Ala-Lys-Tyr	III	3_{10} helix
209–212	Ala-Lys-Tyr-Gly	I	
216–219	His-Tyr-Ser-Lys	III	3_{10} helix
217–220	Tyr-Ser-Lys-Arg	I	
229–232	Gly-Val-His-Ile	I	
232–235	Ile-Asn-Ser-Gly	III	*N*-end, α helix
233–236	Asn-Ser-Gly-Ile	III	
260–263	Arg-Asp-Lys-Leu	I	*N*-end, α helix
276–279	Thr-Pro-Thr-Ser	I	
297–299	Gly-Ser-Thr-Ser	I	
311–314	Asp-Ala-Val-Gly	I	*C*-end, α helix

The following dihedral angles specify the theoretical conformation of the 2nd and 3rd residue for each bend type (26).

Type	ϕ_2	ψ_2	ϕ_3	ψ_3
I	−60	−30	−90	0
III	−60	−30	−60	−30
II′	60	120	−80	0
I′	60	30	90	0
III′	60	30	60	30
II	−60	120	80	0

(3_{10}) bends in thermolysin (Table VIII) with average conformation (−57°, −25°), and (−67°, −26°), Chandrasekaran *et al.* (29) do not find any energy minimum for dihedral angles close to these. (The closest minimum is that already quoted above.) These discrepancies demonstrate quite clearly that significant deformations of the polypeptide backbone must occur relatively frequently in proteins, and highlight the need for allowing for such distortions in empirical energy calculations.

Conformation of Individual Residues—The backbone dihedral angles for thermolysin are listed in Table IV and plotted in Fig. 4*a*. As discussed above, the probable error in these angles due to inaccuracies in measuring the atomic coordinates is about ±10°. This estimate does not include other, possibly more serious, errors which might arise through the misinterpretation of the electron density map. Notwithstanding the unusual thermostability of thermolysin, the over-all distribution of conformations is quite similar to those observed for lysozyme (31), carboxypeptidase A (16), ribonuclease S (32), and α-chymotrypsin (21), for example. As is clear from the figure, the conformations of most of the residues lie within or near the allowed regions for an empirical hardsphere model. On the other hand, we do not believe that the apparent incursions of some residues into disallowed regions can be attributed entirely to experimental errors. The electron density map in the vicinity of Thr 152, Ser 107, Asn 159, and Thr 26 *e.g.*, is well defined, and we do not think it likely that the conformations of these residues are in error by more than about 20°. As mentioned above, Thr 26 participates in a γ turn (28).

For comparison we have included in Fig. 4 the conformational limits for alanine determined by the quantum mechanical technique of Pullman and colleagues (33). In this case, only Thr 152 appears to have a conformation which may differ significantly from allowed.

Of the 39 asparagine residues in thermolysin, 7 have ϕ near 60°. The tendency for asparagine to have this conformation has been noted elsewhere (24).

The electron density at Pro 51 can only be explained satisfactorily by assuming this residue to be in the *cis* conformation.

Calcium Binding Sites—For optimal thermostability, thermolysin requires the presence of calcium ion (34), although the metal is not involved directly with catalytic activity (3, 35). The x-ray analysis of thermolysin was carried out in the presence of 0.01 M Ca^{2+}, and under these conditions the enzyme binds four calcium ions, apparently with essentially full occupancy in each case.

The stereochemistry of the four calcium binding sites is illustrated in Fig. 6, and the distances to the closest ligands are listed in Table IX. It must be emphasized that these distances were calculated from the preliminary atomic coordinates listed in Table V, and have an estimated uncertainty of about ±0.4 A. As judged by the appearance of the thermolysin electron density map, the primary calcium ligands at Site 1 are $O^{\delta 2}$ of Asp 185, $O^{\epsilon 1}$ of Glu 177, $O^{\epsilon 1}$ of Glu 190, $O^{\delta 1}$ of Asp 138, the backbone carbonyl oxygen of Glu 187, and an internal water molecule. In addition, $O^{\epsilon 2}$ of Glu 177 and $O^{\epsilon 2}$ of Glu 190 are sufficiently close to Ca 1 that they possibly contribute to the over-all coordination of this ion. The latter 2 oxygens appear to interact more strongly with Ca 2, which also has as ligands $O^{\delta 1}$ of Asp 185, the backbone carbonyl oxygen of Asn 183, and 2 surface water molecules. It might be noted that there are examples, for instance in the structure of $CaCl_2$·glycylglycylglycine (36), where 2 calcium ions share the coordination of a single oxygen atom. Depending on the extent to which such sharing occurs at the

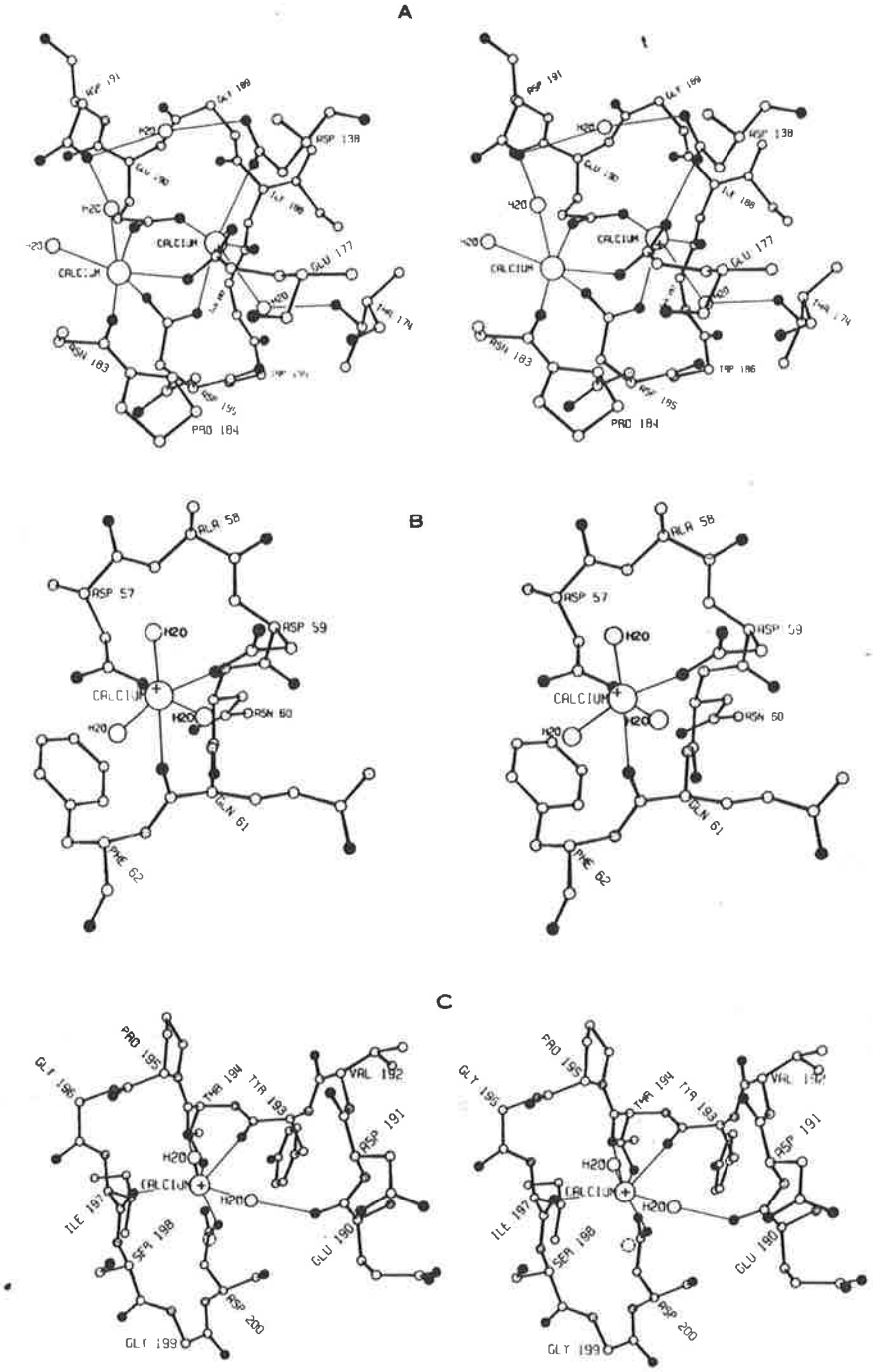


FIG. 6. Stereodrawings illustrating the calcium binding sites. Carbon atoms are represented by *open circles*, oxygen by *solid circles*, and nitrogen atoms are drawn *stippled*. Water molecules and calcium ions, indicated by *larger circles*, are labeled. The calcium ligands are indicated by *thin lines*, and the *crosses* show the sites occupied by europium ions. *a*, Sites 1 and 2; *b*, Site 3; and *c*, Site 4.

double calcium site in thermolysin, the coordination number of Ca1 could range from 6 to 8. The coordination number of Ca 2 is 6. These numbers are consistent with the studies of the structure of phosphates, silicates, and other calcium complexes where calcium ions have been found to have coordination numbers ranging from 6 to 9, and calcium-oxygen distances from 2.3 to 3.0 A (37).

When europium ion is used to displace calcium from thermolysin, both Ca 1 and Ca 2 are lost, and a single Eu^{3+} ion binds about 0.3 A from the Ca 1 site, as shown in Fig. 6*a* (7). For comparison, we have included in Table IX the presumed ligand distances calculated for the europium ion, assuming the protein structure to be unchanged. Judging by these distances it seems likely that the Eu^{3+} ion bound at Site 1 is very likely coordinated by the eight potential Ca 1 ligands listed above, and possibly also by $O^{\delta 1}$ of Asp 185, to give an over-all coordination number of 8 or 9. Because no europium ion is bound at Site 2, it is reasonable to suppose that one or more of the side chains of Glu 177, Glu 190, and Asp 185 adjust slightly so that the coordination distances from their respective oxygen atoms to the bound europium ion are more nearly equal than indicated in Table IX.

There is some similarity between the double calcium binding site in thermolysin, and the $Ca^{2+} - Mn^{2+}$ binding site in con-

canavalin A (38, 39), although the comparison cannot be pressed too closely. In concanavalin A, the distance between the Ca^{2+} and Mn^{2+} is about 5 A (38, 39) compared with 3.8 A between the two Ca^{2+} ions in thermolysin. In both proteins the metal ligands include bridging acid groups, two for concanavalin A (38, 39) and three for thermolysin. Edelman *et al.* (38) describe the coordination of both Ca^{2+} and Mn^{2+} as roughly octahedral, and this is also true for the closest ligands at the double calcium binding site in thermolysin (Fig. 6*a*). On the other hand, the folding of the polypeptide chain which results in the respective double metal binding sites of concanavalin A and thermolysin seem to be quite different. The Mn^{2+} ligands in concanavalin A include a histidine imidazole, whereas the calcium ligands as in thermolysin are restricted to acid oxygens, backbone carbonyl oxygens, and water molecules.

TABLE IX

Calcium and europium ligand distances

Ligand	Distance (A)		
	Ca 1	Eu 1	Ca 2
Sites 1 and 2			
Asp 185 OD1	3.6	3.3	2.4
OD2	2.7	2.4	
Glu 177 OE1	2.1	2.1	
OE2	2.6	2.4	2.1
Glu 190 OE1	1.9	2.1	
OE2	2.6	2.6	2.1
Glu 187 O	2.0	2.1	
Asn 183 O			2.1
Asp 138 OD1	3.0	3.3	
H_2O 1	2.9	2.7	
H_2O 2			2.9
H_2O 3			2.6
	Ca 3	Eu 3	
Site 3			
Asp 57 OD1	3.5	3.6	
OD2	2.6	3.0	
Asp 59 OD1	2.0	2.4	
OD2	3.4	3.3	
Glu 61 O	2.6	3.2	
H_2O	2.3	2.0	
H_2O	2.1	1.8	
H_2O	2.1	1.7	
	Ca 4	Eu 4	
Site 4			
Tyr 193 O	2.5	2.4	
Thr 194 O	1.8	1.9	
OG1	2.2	2.7	
Ile 197 O	3.0	2.9	
Asp 200 OD1	2.5	2.8	
H_2O	2.5	2.2	
H_2O	2.2	1.8	
H_2O		2.8	

The geometry of the third calcium binding site in thermolysin is illustrated in Fig. 6*b*. In this case, the six closest ligands (Table IX) are in roughly octahedral arrangement, but it is quite likely that the second δ-oxygen of both Asp 57 and Asp 59 may contribute to the calcium coordination. For comparison, the calculated ligand distances for a europium ion bound at Site 3 are also shown in Table IX. The calculated europium distances should be regarded with caution, first because there is evidence for a slight movement in the vicinity of Asp 59 when Eu^{3+} replaces Ca^{2+} (7), and second because the 3 water molecules almost certainly do not occupy exactly the same positions after the europium substitution.

The coordination at the fourth calcium site appears to be 7-fold (Fig. 6*c*) and includes a threonine hydroxyl. On europium substitution, an additional water molecule increases the coordination number to 8 (7). It may be noted that as for the double calcium binding site, Site 4 is located in the "waist" region of the molecule. The respective binding sites are about 12 A apart, and the relation between them may be visualized by making use of the fact that Glu 190 and Asp 191, which appear in Fig. 6*a*, also appear in Fig. 6*c*. Both sites are within about 10 A of the hydrophobic pocket in the active site, and may help in part to stabilize this region of the molecule.

Binding sites for Ca 3 and Ca 4 both consist of essentially a single loop of polypeptide chain. It could be argued that the calcium tends to stabilize such loops, and prevent autolysis, although there are a number of surface loops in the polypeptide chain which do not, apparently, require calcium for stabilization.

As is clear from Fig. 6, *b* and *c* the over-all folding of the polypeptide chain in the vicinity of Ca 3 and Ca 4 is quite dissimilar; neither does there seem to be any over-all similarity between these sites and the two calcium binding sites of the calcium binding-carp muscle binding protein studied by Kretsinger and Nockolds (40, 41), or the calcium site in the inhibitor complex of *Staphylococcus aureus* nuclease, analyzed by Cotton and colleagues (42, 43).

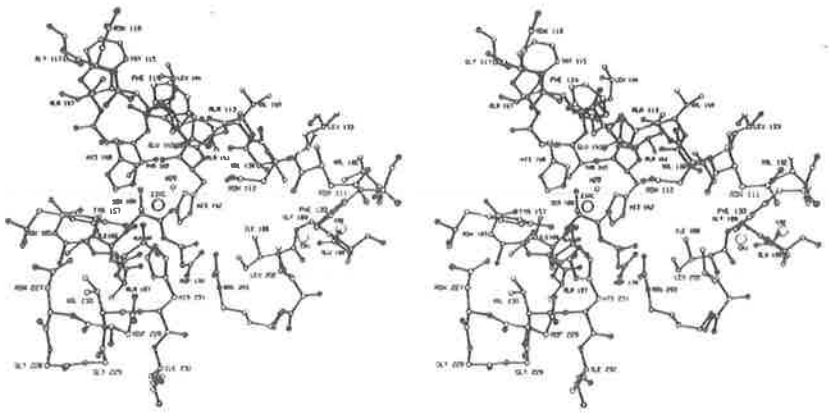


FIG. 7. Active site of thermolysin. Atoms are identified as in Fig. 6.

Active Site—The general stereochemistry of the active site region of the molecule is illustrated in Fig. 7. The zinc ion is liganded by N^{ϵ} of His 142, N^{ϵ} of His 146, O^{ϵ} of Glu 166, and an apparent water molecule, in an approximately tetrahedral arrangement. Asp 226, which interacts ionically with His 231, is one of the residues in the single turn of the left-handed helix discussed above. To the right of the zinc ion, as seen in Fig. 7, a number of nonpolar residues form a well defined hydrophobic pocket.

In experiments which are underway at the present time, we have shown by difference Fourier methods that a number of competitive inhibitors bind in the general vicinity of the zinc ion, and in part occupy the hydrophobic pocket. However, the detailed interpretation of the binding of these inhibitors is complicated by the finding that some analogous inhibitors apparently exhibit radically different modes of binding. For example, preliminary experiments suggest that L-alanyl-L-phenylalanine and possibly β-phenylpropionyl-L-phenylalanine (44) bind in a "normal" mode with the phenylalanyl side chain in the specificity pocket, but that benzyloxycarboxyl-L-phenylalanine (45) binds "backwards" with the acyl group in the pocket. Independent evidence of the existence of different modes of interaction for these thermolysin inhibitors has been provided recently by a nuclear magnetic resonance study of their binding to the manganese-substituted enzyme (46).

In view of the fact that the crystallographic experiments on inhibitor binding are not yet definitive, we will defer discussion of these results, together with a detailed comparison of the active sites of thermolysin and carboxypeptidase A, to a subsequent publication.

Thermal Stability—With the possible exception of the four calcium binding sites there is nothing particularly unusual in the thermolysin conformation to which one might attribute the thermostability of the molecule. In fact, the absence of unusual features might be taken as evidence that the resistance of thermolysin to thermal denaturation might be due, not to a single determinant, but to a combination of factors including calcium binding, hydrophobic interactions, hydrogen bonding, and ionic stabilization.

The fact that thermolysin is stabilized by calcium has been recognized since the original isolation of the enzyme (34). For maximum stability at elevated temperatures the enzyme requires at least 0.1 M $CaCl_2$, while in the absence of calcium it is quite unstable (35).

Stabilization and protection of macromolecules by metal ions has been recognized for many years, and many examples are known (*e.g.* see Refs. 47–54). Also, a survey of the literature suggests that many of the proteins from thermophilic organisms are stabilized or activated (or both) by metal ions. (The recent review article by Singleton and Amelunxen (55) may be consulted for references.) It is therefore tempting to speculate that metal stabilization might provide a general mechanism for thermostability. On the other hand, the existence of proteins which apparently do not require metal ions for thermostability suggests that stabilization can be achieved by other means. Hsiu *et al.* (49) have proposed that the alkaline earth metals may stabilize protein structures by forming intramolecular bridges or cross-links analogous to disulfide bonds. Although the double calcium site in thermolysin could be envisaged readily as "cross-linking" the two lobes of the structure, the roles of Ca 3 and Ca 4 seem to be different. Both of these ions are bound in loops of the polypeptide chain on the surface of the molecule, and it might be speculated that in these instances the function of the calcium ions is to protect these loops against autolysis. In the vicinity of Ca 4, in particular, the polypeptide backbone extends into the solvent and seems to be quite exposed. Also, it may be noted from Fig. 5 that the site of binding of Ca 4 is within the most extended region of irregular conformation in the thermolysin molecule, and it seems reasonable to suppose that in the absence of calcium this portion of the molecule would become a likely site of autolytic attack.

It has also been proposed that thermostability might be attributed to more numerous and stronger hydrogen or hydrophobic bonds (or both) in thermophilic proteins relative to their mesophilic counterparts (56). The contribution of the hydrophobic bonds is of particular interest because the strength of the hydrophobic interactions increases with temperature up to about 60–70°, whereas the contribution of hydrogen bonding to the free energy remains approximately constant with increasing temperature (57, 58). However, extensive comparisons of the proportion of hydrophobic residues in thermophilic proteins and their mesophilic counterparts do not reveal any definite correlation between hydrophobicity and thermal stability (55, 59). Of course, such comparisons are necessarily crude, and do not measure the actual hydrophobic interactions within a given protein, but the negative result does suggest that the differences in hydrophobic bonding between thermophilic and mesophilic proteins may be quite subtle. The conclusion that the three-dimensional structures of thermophilic and mesophilic proteins may be very similar is supported most convincingly by a comparison of the amino acid sequences of similar molecules from thermophilic and mesophilic sources. For example, it was found that the primary structure of the heat-stable ferredoxin from the thermophile *Clostridium tartarivorum* was obviously similar to those of four mesophilic bacterial ferredoxins (60), suggesting that the respective three-dimensional structures were also closely related. The same conclusion was reached from a comparison of the amino acid sequences of glyceraldehyde 3-phosphate dehydrogenases from *Bacillus stearothermophilus* and yeast, where an over-all homology of approximately 70% was observed (61). Also, preliminary results indicate that a neutral protease from *Bacillus subtilis*, much less stable than thermolysin, has a homologous amino acid sequence (62, 63). These findings all indicate that the three-dimensional structures of these related enzymes must be quite similar, and suggest that the difference between thermostability and thermolability in a given instance may be due to rather subtle changes in hydrophobic character, metal binding, hydrogen bonding, ionic interactions, or a combination of all of these.

The notion that thermostable proteins might exist as randomly coiled molecules, yet still be active, as proposed by Manning *et al.* (64), is very unlikely on general grounds, and is certainly not the case for thermolysin. Also, the inability of Pfueller and Elliott (65) to reproduce the results of Manning and Campbell (64, 66) places the earlier work in question, although it is stated (67) that the strains of bacteria used for the respective studies were not identical.

It has been reported that a number of thermostable proteins undergo conformational changes as a function of temperature; however, experiments with crystals of thermolysin suggest that in this case no such changes occur up to at least 65°. In the first place, crystals of thermolysin can be heated in water to temperatures of 70° without cracking or undergoing any macroscopic change indicative of a major rearrangement of the structure. Also, photographs of thermolysin crystals kept at constant temperatures up to 65° during x-ray exposure do not show any

detectable changes in the diffracted intensities other than a general attenuation due to an increased "thermal factor." These experiments show that at temperatures up to 65°, and possibly higher, the structure of the thermolysin molecule in the crystal does not alter by more than a few tenths of an Angstrom. Furthermore, because the protein in the crystal is in a low salt or salt-free environment, similar to that in solution, and because the lattice stabilization energy for protein crystals is relatively small (estimated, following Perutz (68) as 6 Cal per mole for thermolysin), it can be assumed reasonably that thermolysin in solution does not undergo a conformational change as a function of temperature, at least up to about 70°.

Above 70°, crystals of thermolysin begin to develop slight cracks which become severe at about 85°, possibly indicating a conformational change, or partial denaturation of the molecules in the crystal. These observations are consistent with the conclusion of Ohta *et al.* (69) that after incubation at 80° for 1 hour, thermolysin undergoes a conformational change. On the other hand, the three-dimensional structure of thermolysin does not support the conclusion of Ohta (70) that an abundance of abnormally ionizing tyrosine residues plays an important role in stabilizing the enzyme through hydrogen bonding of the phenolic hydroxyls to carboxy and carbonyl groups, and through hydrophobic bonds involving the aromatic ring. Of the 28 tyrosine side chains, two interact with acidic groups, five with basic groups, and an additional four with both a basic and an acidic group. The fact that only 2 tyrosine residues are internal (4) does not suggest that tyrosine plays an unusually important role in hydrophobic stabilization.

Acknowledgments—We wish to acknowledge the substantial contributions of Drs. P. M. Colman and J. N. Jansonius during the previous phases of this study. Also, we are grateful to a number of colleagues including Drs. F. W. Dahlquist and L. F. TenEcyk, and Mr. S. J. Remington and J. O. Selzer for their helpful conversations and for a variety of assistance.

REFERENCES

1. Titani, K., Hermodson, M. A., Ericsson, L. H., Walsh, K. A., and Neurath, H. (1972) *Nature New Biol.* **238**, 35
2. Matthews, B. W., Jansonius, J. N., Colman, P. M., Schoenborn, B. P., and Dupourque, D. (1972) *Nature New Biol.* **238**, 37
3. Matthews, B. W., Colman, P. M., Jansonius, J. N., Titani, K., Walsh, K. A., and Neurath, H. (1972) *Nature New Biol.* **238**, 41
4. Colman, P. M., Jansonius, J. N., and Matthews, B. W. (1972) *J. Mol. Biol.* **70**, 701
5. Hart, R. G. (1901) *Acta Crystallogr.* **14**, 1188
6. Colman, P. M., Weaver, L. H., and Matthews, B. W. (1972) *Biochem. Biophys. Res. Commun.* **46**, 1999
7. Matthews, B. W., and Weaver, L. H. (1974) *Biochemistry* **13**, 1719
8. Broderson, K., Thiele, G., and Holle, B. (1969) *Z. Anorg. Allg. Chem.* **369**, 154
9. Lipscomb, W. N., Hartsuck, J. A., Reeke, G. N., Jr., Quiocho, F. A., Bethge, P. H., Ludwig, M. L., Steitz, T. A., Muirhead, H., and Coppola, J. C. (1968) *Brookhaven Symp. Biol.* **21**, 24
10. Waara, I., Lövgren, S., Liljas, A., Kannan, K. K., and Bergstén, P.-C. (1972) in *Hemoglobin and Red Cell Structure and Function, Advances in Experimental Medicine and Biology* (Brewer, G. J., ed) p. 169, Plenum Publishing Corp., New York
11. Matthews, B. W., Klopfenstein, C. E., and Colman, P. M. (1972) *J. Sci. Instrum.* **5**, 353
12. Blow, D. M., and Crick, F. H. C. (1959) *Acta Crystallogr.* **12**, 794
13. Matthews, B. W. (1966) *Acta Crystallogr.* **20**, 82
14. Matthews, B. W. (1969) in *Crystallographic Computing* (Ahmed, F. R., ed) p. 146, Munksgaard, Copenhagen
15. Richards, F. M. (1968) *J. Mol. Biol.* **37**, 225
16. Lipscomb, W. N., Reeke, G. N., Jr., Hartsuck, J. A., Quiocho, F. A., and Bethge, P. H. (1970) *Phil. Trans. Roy. Soc. Lond. Ser. B Biol. Sci.* **257**, 177
17. Diamond, R. (1966) *Acta Crystallogr.* **21**, 253
18. IUPAC-IUB Commission on Biochemical Nomenclature (1970) *J. Biol. Chem.* **245**, 6489
19. Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963) *J. Mol. Biol.* **7**, 95
20. Ramachandran, G. N., and Sasisekharan, V. (1968) *Advan. Protein Chem.* **23**, 283
21. Birktoft, J. J., and Blow, D. M. (1972) *J. Mol. Biol.* **68**, 187
22. Protein Data Bank (1973) *J. Mol. Biol.* **78**, 587
23. Phillips, D. C. (1970) in *British Biochemistry Past and Present* (Goodwin, T. W., ed) Biochemical Society Symposium No. 30, p. 11, Academic Press, New York
24. Matthews, B. W. (1975) in *The Proteins* (Neurath H., and Hill, R. L., eds) 3rd Ed, in press, Academic Press, New York
25. Némethy, G., Phillips, D. C., Leach, S. J., and Scheraga, H. A. (1967) *Nature* **214**, 363
26. Venkatachalam, C. M. (1968) *Biopolymers* **6**, 1425
27. Némethy, G., and Printz, M. P. (1972) *Macromolecules* **5**, 755
28. Matthews, B. W. (1972) *Macromolecules* **5**, 818
29. Chandrasekaran, R., Lakshminarayanan, A. V., Pandya, U. V., and Ramachandran, G. N. (1973) *Biochim. Biophys. Acta* **303**, 14
30. Ramachandran, G. N., Chandrasekaran, R., and Chidambaram, R. (1971) *Proc. Indian Acad. Sci. Sect. A* **74**, 270
31. Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C., and Sarma, V. R. (1967) *Proc. Roy. Soc. London Ser. B Biol. Sci.* **167**, 365
32. Richards, F. M., and Wyckoff, H. W. (1971) in *The Enzymes* (Boyer, P. D., ed) 3rd Ed, Vol. 4, p. 647, Academic Press, New York
33. Pullman, B., Maigret, B., and Perahia, D. (1970) *Theor. Chim. Acta* **18**, 44
34. Endo, S. (1962) *J. Ferment. Technol.* **40**, 346
35. Feder, J., Garrett, L. R., and Wildi, B. S. (1971) *Biochemistry*, **10**, 4552
36. Van der Helm, D., and Willoughby, T. V. (1969) *Acta Crystallogr. Sect. B* **25**, 2317
37. *International Tables for X-ray Crystallography* (1962) (MacGillavry, C. H., and Rieck, G. D., eds) Vol. 3, p. 260, Kynoch Press, Birmingham, England
38. Edelman, G. M., Cunningham, B. A., Reeke, G. N., Jr., Becker, J. W., Waxdal, M. J., and Wang, J. L. (1972) *Proc. Nat. Acad. Sci. U. S. A.* **69**, 2580
39. Hardman, K. D., and Ainsworth, C. F. (1972) *Biochemistry* **11**, 4910
40. Kretsinger, R. H. (1972) *Nature New Biol.* **240**, 85
41. Kretsinger, R. H., and Nockolds, C. E. (1973) *J. Biol. Chem.* **248**, 3313
42. Arnone, A., Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. E., Jr., Richardson, D. C., Richardson, J. S., and Yonath, A. (1971) *J. Biol. Chem.* **246**, 2302
43. Cotton, F. A., Bier, C. J., Day, V. W., Hazen, E. E., Jr., and Larsen, S. (1971) *Cold Spring Harbor Symp. Quant. Biol.* **36**, 243
44. Blumberg, S., Holmquist, B., and Vallee, B. L. (1973) *Biochem. Biophys. Res. Commun.* **51**, 987
45. Burstein, Y., Walsh, K. A., and Neurath, H. (1974) *Biochemistry* **13**, 205
46. Bigbee, W. L., and Dahlquist, F. W. (1974) *Biochemistry* **13**, 3542
47. Gorini, L. (1951) *Biochim. Biophys. Acta* **7**, 318
48. Green, N. M., Gladner, J. A., Cunningham, L. W., Jr., and Neurath, H. (1952) *J. Amer. Chem. Soc.* **74**, 2122
49. Hsiu, J., Fischer, E. H., and Stein, E. A. (1964) *Biochemistry* **3**, 61
50. Shapiro, B. M., and Ginsburg, A. (1968) *Biochemistry* **7**, 2153
51. McConn, J. D., Tsuru, D., and Yasunobu, K. T. (1964) *J. Biol. Chem.* **239**, 3706
52. Millet, J. (1969) *Bull. Soc. Chim. Biol.* **51**, 61

53. Nomoto, M., Narahashi, Y., and Murakami, M. (1960) *J. Biochem. (Tokyo)* **48**, 453
54. Levy, J., Rialdi, G., and Biltonen, R. (1972) *Biochemistry* **11**, 4138
55. Singleton, R., Jr., and Amelunxen, R. E. (1973) *Bacteriol. Rev.* **37**, 320
56. Koffler, H., Mallett, G. E., and Adye, J. (1957) *Proc. Nat. Acad. Sci. U. S. A.* **43**, 464
57. Brandts, J. F. (1967) in *Thermobiology* (Rose, A. H., ed) p. 25, Academic Press, New York
58. Scheraga, H. A. (1963) in *The Proteins* (Neurath, H., ed) 2nd Ed., Vol. 1, p. 477, Academic Press, New York
59. Goldsack, D. E. (1970) *Biopolymers* **9**, 247
60. Tanaka, M., Naniu, M., Matsueda, G., Yasunobu, K. T., Himes, R. H., Akagi, J. M., Barnes, E. M., and Devanathan, T. (1971) *J. Biol. Chem.* **246**, 3953
61. Bridgen, J., and Harris, J. I. (1973) *Ninth Int. Congr. Biochem.* Stockholm, July
62. Pangburn, M. K., Burstein, Y., Morgen, P. H., Walsh, K. A., and Neurath, H. (1973) *Biochem. Biophys. Res. Commun.* **54**, 371
63. Pangburn, M. K. (1974) *Ph.D. thesis*, University of Washington
64. Manning, G. B., Campbell, L. L., and Foster, R. J. (1961) *J. Biol. Chem.* **236**, 2958
65. Pfueller, S. L., and Elliott, W. H. (1969) *J. Biol. Chem.* **244**, 48
66. Manning, G. B., and Campbell, L. L. (1961) *J. Biol. Chem.* **236**, 2952
67. Farrell, J., and Campbell, L. L. (1969) *Advan. Microbial Physiol.* **3**, 83
68. Perutz, M. F. (1965) *J. Mol. Biol.* **13**, 646
69. Ohta, Y., Ogura, Y., and Wada, A. (1966) *J. Biol. Chem.* **241**, 5919
70. Ohta, Y. (1967) *J. Biol. Chem.* **242**, 509

[Reprinted from Biochemistry, (1977) 16, 2506.]
Copyright 1977 by the American Chemical Society and reprinted by permission of the copyright owner.

Crystallographic Study of the Binding of Dipeptide Inhibitors to Thermolysin: Implications for the Mechanism of Catalysis†

W. R. Kester and B. W. Matthews*

ABSTRACT: The mode of binding to thermolysin of a number of dipeptide inhibitors including β-phenylpropionyl-L-phenylalanine and carbobenzoxy-L-phenylalanine has been determined by x-ray crystallography to a nominal resolution of 2.3 Å. These results indicate the probable mode of binding of extended substrates to thermolysin and also suggest a mechanism of action for this neutral metalloendopeptidase which is similar in a number of respects to one of the two alternate mechanisms for peptide hydrolysis by carboxypeptidase A proposed by Lipscomb and co-workers (Lipscomb, W. N., et al. (1968), *Brookhaven Symp. Biol. 21*, 24). The crystallographic results suggest that substrates bind to thermolysin with the carbonyl oxygen of the scissile peptide bond displacing a water molecule and becoming the fourth zinc ligand. Also, in the case of β-phenylpropionyl-L-phenylalanine, the side chain of Glu-143 and a neighboring water molecule are seen to be close to the carbonyl carbon of the scissile peptide, as is the imidazole of His-231 to the peptide nitrogen, suggesting that these groups may participate in catalysis. It is proposed that following substrate binding, Glu-143, acting as a general base, promotes the attack of a water molecule on the carbonyl carbon of the scissile peptide bond. Concurrent with, or following this step, His-231 donates a proton to the peptide nitrogen, forming a tetrahedral intermediate which is stabilized by both hydrogen bonds and hydrophobic interactions. Finally, the carbon–nitrogen bond of the intermediate breaks to yield the products. Thus, Glu-143 in thermolysin appears to be a counterpart of Glu-270 in carboxypeptidase A, while the role of His-231 as a proton donor is analogous to that of Tyr-248. The proposed mechanism for thermolysin is similar to that suggested by Lipscomb for carboxypeptidase A in which Glu-270 acts as a general base. In contrast, however, in the case of thermolysin, stereochemical restrictions appear to exclude the possibility of direct nucleophilic attack of the carbonyl carbon by the acid, as occurs in the alternative mechanism of action for carboxypeptidase A proposed by Lipscomb and co-workers.

Thermolysin (TLN)[1] is a heat-stable extracellular endopeptidase of molecular weight 34 600 isolated from *Bacillus thermoproteolyticus* (Endo, 1962). The enzyme binds one zinc ion, required for activity (Latt et al., 1969), and four calcium ions which are necessary for optimal thermostability (Feder et al., 1971). Kinetic studies have shown that TLN specifically hydrolyzes peptide bonds on the imino side of large hydrophobic residues such as leucine, isoleucine, and phenylalanine and that hydrolysis of peptide bonds between two successive hydrophobic residues is most efficient.

The amino acid sequence (Titani et al., 1972) and the x-ray structure to a nominal resolution of 2.3 Å have been determined (Matthews et al., 1972a,b; Colman et al., 1972). Preliminary atomic coordinates are available and are currently being refined (Matthews et al., 1974). In this communication, we describe the mode of binding of several dipeptide inhibitors to the crystalline enzyme and present a mechanism of action which is suggested by these results.

Experimental Section

Materials. Thermolysin, three times recrystallized, was obtained from Calbiochem. Inhibitors, the companies from which they were purchased, and the abbreviations used in this paper follow: L-alanyl-L-phenylalanine (Ala-Phe), Cyclo Chemical; *N*-carbobenzoxy-L-phenylalanine (Cbz-Phe), Cyclo Chemical; L-phenylalanyl-L-phenylalanylamide (PPN), Fox Chemical Co.; L-phenylalanine (L-Phe), Cyclo Chemical. β-Phenylpropionyl-L-phenylalanine (β-PPP) was a generous gift of Dr. B. L. Vallee.

Methods. Native crystals suitable for x-ray diffraction were prepared as described previously (Colman et al., 1972), and kept in a solution of 0.01 M calcium acetate, 0.01 M Tris–acetate, and 5% by volume dimethyl sulfoxide, pH 7.3, used as standard mother liquor in the determination of the native structure. To prepare enzyme–inhibitor derivatives, native crystals were soaked at room temperature in standard mother liquor to which saturating, or nearly saturating, concentrations of the various inhibitors were added and adjusted to pH 7.0. Since binding appeared to be complete after soaking overnight, soaking times were from 2 to 8 days, with the solutions refreshed every 2 days. The concentration used for x-ray analysis and approximate K_I of each inhibitor are given in Table I.

Thermolysin crystallizes in space group $P6_122$, with cell axes $a = b = 94.2$ Å, $c = 131.4$ Å. Data were collected by the photographic method using Enraf-Nonius precession cameras in the same manner as isomorphous derivatives for phase determination. The x-ray source was an Elliot GX-6 rotating anode generator run at 40 kV, 40 mA, and crystals were exposed for 40 h. A summary of the data collection statistics and changes in unit cell parameters is given in Table II.

It was found that the different inhibitors could not be reliably screened using *h0l* projection data alone, because of the small differences observed when inhibitors were bound. Therefore 2.3 Å resolution "subsets" were collected, consisting of 6 planes: *h0l*, *h*1*l*, *hhl*, *h*,*h* − 1,*l*, *hk*0, and *hk*1. Compared with a "full" data set, such a subset contains about a third of

† From the Institute of Molecular Biology and the Departments of Chemistry and Physics, University of Oregon, Eugene, Oregon 97403. *Received December 22, 1976.* This work was supported in part by the National Institutes of Health (GM 20066; GM 15423), by the National Science Foundation (PCM 74-18407), and by a Career Development Award (GM 70585) from the National Institute of General Medical Science to B.W.M.

[1] Abbreviations used: TLN, thermolysin; Ala-Phe, L-alanyl-L-phenylalanine; Cbz-Phe, *N*-carbobenzoxy-L-phenylalanine; PPN, L-phenylalanyl-L-phenylalanylamide L-Phe, L-phenylalanine; β-PPP, β-phenylpropionyl-L-phenylalanine; Tris, tris(hydroxymethyl)aminomethane; rms, root mean square; NMR, nuclear magnetic resonance; DMA, dimercury acetic acid.

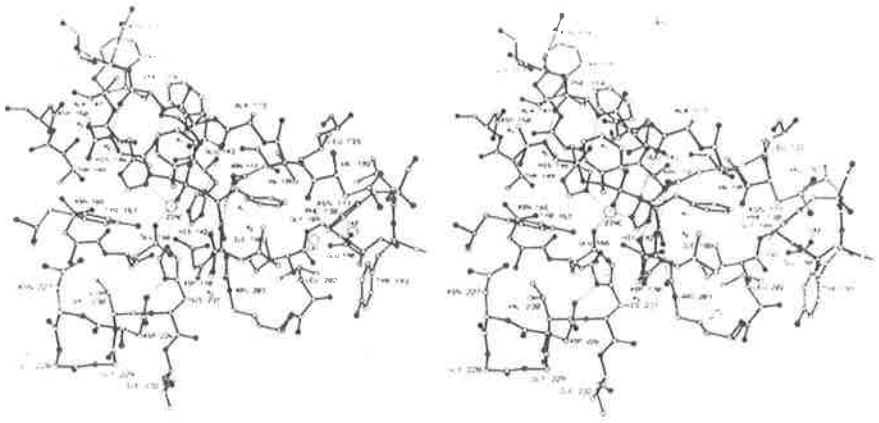


FIGURE 11: Stereodiagram showing the binding of the presumed tetrahedral transition state intermediate to thermolysin.

The alternative schemes outlined above are clearly similar to those postulated for carboxypeptidase A (Lipscomb et al., 1968, 1969, 1970). Glu-143 of TLN appears to be a counterpart of Glu-270 in carboxypeptidase A, as anticipated previously (Colman et al., 1972), while the role of His-231 in TLN as a proton donor is analogous to the role of Tyr-248 in carboxypeptidase A, as also suggested by Pangburn and Walsh (1975). A detailed comparison of the active sites of TLN and carboxypeptidase A will be presented elsewhere, but for the moment it may be noted that the binding of dipeptides to TLN is very similar to that of, for example, glycyltyrosine to carboxypeptidase A, although no major movements are observed in any thermolysin side chains, as is the case with Tyr-248 and Arg-145 of carboxypeptidase A.

Although it has not been possible to date to differentiate between the mechanisms of Figures 9 and 10 above as favored mechanisms for the metalloproteases, the following structural considerations suggest that the mechanism of Figure 9 is more plausible for thermolysin. As mentioned earlier, when β-PPP is bound to TLN, the closest oxygen of Glu-143 is 3.9 Å from the carbonyl carbon of the scissile bond. Inspection of the complex (Figure 8) shows that Glu-143 cannot be brought closer by rotations about side-chain bonds. Also the backbone of Glu-143, being in an internal helix, is not free to move. On attempting to build models of intermediates which would be formed in the case of the mechanism of Figure 9, it was found that the tetrahedral intermediate (A) could be made to fit very reasonably in the active site with a minimum of change from the positions proposed for substrate binding (Figure 11). Each of the hydrogen bonds can be preserved, and there appear to be no prohibitive van der Waals contacts. Also, the proposed hydrogen bond with Glu-143 has a length of 2.5–2.7 Å and the appropriate stereochemistry. The change from planar to tetrahedral geometry at the carbonyl carbon causes the side chain of R_1 to move closer to Phe-114, into approximately the same position as the equivalent group in the β-PPP structure, "stacking" next to the Phe-114 side chain. Also, when the nitrogen of the scissile peptide becomes tetrahedral, it moves ~0.7 Å closer to His-231. In summary there is no stereochemical reason to exclude the mechanism in Figure 9.

On the other hand, model building suggests that the mechanism in Figure 10 is implausible. In order to form intermediate (B), the tetrahedral carbon of the hydrolyzed bond has to be brought 1–1.5 Å closer to Glu-143, resulting in a number of close contacts. The R_1 phenyl group would have to undergo a major reorientation in order to prevent contacts of 2–3 Å with the Phe-114 side chain, and also the hydrogen bond between the R_2 carbonyl and the Trp-115 nitrogen would be shortened to about 2 Å. Also, coordination around the zinc would become very distorted, and the nitrogen of the scissile bond would move farther away from His-231, to a separation of about 5 Å. Similarly, model building of the anhydride intermediate (C) indicates that it would be in a very high-energy conformation. The zinc–oxygen ligand distance would be stretched to about 2.8 Å at the minimum, and the R_1 α carbon would be only 2 Å from C^β of Phe-114. Since all of the close contacts seen for intermediates B and C are with relatively immobile main chain and C^β atoms, a major change in the conformation of the active site would have to be postulated for the mechanism in Figure 10 to be possible. For this reason we favor the mechanism in Figure 9.

The binding of dipeptide inhibitors and the model-building studies based on them suggest that a number of groups participate in substrate binding and may also contribute significantly toward stabilization of the transition state. Figures 7, 8, and 11 illustrate the presumed hydrogen-bonded interactions of the substrate upon initial binding and in the transition state. All of the available hydrogen-bond donors and acceptors of the substrate appear to interact with the enzyme. Three of these bonds are with main-chain atoms of the enzyme and four with side chains.

Hydrophobic interactions also are important in forming the Michaelis complex and possibly in stabilizing the transition state. The largest contribution seems to come from the binding of the R_1' side chain in the hydrophobic pocket. At this site, the interactions appear to be substantially the same for the initial substrate binding mode and for the transition state. On the other hand, the hydrophobic interaction between the R_1 side chain and Phe-114 may increase in the transition state. The β-PPP–thermolysin complex indicates that the most favorable orientation of R_1 is "stacked" next to Phe-114. When attempts were made to model-build an extended substrate, it was found that the additional stereochemical restrictions imposed by adding residue R_2 necessitated a shift of the R_1 side chain away from Phe-114. This shift was also indicated in the subset difference map for the Phe-Phe-amide where, in contrast to β-PPP, R_1 has the amino group present. When the transition state model was fit into the active site, the altered geometry of the peptide bond was seen to allow R_1 to occupy the stacked positon next to Phe-114. This interaction might lower the free-energy barrier of the transition state for substrates in which R_1 is hydrophobic and bulky.

Upon the binding of substrates (Figure 8), Glu-143 becomes completely buried in a nonpolar environment, and we assume that a water molecule must also be buried along with it. At neutral pH, Glu-143 is expected to be charged, and the high

free-energy required to bury this group must be provided by the favorable interactions described above. Under these conditions, Glu-143 becomes a strong base capable of mediating the attack of the water molecule upon the carbonyl carbon of the susceptible peptide bond. Formation of the tetrahedral transition state would be promoted by a lowering of the free energy of Glu-143 on neutralization and the transfer of the charge to the zinc-bound oxygen. Thus, the initial "burial" of Glu-143 may be one of the driving forces of catalysis, as has also been postulated for carboxypeptidase (Lipscomb et al., 1969).

The stereochemistry of β-PPP binding indicates that His-231 is in a favorable position to act as a proton donor, N^{ϵ} of the imidazole being about 4.2 Å from the nitrogen of the scissile bond. Also, in both the native structure and the inhibited complexes, N^{δ} of this histidine hydrogen-bonds to the carboxyl of Asp-226 to form a complex reminiscent of the "charge relay system" observed in the serine proteases (Blow et al., 1969). One might therefore expect the pK of this group to be somewhat higher than that of a "normal" histidine, so that it should be protonated at neutral pH. However, one cannot press too closely the analogy of the "charge relay system", as in thermolysin both the imidazole and aspartate are largely exposed to solvent. The distance of 4.2 Å from His-231 to the peptide nitrogen indicates that a hydrogen bond does *not* exist between these groups. This is as expected, since an amide nitrogen cannot act as an acceptor of a hydrogen bond (Sundberg and Martin, 1974). Model building of the transition state intermediate suggests that as the carbonyl carbon and the nitrogen of the scissile bond both become tetrahedral, the nitrogen moves to within about 3.5 Å of $N^{\epsilon 2}$ of His-231. The geometry is such that a small movement of His-231 can then bring $N^{\epsilon 2}$ 0.5 Å closer to the substrate with the $N^{\epsilon 2}$–proton bond pointed directly at the "lone pair" of the newly tetrahedral nitrogen. Therefore, on stereochemical grounds, direct proton transfer seems likely, although one cannot exclude the possibility of proton transfer via a water molecule.

The role of the zinc in thermolysin appears to be very similar to that of the metal in carboxypeptidase A. The carbonyl oxygen of the scissile peptide bond binds directly to the zinc at a distance of ~2.1 Å to make the zinc coordination approximately tetrahedral. As suggested earlier by Lipscomb and co-workers in discussions of the carboxypeptidase A mechanism, the role of the zinc could be not only to align the peptide in the optimal orientation, but also to aid in polarizing the carbonyl bond (Lipscomb et al., 1968, 1969, 1970; Quiocho and Lipscomb, 1971). The binding of the carbonyl may be seen as a "burial" of the charged zinc ion in a relatively nonpolar medium, as compared with the binding of a water molecule and associated solvent in the absence of substrate. The zinc could then act as a Lewis acid and in concert with Glu-143 cause the formation of the tetrahedral carbon. The zinc may also be particularly important in stabilizing the transition state by an electrostatic interaction with the negatively charged oxygen.

In Figure 9 the mechanism is illustrated as one in which Glu-143, His-231, and the zinc act in a concerted fashion. Alternatively, one may conceive of the mechanism as a stepwise one in which, first, the zinc polarizes the carbonyl, second, the carbonyl carbon becomes tetrahedral by the addition of the water molecule, and finally, His-231 donates a proton to the nitrogen. The present experiments do not resolve this ambiguity.

pH Dependence of Activity. The mechanism for thermolysin proposed above is based primarily on the crystallographic observations but is also consistent with a variety of chemical data which will be summarized in this and the following sections.

The optimum activity for thermolysin and the related neutral proteases has been shown by a number of investigators to occur near pH 7.0 (e.g., see Tsuru et al., 1966; Feder and Schuck, 1970; Pangburn and Walsh, 1975). Pangburn and Walsh (1975) have shown that thermolysin activity depends on two groups, an unprotonated species with a pK_a of 5.9 which they propose to be Glu-143, and a protonated species with a pK_a of 7.5 which they propose to be His-231. The identification of His-231 as a proton donor rests primarily on studies of the inhibition of thermolysin with ethoxyformic anhydride (Blumberg et al., 1973, 1974; Burstein et al., 1974). Although direct proof is lacking, a variety of evidence indicates that this inhibitor acylates His-231 (crystallographic experiments indicated that ethoxyformic anhydride reacts with His-216, 25 Å from the active site, but did not reveal the site responsible for inhibition, partly due to the fact that a maximum of 25% inhibition could be obtained with large crystals suitable for x-ray analysis (S. J. Remington and B. W. Matthews, unpublished results)). The identification of Glu-143 as the unprotonated species with pK_a of 5.9 rests exclusively on its location in the thermolysin active site.

On the basis of the above considerations Pangburn and Walsh have already proposed the general base mechanism, which we favor, or direct nucleophilic attack by Glu-143, as alternative mechanisms for the action of thermolysin.

Inhibitions by Metal Ions. A number of metal ions, including Hg^{2+}, Zn^{2+}, and Ag^{+}, have been shown to inhibit thermolysin (Endo, 1962; Holmquist and Vallee, 1974; Pangburn and Walsh, 1975).

The binding sites for Zn^{2+} and Ag^{2+} were determined crystallographically by calculating difference Fourier projections (Figure 12) for crystals soaked respectively in 0.01 M $ZnCl_2$, pH 7.0, and 0.005 M $AgNO_3$, pH 6.0, in standard mother liquor. In addition, the sites of binding of free mercury and of dimercury acetic acid (DMA) had been determined during the thermolysin structure refinement.

Each of the metals is complexed by His-231, and in the case of Zn^{2+} and Hg^{2+}, apparently also by the phenolic oxygen of Tyr-157. As mentioned previously (Colman et al., 1972), Ag^{+} also binds with variable occupancy at His-88. Zn^{2+} appears to occupy two alternative positions, one similar to that for Hg^{2+}, and the second such that the metal is complexed by His-231 and Asn-112. For DMA, Hg^{2+}, Zn^{2+}, and Ag^{+}, the position occupied by the metal is in each case within no more than 3 Å from the presumed position of the R_1 α-carbon atom of a substrate, almost certainly resulting in competitive binding of inhibitor and substrate. Thus, inhibition on the binding of these metals to His-231 is consistent with this group participating in catalysis, but does not constitute proof.

Inhibition by Dipeptides. A number of studies have shown that thermolysin is inhibited by substrates containing either a free amino terminus at R_1 or a free carboxyl terminus at R_1' (Matsubara, 1966; Morihara, 1967; Morihara and Tsuzuki, 1970; Feder et al., 1974, 1976). While substrates with a free carboxyl group may be hydrolyzed slowly, no cleavage at all is observed in the presence of a free amino terminus.

The difference maps calculated with partial data sets for Ala-Phe and Phe-Phe-amide indicate that these dipeptides bind similarly to β-PPP, i.e., as an extended substrate (cf. Figure 3). Assuming this to be the case, the free amino group of residue R_1 would be in an ideal position to hydrogen-bond to the water molecule bound to Glu-143, or might interact directly

with this acid group. In either event, hydrolysis of the substrate would not occur since Glu-143 and its associated water molecule would not be "buried" as for an extended substrate. A similar interference with Glu-270 in carboxypeptidase A is thought to cause glycyl-L-tyrosine to be very poor substrate for this enzyme (Hartsuck and Lipscomb, 1971). For thermolysin substrates, blocking of the N terminus with an acetyl group does result in hydrolysis (Morihara et al., 1968), suggesting that this group is sufficient to provide the correct environment for Glu-143, and also that the formation of the hydrogen bond from the R_2 carbonyl oxygen to the protein backbone at Trp-115 may be essential for catalysis.

Reduction in the rate of catalysis by a free carboxyl group could be due to a number of causes. For example, the carboxyl group would be close enough to His-231 that there could be an electrostatic interaction between them hindering the action of this group. The presumed interaction between His-231 and the free carboxyl could either be direct, or perhaps, via solvent, as we do see apparent solvent density adjacent to the carboxyl group of β-PPP (Figure 2). Also, the geometry of interaction of the free carboxyl with Arg-203 and Asn-112 is presumably slightly different than for an extended substrate, possibly contributing to nonproductive alignment.

Specificity. The specificity of thermolysin toward both protein and synthetic substrates has been analyzed by a number of investigators (e.g., see Matsubara et al., 1966; Matsubara, 1966; Morihara et al., 1968; Feder and Schuck, 1970).

Morihara and Tsuzuki (1970) carried out a detailed kinetic analysis of a series of synthetic substrates and concluded that at least three residues on the N-terminal side and two residues on the C-terminal side of the scissile bond affect catalysis. This is consistent with the extended substrate binding site shown in Figure 8. TLN is most specific for residue R_1', Leu, Ile, and Phe being most efficiently catalyzed at this position. A substrate containing Tyr at R_1' has about the same K_m as Phe, but k_{cat} is greatly reduced, and, as a result, Tyr is catalyzed at a rate only slightly faster than Ala. Catalysis with Trp at R_1' is negligible.

This apparent size restriction at the site of R_1' binding is consistent with the manner in which Phe binds in the hydrophobic pocket (Figures 3 and 8). The ring fits snugly into this depression with its C^ζ only 3.8 Å from C^α of Gly-189. Substitution of Tyr for this side chain would place the phenolic oxygen within 2.6 Å of Gly-189. Since a rotation about the C^α–C^β bond would also create further bad contacts, it appears that such a substrate must bind 0.5–1.0 Å further away from the pocket, and that this translation probably changes the substrate's interaction with the zinc and also moves the nitrogen of the scissile bond away from His-231. Substitution of Trp at R_1' would result in an even larger shift, if such a substrate could bind at all. Also, inspection of the model indicates that a substrate with a D-amino acid at R_1' must bind in a completely different orientation, if at all, consistent with the observation that such substrates are not hydrolyzed (Morihara and Ebata, 1966; Morihara, 1967; Morihara et al., 1968; Morihara and Tsuzuki, 1970).

In contrast to the limitation on R_1', the kinetic analysis of Morihara et al. (1968) suggests that there is no such size restriction for R_1. Substrates with Phe, Tyr, or Trp at R_1 are catalyzed most efficiently, and, from the results of Morihara and co-workers and the binding studies of dipeptide inhibitors by Feder et al. (1976), it is evident that Phe (the only aromatic side chain tested) binds most strongly of all side chains at this position. This is consistent with the observation that, in the

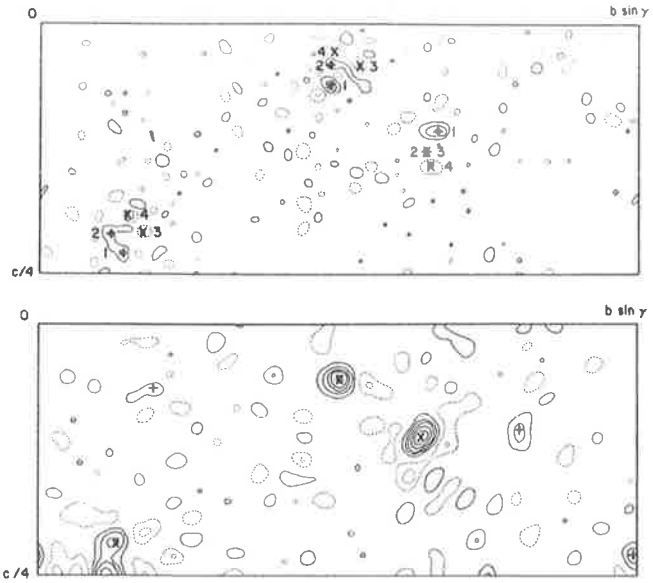


FIGURE 12: Difference Fourier projections showing the binding of zinc and silver ions to thermolysin. Each binding site appears three times within a given projection. (a, top) Zinc thermolysin minus native thermolysin map at 2.3 Å resolution. The refined coordinates for the two alternate zinc binding sites (+) are labeled 1 and 2, and two negative features due to loss of solvent (X) are labeled 3 and 4. (b, bottom) Silver thermolysin minus native thermolysin at 4.0 Å resolution. The refined position for the principal silver binding site at His-231 is marked X, and the secondary site at His-88 is marked +.

binding of β-PPP, the planar aromatic side chain "stacks" next to Phe-114, giving added stability to this complex. Again, the substitution of a D-amino acid such as D-alanine at this position completely abolishes activity (Morihara et al., 1968; Morihara and Tsuzuki, 1970). In this case the D residue could not be bound in a productive mode because of close approaches to C^β of Phe-114.

Morihara and Tsuzuki (1970) have also shown that changes in R_2 result in more complex kinetic effects. Values of K_m for a set of synthetic substrates containing Gly, Ala, and Phe at this position were 12.2, 8.6, and 0.91 mM, respectively. However, k_{cat} for the same substrates was 362.0, 5208.0, and 446.4 s^{-1}, suggesting that Phe binds the most tightly, but Ala binds in a manner most efficient for catalysis. Model building with Ala at position R_2 indicates that this side chain binds close to the imidazole of His-146 (Figure 8). It appears that the more bulky phenylalanyl side chain could be accommodated in a shallow depression near Asp-150, Asn-165, and Tyr-157, although it would not be surprising if the presence of this larger group resulted in a slight change in the overall geometry of the substrate. In this case, it appears that, with minor alterations in the alignment of the substrate, D-alanine could be bound at this site, consistent with the observation that Cbz-D-Ala-Gly-Leu-Ala is hydrolyzed (Morihara and Tsuzuki, 1970).

At R_2', there appears to be a definite preference for a hydrophobic residue. The K_m's of substrates with Leu or Phe at this position are about four times smaller than those of Ala or Gly. As before, however, Ala at R_2' gives the highest value for k_{cat}. The presumed binding of substrates with Leu at R_2', which have the largest value for k_{cat}/K_m, is shown in Figure 8 and was inferred in part from the observed binding of phosphoramidon (R_2' = Trp) (Weaver et al., 1977). As seen in the figure, there are several hydrophobic residues including Leu-202, Phe-130, and Tyr-193 which can contribute to the binding of bulky hydrophobic side chains.

Morihara and Tsuzuki (1970) also present kinetic data for

sites R_3 and R_3'. The largest effects of amino acid substitutions were on k_{cat}, with very little effect on the values of K_m. We can offer little comment on these results except to say that the binding sites for R_3 and R_3' are not well defined. As can be seen in Figure 8, a bulky side chain at R_3 would be expected to interact with the protein, whereas little if any interaction would be expected for a side chain at R_3'.

Recently Vallee and co-workers have shown that thermolysin acylated with amino acid *N*-hydroxysuccinimide esters becomes "superactive" toward certain substrates (Blumberg et al., 1973, 1974; Blumberg and Vallee, 1975). The degree of superactivation depends both on the nature of the acyl group incorporated and on the structure of the substrate. For relatively poor thermolysin substrates, acylation of the enzyme with bulky aromatic derivatives can result in a 400-fold enhancement in activity, whereas for the best substrates tested little if any enhancement was observed. Blumberg and Vallee (1975) show that the site of acylation is probably a tyrosine residue, and the most likely candidate appears to be Tyr-157. The role played by this residue in thermolysin catalysis is not yet clear. In the electron density map for the native enzyme, the density for the side chain of Tyr-157 is somewhat weak, suggesting that some rotation about the C^{α}–C^{β} bond occurs in the crystal.

The phenolic oxygen is in a position to interact either with the zinc, with Glu-166, or possibly to form a hydrogen bond to N^{ϵ} of His-231, although the bond geometry is far from ideal. When a substrate is bound, a movement of about 0.5 Å would bring the phenolic oxygen within hydrogen-bonding distance of the imino nitrogen of residue R_1, as shown in Figure 8. Because of its freedom to rotate away from the active site cleft, Tyr-157 could still be acylated in the presence of the competitive inhibitors β-PPP or Zn^{2+}, as found by Vallee and coworkers. Also, it would be reasonable to speculate that modification of Tyr-157 would alter the "specificity" of thermolysin with respect to residues R_1' and R_2', resulting in "superactive" cleavage of selected substrates; however, a definitive explanation of thermolysin superactivation must await further experimental evidence.

Acknowledgments

We are grateful to Dr. B. L. Vallee for a gift of β-phenylpropionyl-L-phenylalanine and thank Dr. Lynn F. Ten Eyck for help with a number of computational problems and Drs. F. W. Dahlquist and R. S. Roche for several helpful discussions.

References

Bigbee, W. L., and Dahlquist, F. W. (1974), *Biochemistry 13,* 3542.

Blow, D. M., Birktoft, J. J., and Hartley, B. S. (1969), *Nature (London) 221,* 337.

Blumberg, S., Holmquist, B., and Vallee, B. L. (1973), *Biochem. Biophys. Res. Commun. 51,* 987.

Blumberg, S., Holmquist, B., and Vallee, B. L. (1974), *Isr. J. Chem. 12,* 643.

Burstein, Y., Walsh, K. A., and Neurath, H. (1974), *Biochemistry 13,* 205.

Colman, P. M., Jansonius, J. N., and Matthews, B. W. (1972), *J. Mol. Biol. 70,* 701.

Endo, S. (1962), *J. Ferment. Tech. 40,* 346.

Feder, J., Auferheide, N., and Wildi, B. S. (1976), in Enzymes and Proteins from Thermophilic Microorganisms, Zuber, H., Ed., Basel, Switzerland, Birkhäuser Verlag, p 31.

Feder, J., Brougham, L. R., and Wildi, B. S. (1974), *Biochemistry 13,* 1186.

Feder, J., Garrett, L. R., and Wildi, B. S. (1971), *Biochemistry 10,* 4552.

Feder, J., and Schuck, J. M. (1970), *Biochemistry 9,* 2784.

Hartsuck, J. A., and Lipscomb, W. N. (1971), *Enzymes, 3rd Ed. 3,* 1.

Holmquist, B., and Vallee, B. L. (1974), *J. Biol. Chem. 249,* 4601.

Latt, S. A., Holmquist, B., and Vallee, B. L. (1969), *Biochem. Biophys. Res. Commun. 37,* 333.

Lipscomb, W. N., Hartsuck, J. A., Quiocho, F. A., and Reeke, G. N., Jr. (1969), *Proc. Natl. Acad. Sci. U.S.A. 64,* 28.

Lipscomb, W. N., Hartsuck, J. A., Reeke, G. N., Jr., Quiocho, F. A., Bethge, P. H., Ludwig, M. L., Steitz, T. A., Muirhead, H., and Coppola, J. C. (1968), *Brookhaven Symp. Biol. 21,* 24.

Lipscomb, W. N., Reeke, G. N., Jr., Hartsuck, J. A., Quiocho, F. A., and Bethge, P. H. (1970), *Phil. Trans. R. Soc. London, Ser. B 257,* 177.

Matsubara, H. (1966), *Biochem. Biophys. Res. Commun. 24,* 427.

Matsubara, H., Sasaki, R., Singer, A., and Jukes, T. H. (1966), *Arch. Biochem. Biophys. 115,* 324.

Matthews, B. W., Colman, P. M., Jansonius, J. N., Titani, K., Walsh, K. A., and Neurath, H. (1972a), *Nature (London), New Biol. 238,* 41.

Matthews, B. W., Jansonius, J. N., Colman, P. M., Schoenborn, B. P., and Dupourque, D. (1972b), *Nature (London), New Biol. 238,* 37.

Matthews, B. W., Klopfenstein, C. K., and Colman, P. M. (1972c), *J. Sci. Instrum. 5,* 353.

Matthews, B. W., Weaver, L. H., and Kester, W. R. (1974), *J. Biol. Chem. 249,* 8030.

Morihara, K. (1967), *Biochem. Biophys. Res. Commun. 26,* 656.

Morihara, K., and Ebata, M. (1966), *J. Biochem. (Tokyo) 61,* 149.

Morihara, K., and Tsuzuki, H. (1970), *Eur. J. Biochem. 15,* 374.

Morihara, K., Tsuzuki, H., and Oka, T. (1968), *Arch. Biochem. Biophys. 123,* 572.

Pangburn, M. K., and Walsh, K. A. (1975), *Biochemistry 14,* 4050.

Quiocho, F. A., and Lipscomb, W. N. (1971), *Adv. Protein Chem. 25,* 1.

Richards, F. M. (1968), *J. Mol. Biol. 37,* 225.

Sober, H. A., Ed., (1970), Handbook of Biochemistry: Selected Data for Molecular Biology, Cleveland, Ohio, Chemical Rubber Publishing Co., pp J4–5.

Sundberg, R. J., and Martin, R. B. (1974), *Chem. Rev. 74,* 471.

Ten Eyck, L. F., Weaver, L. H., and Matthews, B. W. (1976), *Acta Crystallogr., Sect. A, 32,* 349.

Titani, K., Hermodson, M. A., Ericsson, L. H., Walsh, K. A., and Neurath, H. (1972), *Nature (London), New Biol. 238,* 35.

Tsuru, D., Yamamoto, T., and Fukumoto, J. (1966), *Agric. Biol. Chem. 30,* 651.

Weaver, L. H., Kester, W. R., and Matthews, B. W. (1977), *J. Mol. Biol.* (in press).

J. Mol. Biol. (1977) **114**, 119–132

A Crystallographic Study of the Complex of Phosphoramidon with Thermolysin. A Model for the Presumed Catalytic Transition State and for the Binding of Extended Substrates

L. H. Weaver†, W. R. Kester and B. W. Matthews

Institute of Molecular Biology and Departments of Physics and Chemistry
University of Oregon, Eugene, Ore. 97403, U.S.A.

(Received 4 January 1977)

The structure of the thermolysin inhibitor phosphoramidon (*N*-(α-L-rhamnopyranosyl-oxyhydroxyphosphinyl)-L-leucyl-L-tryptophan) bound to the crystalline enzyme has been determined to a resolution of 2·3 Å by X-ray crystallography. The study shows that the complex of phosphoramidon with thermolysin resembles that of the presumed catalytic transition state inferred from the geometry of binding of dipeptide inhibitors. Also, the study reveals the mode of binding of thermolysin substrates extended on the imino side of the scissile peptide bond.

The crystallographic results are consistent with a variety of other studies on the catalytic activity of thermolysin, and suggest a mechanism of action which is analogous to one of the two alternative mechanisms proposed by Lipscomb and co-workers (1968) for carboxypeptidase A. Key features of the proposed mechanism are that the substrate is initially bound to the enzyme with the carbonyl oxygen of the scissile peptide liganded to the zinc; that Glu143 promotes the nucleophilic attack of a buried water molecule on the carbonyl carbon, forming a tetrahedral intermediate; and that His231 acts as a proton donor. The observed binding of phosphoramidon to thermolysin provides further evidence supporting the mechanism in which Glu143 acts as a general base, promoting the attack of water on the carbonyl carbon, rather than the alternative mechanism in which Glu143 attacks the carbonyl carbon directly, forming an anhydride intermediate.

1. Introduction

Thermolysin is an endopeptidase isolated from the thermophilic bacterium *Bacillus thermoproteolyticus* (Endo, 1962). It is one of a family of extracellular neutral proteases which require zinc for activity, and are not inhibited by diisopropylfluorophosphate, by chloroketones or by sulfhydryl reagents. The amino acid sequence of thermolysin is known (Titani *et al.*, 1972), and the three-dimensional structure has been determined to a resolution of 2·3 Å by X-ray crystallography (Matthews *et al.*, 1974).

In this communication we describe the structure of the complex of the thermolysin inhibitor phosphoramidon with the crystalline enzyme, determined by X-ray crystal-

† Present address: Biozentrum der Universität Basel, CH.-4056 Basel, Klingelbergstrasse 70, Switzerland.

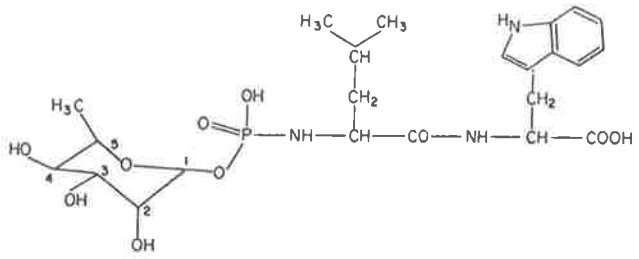


FIG. 1. Structure of phosphoramidon.

lography to a resolution of 2·3 Å. Phosphoramidon (*N*-(α-L-rhamnopyranosyloxyhydroxyphosphinyl)-L-leucyl-L-tryptophan), shown in Figure 1, is a potent inhibitor of thermolysin ($K_I = 2{\cdot}8 \times 10^{-8}$ M), which has little if any inhibitory effect toward trypsin, α-chymotrypsin, papain or pepsin (Suda *et al.*, 1973; Komiyama *et al.*, 1975).

The complex of phosphoramidon with thermolysin is shown to resemble that of the presumed catalytic transition state inferred from the geometry of binding of dipeptide inhibitors to the enzyme (Kester & Matthews, 1977). In addition, the present study also reveals the mode of binding to thermolysin of substrates extended on the imino side of the scissile peptide bond.

The observed binding of phosphoramidon to thermolysin, taken together with studies of the binding of dipeptides, and the results of others, suggests that three groups on the enzyme are of prime importance in catalysis. (1) Glu143, which promotes the nucleophilic attack of a water molecule on the carbonyl carbon of the scissile peptide bond. (2) The zinc ion, which helps align the substrate and polarize the carbonyl group. (3) His231, which acts as a proton donor. The proposed mechanism for thermolysin is, therefore, analogous to one of the two alternative mechanisms proposed by Lipscomb *et al.* (1968) for carboxypeptidase A.

2. Experimental Methods

Thermolysin, obtained from Calbiochem, was crystallized as described previously, and the crystals equilibrated with a solution containing 0·01 M-calcium acetate, 0·01 M-Tris acetate, and 5% (v/v) dimethyl sulfoxide, pH 7·3 (Colman *et al.*, 1972).

Phosphoramidon is readily soluble in this solution, and crystals of thermolysin soaked in the presence of varying concentrations of the inhibitor showed slight surface cracks but were otherwise unaffected. Binding in the crystal was first monitored by calculating (*h*0*l*) difference Fourier projections. In the thermolysin space group, $P6_122$, this projection allows one to determine the 3 co-ordinates specifying the site of binding in 3 dimensions. The location of the inhibitor was indicated by a prominent peak in the thermolysin active site, attributed to the electron-dense phosphinyl group.

Low concentrations of phosphoramidon were found to be sufficient to cause the inhibitor to bind with high occupancy in the crystal. For example a projection difference map, obtained for a crystal of thermolysin soaked in 0·011 mM-phosphoramidon for 3 days, clearly showed that the inhibitor binds at this concentration, and indicated that phosphoramidon retains its high affinity for thermolysin in the crystal phase.

The conditions adopted for the collection of 3-dimensional data were to soak the crystals in 0·1 mM-phosphoramidon for at least 2 days. Data were collected to a resolution of 2·3 Å using a set of 22 conventional precession photographs (Matthews *et al.*, 1972*b*; Colman *et al.*, 1972). Intensity statistics are summarized in Table 1.

TABLE 1

Intensity statistics for phosphoramidon

	Native	Phosphoramidon
Films	24	22
Av. $R_{sym.}$†	0·054	0·065
Av. $R_{sca.}$	0·053	0·041
R_{merge}	0·020	0·024
Av. isom. diff.	—	11·6%
Reflections	14,210	13,730
Cell dimensions		
a,b	94·2 Å	94·1 Å
c	131·4 Å	131·4 Å

† The respective R values are as defined previously (Colman *et al.*, 1972; Matthews *et al.*, 1972*b*).

3. Results

A difference electron density map calculated with amplitudes ($F_{Pam} - F_{Nat}$) (Pam = Phosphoramidon) and phases determined by isomorphous replacement (Matthews *et al.*, 1974) showed a prominent peak about 3 Å from the zinc, obviously indicating the location of the phosphinyl moiety, and other positive features indicating the approximate locations of other portions of the inhibitor. In order to illustrate the general quality of this map Figure 2 shows part of the section $z = -9/180$, which intersects the zinc site and includes the density corresponding to the phosphinyl group. Selected parts of the thermolysin molecule and of the bound inhibitor which are close to this section are also shown. These include the imidazole of His231, and the polypeptide backbone between Phe114 and Trp115, both of which are intersected by the section, and the carboxyl of Glu143 which is about 2 Å "in front" of the section. The directions of the three zinc ligands, His142, His146 and Glu166, are also shown. The rhamnose of phosphoramidon is about 3 Å "behind" $z = -9/180$, with $O_{(2)}$ and $O_{(3)}$ extending to within about 1 Å of the section, and the carboxyl terminus of the inhibitor is about 1 Å in front of the section. The difference electron density at the center of the phosphinyl group, seen in Figure 2, is seven times the height of any other peak away from the active site region and 22 times the root-mean-square background. Also, density corresponding to other portions of the inhibitor extends up to 10σ; however the difference map also contains some negative features which we interpret as arising from the displacement of ordered solvent in the active site. Similar features were observed in difference maps showing the binding of dipeptide inhibitors to thermolysin (Kester & Matthews, 1977).

For this reason the map used for model fitting was one calculated with coefficients ($2F_{Pam} - F_{Nat}$) (Steitz *et al.*, 1969). This map, illustrated in Figure 3, showed the orientation of the inhibitor quite clearly, although the density for the distal portion of the tryptophan ring is rather weak, indicating that this portion of the inhibitor is somewhat disordered in the crystal. The conformation of the inhibitor complexed to the enzyme was determined by first matching a brass model (Cambridge Repetition Engineers) to the electron density map in an optical comparator (Richards, 1968; Colman *et al.*, 1972). Approximate co-ordinates were then obtained by placing markers in the electron density map at positions compatible both with the brass

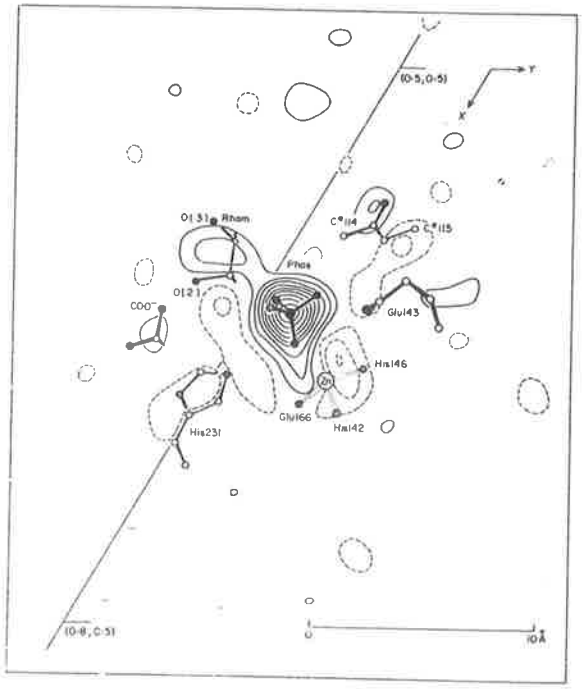
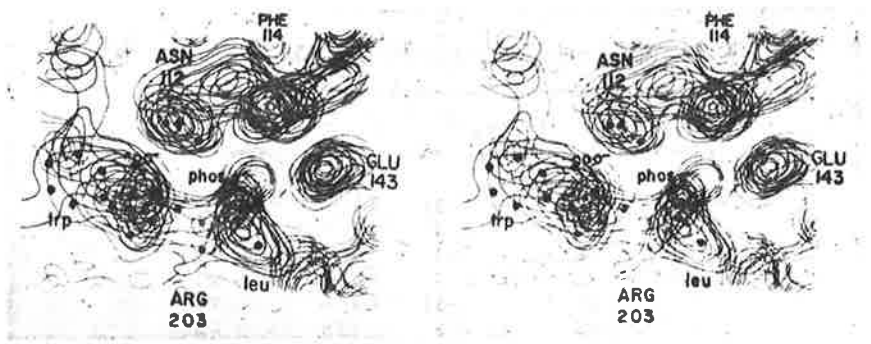


FIG. 2. Portion of the section $z = -9/180$ of the 2·3 Å resolution difference electron density map showing the difference between phosphoramidon-substituted and native thermolysin. Contours are drawn at intervals equal to about $2{\cdot}3\sigma$, positive contours solid and negative contours broken. The positions of portions of the protein structure close to this section are shown. Also indicated is the phosphinyl group of the inhibitor (Phos) which coincides with the section, portion of the rhamnose moiety (Rham) which extends to within 1 Å of the section, and the carboxyl terminus of the inhibitor (COO) which is about 1 Å from the opposite side of the section.

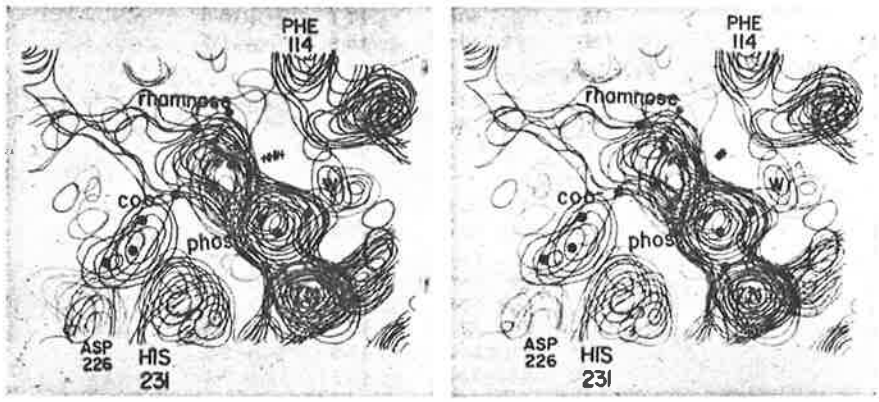
model and with the electron density (the markers can be seen in Fig. 3). These approximate co-ordinates were then idealized by a procedure in which the model is adjusted to conform to the known stereochemical limits on bond lengths and angles, but at the same time to fit as closely as possible to the original guide co-ordinates (Ten Eyck *et al.*, 1976). The idealized co-ordinates for phosphoramidon, given in Table 2, differ from the guide co-ordinates by 0·15 Å. It is difficult to assess the accuracy of these co-ordinates, but the maximum error is unlikely to exceed 0·5 Å, and the average error is estimated to be about 0·2 Å.

The binding of phosphoramidon in the extended thermolysin active site is illustrated in Figure 4. Table 3 illustrates the nomenclature used to identify the different subsites within the extended substrate-binding region of the enzyme, and also summarizes the modes of binding which have been observed for a number of thermolysin inhibitors. The identification of the respective subsites follows that introduced by Schecter & Berger (1967).

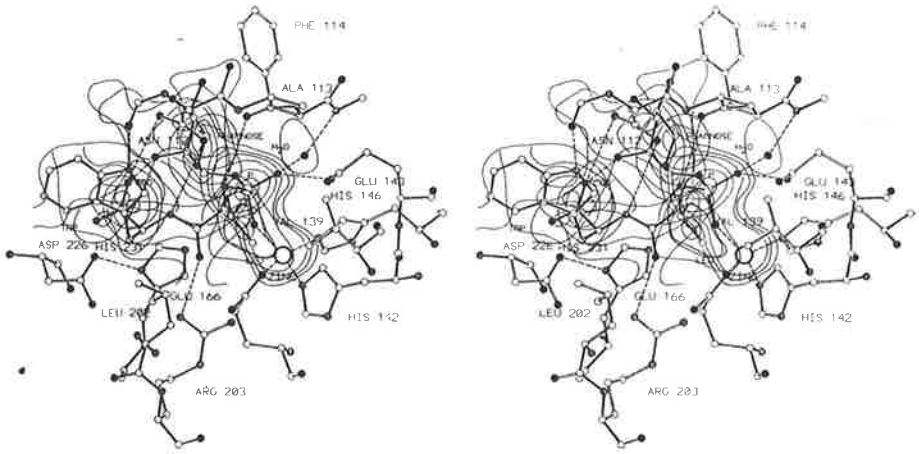
Phosphoramidon binds to thermolysin with one of the phosphite oxygens 2·0 Å from the zinc, displacing a water molecule bound to the native enzyme. The zinc co-ordination is approximately tetrahedral. The hydroxyl oxygen is 2·6 Å from one of the carboxyl oxygens of Glu143, forming an apparent strong hydrogen bond. There is also a water molecule hydrogen-bonded to the phosphite hydroxyl and the peptide nitrogen of Trp115. The distance between the phosphoramide nitrogen and $N^{\varepsilon 2}$ of



(a)



(b)



(c)

FIG. 3. Electron density corresponding to the phosphoramidon–thermolysin complex. (a) Sections $z = -1/80$ to $-8/180$. (b) Sections $z = -8/180$ to $-15/180$. (c) Key to (a) and (b) showing the inhibitor and nearby protein side chains. For simplicity, in (c) the electron density is shown only for every second section, and in addition the contour interval is twice that used in (a) and (b). Also, the 3 highest contours at the phosphorous have been omitted as have 5 contours at the zinc site. Carbon atoms are indicated by open circles, oxygen atoms by solid circles, and nitrogen atoms are drawn hatched. Bonds for the inhibitor and for the protein backbone are drawn solid.

TABLE 2

Idealized co-ordinates for phosphoramidon

	X	*Y*	*Z*
Rhamnose			
C1	51·2	17·0	−8·6
O1	51·4	18·3	−8·0
C2	50·3	16·3	−7·7
O2	50·6	14·9	−7·7
C3	48·8	16·6	−7·9
O3	47·9	15·6	−7·3
C4	48·4	16·5	−9·5
O4	47·1	17·1	−9·7
C5	49·3	17·2	−10·2
O5	50·8	17·1	−10·0
C6	49·2	16·8	−11·7
Phosphite			
P	52·0	18·9	−6·6
O1	53·3	19·1	−6·6
O2	51·2	20·1	−6·6
Leu			
N	51·7	18·0	−5·5
CA	52·8	17·8	−4·4
C	53·6	16·6	−4·6
O	54·8	16·7	−4·8
CB	52·0	17·7	−3·1
CG	53·2	17·9	−1·8
CD1	54·3	18·9	−2·2
CD2	52·5	18·5	−0·5
Trp			
N	53·0	15·5	−4·7
CA	53·7	14·1	−4·9
C	52·9	13·3	−5·8
O1	51·5	13·5	−6·0
O2	53·2	12·1	−6·2
CB	54·2	13·5	−3·7
CG	52·9	13·3	−2·9
CD1	52·1	14·1	−2·4
NE1	51·0	13·6	−1·6
CE2	51·3	12·1	−1·8
CZ2	50·6	11·0	−1·3
CEE2	51·1	9·7	−1·5
CZ3	52·2	9·5	−2·2
CE3	52·9	10·6	−2·9
CD2	52·4	12·0	−2·6
Water	50·4	21·4	−8·0
Leu 202			
CG	56·5	14·9	0·9
CD1	56·8	14·4	−0·6
CD2	56·0	13·6	1·6
Asn 112			
CB	48·4	15·4	−2·1
CG	49·0	15·5	−3·4
ND	48·7	14·5	−4·3
OD	49·7	16·4	−3·8

Co-ordinates are in ångströms relative to the orthogonal co-ordinate frame used previously (Matthews *et al.*, 1974).

TABLE 3

Binding of thermolysin inhibitors determined crystallographically

Inhibitor	Subsite S_2	S_1	S_1'	S_2'
Phosphoramidon		Rham	Leu	Trp
β-Phenylpropionyl-L-phenylalanine		β-Phe	Phe	
L-Phenylalanyl-L-phenylalanyl amide		Phe	Phe	NH_2
L-Alanyl-L-phenylalanine		Ala	Phe	
L-Phenylalanine		(Phe)	(Phe)	
Carbobenzoxy-L-phenylalanine		Phe	Cbz	

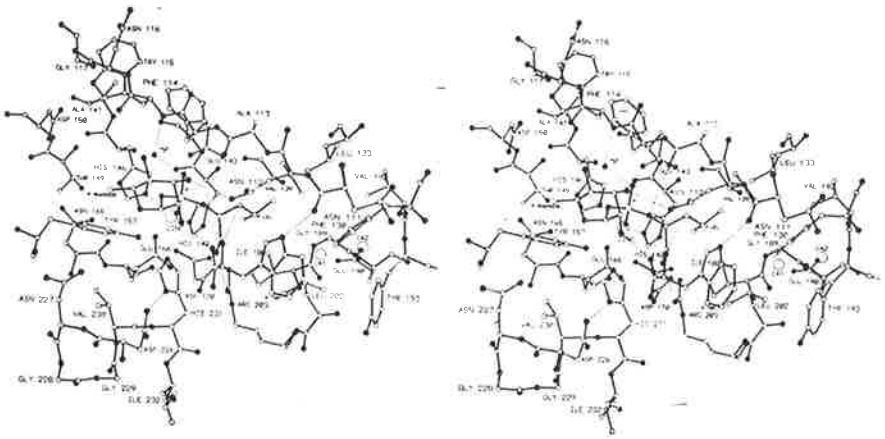


Fig. 4. View of the extended thermolysin active site with phosphoramidon bound. The view is approximately from the left of Fig. 3.

His231 is 4·1 Å, indicating that the amide nitrogen does not accept a hydrogen bond. It does, however, donate a hydrogen bond to the backbone carbonyl oxygen of Ala113. The remainder of the peptide backbone of phosphoramidon forms several hydrogen bonds, shown in Figure 4, to both the protein backbone and the side chains of Asn112 and Arg203. Except for the free carboxyl terminus, all potential hydrogen-bonding groups in the inhibitor backbone are satisfied.

The specificity of thermolysin is directed primarily toward the residue R_1', on the imino side of the scissile bond, but the rate of catalysis is considerably enhanced if R_1 is bulky and hydrophobic (Morihara & Tsuzuki, 1970; Feder & Schuck, 1970). Also, thermolysin inhibitors are more effective if they have, say, Phe or Leu at R_1 rather than a smaller group. In the binding of β-phenyl propionyl-L-phenylalanine (-βPPP) to thermolysin (Kester & Matthews, 1977) the β-phenyl group binds adjacent to the side chain of Phe114, and in the binding of phosphoramidon the rhamnose occupies essentially the same site.

The leucyl side chain of the inhibitor binds in the hydrophobic pocket described previously (Colman *et al.*, 1972). As mentioned above, the specificity of thermolysin is directed toward the residue at this position (R_1') and substrates with Leu (cf. phosphoramidon), Ile or Phe at R_1' are preferentially hydrolyzed.

The tryptophan of phosphoramidon binds in what appears to be a secondary, less specific, hydrophobic site which includes the side chains of Leu202 and Phe130. In addition, there appears to be a hydrogen bond formed from the indole nitrogen to the protein backbone at Asn111. This hydrogen bond could presumably occur for substrates having Trp, His or Gln at R_2', and possibly others.

There is one intramolecular hydrogen bond (2·4 Å), between the carboxyl terminal of the inhibitor and the $C_{(2)}$ hydroxyl of the sugar. This hydrogen bond may help "lock" the inhibitor in a conformation optimal for binding to the enzyme, although it is debatable whether this interaction would be preserved for the inhibitor in solution.

The difference map (Fig. 2) shows clearly that the thermolysin molecule undergoes several localized conformational changes on binding phosphoramidon; however, the movements which do occur are restricted to the active site region and generally involve shifts of at most a few tenths of an ångström. The side chain of Asn112 appears to rotate about its C^{α}–C^{β} and C^{β}–C^{γ} bonds in order to form a pair of hydrogen bonds to the inhibitor backbone (Fig. 4), and also the side chain of Leu202, which forms part of the specificity pocket, rotates about its C^{α}–C^{β} bond by about 120° as it does when dipeptide inhibitors are bound. Also, the protein backbone between Phe114 and Asn116 undergoes a slight movement ("upwards" in Figs 2 and 3) away from the zinc. In concert with this shift, the carboxyl of Glu143 moves slightly in the same direction. There is also an indication in the difference map (Fig. 2) that the zinc moves slightly (probably less than 0·1 Å) away from His142 and His146.

4. Discussion

The observed binding of phosphoramidon to crystalline thermolysin is analogous to the binding of dipeptide inhibitors (Kester & Matthews, 1977). Both studies suggest that the active site cleft of thermolysin contains at least foursubsites, S_2, S_1, S_1' and S_2', which participate in the binding of extended substrates. On the amino-terminal side of the scissile bond, subsite S_2 is near Trp115, and subsite S_1 is adjacent to the phenyl ring of Phe114. On the carboxyl-terminal side of the scissile bond, subsite S_1', the "specificity" site, consists of a pocket lined by the hydrophobic side chains of Val139, Leu133, Phe130, Leu202, Gly189, Val192 and Ile188, whereas subsite S_2' is adjacent to the side chains of Leu202 and Phe130.

The nature of the respective subsites inferred from the crystallographic results is consistent with a number of other studies, including kinetic analyses of the hydrolysis of synthetic substrates, determination of the relative inhibition by different dipeptides and studies of the hydrolysis of other proteins by thermolysin (Kester & Matthews, 1977).

As can be seen in Figure 4, the direction of the inhibitor backbone is antiparallel to that of the protein backbone between Asn111 and Asn116. If one attempts to model-build the extended substrate Gly-Ala-Phe↓Phe-Leu-Gly into the thermolysin active site, using the co-ordinates of β-phenyl propionyl-L-phenylalanine and phosphoramidon to first place -Phe↓Phe-Leu, then the substrate backbone at R_2 appears to form a pair of "antiparallel β-sheet" hydrogen bonds with the protein backbone at Trp115. The presumed mode of binding for such a substrate is shown diagrammatically in Figure 5. The relative arrangement of the substrate and the protein, as in an

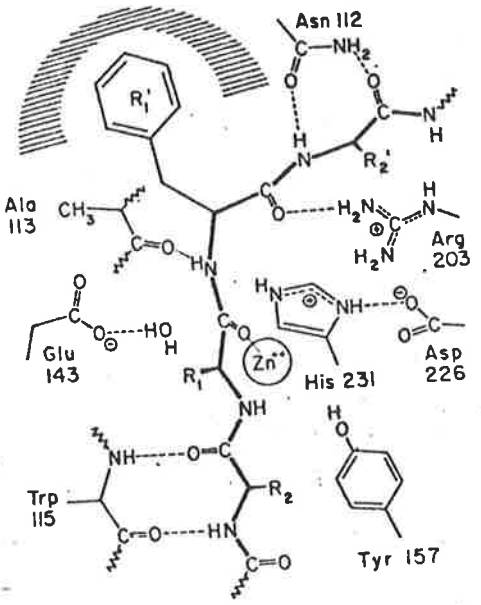


FIG. 5. Schematic diagram illustrating the presumed mode of binding to thermolysin of an extended substrate.

antiparallel β-sheet, is reminiscent of the binding of extended polypeptide substrates to the pancreatic serine proteases (Segal *et al.*, 1971) and to subtilisin (Robertus *et al.*, 1972*a*,*b*), although in the serine proteases it is R_3 of the inhibitor which forms a pair of backbone hydrogen bonds to the protein, whereas in thermolysin it is R_2. In this respect the binding of extended substrates to thermolysin is quite different from that observed for carboxypeptidase A (Hartsuck & Lipscomb, 1971), although in the immediate vicinity of the scissile bond there are a number of similarities.

From the mode of binding observed for thermolysin inhibitors, it can be reasonably inferred that thermolysin substrates bind to the enzyme with the carbonyl oxygen of the scissile peptide displacing a water molecule from the zinc, and becoming the fourth metal ligand. That is, the apparent binding of substrates to thermolysin is similar to the presumed binding of substrates to the pancreatic metallopeptidase carboxypeptidase A (Lipscomb *et al.*, 1968).

Also, as has been pointed out previously (Matthews *et al.*, 1972*a*; Colman *et al.*, 1972), Glu143 in thermolysin is in a position analogous to that of Glu270 in carboxypeptidase A. Glu270 has been implicated by Lipscomb *et al.* (1968) as a nucleophile in catalysis. For carboxypeptidase, Tyr248 undergoes a large conformational movement on binding substrates, and has been proposed as the proton donor. The active sites of thermolysin and carboxypeptidase A will be compared in detail elsewhere, but for the moment it will suffice to say that there is no tyrosine in the thermolysin active site at a position corresponding to that of Tyr248 in carboxypeptidase. Also it may be noted that the largest conformational change observed on binding any inhibitor to thermolysin is a rotation of Leu202 about its C^{α}–C^{β} bond by about 120°. In some cases the side chain of Asn112 rotates somewhat, and also the protein backbone between Phe114 and Asn112 appears to move upwards (in Fig. 4) by perhaps 0·2 Å, but none of these changes are comparable with the shifts of 2 Å or more

observed for the side chains of Arg145, Glu270 and Tyr248 on binding Gly–Tyr to carboxypeptidase A (Lipscomb *et al.*, 1968).

Several lines of evidence suggest that in the case of thermolysin, the proton donor is His231. This residue is located in the active site and interacts with Asp226 in a manner reminiscent of the "charge relay system" of the serine proteases (Blow *et al.*, 1969). It has been shown that several metal ions including Ag^+, Hg^{2+} and Zn^{2+}, which inactivate thermolysin bind to this residue (Colman *et al.*, 1972; Kester & Matthews, 1977), suggesting that it may be involved in catalysis. In addition ethoxyformic anhydride, which reversibly inactivates thermolysin (Blumberg *et al.*, 1973,1974; Burstein *et al.*, 1974) is thought to acylate an essential histidine residue, probably His231. Furthermore, Pangburn & Walsh (1975) have pointed out that the pH dependence of the rate of inhibition by ethoxyformic anhydride correlates with inactivation of a single group with a pK_a of 7·6, and that the same group, presumably His231, acting as a proton donor in catalysis, would account for the apparent pK_a of 7·5 observed in thermolysin catalysis at acid pH.

The pH activity profile for thermolysin-catalyzed hydrolysis of a number of synthetic substrates is bell-shaped with a rather sharp maximum at about pH 7·0 (Feder & Schuck, 1970; Morihara & Tsuzuki, 1970). Morihara & Tsuzuki have shown that it is primarily k_{cat} rather than K_m which changes with pH, suggesting that catalysis depends upon two ionizable groups with pK values close to neutrality, and Pangburn & Walsh (1975) have recently estimated that for thermolysin-catalyzed hydrolysis of furylacryloylglycyl-L-leucinamide the respective pK_a values of these two groups are 5·9 and 7·5. For hydrolysis of casein, the pH profile has a broader maximum, extending from about pH 7·0 to 8·5 (Endo, 1962), but this could in part reflect changes in binding affinity at several of the binding subsites.

Taken together, the above observations suggest that the mechanism for thermolysin-catalyzed peptide hydrolysis includes the following features.

(1) The substrate is initially bound to the enzyme with the carbonyl oxygen of the scissile peptide liganded to the zinc. Formation of the enzyme–substrate complex "buries" the zinc, Glu143 and an associated water molecule.

(2) Glu143 promotes the nucleophilic attack of the buried water molecule on the carbonyl carbon, forming a tetrahedral intermediate, as shown in Figure 6.

(3) His231 donates a proton to the newly tetrahedral nitrogen of the scissile bond. (This step could follow step (2) or could be concerted with it.)

Unfortunately, it has not been possible to prove directly by chemical modification that Glu143 is essential for thermolysin catalysis. Neither is the pK of this group known. Nevertheless, the location of Glu143 in the thermolysin active site (Fig. 4) strongly suggests that it participates in catalysis, and it can reasonably be inferred that this acid group, acting as a nucleophile, accounts for the apparent pK_a of 5·9 seen in the acid limb of the pH dependence of thermolysin (Pangburn & Walsh, 1975). The proposed steps in peptide hydrolysis are illustrated in Figure 7.

Steps (1) and (2) are clearly analogous to the first steps in one of the two alternative mechanisms for peptide hydrolysis by carboxypeptidase, proposed by Lipscomb and co-workers (Lipscomb *et al.*, 1968,1970; Lipscomb, 1974).

In the case of carboxypeptidase A, it has not been possible to distinguish between the mechanism in which Glu270 acts as a general base, promoting the attack of water on the carbonyl carbon, and an alternative mechanism in which Glu270 attacks the

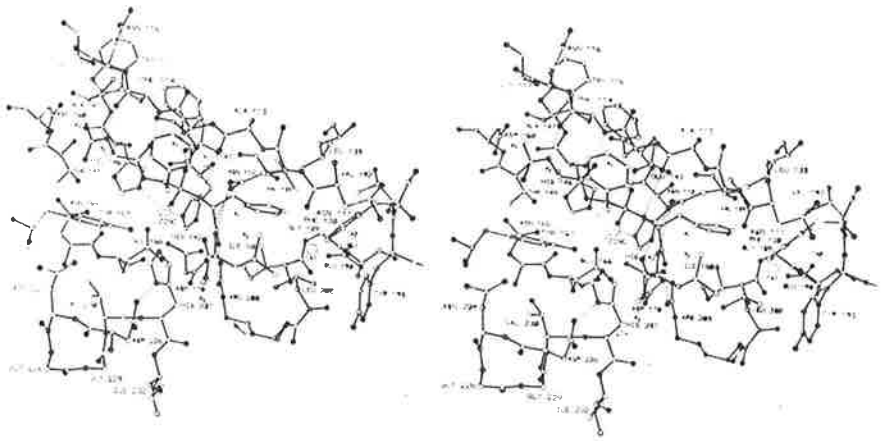


FIG. 6. Stereo view showing the presumed mode of binding of the transition state of the extended substrate shown in Fig. 5.

carbonyl carbon directly, forming an anhydride intermediate. While the mechanism proceeding *via* the anhydride intermediate is the simplest one that accounts for all the data available (Kaiser & Kaiser, 1972), and is also preferred on stereochemical grounds (Lipscomb *et al.*, 1968), the general base mechanism cannot be ruled out.

For thermolysin, there are two reasons for favoring the general base mechanism. In the first place, formation of the anhydride intermediate seems to be stereochemically improbable. Attempts to model-build such an intermediate into the thermolysin active site result in a number of bad contacts due to the fact that the substrate backbone between R_1 and R_1' has to be brought at least 1 Å closer to Glu143 than shown in Figure 6. As an indication of the difference between carboxypeptidase and thermolysin in this respect, Gly–Tyr binds to carboxypeptidase A with its carbonyl carbon about 2·5 Å from Glu270 (Lipscomb *et al.*, 1968), suggesting direct nucleophilic attack, whereas β-phenyl-propionyl phenylalanine binds to thermolysin with a corresponding distance of about 3·9 Å.

The mode of binding observed for phosphoramidon to thermolysin provides

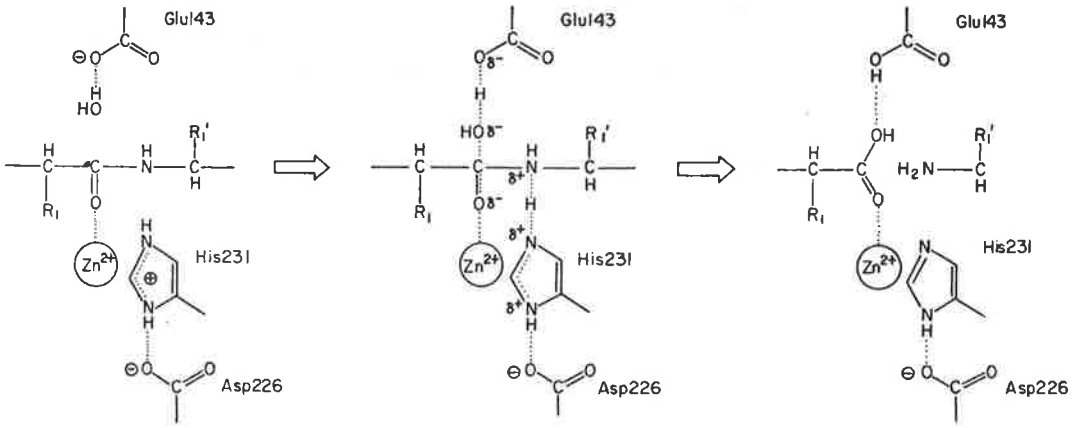


FIG. 7. Proposed mechanism for the hydrolysis of peptides by thermolysin.

additional support for the general-base mechanism. As can be seen in Figure 4, the geometry of binding is just what would be expected for a tetrahedral intermediate resulting from the attack of a water molecule on the carbonyl carbon of a peptide substrate. Apart from the fact that the bond lengths to the tetrahedral phosphorous are about 0·1 Å longer than would be the case for a tetrahedral carbon, phosphoramidon provides a very good transition state analog. As would be expected for such an intermediate, the affinity of binding of phosphoramidon to thermolysin is very high ($K_I = 2{\cdot}8 \times 10^{-8}$ M), compared to the best dipeptide inhibitors such as carbobenzoxy phenylalanine (Feder *et al.*, 1976) or β-phenylalanyl phenylalanine (Holmquist & Vallee, 1974), which have binding constants of about 10^{-3} M. It could be argued that the high affinity for phosphoramidon is due to the presence of the sugar, but inspection of the mode of binding of this group does not suggest that this is likely. The rhamnose is located adjacent to Phe114, as expected for a bulky side chain at R_1, and appears to form one hydrogen bond to the protein (to Asn112), but there is nothing in the observed mode of binding to suggest that the sugar is bound with unusual avidity. Furthermore, Komiyama *et al.* (1975) have recently shown that the phosphoramidon analog *N*-phosphoryl-L-leucyl-L-tryptophan, which lacks the rhamnose moiety, binds to thermolysin even more tightly than does phosphoramidon.

Presumably the tryptophan of phosphoramidon also contributes to the binding, but again the binding is not obviously more specific than would be the case for other residues at this site. Although it was noted by Matsubara *et al.* (1969) that certain peptides of the form -X-Y-Trp- were not hydrolyzed by thermolysin at the X–Y bond, there are now examples where this is observed, for example the cleavage -Asn-Leu↓Leu-Trp- observed in the thermolytic hydrolysis of pepsin fragments (Marciniszyn *et al.*, 1975). Therefore there is no reason to believe that the binding of the -Leu-Trp portion of phosphoramidon does not correspond to that which occurs for a good substrate. Also the identity of the residue R_2' in an extended substrate does not in general have a marked effect on the efficiency of catalysis, unless this residue be proline. The hydrolysis of potential thermolysin substrates including a proline residue at one or other of the respective subsites has been investigated by Ohta & Ogura (1965), Morihara & Tsuzuki (1970) and Matsubara *et al.* (1969). It has been found that catalysis is possible for substrates with Pro at R_1, but not at R_1' or R_2'. Inspection of Figure 6 suggests that a proline could be accommodated at R_1, and that the main result would be a rotation of the plane of the peptide between R_1 and R_2 of 30° or so, which might result in non-optimal alignment of the substrate, but would still permit catalysis. On the other hand it is not surprising that no hydrolysis is observed with a proline at R_1', both because of the modification of the peptide group itself, and also because the proline side chain could not be accommodated without major changes in position of the substrate backbone. Similarly, the dihedral angles observed for residue R_2' in the phosphoramidon–thermolysin complex are such that a proline could not be accommodated at this site unless major changes in the geometry of substrate binding were made. This accounts for the rather striking observation, first pointed out by Matsubara *et al.* (1969), that for polypeptides of the form -X-Y-Pro-, thermolysin never cleaves the X–Y bond.

It seems likely that the structures and the mechanism of action of the neutral metalloendopeptidases from mesophilic bacteria are very similar to that of thermolysin. Not only is the pH profile and the specificity of the neutral protease from *Bacillus subtilis* very similar to that of thermolysin (Morihara *et al.*, 1969; Feder &

Schuck, 1970), but also Pangburn *et al.* (1976) have shown that the amino acid sequence of neutral protease A from *B. subtilis* is homologous with that of thermolysin. Insofar as the amino acid sequence has been determined, all the residues implicated above in the mechanism of thermolysin are unchanged in neutral protease A, and in addition the R_1' specificity pocket appears to be virtually identical in the two enzymes.

On the other hand there does not appear to be an evolutionary relationship between the neutral metalloendopeptidases and carboxypeptidase A. While there are obvious similarities, described above, between the active sites of carboxypeptidase A and thermolysin, and in the respective modes of binding of dipeptide inhibitors, there are also differences, not only in the replacement of tyrosine by histidine as the proton donor, and in the geometry of binding of extended substrates, but also, possibly, in the details of the mechanism of action. Certainly, the degree of correspondence between the active sites of thermolysin and carboxypeptidase A does not approach that observed for subtilisin and the pancreatic serine proteases (Robertus *et al.*, 1972*a*,*b*). Thus it seems that carboxypeptidase A and thermolysin provide an example of convergent evolution, but in contrast to the serine proteases, where evolution has led to virtually *identical* active sites, the active sites of the zinc metallopeptidases do have significant differences.

We are most grateful to Dr H. Neurath for bringing to our attention the isolation of phosphoramidon by Suda, Aoyagi, Takeuchi and Umezawa and for sharing with us a sample of the inhibitor generously provided by Dr T. Aoyagi. We also thank Dr L. F. Ten Eyck for help with the computing.

This study was supported in part by the National Institutes of Health (GM20066; GM15423), by the National Science Foundation (PCM74-18407) and by a Career Development Award (GM70585) from the National Institutes of General Medical Science to one of the authors (B. W. M.).

REFERENCES

Blow, D. M., Birktoft, J. J. & Hartley, B. S. (1969). *Nature (London)*, **221**, 337–340.

Blumberg, S., Holmquist, B. & Vallee, B. L. (1973). *Biochem. Biophys. Res. Commun.* **51**, 987–992.

Blumberg, S., Holmquist, B. & Vallee, B. L. (1974). *Isr. J. Chem.* **12**, 643–649.

Burstein, Y., Walsh, K. A. & Neurath, H. (1974). *Biochemistry*, **13**, 205–210.

Colman, P. M., Jansonius, J. N. & Matthews, B. W. (1972). *J. Mol. Biol.* **70**, 701–724.

Endo, S. (1962). *J. Fermentation Tech.* **40**, 346–353.

Feder, J. & Schuck, J. M. (1970). *Biochemistry*, **9**, 2784–2791.

Feder, J., Aufderheide, N. & Wildi, B. S. (1976). In *Enzymes and Proteins from Thermophilic Microorganisms* (Zuber, H., ed.), Birkhäuser Verlag, Basel. *Experientia Suppl.* **26**, 41–54.

Hartsuck, J. A. & Lipscomb, W. N. (1971). *The Enzymes*, third edit., vol. 3 (Boyer, P. D., ed.), pp. 1–56, Academic Press, New York.

Holmquist, B. & Vallee, B. L. (1974). *J. Biol. Chem.* **249**, 4601–4607.

Kaiser, B. L. & Kaiser, E. T. (1972). *Accounts Chem. Res.* **5**, 219–224.

Kester, W. L. & Matthews, B. W. (1977). *Biochemistry*, in the press.

Komiyama, T., Suda, H., Aayagi, T., Takeuchi, T., Umezawa, H., Fujimoto, K. & Umezawa, S. (1975). *Arch. Biochem. Biophys.* **171**, 727–731.

Lipscomb, W. N. (1974). *Tetrahedron*, **30**, 1725–1732.

Lipscomb, W. N., Hartsuck, J. A., Reeke, G. N., Quiocho, F. A., Bethge, P. H., Ludwig, M. L., Steitz, T. A., Muirhead, H. & Coppola, J. C. (1968). *Brookhaven Symp. Biol.* **21**, 24–90.

Lipscomb, W. N., Reeke, G. N., Hartsuck, J. A., Quiocho, F. A. & Bethge, P. H. (1970). *Phil. Trans. Roy. Soc. ser. B*, **257**, 177–214.

Marciniszyn, J. Jr, Sepulveda, P., Huang, W-Y., Lanier, J. P. & Tang, J. (1975). *J. Biol. Chem.* **250**, 5076–5081.

Matsubara, H., Singer, A. & Sasaki, R. M. (1969). *Biochem. Biophys. Res. Commun.* **34**, 719–724.

Matthews, B. W., Colman, P. M., Jansonius, J. N., Titani, K., Walsh, K. A. & Neurath, H. (1972*a*). *Nature New Biol.* **238**, 41–43.

Matthews, B. W., Klopfenstein, C. E. & Colman, P. M. (1972*b*). *J. Phys. E. Sci. Instrum.* **5**, 353–359.

Matthews, B. W., Weaver, L. H. & Kester, W. R. (1974). *J. Biol. Chem.* **249**, 8030–8044.

Morihara, K. & Tsuzuki, H. (1970). *Eur. J. Biochem.* **15**, 374–380.

Morihara, K., Oka, T. & Tsuzuki, H. (1969). *Arch. Biochem. Biophys.* **132**, 489–501.

Ohta, Y. & Ogura, Y. (1965). *J. Biochem.* **58**, 607–608.

Pangburn, M. K. & Walsh, K. A. (1975). *Biochemistry*, **14**, 4050–4054.

Pangburn, M. K., Levy, P. L., Walsh, K. A. & Neurath, H. (1976). In *Enzymes and Proteins from Thermophilic Microorganisms* (Zuber, H., ed.), Birkhäuser Verlag, Basel. *Experientia Suppl.* **26**, 19–30.

Richards, F. M. (1968). *J. Mol. Biol.* **37**, 225–230.

Robertus, J. D., Alden, R. A., Birktoft, J. J., Kraut, J., Powers, J. C. & Wilcox, P. E. (1972*a*). *Biochemistry*, **11**, 2439–2449.

Robertus, J. D., Kraut, J., Alden, R. A. & Birktoft, J. J. (1972*b*). *Biochemistry*, **11**, 4293–4303.

Schecter, I. & Berger, A. (1967). *Biochem. Biophys. Res. Commun.* **27**, 157–162.

Segal, D. M., Powers, J. C., Cohen, G. C., Davies, D. R. & Wilcox, P. E. (1971). *Biochemistry*, **10**, 3728–3737.

Steitz, T. A., Henderson, R. & Blow, D. M. (1969). *J. Mol. Biol.* **46**, 337–348.

Suda, H., Aoyagi, T., Takeuchi, T. & Umezawa, H. (1973). *J. Antibiotics*, **26**, 621–623.

Ten Eyck, L. F., Weaver, L. H. & Matthews, B. W. (1976). *Acta Crystallogr. sect. A*, **32**, 349–350.

Titani, K., Hermodson, M. A., Ericsson, L. H., Walsh, K. A. & Neurath, H. (1972). *Nature New Biol.* **238**, 35–37.

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 252, No. 21, Issue of November 10, pp. 7704-7710, 1977
Printed in U.S.A.

Comparison of the Structures of Carboxypeptidase A and Thermolysin*

(Received for publication, April 4, 1977)

WILLIAM R. KESTER AND BRIAN W. MATTHEWS

From the Institute of Molecular Biology and Department of Chemistry and Physics, University of Oregon, Eugene, Oregon 97403

Using the known three-dimensional structures of the zinc metallopeptidases carboxypeptidase A and thermolysin, the two enzymes are compared by superimposing one active site upon the other and identifying corresponding structural and functional elements. The overall folding of the respective enzymes is quite different, but the active sites have several features in common, and the modes of binding of dipeptide inhibitors are very similar. The known differences in specificity of the two enzymes are readily explained in terms of the observed differences between the respective substrate binding sites. Also, it is clear that the similarities between carboxypeptidase A and thermolysin are due to convergent rather than divergent evolution. The close structural correspondence in the immediate vicinity of the zinc suggests that these two neutral proteases from mammalian and bacterial sources may have similar mechanisms for peptide hydrolysis. For both thermolysin and carboxypeptidase A, peptide substrates bind to the enzyme with the carbonyl oxygen of the scissile peptide displacing a bound water molecule and becoming the fourth zinc ligand. Evidence is presented that for both enzymes, hydrolysis proceeds by a general base mechanism with Glu 143 of thermolysin, or Glu 270 of carboxypeptidase A, promoting the attack of a water molecule on the carbonyl carbon of the substrate. There is no tyrosine in the thermolysin active site which is analogous to Tyr 248 of carboxypeptidase A. In contrast, the imidazole of His 231 is the presumptive proton donor. The comparison of the active sites of thermolysin and carboxypeptidase A does not obviously confirm that in carboxypeptidase A Tyr 248 acts as a proton donor in peptide hydrolysis, but neither does it rule out this possibility.

Much has been written recently concerning the evolutionary relatedness of several groups of proteins, in particular in cases where the x-ray structures of such proteins have been determined to high resolution. Most of the discussion has centered on proteins which may have diverged from a common ancestor, for example, the nucleotide binding proteins (1). The one clear example of x-ray crystallographic evidence supporting the convergent evolution of two proteins is provided by chymotrypsin and subtilisin (2, 3). As is discussed below, it is now clear that the two proteases thermolysin and carboxypeptidase A also belong in the latter classification.

Prior to the determination of their respective structures, it was thought that thermolysin and carboxypeptidase A might be evolutionarily related. Both are zinc metalloenzymes with similar specificities, hydrolyzing peptide bonds on the amino side of large hydrophobic residues, and both are composed of a single polypeptide chain of molecular weight about 34,000. However, comparison of the respective amino acid sequences (4, 5) indicated that they were not homologous. Furthermore, inspection of the two x-ray structures revealed that the courses of the polypeptide chains were almost completely unrelated (6). On the other hand, an examination of the active sites of both enzymes revealed a number of marked structural similarities, and this similarity has recently been shown to extend to the binding of peptide inhibitors, and, by inference, of peptide substrates (7, 8).

The purpose of the report is to put the comparison of these two enzymes on a more quantitative basis, to explore the similarities of the respective active sites, and to evaluate the relationship of groups which may have equivalent mechanistic roles.

EXPERIMENTAL PROCEDURES

The "best" transformation between the carboxypeptidase A and thermolysin active sites is defined as that which minimizes the root mean square difference between two sets of equivalent atomic positions (*e.g.* see Refs. 9 to 11). The transformation of carboxypeptidase A orthogonal coordinates into the thermolysin orthogonal coordinate system was accomplished by the use of a rotational plus a translational matrix where the rotational matrix was defined by three Eulerian angles (ϕ, θ, ψ) using the conventions of Goldstein (12). The translational part of the transformation was determined first, by superimposing the centers of gravity of the respective coordinate sets. Then, the rotational matrix was refined by a simple stepwise process of incrementing each angular variable until ΔR_{RMS},[1] the sum of the root mean square distances between equivalent points, reached a minimum. Since ϕ, θ, and ψ are not independent in their effects on ΔR_{RMS}, it is necessary to repeat the refinement through several cycles. Convergence upon the true minimum of ΔR_{RMS} is realized when the value of all angles is unchanged at the end of a cycle.

* This work was supported in part by grants from the National Science Foundation (PCM 74-18407) and the National Institutes of Health (GM 20066) and by a Public Health Service Career Development Award to B.W.M. (GM 70585) from the Institutes of General Medical Sciences. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] The abbreviations used are: RMS, root mean square; CPA, carboxypeptidase A; $PhCH_2CH_2COPhe$, β-phenylpropionyl-L-phenylalanine.

Orthogonal coordinates for the active conformation of carboxypeptidase A and the inhibitor glycyl-L-tyrosine are those quoted by Quiocho and Lipscomb (13). Coordinates for those carboxypeptidase A side chains which move when Gly-Tyr is bound were taken from Lipscomb *et al.* (14). Native thermolysin coordinates used here are the raw coordinates listed by Matthews *et al.* (15), while coordinates for the inhibitor β-phenylpropionyl-L-phenylalanine ($PhCH_2CH_2COPhe$) and protein side chains which move upon its binding are as given by Kester and Matthews (7). Since the $PhCH_2CH_2COPhe$ coordinates had previously been idealized by a least square procedure (16), the coordinates for Gly-Tyr were also idealized by the same treatment. The shifts in the Gly-Tyr coordinates introduced by this procedure were 0.04 Å root mean square overall, with no single shift larger than 0.10 Å. No attempt was made to idealize any other carboxypeptidase A coordinate. The estimated uncertainty in the coordinates is about 0.4 Å for carboxypeptidase A (13) and about 0.2 Å for thermolysin (7).

RESULTS

There are several sets of atomic positions which could potentially be used to determine the transformational matrix which gives the "best" superposition of the thermolysin and carboxypeptidase A active sites. One could compare the two enzymes in the native or in the inhibited conformation, or the refinement could be performed using the coordinates of the inhibitors only, rather than those of enzyme side chains. The problem is to determine which choice of atoms gives the most meaningful results. A refinement using only native coordinates is of doubtful validity since the important residues in the carboxypeptidase A active site rearrange when an inhibitor is bound. If, instead, the coordinates of the inhibitors alone were used, one might expect the comparison to more accurately reflect the relative orientations of the side chains in the active sites as viewed by a substrate. The assumption, in this case, is that the respective inhibitors bind in equivalent, although nonproductive, modes. Ideally, one would like to compute the transformation using the coordinates of the scissile peptide bond of a normal substrate, but these are not, at present, accessible. The procedure adopted, which was a compromise, was to use coordinates of the respective bound inhibitors, Gly-Tyr and $PhCH_2CH_2COPhe$, together with selected coordinates of those enzyme side chains which occupy similar relative positions, and may have equivalent roles in hydrolyzing substrates.

In order to indicate those amino acid side chains in the carboxypeptidase A and thermolysin active sites which may be considered "equivalent," Table I lists a series of corresponding distances from the respective zinc ions. As can be seen in the table, the distances for Glu 143 (thermolysin) and Glu 270 (carboxypeptidase A) are quite similar, especially when carboxypeptidase A is in its native conformation. In thermolysin, Glu 143 remains in essentially the native position when $PhCH_2CH_2COPhe$ is bound. However, when Gly-Tyr binds to carboxypeptidase A, Glu 270 moves ~2 Å and interacts with a water molecule which is bound to the free amino group of the dipeptide (14). This interaction is thought to be responsible for the low activity of carboxypeptidase A toward Gly-Tyr, and not to occur during productive binding (17). Therefore, the position assumed for Glu 270 in the refinement was that of the native conformation.

It will be noticed in Table I that the guanidinium group of Arg 203 (thermolysin) is 2 to 3 Å closer to the zinc than that of Arg 145 (carboxypeptidase A). As is well known, carboxypeptidase A cleaves specifically the peptide bond adjacent to a free COOH-terminal carboxyl group, and amidation of the carboxyl prevents catalysis (18). Thermolysin, on the other hand, is essentially inactive toward dipeptides with a free carboxyl group at this point and instead hydrolyzes substrates with the carboxyl amidated. In the binding of Gly-Tyr to carboxypeptidase A, Arg 145 is observed to interact with the carboxyl group and is thought to be responsible for the COOH-terminal specificity. Although Arg 203 of thermolysin occupies approximately the same relative position as Arg 145 in carboxypeptidase A, it is clear from the different specificities described above that these 2 residues cannot play exactly the same role, and for this reason they were not included in the determination of the "best" transformation.

TABLE I

Distances from Zn^{2+} in thermolysin and carboxypeptidase A

Function	Thermolysin			Carboxypeptidase A		
	Atom		Distance	Distance	Atom	
			Å	Å		
Zn^{2+} ligand						
	Glu 166	OE1	1.6	2.2	His 69	ND1
	His 142	NE2	2.1	2.2	His 196	ND1
	His 146	NE2	2.4	2.2	Glu 72	OE1
General base or nucleophile:						
Native conformation	Glu 143	CA	5.9	6.9	Glu 270	CA
		CD	4.8	5.2		CD
		OE1	3.8	4.9		OE2
		OE2	5.0	4.7		OE1
With inhibitor bound		OE1	3.8	5.8		OE1
		OE2	5.0	5.9		OE2
Binding						
Native conformation	Arg 203	CA	11.8	10.7	Arg 143	CA
		NEE1	6.3	8.0		NEE1
		NEE2	5.4	9.9		NEE2
With inhibitor bound		NEE1	6.3	6.7		NEE1
		NEE2	5.4	8.7		NEE2
Proton donor						
	His 231	NE2	4.3	7.0	Tyr 248	OEE
	(Asn 112)	ND	7.7			

Another pair of amino acids which may be related is Tyr 248 of carboxypeptidase A and His 231 of thermolysin. The phenolic oxygen of Tyr 248 is almost 3 Å farther from the zinc than N^{ϵ} of His 231, and actually appears to be closer to the position of Asn 112 in thermolysin. Since one of the questions to be addressed in this comparison is whether these two groups do indeed have the same role in catalysis, their positions were not included in the refinement of the transformation.

In summary, the atoms used to refine the transformation include the following: both zinc ions, the three atoms liganded to the zinc (note that the respective glutamic acids are not in equivalent positions), the carboxyl groups of Glu 143 and Glu 270, and all common atoms of the inhibitors $PhCH_2CH_2COPhe$ and Gly-Tyr. The refined value of the resultant transformation relating the thermolysin and carboxypeptidase A active sites is:

$$X_{TLN} = 0.3095\,X_{CPA} - 0.9178\,Y_{CPA} - 0.2488\,Z_{CPA} - 78.76$$

$$Y_{TLN} = -0.9428\,X_{CPA} - 0.3303\,Y_{CPA} + 0.0455\,Z_{CPA} + 26.26$$

$$Z_{TLN} = -0.1240\,X_{CPA} + 0.2205\,Y_{CPA} - 0.9675\,Z_{CPA} - 18.90$$

The root mean square distance between the 22 equivalent atoms included in the refinement is 1.3 Å and the distances between these and other pairs of potentially related atoms are given in Table II.

TABLE II

Distances between corresponding atoms in superimposed thermolysin and carboxypeptidase active sites

Thermolysin		Distance: Native conformation	Distance: With inhibitor	Carboxypeptidase A	
		A			
	Zn^{2+}	0.6[a]	0.6		Zn^{2+}
Glu 166	OE1	0.7[a]	0.7	His 69	ND1
His 142	NE2	0.4[a]	0.4	His 196	ND1
His 146	NE2	1.2[a]	1.2	Glus 72	OE1
Glu 143	CA	2.2	2.2	Glu 270	CA
	CB	2.4	1.9		CB
	CG	1.6	1.7		CG
	CD	1.5[a]	2.1		CD
	OE1	1.9[a]	3.0		OE2
	OE2	1.3[a]	2.1		OE1
Arg 203	CA	3.8	3.8	Arg 145	CA
	CB	4.8	4.7		CB
	CG	4.7	4.6		CG
	CD	5.5	3.5		CD
	NE	7.1	5.8		NE
	CZ	6.4	4.9		CZ
	NEE1	4.2	3.0		NEE1
	NEE2	8.1	6.5		NEE2
His 231	NE2		4.3	Tyr 248	OEE
Asn 112	OD		2.5	Tyr 248	OEE
β-Phenylpropionyl-			Inhibitor	Glycyl-	
	CA		1.3[a]		CA
	C		0.9[a]		C
	O		0.5[a]		O
-phenylalanine				-tyrosine	
	N		1.4[a]		N
	CA		0.9[a]		CA
	C		0.8[a]		C
	O1		0.9[a]		O1
	O2		0.7[a]		O2
	CB		0.6[a]		CB
	CG		0.5[a]		CG
	CD1		2.5[a]		CD2
	CD2		1.7[a]		CD1
	CE1		2.9[a]		CE2
	CE2		1.3[a]		CE1
	CZ		1.3[a]		CZ

[a] Pairs of atoms included in the refinement of the transformation matrix.

DISCUSSION

Figs. 1 and 2 show the active sites of thermolysin and carboxypeptidase A, viewed from the same orientation, as determined by the transformational refinement. Fig. 3 superimposes the two structures. It is interesting to note that even though the respective polypeptide chains follow completely different courses, the general topology of the two sites are remarkably similar. In the following discussion the two active sites are compared in detail.

The respective zinc ions are transformed to within 0.6 Å of each other, with the tetrahedral coordination around the zinc in carboxypeptidase A being more distorted than in thermolysin. For thermolysin, the carbonyl bond of the scissile peptide points almost directly at the zinc, while for carboxypeptidase A the same group is considerably tilted. In general, the amino acid side chains which serve as zinc ligands are closely aligned only at the liganding atoms themselves, but not further away from the zinc. Although for each enzyme the zinc ligands constitute 2 histidines and a glutamic acid, the respective locations of these residues do not coincide. His 142, His 146, and Glu 166 of thermolysin align, respectively, with His 196, Glu 72, and His 69 of carboxypeptidase A. If the chemically identical zinc ligands of the respective enzymes were forced to coincide, then the active site of one enzyme would be rotated 120° relative to the other, resulting in very bad alignment of the respective substrates and catalytic elements.

There are significant differences between the positions of the presumed nucleophiles Glu 143 (thermolysin) and Glu 270 (carboxypeptidase A) relative to the scissile bond. In Fig. 3*B*, Glu 270 can be seen to be closer to the scissile bond than in Glu 143. It has already been pointed out that at least some of this difference is due to the nonproductive interaction of Glu 270 with the free amino group of the inhibitor; however, in its conformation in native carboxypeptidase A, Glu 270 would be even closer to the scissile bond (Fig. 4).

As expected, in view of the fact that carboxypeptidase A is an exopeptidase and thermolysin an endopeptidase, the positions of Arg 203 (thermolysin) and Arg 145 (carboxypeptidase A), relative to the carboxyl group of the respective inhibitors are quite different. Arg 203 is in an orientation such that the binding of either a free carboxyl group or an extended peptide chain would be allowed. The position of Arg 145, on the other hand, is such that it would sterically interfere with the binding of an extended polypeptide substrate. Arg 203 actually appears to correspond more closely to the position of Asn 144 of carboxypeptidase A than of Arg 145.

Fig. 3*A* illustrates the differences between the two hydrophobic pockets. One portion of the pocket, that formed by Gly 252-Ser 254 of carboxypeptidase A and Ile 188, Gly 199, and Leu 202 of thermolysin, appears to be quite similar, although elsewhere the overlap is not so good, resulting in the carboxypeptidase A hydrophobic pocket being somewhat larger than that of thermolysin. At the back of the carboxypeptidase A pocket, there is an asparagine (Asn 256) to which the tyrosyl hydroxyl of Gly-Tyr may hydrogen-bond, whereas the thermolysin pocket is strictly hydrophobic.

The two postulated proton donors, His 231 of thermolysin and Tyr 248 of carboxypeptidase A (in the presence of Gly-Tyr), are seen to be in substantially different locations with the Tyr 248 hydroxyl of carboxypeptidase A approximately midway between His 231 and Asn 112 of thermolysin.

The above comparison reveals the basis of the important functional differences between thermolysin and carboxypeptidase A. The specificity of carboxypeptidase A as an exopeptidase can be attributed to Arg 145, which binds the free carboxyl of the substrate and, with Tyr 248, combines to form a dead-end pocket, incapable of accommodating an extended polypeptide (13). In contrast, in the case of the endopeptidase thermolysin, Arg 203 is placed on the "side" of the substrate rather than at its terminus so that a substrate of any length can be accommodated. The fact that thermolysin does not efficiently catalyze substrates with side chains larger than phenylalanine on the carboxyl side of the scissile bond, whereas carboxypeptidase A hydrolyzes tyrosine- and tryptophan-containing substrates, is also explained by the deeper hydrophobic pocket of the latter enzyme.

The postulated mechanisms of action of the two enzymes can also be evaluated in terms of the relation between the active sites. For each enzyme, two alternative mechanisms for the hydrolysis of peptide substrates have been considered (7, 14). In the first mechanism, a glutamic acid (Glu 143, thermolysin; Glu 270, carboxypeptidase A) acts as a general

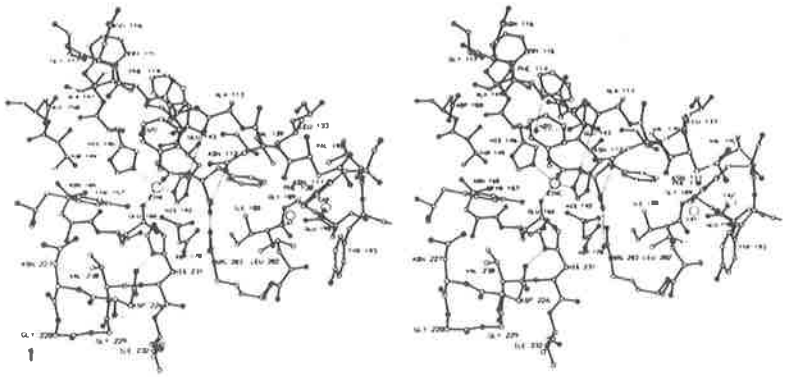


FIG. 1. Stereo view of the active site of thermolysin with the bound inhibitor β-phenylpropionyl-L-phenylalanine. Oxygen atoms are drawn *solid*, nitrogen *hatched*, and carbon as *open circles*. Bonds of the inhibitor and of the protein backbone are drawn *solid*.

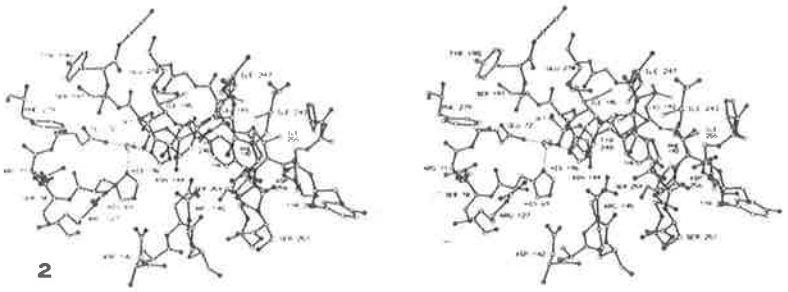
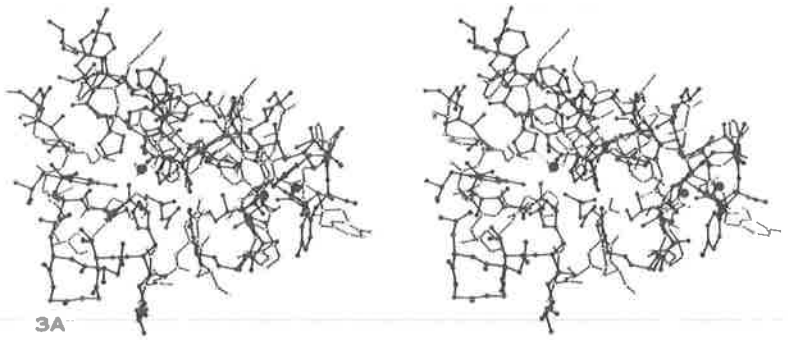


FIG. 2. Active site of carboxypeptidase A, with the bound pseudosubstrate glycyl-L-tyrosine, viewed from the same direction as in Fig. 1. Coordinates are from Lipscomb and coworkers (see text).



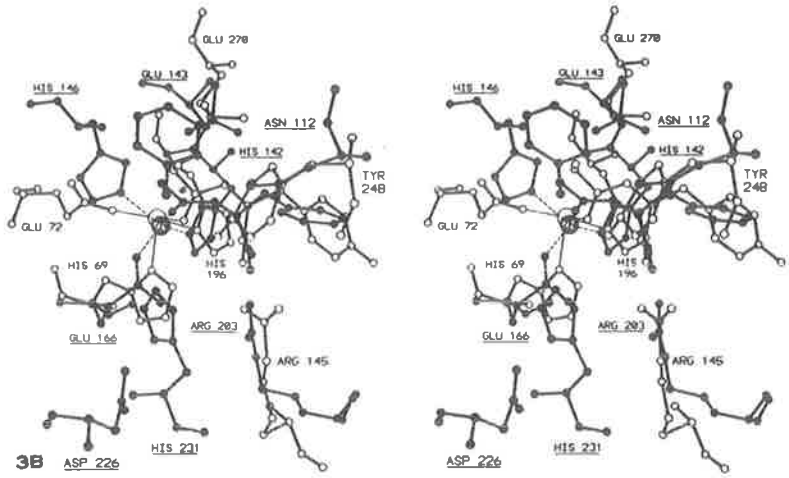


FIG. 3. Superposition of the active site of thermolysin, drawn *solid* (*cf.* Fig. 1), on the active site of carboxypeptidase A, drawn with *open bonds* (*cf.* Fig. 2). *A*, complete active site region; *B*, active site elements in the immediate vicinity of a bound dipeptide. Residue names for thermolysin are *underlined*.

base promoting the attack of a water molecule on the carbonyl carbon of the scissile bond, while in the second mechanism the glutamic acid attacks the carbonyl directly and the reaction proceeds through an anhydride intermediate.

The obvious structural similarity of the active sites of thermolysin and carboxypeptidase A suggests that the mechanisms of action of the respective enzymes are probably closely related. It would certainly be surprising if this were not the case. In the case of thermolysin, there are two reasons for favoring the general base mechanism of peptide hydrolysis. Firstly, model building experiments, based on the observed mode of binding of dipeptides (7), indicate that direct nucleophilic attack of Glu 143 on the carbonyl carbon of the substrate is stereochemically implausible. As can be seen in Fig. 3, the

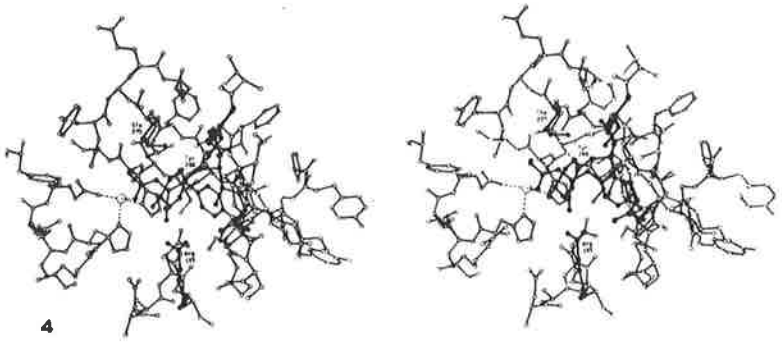


FIG. 4. Active site of carboxypeptidase A showing the conformation of the native structure (*open bonds*) and the structural changes which occur (*solid bonds*) on binding glycyl-L-tyrosine (also drawn *solid*). Coordinates are from Lipscomb and co-workers (see text). The direction of view is the same as in Fig. 2.

critical location at which steric restrictions prevent the close approach of the scissile peptide to Glu 143 is at the α and β carbon atoms of Phe 114. In the case of carboxypeptidase A there is no such stereochemical restriction, and the carbonyl carbon of the pseudosubstrate glycyl-tyrosine binds much closer (2.5 Å *cf.* 3.9 Å in thermolysin) to the glutamate oxygen. Because of this close approach, the sterically favored mechanism for carboxypeptidase appears to be that in which Glu 270 attacks the carbonyl carbon directly; however, as Lipscomb and co-workers point out, their data do not rule out attack by a water molecule (13, 14, 19). Our results suggest that the observed mode of binding of glycyl-tyrosine to carboxypeptidase A is nonproductive, and that in the productive complex a water molecule is trapped between the substrate and Glu 270.

The second reason for favoring the general base mechanism for thermolysin follows from the observed mode of binding of the competitive inhibitor phosphoramidon ($K_I = 2.8 \times 10^{-8}$ M) (20, 21). On the basis of a 2.3-Å resolution electron density map of the thermolysin·phosphoramidon complex, we have proposed that phosphoramidon is a transition state analog (8). The tetrahedral phosphoramide moiety is very similar to the intermediate which would be formed by attack of a water molecule, promoted by Glu 143, on the carbonyl carbon of the scissile bond. If peptide hydrolysis by carboxypeptidase A proceeds by the same mechanism as for thermolysin, then it would be expected that an analog of phosphoramidon appropriate to the specificity requirements of carboxypeptidase A, *i.e.*, an *N*-phosphorylated amino acid with a free carboxyl group, ought to be a potent inhibitor of the enzyme. In fact, J. C. Powers and co-workers[2] (Ref. 22), have recently shown that phosphoryl phenylalanine is a very good inhibitor of carboxypeptidase A ($K_I = 5 \times 10^{-6}$ M). Conversely, we have shown that the "by-product analog" of carboxypeptidase A, L-benzyl succinic acid, which is an excellent inhibitor ($K_I = 4 \times 10^{-7}$ M) (23, 24) also binds with high affinity in the active site of crystalline thermolysin.[3]

In addition to these structural results, independent evidence for the general base mechanism for carboxypeptidase A has been obtained recently by Breslow and Wernick (25, 26) from oxygen-18 exchange studies. Breslow and Wernick suggest that carboxypeptidase A-catalyzed hydrolysis of peptides proceeds by a general base mechanism, but that ester hydrolysis proceeds by a nucleophilic pathway with Glu 270 forming an anhydride intermediate. The latter mechanism is consistent with the recent results of Makinen *et al.* (27) who, using low temperature techniques, were able to trap a covalent anhydride intermediate resulting from the carboxypeptidase A-catalyzed hydrolysis of the ester substrate *O*-(*trans-p*-chlorocinnamoyl)-L-β-phenyllactate. While thermolysin will not catalyze the hydrolysis of the range of esters that carboxypeptidase A does, it will hydrolyze ester analogs of suitable dipeptides, such as hippuryl-L-β-phenyllactyl-L-alanine.[4] To date, no crystallographic data have been obtained for ester binding to either carboxypeptidase A or to thermolysin, so that direct structural support for the mechanism of ester hydrolysis is lacking.

There has been some controversy over the role of Tyr 248 in the mechanism of action of carboxypeptidase A. When glycyl-tyrosine is bound to the crystalline enzyme, Tyr 248 undergoes a large conformational change (Fig. 4) so that its phenolic oxygen interacts with the nitrogen of the scissile peptide, leading Lipscomb and co-workers to postulate that Tyr 248 acts as a proton donor in the hydrolysis of peptides (14, 17, 19). In contrast, Vallee and co-workers (28–31) have shown that in arsanilazotyrosine-248 carboxypeptidase A, Tyr 248 interacts with the zinc in the native enzyme and moves away from the metal when glycyl-tyrosine is bound. Subsequently, Lipscomb (32) has confirmed that in 15 to 25% of the crystalline enzyme used for the structure determination of carboxypeptidase A Tyr 248 approaches close to the zinc. It has also been shown by Vallee and co-workers for carboxypeptidase Aβ and Aγ that the conformation of arsanilazo-Tyr 248 is different in solution from that in the crystal (28–31). Quiocho and co-workers (32, 33) have argued that this difference can be attributed to intermolecular contacts which restrict the motion of the modified tyrosine. In contrast, in crystals of carboxypeptidase Aα, the form used for the structure analysis, Tyr 248 is not restricted and can assume either the "up" or "down" conformation. Thus, the apparently anomalous behavior of carboxypeptidase A in the crystal and in solution can be rationalized, but the role of Tyr 248 in catalysis remains uncertain.

In the case of thermolysin, there are several lines of evidence, summarized elsewhere (7, 8), suggesting that His 231 acts as a proton donor. When dipeptide inhibitors bind to crystalline thermolysin, (Fig. 1) the distance from $N^{\epsilon 2}$ of His 231 to the peptide nitrogen of the inhibitor is about 4.2 Å. Although this distance may appear too large to be consistent with proton transfer it should be noted that a hydrogen bond distance is not to be expected from a protonated imidazole in the Michaelis complex, since an amide nitrogen cannot accept a hydrogen bond (34). Model building suggests that in the transition state the newly tetrahedral nitrogen of the scissile bond moves to within about 3.5 Å of $N^{\epsilon 2}$ of His 231. A small movement (0.5 Å) of His 231 would then bring the imidazole into the expected position for direct proton transfer. However, it has to be noted that these model-building experiments do

[2] C. M. Kam and J. C. Powers, personal communication.

[3] M. Bolognesi and B. W. Matthews, unpublished results.

[4] B. Holmquist and B. L. Vallee, personal communication.

not rule out proton transfer via a water molecule.

If Tyr 248 of carboxypeptidase A and His 231 of thermolysin are assumed to have similar roles in catalysis, *i.e.* to be proton donors, then, in the presence of substrate, they might reasonably be expected to occupy similar positions. The comparison of the respective sites (Fig. 3) suggests that this is not obviously the case. As indicated in Table II, O^{η} of Tyr 248 and $N^{\epsilon 2}$ of His 231 are 4.3 Å apart. Because of the freedom of motion of Tyr 248 in carboxypeptidase A, and also because the —OH group has a positional uncertainty of about 1 Å in the Gly·Tyr complex (17), it is difficult to judge the significance of this discrepancy. In view of the flexibility of Tyr 248, it could reasonably be argued that its —OH group could move somewhat closer to the position occupied by $N^{\epsilon 2}$ of His 231, but the converse seems unlikely, since the active site of thermolysin appears to be too rigid to allow large movement of His 231.

It can be asked whether there are any other residues in the thermolysin active site, in addition to His 231, which might potentially be counterparts of Tyr 248. As can be seen in Fig. 3, Tyr 248 does approximately overlap Asn 112 of thermolysin, the respective side chain oxygens being 2.5 Å apart. Asn 112 participates in the binding of thermolysin substrates, forming a pair of hydrogen bonds to the two peptides following the one to be cleaved (see Figs. 7 and 8 of Ref. 7). The only tyrosine in the thermolysin active site is residue 157 which, in the native enzyme, appears to be close to His 231 and the zinc. When an extended substrate binds, its O^{η} would be ideally placed to accept a hydrogen bond from the peptide nitrogen preceding the scissile bond, although this remains to be confirmed experimentally. Alternatively, Tyr 157 would be free to rotate about its C^{α}—C^{β} bond, a movement which would take its —OH group away from the substrate and into essentially free solution (see Fig. 1). The location of Tyr 157 in the thermolysin active site is "below" the substrate, whereas Tyr 248 of carboxypeptidase A is "above," so that it is unlikely that these 2 residues could fulfill a similar role. Furthermore, Blumberg *et al.* (35) have shown that simple acylating agents such as *N*-acetylimidazole, known to modify tyrosyl residues, affect the activity of thermolysin only slightly.

The fact that there is no tyrosine in thermolysin which is a counterpart to Tyr 248 of carboxypeptidase A, nor, conversely, a histidine in carboxypeptidase A as a counterpart to His 231, indicates that there is no absolute requirement for the involvement of a histidine or a tyrosine in hydrolysis catalyzed by the zinc neutral proteases. In contrast, the close structural correspondence of the zinc and the respective nucleophilic acid groups (Glu 270 in carboxypeptidase A; Glu 143 in thermolysin) suggests that this stereochemical juxtaposition is vital to catalysis. Thus, the comparison of the active sites of carboxypeptidase A and thermolysin suggests that Tyr 248 participates in the binding of peptide substrates to carboxypeptidase A and could facilitate correct alignment of the scissile peptide. The active site comparison does not obviously confirm that Tyr 248 acts as a proton donor, but neither does it exclude this possibility.

In view of the similarities between the active sites of thermolysin and carboxypeptidase A, and the mode of binding peptide substrates, one may ask whether these two enzymes have a common evolutionary precursor, or on the other hand, whether the observed similarities arise from a convergent process, dictated by common functional requirements. The sequences of the two enzymes are not homologous (4, 5) and the courses of the polypeptide backbones have almost no relation to each other. Carboxypeptidase A contains, as its dominant structural feature, an eight-stranded wall of β structure, making up the "core" of the molecule, while thermolysin is a bi-lobed structure with a helix running through the center of the molecule.

In thermolysin, two of the zinc ligands, His 142 and His 146 occur in an internal helix, whereas in carboxypeptidase A the corresponding ligands, His 196 and Glu 72 are located, respectively, in the extended sheet of internal β structure, and in a segment of "random" chain. Furthermore, the identities of the zinc ligands at the three corresponding tetrahedral loci do not correspond in the two enzymes. Also the presumed nucleophile in carboxypeptidase A, Glu 270, is located in the extended β sheet mentioned above, whereas Glu 143, the corresponding residue in thermolysin, is in the internal helix which donates two of the zinc ligands. One requiremnt for two proteins to have evolved from a common precursor is that the ordering of functionally equivalent residues in the amino acid sequences ought to be preserved, since mutational events such as insertions, deletions, and duplications would not affect this sequence. From the comparison of the active sites of thermolysin and carboxypeptidase A presented here, it is possible to identify a number of pairs of functionally equivalent residues. The most obvious of these are the three zinc ligands and the presumed nucleophiles, but at a lower level of confidence one can also include the presumed proton donors, the residues constituting the "specificity pockets," and Arg 203 (thermolysin) and Arg 145 (carboxypeptidase A). The relative alignment of these "equivalent" residues is shown in Fig. 5. Even allowing for some "wrong connections," Fig. 5 provides compelling evidence that the similarities of thermolysin and carboxypeptidase A arose through convergent rather than divergent evolution.

The spatial agreement between corresponding residues in thermolysin and carboxypeptidase A is striking, but it is neither as extensive nor as exact as for the serine proteases α-chymotrypsin and subtilisin. Robertus *et al.* (3) found that 27 pairs of corresponding atoms from the α-chymotrypsin and subtilisin active sites, not including substrate atoms, could be

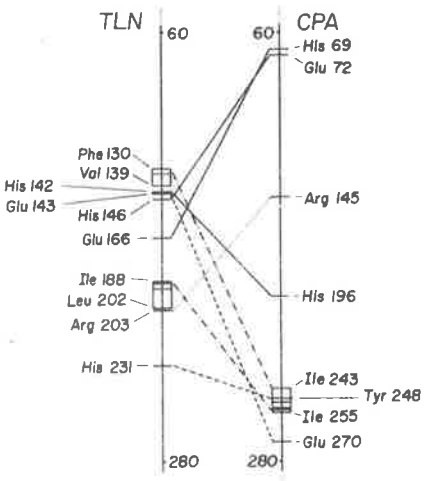


FIG. 5. Sequential alignment of equivalent residues in thermolysin and carboxypeptidase A. The respective zinc ligands are connected by *solid lines*, the presumed nucleophiles and proton donors by *broken lines*, and the respective arginines involved in substrate binding are connected by a *dotted line*. In thermolysin the primary specificity pocket is formed mainly by residues between Phe 130-Val 139 and Ile 188-Leu 202; in carboxypeptidase A the pocket is formed largely by residues between Ile 243-Ile 255. These residues, indicated by the *boxes*, are connected by chain lines.

superimposed within 0.8 Å. In the case of thermolysin and carboxypeptidase A, 22 atoms, including substrate, superimposed within 1.3 Å. One factor contributing to the poorer agreement in the latter case might be the difference in specificity—*i.e.* exo-*versus* endopeptidase. A second potential reason for differences is that Gly-Tyr is a poor substrate for carboxypeptidase, and $PhCH_2CH_2COPhe$ is an inhibitor of thermolysin, so that the coordinates of these bound dipeptides will, at least to some extent, differ from those of good substrates of the respective enzymes. As we have discussed above, the close structural similarity in the immediate location of the zinc and in the alignment of bound dipeptides does suggest that thermolysin and carboxypeptidase A have similar mechanisms for peptide hydrolysis. However, in the case of the zinc neutral proteases, convergent evolution has led to enzymes which have similar spatial arrangement of a few crucial catalytic elements, but do not have the degree of structural correspondence observed in the pancreatic and bacterial serine proteases.

Acknowledgments—We would like to thank Drs. P. Argos, R. M. Garavito, and M. G. Rossman, who have independently compared the active sites of thermolysin and carboxypeptidase A, for a number of helpful comments on our original manuscript.

REFERENCES

1. Rossman, M. G., Liljas, A., Brändén, C.-I. and Banaszak, L. J. (1975) in *The Enzymes* (Boyer, P. D., ed) 3rd Ed, Vol. 11, pp. 62–102, Academic Press, New York
2. Alden, R. A., Wright, C. S. and Kraut, J. (1970) *Phil. Trans. R. Soc. Lond. B. Biol. Sci.* **257**, 119–124
3. Robertus, J. D., Alden, R. A., Birktoft, J. J., Kraut, J., Powers, J. C., and Wilcox, P. E. (1972) *Biochemistry* **11**, 2439–2449
4. Bradshaw, R. A., Ericsson, L. H., Walsh, K. A., and Neurath, H. (1969) *Proc. Natl. Acad. Sci. U. S. A.* **63**, 1389–1394
5. Titani, K., Hermodson, M. A., Ericsson, L. H., Walsh, K. A., and Neurath, H. (1972) *Nature New Biol.* **238**, 35–37
6. Colman, P. M., Jansonius, J. N., and Matthews, B. W. (1972) *J. Mol. Biol.* **70**, 701–724
7. Kester, W. R., and Matthews, B. W. (1977) *Biochemistry* **16**, 2506–2516
8. Weaver, L. H., Kester, W. R., and Matthews, B. W. (1977) *J. Mol Biol.* **114**, 119–132
9. Matthews, B. W., Cohen, G. H., Silverton, E. W., Braxton, H., and Davies, D. R. (1968) *J. Mol. Biol.* **36**, 179–183
10. Kraut, J., Robertus, J. D., Birktoft, J. J., Alden, R. A., Wilcox, P. E., and Powers, J. C. (1971) *Cold Spring Harbor Symp. Quant. Biol.* **36**, 117–123
11. Huber, R., Eppe, O., Steigemann, W., and Formanek, H. (1971) *Eur. J. Biochem.* **19**, 42–50
12. Goldstein, H. (1959) *Classical Mechanics*, pp. 107–109, Addison-Wesley, Reading, Mass.
13. Quiocho, F. A., and Lipscomb, W. N. (1971) *Adv. Protein Chem.* **25**, 1–53
14. Lipscomb, W. N., Hartsuck, J. A., Reeke, G. N., Jr., Quiocho, F. A., Bethge, P. H., Ludwig, M. L., Steitz, T. A., Muirhead, H., and Coppola, J. C. (1968) *Brookhaven Symp. Biol.* **21**, 24–90
15. Matthews, B. W., Weaver, L. H., and Kester, W. R. (1974) *J. Biol. Chem.* **249**, 8030–8044
16. Ten Eyck, L. F., Weaver, L. H., and Matthews, B. W. (1976) *Acta Crystallog.* **A32**, 349–350
17. Hartsuck, J. A., and Lipscomb, W. M. (1971) in *The Enzymes* (Boyer, P. D., ed) 3rd Ed, pp. 1–56, Academic Press, New York
18. Hofmann, K., and Bergmann, M. (1940) *J. Biol. Chem.* **134**, 225–235
19. Lipscomb, W. N. (1974) *Tetrahedron* **30**, 1725–1732
20. Suda, H., Aoyagi, T., Takenchi, T., and Umezawa, H. (1973) *J. Antibiot. (Tokyo)* **26**, 621–623
21. Komiyama, T., Suda, H., Aayagi, T., Takenchi, T., Umezawa, H., Fujimoto, K., and Umezawa, S. (1975) *Arch. Biochem. Biophys.* **171**, 727–731
22. Kam, C.-M., and Powers, J. C. (1977) *Fed. Proc.* **36**, 766
23. Byers, L. D., and Wolfenden, R. (1972) *J. Biol. Chem.* **247**, 606–608
24. Byers, L. D., and Wolfenden, R. (1973) *Biochemistry* **12**, 2070–2078
25. Breslow, R., and Wernick, D. (1976) *J. Am. Chem. Soc.* **98**, 259–261
26. Breslow, R., and Wernick, D. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 1303–1307
27. Makinen, M. W., Yamamura, K., and Kaiser, E. T. (1976) *Proc. Natl. Acad. Sci. U. S. A.* **73**, 3882–3886
28. Johansen, J. T., and Vallee, B. L. (1971) *Proc. Natl. Acad. Sci. U. S. A.* **68**, 2532–2535
29. Johansen, J. T., and Vallee, B. L. (1973) *Proc. Natl. Acad. Sci. U. S. A.* **70**, 2006–2010
30. Johansen, J. T., and Vallee, B. L. (1975) *Biochemistry* **14**, 649–660
31. Spilburg, C. A., Bethune, J. L., and Vallee, B. L. (1974) *Proc. Natl. Acad. Sci. U. S. A.* **71**, 3922–3926
32. Lipscomb, W. N. (1973) *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3797–3801
33. Quiocho, F. A., McMurray, C. H., and Lipscomb, W. N. (1972) *Proc. Natl. Acad. Sci. U. S. A.* **69**, 2850–2854
34. Sundberg, R. J., and Martin, R. B. (1974) *Chem. Rev.* **74**, 471–517
35. Blumberg, S., Holmquist, B., and Vallee, B. L. (1974) *Israel J. Chem.* **12**, 643–649

Biochemistry **1981**, *20*, 6912–6920

Binding of Hydroxamic Acid Inhibitors to Crystalline Thermolysin Suggests a Pentacoordinate Zinc Intermediate in Catalysis†

M. A. Holmes and B. W. Matthews*

ABSTRACT: The mode of binding of a series of hydroxamic acid derivatives, shown by Nishino and Powers [Nishino, N., & Powers, J. C. (1978) *Biochemistry 17*, 2846–2850] to be potent inhibitors of the zinc endopeptidase thermolysin, has been determined by X-ray crystallography. Carbobenzoxy-Gly-Gly-L-Leu-NHOH (K_i = 39 μM) was found to be slowly hydrolyzed in the crystals to yield L-Leu-NHOH, which is itself a good inhibitor (K_i = 190 μM). The structures of the complexes of L-Leu-NHOH and also HONH-benzylmalonyl-L-Ala-Gly-*p*-nitroanilide with crystalline thermolysin were determined at a resolution of 2.3 Å and refined to give crystallographic *R* values of 16.9% and 17.9%, respectively. L-Leu-NHOH binds "backward" in the active site with the hydroxamic acid moiety complexed with the zinc and the leucyl side chain occupying the hydrophobic specificity pocket at $R_{1'}$. The longer benzylmalonyl inhibitor binds in the "normal" mode, occupying sites $R_{1'}$, $R_{2'}$, and $R_{3'}$, with the hydroxamic group complexed with the zinc and the benzyl group in the specificity pocket. For both inhibitors, the hydroxamate group appears to bind in the anionic form, occupying very similar positions in each case. For both inhibitors, the hydroxamate moiety forms a novel bidentate complex with the zinc, with the carbonyl oxygen and the hydroxyl oxygen each approximately 2.0 Å from the metal ion. Thus, in both hydroxamate–thermolysin complexes, the zinc is pentacoordinate, with three ligands from the protein (His-142, His-146, and Glu-166) and two from the hydroxamate group. This is the first case in which a pentacoordinate complex has been observed for thermolysin, and the complex provides the first direct evidence supporting the idea that a pentacoordinate intermediate may participate in the mechanism of catalysis. It is particularly noteworthy that the *N*-hydroxyl group of the inhibitor forms a hydrogen bond to a carboxyl oxygen of Glu-143 and is close to the position which would be expected to be occupied by the water molecule which has been proposed previously to participate in general base catalysis.

During the past several years, it has become increasingly apparent that zinc metalloproteases occur widely and participate in a number of important biological and physiological processes. The role of carboxypeptidases A and B in digestion has long been recognized. Other metallopeptidases include collagenase, thought to be the destructive agent in arthritis (Harris & Krone, 1974), and the angiotensin-converting enzyme, which is important in regulating blood pressure (Peach, 1977). We have previously shown that the active site of thermolysin has striking similarities to that of carboxypeptidase A (Kester & Matthews, 1977a) even though the backbone conformations of the respective enzymes are otherwise very different. Also, known inhibitors of one of the zinc metalloproteases, when modified to take into account the differences in specificity of the respective enzymes, are often found to be equally good inhibitors of other members of the family [e.g., see Kam et al. (1979), Bolognesi & Matthews (1979), and Holmquist & Vallee (1979)]. Taken together, these observations suggest that all these zinc metalloproteases may have similarities in their active sites, and in their respective mechanisms of action (although not necessarily in their overall three-dimensional structures). For this reason, the crystallographic studies on thermolysin which we report here can be seen as part of a broad effort to understand the structure and function of a series of physiologically important enzymes.

Thermolysin is an extracellular, thermostable proteolytic enzyme of molecular weight 34 600 isolated from *Bacillus thermoproteolyticus* (Endo, 1962). The enzyme contains one catalytically essential zinc ion (Latt et al., 1969) and is a neutral endopeptidase specific for peptide bonds on the imino side of hydrophobic residues [e.g., see Matsubara et al. (1966)].

Crystallographic studies of the binding of inhibitors to thermolysin (Weaver et al., 1977; Kester & Matthews, 1977b; Bolognesi & Matthews, 1979) have provided evidence for the mode of binding of extended substrates to the enzyme and have suggested a mechanism of action for peptide hydrolysis.

In this paper, we describe the mode of binding to crystalline thermolysin of hydroxamate inhibitors. These compounds were introduced by Nishino & Powers (1978) and were shown to be potent inhibitors of thermolysin, with inhibition constants as low as 10^{-6} M. The three inhibitors studied crystallographically are carbobenzoxy-Gly-Gly-L-Leu-NHOH (hereafter ZGGL-NHOH)[1] (K_i = 39 μM), HONH-benzylmalonyl-L-Ala-Gly-*p*-nitroanilide (HONH-BzmAGNA; K_i = 0.43 μM), and L-Leu-NHOH (K_i = 190 μM).

It is shown that these inhibitors bind to the enzyme with the hydroxamate group complexed to the zinc in a novel 5-coordinate arrangement. This is the first instance in which a 5-coordinate zinc complex of thermolysin has been observed and provides evidence supporting the idea that a 5-coordinate complex involving the carbonyl oxygen of the substrate and a water molecule may be an intermediate in catalysis.

Materials and Methods

Thermolysin, 3 times recrystallized, was obtained from Calbiochem. The inhibitors ZGGL-NHOH, HONH-

† From the Institute of Molecular Biology and Departments of Chemistry and Physics, University of Oregon, Eugene, Oregon 97403. *Received May 20, 1981.* This research was supported in part by grants from the National Science Foundation (PCM-8014311), the National Institutes of Health (GM20066, GM28138, and 5T32-GM07759), and the M. J. Murdock Charitable Trust.

[1] Abbreviations used: Z, *N*-carbobenzoxy [$C_6H_5CH_2OC(O)$–]; Bzm, benzylmalonyl [–$COCH(CH_2C_6H_5)CO$–]; HONH-BzmAGNA, HONH-benzylmalonyl-L-alanylglycine-*p*-nitroanilide; ZGGL-NHOH, carbobenzoxyglycylglycyl-L-leucyl-NHOH; Tris–acetate, 2-amino-2-(hydroxymethyl)-1,3-propanediol acetate; FAGLA, furylacryloylglycyl-leucinamide.

Reprinted from Biochemistry, 1981, *20*, 6912.
Copyright © 1981 by the American Chemical Society and reprinted by permission of the copyright owner.

Table III: Refined Inhibitor Coordinates[a]

	X	Y	Z	B
		L-Leu-NHOH		
N	52.91	16.55	−4.35	29
CA	52.00	17.68	−4.75	29
C	52.57	18.76	−5.67	30
O	53.63	18.70	−6.37	29
N2	51.84	19.83	−5.65	29
N2O	52.48	20.73	−6.49	28
CB	52.03	18.47	−3.35	27
CG	52.92	17.72	−2.37	27
CD1	54.28	17.67	−3.14	26
CD2	53.03	18.25	−0.95	24
H_2O	50.04	21.24	−8.04	21
H_2O	50.36	17.86	−7.88	31
		HONH-BzmAGNA		
OH1	52.31	20.21	−6.04	21
N1	52.09	19.38	−4.87	23
C1	53.12	18.53	−4.89	24
O1	54.03	18.60	−5.93	23
C2	53.78	16.28	−3.67	25
O2	55.04	16.16	−3.61	27
CA2	53.26	17.58	−3.76	24
CB2	52.73	18.00	−2.60	24
CG2	53.45	18.12	−1.27	23
CD12	52.71	17.71	−0.21	22
CD22	54.70	18.63	−1.13	22
CE12	53.26	17.80	1.00	23
CE22	55.25	18.70	0.07	22
CZ2	54.53	18.28	1.15	23
N3	52.76	15.31	−3.67	27
C3	53.04	13.25	−4.99	29
O3	52.25	14.07	−5.61	30
CA3	53.26	13.79	−3.59	28
CB3	52.90	13.11	−2.48	28
N4	53.64	12.11	−5.50	31
C4	53.90	10.22	−6.79	32
O4	53.02	9.96	−5.80	34
CA4	53.36	11.81	−6.87	32
H_2O	50.04	21.24	−8.04	21
H_2O	50.36	17.86	−7.88	31

[a] The coordinates are in angstroms relative to a right-handed coordinate system, with the origin at the crystallographic origin and X parallel to a^*, Y parallel to b, and Z parallel to c.

group appeared to be hydrogen bonded to the carbonyl oxygen of Ala-113. The phenyl ring of the benzylmalonyl group was in the hydrophobic pocket with the second carbonyl oxygen hydrogen bonded to the side chain of Arg-203. When the wire model of the inhibitor was built, the D stereoisomer was used, as it gave a better fit to the map. The $R_{2'}$ residue of the inhibitor, i.e., alanine, formed two main-chain hydrogen bonds with the side chain of Asn-112. Residue $R_{3'}$, glycine, had no interactions with the enzyme, and the nitroaniline group appeared to assume so many different conformations that there was no density for it in the map.

Examination of $F_{complex} - F_{nat}^{calcd}$ and $2F_{complex} - F_{nat}^{calcd}$ maps for HONH-BzmAGNa yielded the same conclusions as had been reached for the inhibitor L-Leu-NHOH, namely, that the hydroxamic acid group was in its anionic form and that there was no hydrogen bond between the nitrogen of the hydroxamic acid group and Ala-113 (Figure 3B).

Refinement of the Inhibited Complexes. The model coordinates for the inhibitors in the two thermolysin–inhibitor complexes were stereochemically idealized and refined to crystallographic R values of 16.9% and 17.9%, respectively (Table II). Table III gives the two sets of refined inhibitor coordinates, and the complete lists of refined coordinates for the protein–inhibitor complexes have been deposited at the Protein Data Bank (Bernstein et al., 1977). The estimated

Table IV: Thermolysin–Inhibitor Distances

protein	L-Leu-NHOH	distance (Å)
Asn-112 OD1	N	3.3
Ala-113 O	N2[a]	3.5
Glu-143 OE1	OH2[a]	2.7
His-231 NE2	O[a]	3.0
protein	**HONH-BzmAGNA**	**distance (Å)**
Asn-112 OD1	N3	3.0
Asn-112 ND2	O3	3.8
Ala-113 O	N1[a]	3.3
Glu-143 OE1	OH1[a]	2.7
Arg-203 NEE1	O2	2.5
His-231 NE2	O1[a]	2.9

[a] Atoms in the respective hydroxamate groups.

Table V: Zinc–Ligand Distances (Å)

ligand	L-Leu-NHOH	HONH-BzmAGNA
His-142 NE2	2.3	2.2
His-146 NE2	2.0	2.0
Glu-166 OE1	2.0	2.1
inhibitor carbonyl O	2.0	2.1
inhibitor hydroxyl O	2.1	2.3

accuracy of the inhibitor coordinates is about 0.2 Å. Table IV gives selected distances between protein atoms and inhibitor atoms in the two refined complexes, and Figure 4 shows the structures of the complexes in the vicinity of the active site.

In each complex, the distance between the carbonyl oxygen of Ala-113 and the nearby inhibitor nitrogen atom (3.5 and 3.3 Å) is too large for a hydrogen bond to occur (a good hydrogen-bonding distance is 2.8 Å; a distance of 3.0 Å would correspond to a weak hydrogen bond). There is a hydrogen bond between the hydroxyl oxygen of each inhibitor and OE1 of Glu-143 which was not apparent in the difference maps. A weak hydrogen bond exists between OD1 of Asn-112 and a backbone nitrogen of HONH-BzmAGNA, but the other possible hydrogen bond between Asn-112 and the inhibitor does not appear to be made. As seen previously in other inhibitors, a backbone carbonyl oxygen of HONH-BzmAGNA is hydrogen bonded to the guanidinium group of Arg-203. The hydroxamate group of the inhibitors binds to the active-site zinc atom in a bidentate manner (Table V).

For each of the hydroxamate inhibitors, there is no indication that binding occurs to any of the four calcium ions bound to the thermolysin molecule. There is also no indication that the inhibitors remove any of the metal ions.

Discussion

Comparison of Kinetic and Crystallographic Results. Kinetic data on the inhibition of thermolysin by various peptide–hydroxamic acid derivatives are given in Table VI (Nishino & Powers, 1978, 1979). The abilities of these compounds to inhibit the enzyme can be compared in order to ascertain which elements of the compounds contribute to binding, and the inferences from the kinetic data agree well with those obtained from the crystallographic analyses of the two inhibitors ZGGL-NHOH and HONH-BzmAGNA.

For compounds related to ZGGL-NHOH, it can be seen that modification of the hydroxamic acid group either at the hydroxyl oxygen or at the nitrogen, or replacement of the hydroxamic acid group with an amino group, results in a significant loss of inhibitory ability. This suggests that the hydroxamic acid group is an essential element in this series

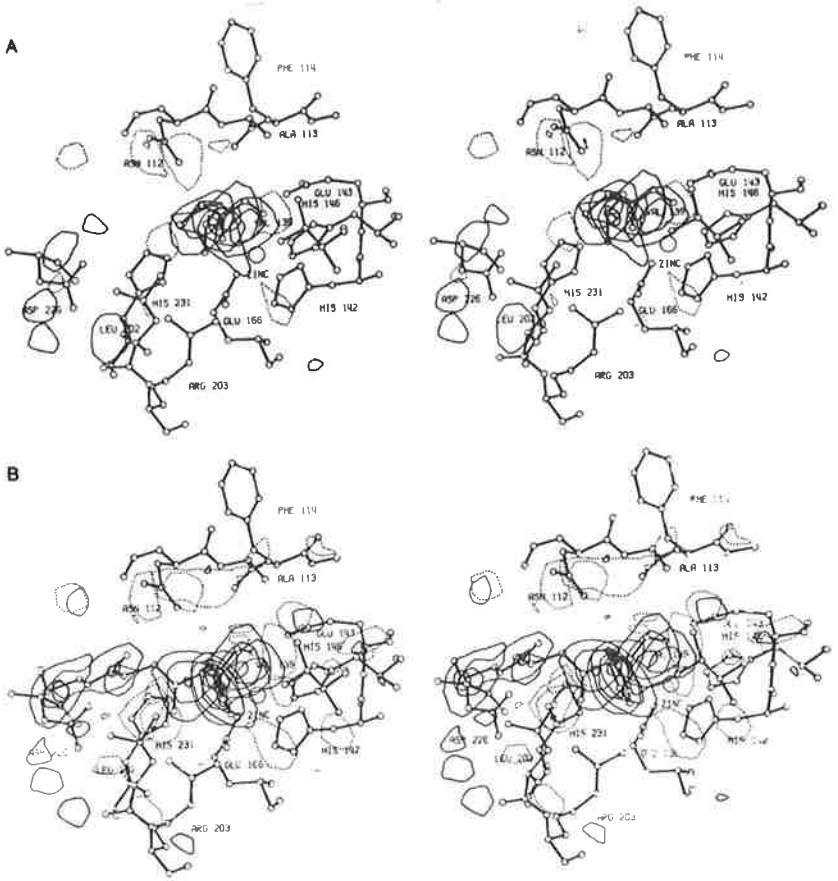


FIGURE 3: Stereo drawings showing the refined inhibitor complexes superimposed on a difference electron density map with coefficients of the form $F_{complex} - F_{nat}^{calcd}$ (see text). (A) L-Leu-NHOH–inhibitor complex; (B) HONH-BzmAGNA–inhibitor complex.

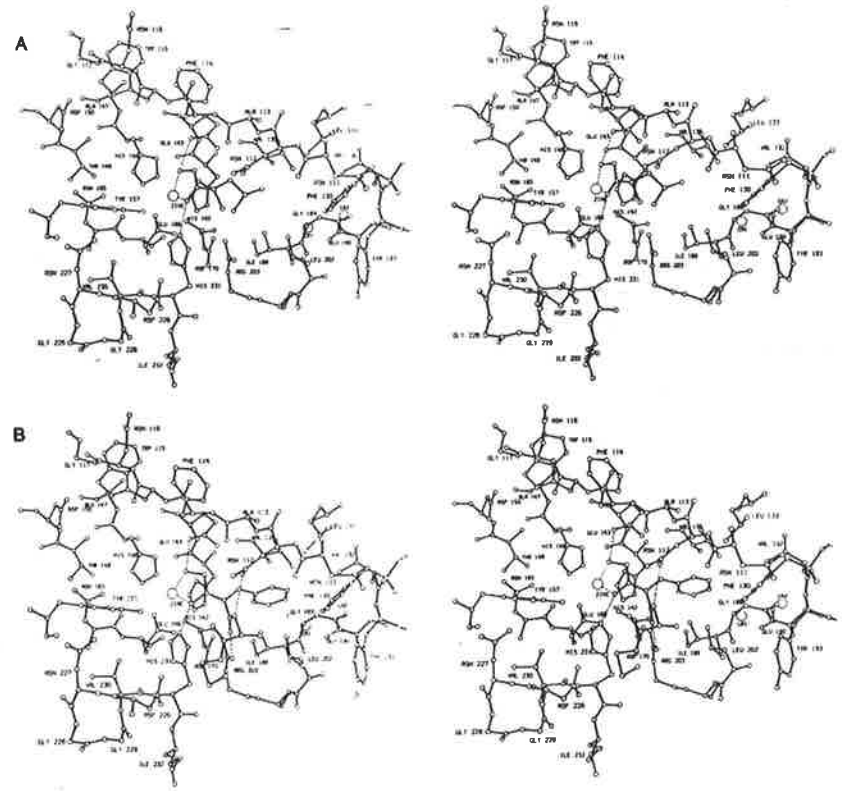


FIGURE 4: Stereoviews of the hydroxamate inhibitors bound to the extended thermolysin active site. The direction of view is from the left of that shown in Figure 3. (A) Thermolysin–L-Leu-NHOH complex; (B) thermolysin–HONH-BzmAGNA complex.

of inhibitors. Similarly, comparison of the K_i's of ZGG-NHOH and ZGL-NHOH shows that the presence of a hydrophobic side chain is important for binding. An examination of the K_i's of ZGL-NHOH and ZGGL-NHOH and their optical antipodes suggests that the inhibitors may bind initially with a normal peptide orientation, since, with a reverse orientation, the D optical isomers would be expected to be as good or better inhibitors. It should be noted that hydrolysis of the

Table VI: Kinetic Data for Inhibitors[a]

	K_i (μM)
L-Leu-NHOH	190
Z-Gly-L-Leu-NH_2	no inhibition[b]
Z-Gly-L-Leu-NHOH	13
Z-Gly-D-Leu-NHOH	59
Z-Gly-L-Leu-N(CH_3)OH	2230
Z-Gly-L-Leu-$NHOCH_3$	no inhibition
Z-Gly–Gly-NHOH	940
Z-Gly-Gly-L-Leu-NHOH	39
Z-Gly-Gly-D-Leu-NHOH	250
HONH-Bzm-OEt	20[c]
HONH-Bzm-L-Ala-Gly-NA	0.43[c,d]
HONH-Bzm-L-Ala-Gly-NH_2	0.66[c]
HONH-Mal-L-Ala-Gly-NH_2	1100
OH-Bzm-L-Ala-Gly-NH_2	420[c]

[a] Taken from Nishino & Powers (1978, 1979). [b] K_m value = 21 000 μM. [c] The benzylmalonyl residue is racemic. [d] I_{50} value at 1 mM substrate concentration.

L-Leu-NHOH derivatives by the enzyme is sufficiently slow that it does not affect the results of the kinetic experiments (Nishino & Powers, 1978).

Inspection of Figure 4A shows why methylation of the hydroxamate nitrogen decreases the affinity of the inhibitor by a factor of 170 (Table VI). As discussed previously, this nitrogen is 3.5 Å from the carbonyl oxygen of Ala-113 so that methylation results in steric interference between the bound inhibitor and the enzyme. Methylation of the hydroxyl oxygen of the hydroxamate moiety no doubt prevents binding because the methyl group would sterically interfere with Glu-143 (Figure 4A).

When compounds related to HONH-BzmAGNA are considered, a comparison of the K_i values of HO-BzmAG-NH_2 and HONH-BzmAG-NH_2 indicates that the hydroxamic acid group again makes a major contribution to the binding. Replacement of the benzylmalonyl group with an unsubstituted malonyl moiety results in a substantial loss of inhibitory activity, suggesting that the presence of a hydrophobic side chain adjacent to the hydroxamic acid is important for binding. The existence of enzyme–inhibitor interactions beyond the $R_{1'}$ residue is indicated by the decrease in K_i upon addition of a dipeptide to the compound HONH-Bzm-OEt. All these conclusions are consistent with the mode of binding of HONH-BzmAGNA to thermolysin which has been deduced from the crystallographic data (Figure 4B).

Our results confirm the main conclusions reached by Nishino & Powers (1978, 1979). In particular, the data given here provide the first definitive evidence that hydroxamate inhibitors of thermolysin bind with the hydroxamic acid moiety complexed with the zinc. Also the crystallographic data show that the mode of hydroxamate binding is not monodentate (Nishino & Powers, 1978), but bidentate (Nishino & Powers, 1979), with the *N*-hydroxycarboxamido group in the cis configuration. It is reasonable to infer that the geometry of hydroxamate binding to other zinc proteases is very similar to that observed for thermolysin.

Hydrolysis of ZGGL-NHOH. Our results show clearly that ZGGL-NHOH is slowly hydrolyzed by crystalline thermolysin. As discussed above, ZGGL-NHOH appears to bind predominantly in a backward mode in the active-site cleft, with the hydroxamate moiety complexed with the zinc. However, the inhibitor must sometimes bind in a mode which permits hydrolysis of the Gly-Leu bond, i.e., with the glycine at R_1 and the leucine at $R_{1'}$ with its side chain in the specificity pocket. Eventually, all of the ZGGL-NHOH is hydrolyzed, leaving the product L-Leu-NHOH, a poorer but still good inhibitor of thermolysin (K_i = 190 μM), to occupy the active site, as is observed in the crystallographic study.

Value of Refinement. In common with other laboratories, we found the crystallographic refinement of the native thermolysin structure and the thermolysin–inhibitor complexes to be particularly helpful in two ways. First, once the native thermolysin structure had been refined, reliable calculated native structure factors, with the contribution of water molecules in the enzyme's active site removed, were available. This made it possible to calculate difference electron density maps which were not confused by the displacement of water molecules. In the case of the inhibitor ZGGL-NHOH, the difference map indicated that the molecules actually binding to thermolysin consisted entirely of L-Leu-NHOH, rather than a mixture of L-Leu-NHOH and Gly-L-Leu-NHOH, as had been indicated by the initial map calculated with the isomorphous replacement phases.

Also, in the $2F_{complex} - F_{nat}^{calcd}$ difference maps, there was no density indicating an interaction between the carbonyl oxygen of Ala-113 and the hydroxamic acid group. The apparent absence of this hydrogen bond helped confirm the idea that it is the hydroxamate form of the inhibitors which binds to thermolysin.

The second benefit of the refinement process was that it provided reliable atomic coordinates from which interatomic distances could be calculated. From these distances, it was quite clear which protein atoms were interacting with which inhibitor atoms. In the process of refinement, some pairs of atoms which had been thought to interact moved apart, while some unsuspected interactions became apparent. For example, a tentative hydrogen bond between a carbonyl oxygen atom of HONH-BzmAGNA and the side-chain nitrogen atom of Asn-112 is not supported by the refined coordinates (Table IV). Also, the lack of a hydrogen bond between the hydroxamate nitrogen and Ala-113 is confirmed in the refined complex structures (Table IV). On the other hand, the interaction between the hydroxamate hydroxyl oxygen and the Glu-143 side chain which is seen in both refined structures was not apparent until refinement took place.

Mechanism of Catalysis. A catalytic mechanism for thermolysin has been proposed by Kester & Matthews (1977b) [see also Pangburn & Walsh (1975)]. In the first step, the substrate is bound to the enzyme in the manner shown in Figure 1. In the next step, Glu-143 promotes the nucleophilic attack of a water molecule on the carbonyl carbon of the scissile peptide bond, which has been somewhat polarized by the zinc ion. Concurrently, His-231 donates a proton to the nitrogen of the scissile bond. This intermediate is tetrahedral at both the carbon and nitrogen atoms and is stabilized by hydrogen bonds with Glu-143 and His-231 and an electrostatic interaction with the zinc ion. In the final step, the carbon–nitrogen bond of the intermediate breaks to yield the two product peptides. The burial of the charged groups Glu-143 and Zn^{2+} in a nonpolar environment upon binding of substrate to the enzyme may be one of the driving forces of catalysis. It is presumed that the favorable interactions between enzyme and substrate discussed above provide the free energy required to bury the charged groups. Once buried, Glu-143 is a strong base capable of mediating the attack of a water molecule upon the carbonyl carbon, and the zinc ion acts as a Lewis acid to polarize the carbonyl bond. The subsequent neutralization of Glu-143 and Zn^{2+} provides the free energy needed to form the tetrahedral transition state.

When the catalytic mechanism is considered in more detail, the enzyme–substrate complex could be formed in two ways.

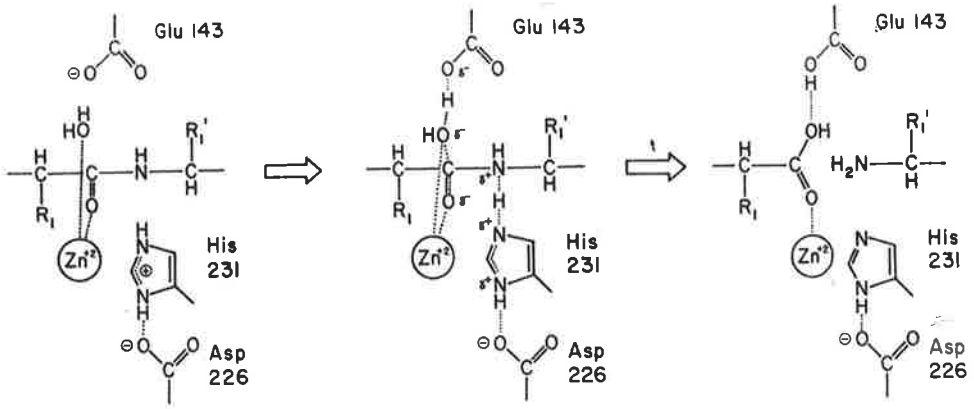


FIGURE 5: Schematic illustration of the proposed mechanism of catalysis for thermolysin.

In the first mechanism, the water molecule would be displaced from the zinc ion, leaving an empty ligand site, which the substrate carbonyl oxygen would then fill. The second and more likely mechanism has the zinc ion becoming temporarily 5-coordinate by interacting with both the water molecule and the incoming substrate. Model binding of this intermediate shows that this change in coordination geometry would cause the water molecule to move in a direct line toward Glu-143. The appealing aspect of this second mechanism is that it avoids postulating the existence of an empty ligand site at the zinc ion. Also, the displaced water molecule is translocated to the vicinity of Glu-143 much more efficiently than in the first mechanism.

This 5-coordinate zinc mechanism, illustrated in Figure 5, combines elements of a general base-catalyzed reaction with a direct catalytic role for zinc as suggested by studies of model compounds [e.g., see Buckingham et al. (1974), Woolley (1975), Fife & Squillacote (1977, 1978), Wells & Bruice (1977), and Groves & Dias (1979)]. In the case of thermolysin, the native enzyme has a water molecule, or possibly a hydroxide ion, coordinated tetrahedrally to the zinc. As the substrate enters the active-site cleft, and begins to form a favorable complex with the enzyme, the zinc-bound water molecule is translocated toward Glu-143, and, at the same time, the carbonyl oxygen of the scissile peptide bond begins to coordinate with the metal. As the water molecule approaches Glu-143, the negative charge enhances its nucleophilicity. Thus, the zinc ion not only polarizes the substrate carbonyl bond but also helps determine the correct alignment of the carbonyl bond and the attacking nucleophile.

The 5-coordinate zinc mechanism for the binding of substrate to thermolysin is supported by experiments in which the active-site zinc ion has been replaced by other metal ions. Zinc and cobalt give the highest enzyme activities toward the chromophoric substrate FAGLA, 100% and 200%, respectively, while Mn^{2+}-substituted thermolysin has 10% activity (Holmquist & Vallee, 1974; Bigbee & Dahlquist, 1974). The substitution of Cu^{2+}, Cd^{2+}, or Hg^{2+} into the thermolysin active site results in inactive forms of the enzyme (Holmquist & Vallee, 1974). Of these metals, zinc and cobalt are the ones most often observed with trigonal bipyramidal coordination (Cotton & Wilkinson, 1972). Zinc and cobalt are also able to change their coordination state relatively easily, compared to other metals (Lindskog, 1970; Vallee & Williams, 1968; Prince & Woolley, 1972). Furthermore, the enzymatic activity of the various metal-substituted thermolysins does *not* correlate with the "hardness" of the metals as Lewis acids. Zn^{2+}, Co^{2+}, and Cu^{2+} are all classified as intermediate Lewis acids (Pearson, 1966), yet the respective metal-substituted thermolysins have activities ranging from maximum to zero. Mn^{2+} is a hard Lewis acid but yields an active enzyme. Also there is no correlation between enzymatic activity and the size of the metal ion (Cu^{2+} and Zn^{2+}, for example, are practically identical in size). Clearly, the role of the metal is more than that of a simple Lewis acid.

The possibility of a pentacoordinate intermediate has been suggested for other enzymes which have a catalytically essential zinc. Woolley (1975) suggested on the basis of studies of model catalysts for carbonic anhydrase that a zinc-bound water molecule could ionize to provide an effective nucleophile in the form of a metal-bound hydroxide. Following their crystallographic studies of the structure of carbonic anhydrase, Kannan et al. (1977) incorporated zinc hydroxide in a mechanism in which the carbon dioxide substrate bound weakly at a fifth zinc coordination site. Also Dworschack & Plapp (1977) and Boiwe & Brändén (1977) have proposed that a pentacoordinate transition state complex with both a hydroxyl ion and a neutral alcohol bound to the zinc may occur in alcohol dehydrogenase. On the basis of structural comparisons of the zinc environments of carbonic anhydrase, alcohol dehydrogenase, carboxypeptidase A, and thermolysin, Rossmann & Argos (1978) have postulated that a 5-coordinate intermediate might occur for each of these zinc-containing enzymes.

In the case of thermolysin, there has been no direct evidence to date for anything other than 4-fold coordination at the active-site zinc. Now, for the first time, the hydroxamate inhibitors provide a clear example of 5-fold coordination. In both hydroxamate complexes, the zinc is 5-coordinate, the ligands being three protein side chains which contribute two nitrogens and one oxygen, and the two oxygens of the hydroxamic acid group of the inhibitor. The respective ligand distances, all of which are close to 2.0 Å, are given in Table V. The coordination is distorted trigonal bipyramidal, with Glu-166 and the hydroxyl oxygen of the inhibitor as the axial ligands and His-142, His-146, and the carbonyl oxygen of the inhibitor as the equatorial ligands (see Figure 6). This structure resembles quite closely the proposed 5-coordinate zinc reaction intermediate, in which the two axial zinc ligands are Glu-166 and the catalytically important water molecule, and one of the other three ligands is the carbonyl oxygen of the substrate. In particular, it is striking to see for both inhibitors that the hydroxamate hydroxyl oxygen forms a strong hydrogen bond (distance 2.7 Å) to the carboxyl oxygen OE1 of Glu-143, just as the water molecule would be expected to do.

In order to show the 5-coordinate geometry relative to the position occupied by a peptide substrate of thermolysin, in Figure 7 we have superimposed the hydroxamate group, as observed for L-Leu-NHOH, on the inhibitor β-phenyl-

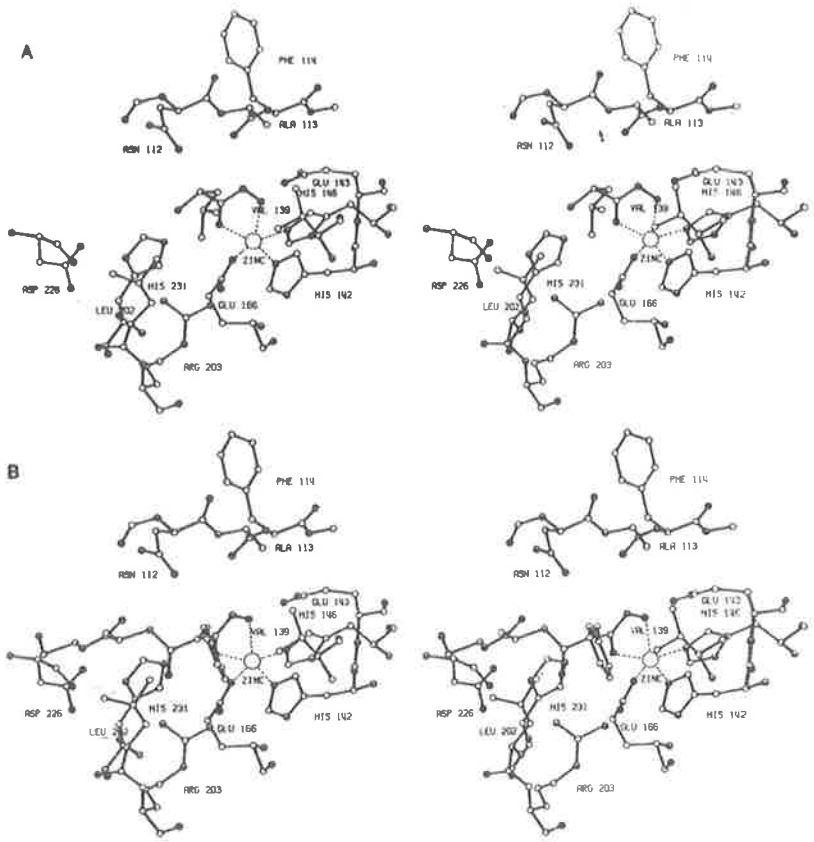


FIGURE 6: Stereoviews showing the similar 5-fold zinc coordination observed in the two hydroxamate–thermolysin complexes. (A) L-Leu-NHOH complex; (B) HONH-BzmAGNA complex.

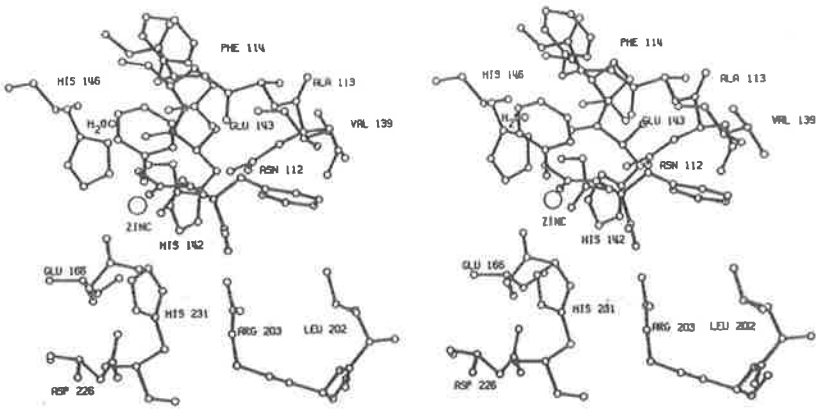


FIGURE 7: Stereo drawing of the hydroxamate group of L-Leu-NHOH (open bonds) superimposed on the inhibitor β-phenylpropionyl-L-phenylalanine (solid bonds) in the thermolysin active site.

propionylphenylalanine (Kester & Matthews, 1977b). As can be seen, the hydroxyl oxygen of the inhibitor is located within the triangle formed by the carbonyl carbon of the "substrate", the zinc ion, and oxygen OE1 of Glu-143. During catalysis, the attacking water molecule would be expected to occupy a very similar position. Of course, the position occupied by the hydroxyl oxygen is constrained by the geometry of the inhibitor and need not coincide exactly with that of the water molecule.

Certainly, the best evidence for the existence of the 5-coordinate zinc intermediate during catalysis would be the direct observation of such an intermediate formed with a true substrate. One way this might be achieved is with the technique of low-temperature crystallography, in which substrate is diffused into the protein crystal at subzero temperatures (the temperature being chosen to give a half-life for the enzyme–substrate complex of at least several days, thus making it possible to collect diffraction data). We have attempted such a study, but have not been successful to date. Crystals of thermolysin can readily be equilibrated with cryosolvents such as, for example, 80% (v/v) ethanol–water or methanol–water and give diffraction patterns at room temperature which are quite similar to those of the native enzyme. However, when these crystals are cooled, they crack at about 0 °C, which is not a sufficiently low temperature to obtain stable enzyme–substrate complexes.

It is interesting to note that the two inhibitors we have studied, one with the hydroxamate group at its C terminus and the other with the hydroxamate at the N terminus, both bind with the hydroxamate moiety complexed to the zinc in a very similar manner, but the peptide portions of the respective inhibitors run in opposite directions. We choose the inhibitor ZGGL-NHOH for crystallographic analysis with the hope that the study would, in part, reveal the mode of binding of an extended thermolysin substrate in the subsites to the "left"

of the zinc, in particular, subsites R_2, R_3, etc. However, no new information of these subsites is provided, and we have to rely on model building to infer the presumed interactions between protein and substrate (i.e., for residue R_2 in Figure 1). In remains unclear whether a suitably designed C-terminal peptide hydroxamate might bind in the "normal" mode in subsites S_1, S_2, etc. One obvious requirement for such an inhibitor is that the residue adjacent to the hydroxamate should not be one, such as leucine, which binds readily in the $R_{1'}$ specificity pocket. Replacing the leucine, as used here, with another residue would, hopefully, prevent the enzymatic hydrolysis of the inhibitor and would also reduce the tendency of this side chain to occupy the $R_{1'}$ subsite, as happens for both inhibitors described here.

Acknowledgments

In writing this paper, we have benefited from the unpublished laboratory notes of Dr. William Kester, in particular his comments concerning a possible 5-coordinate zinc intermediate in the mechanism of thermolysin. We are grateful to Dr. James Powers for generous gifts of hydroxamic acid inhibitors and for continued helpful discussions. Also we thank Dr. F. W. Dahlquist for helpful discussions concerning the mechanism of thermolysin and Drs. L. F. Ten Eyck, M. F. Schmid, and L. H. Weaver for help with implementing the refinement programs.

References

Andrews, A. L., Bishop, R. J., Cleaver, B., Dennard, A. E., Henn, D. E., Loader, B. E., Mackay, A. L., Mendel, H., Morgan, D. J., Owston, P. G., & Phillips, D. C. (1965) in *Tables of Interatomic Distances and Configuration in Molecules and Ions, Supplement* (Mitchell, A. D., Somerfield, A. E., & Cross, L. C., Eds.) The Chemical Society, London.

Bauer, L., & Exner, O. (1974) *Angew. Chem., Int. Ed. Engl. 13*, 376–384.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Jr., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535–542.

Bigbee, W. L., & Dahlquist, F. W. (1974) *Biochemistry 13*, 3542–3549.

Boiwe, T., & Brändén, C. I. (1977) *Eur. J. Biochem. 77*, 173–179.

Bolognesi, M. C., & Matthews, B. W. (1979) *J. Biol. Chem. 254*, 634–639.

Bowen, H. J. M., Donohue, J., Jenkin, D. G., Kennard, O., Wheatley, P. J., & Whiffen, D. H. (1958) in *Tables of Interatomic Distances and Configuration in Molecules and Ions* (Mitchell, A. D., & Cross, L. C., Eds.) The Chemical Society, London.

Buckingham, D. A., Keene, F. R., & Sargeson, A. M. (1974) *J. Am. Chem. Soc. 96*, 4981–4983.

Colman, P. M., Jansonius, J. N., & Matthews, B. W. (1972) *J. Mol. Biol. 70*, 701–724.

Cotton, F. A., & Wilkinson, G. (1972) *Advanced Inorganic Chemistry*, pp 503ff, 845ff, 875ff, 903ff, Interscience, New York.

Dworschack, R. T., & Plapp, B. V. (1977) *Biochemistry 16*, 111–116.

Endo, S. (1962) *J. Ferment. Technol. 40*, 346–353.

Fife, T. H., & Squillacote, V. L. (1977) *J. Am. Chem. Soc. 99*, 3762–3769.

Fife, T. H., & Squillacote, V. L. (1978) *J. Am. Chem. Soc. 100*, 4787–4793.

Groves, J. T., & Dias, R. M. (1979) *J. Am. Chem. Soc. 101*, 1033–1035.

Harris, E. D., & Krone, S. M. (1974) *N. Engl. J. Med. 291*, 557, 605, 652.

Hendrickson, W. A., & Konnert, J. H. (1980) in *Computing in Crystallography* (Diamond, R., Ramaseshan, S., & Venkatesan, K., Eds.) pp 13.01–13.23, Indian Academy of Sciences, Bangalore.

Holmquist, B., & Vallee, B. L. (1974) *J. Biol. Chem. 249*, 4601–4607.

Holmquist, B., & Vallee, B. L. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 6216–6220.

Kam, C.-M., Nishino, N., & Powers, J. C. (1979) *Biochemistry 18*, 3032–3038.

Kannan, K. K., Petef, M., Fridborg, K., Cid-Dresdner, H., & Lougren, S. (1977) *FEBS Lett. 73*, 115–119.

Kester, W. R., & Matthews, B. W. (1977a) *J. Biol. Chem. 252*, 7704–7710.

Kester, W. R., & Matthews, B. W. (1977b) *Biochemistry 16*, 2506–2516.

Latt, S. A., Holmquist, B., & Vallee, B. L. (1969) *Biochem. Biophys. Res. Commun. 37*, 333–339.

Lindskog, S. (1970) *Struct. Bonding (Berlin) 8*, 154–196.

Matsubara, H., Sasaki, R., Singer, A., & Jukes, T. H. (1966) *Arch. Biochem. Biophys. 115*, 324–331.

Nielands, J. B. (1966) *Struct. Bonding (Berlin) 1*, 59–108.

Nishino, N., & Powers, J. C. (1978) *Biochemistry 17*, 2846–2850.

Nishino, N., & Powers, J. C. (1979) *Biochemistry 18*, 4340–4347.

Pangburn, M. K., & Walsh, K. A. (1975) *Biochemistry 14*, 4050–4054.

Peach, M. J. (1977) *Physiol. Rev. 57*, 313–370.

Pearson, R. G. (1966) *Science (Washington, D.C.) 151*, 172–177.

Prince, R. H., & Woolley, P. R. (1972) *Angew. Chem., Int. Ed. Engl. 11*, 408–417.

Richards, F. M. (1968) *J. Mol. Biol. 37*, 225–230.

Rossman, M. G., & Argos, P. (1978) *Mol. Cell. Biochem. 21*, 161–182.

Schechter, I., & Berger, A. (1967) *Biochem. Biophys. Res. Commun. 27*, 157–162.

Vallee, B. L., & Williams, R. J. P. (1968) *Proc. Natl. Acad. Sci. U.S.A. 59*, 498–505.

Weaver, L. H., Kester, W. R., & Matthews, B. W. (1977) *J. Mol. Biol. 114*, 119–132.

Wells, M. A., & Bruice, T. C. (1977) *J. Am. Chem. Soc. 99*, 5341–5356.

Woolley, P. (1975) *Nature (London) 258*, 677–682.

Zalkin, A., Forrester, J. D., & Templeton, D. H. (1966) *J. Am. Chem. Soc. 88*, 1810–1814.

J. Mol. Biol. (1982) **160**, 623–639

Structure of Thermolysin Refined at 1·6 Å Resolution

M. A. Holmes and B. W. Matthews

Institute of Molecular Biology and Department of Physics
University of Oregon, Eugene, OR 79403, U.S.A.

(Received 19 February 1982, and in revised form 26 May 1982)

The structure of the thermostable protease thermolysin has been refined by a restrained least-squares procedure at a nominal resolution of 1·6 Å to a conventional R-value of 21·3% for 34,671 observed reflections (or $R = 19{\cdot}5\%$ for reflections with $F_0 > 3\sigma(F_0)$). The refined structure was constrained to adhere to known stereochemistry, with root-mean-square deviations of 0·021 Å from ideal bond lengths and 2·9° from ideal bond angles. The final model included 173 solvent molecules, which were given unit occupancies. Seven of these are "buried" within the protein. Atoms with the least apparent thermal motion tend to be those that are most deeply buried within the two domains of the structure. The active-site zinc is shown to have approximately tetrahedral co-ordination. Unusual features of the structure, confirmed by the refinement, include a *cis*-proline, a γ-turn, and a single turn of left-handed α-helix. The refinement shows that thermolysin does not contain unusual structures and supports our previous assertion that the thermostability of thermolysin and of thermostable proteins in general is due to a combination of factors which, in different instances, can include hydrophobic interactions, hydrogen bonding, ionic interactions, disulfide linkages, metal binding and other forms of stabilization.

1. Introduction

The crude atomic co-ordinates obtained from fitting a model of an enzyme to a medium resolution (2·5 to 4·0 Å) electron density map can yield much information about the structure of the molecule, including the geometry of the active site and the spatial disposition of amino acid residues implicated in catalysis. However, these initial atomic co-ordinates are not reliable enough to define the changes in enzyme conformation that may occur during catalysis. To examine subtle details of structure, atomic co-ordinates that have been refined against high-resolution data are needed. In this paper, we describe the high-resolution refinement of the structure of native thermolysin and discuss the results of the refinement.

The structure of thermolysin, a heat-stable protease, M_r 34,600, was initially determined at 2·3 Å resolution (Matthews *et al.*, 1972*a*,1974; Colman *et al.*, 1972). The overall structure consists of two roughly spherical domains with a deep cleft across the middle of the molecule, between the two domains. This cleft contains the active site. Crystallographic studies of the binding of inhibitors to thermolysin have revealed the probable mode of binding of extended substrates and have suggested a

0022–2836/82/280623–17 $03.00/0
© 1982 Academic Press Inc. (London) Ltd.

mechanism of action for the enzyme (e.g. see Weaver *et al.*, 1977; Kester & Matthews, 1977; Holmes & Matthews, 1981).

2. Data Collection

(a) *Crystals*

Crystals of thermolysin used for high-resolution data collection were grown as described previously (Matthews *et al.*, 1972*a*) with the modification that the protein solution included 2·5 M-CsCl. The "standard mother liquor" used was 0·01 M-calcium acetate, 0·01 M-Tris, and 7% (v/v) dimethyl sulfoxide, adjusted to pH 7·2.

(b) *Intensity measurement*

The 2·3 Å resolution data set was collected previously by the method of precession photography (Colman *et al.*, 1972). A 1·6 Å resolution data set was obtained using oscillation photography on an Elliot GX-21 rotating anode generator equipped with a graphite monochromator and a collimator with adjustable apertures (Schmid *et al.*, 1981). The generator was run at 5·5 kW on a focal spot measuring 0·5 mm × 5·0 mm. A cylindrical cassette was used in conjunction with a modified Enraf-Nonius precession camera with a stepping motor driven by a microprocessor controller. A pair of reference pins in the cassette eliminated the need for fiducial marks on the film other than the direct beam (Schmid *et al.*, 1981). Two films of Kodak No-Screen X-ray film were included in each oscillation film pack (cf. Matthews *et al.*, 1972*b*). Although a number of strong reflections were "overloaded" on the second film, these reflection intensities were obtained from the previously measured precession data set. The oscillation films were processed using a program based on that of M. G. Rossmann (Rossmann, 1979; Schmid *et al.*, 1981). Statistics are given in Table 1. The merged high-resolution data set contains approximately 95% of the data to 1·7 Å resolution, but beyond this point the data become very weak and only 27% of the reflections between 1·7 Å and 1·6 Å were recorded.

TABLE 1

Statistics for intensity data

Oscillation data set	
Number of film packs	48
Total number of measurements	95,806
Number of unique reflections	32,660
Average R_{sym} (%)†‡	5·6
R_{merge} (%) including partial reflections§‖	8·8
R_{merge} (%) without partial reflections	7·1
Merged oscillation/precession data set	
R_{merge} (%) between oscillation and precession data	6·6
Number of unique reflections¶	34,971

† $R = \Sigma|I - \bar{I}|/\Sigma I$.

‡ R_{sym} is the R-value for symmetry-related intensities recorded on the same film.

§ R_{merge} is the R-value for intensities measured on different films or from different data sets.

‖ Partial reflections are measured by summing 2 or more components recorded on adjacent films as described by Schmid *et al.* (1981).

¶ The additional reflections relative to the oscillation data alone are mainly intense, low-resolution, reflections that were "overloaded" on the oscillation films.

3. Refinement

(a) *Method of refinement*

The refinement of thermolysin was carried out by the restrained least-squares procedure of Hendrickson & Konnert (1980: see also Sielecki *et al.*, 1979). The original program was separated into two parts: PROLSF, which produces the matrix elements for the structure factor derivatives, and PROLSO, which completes the parameter shifts. This division of the single program into two parts increases efficiency for the VAX computer (P. Briley, personal communication).

One change was made to the "standard groups dictionary" that defines the ideal geometry used in the stereochemical restraints. On the basis of small molecule geometry, the arginine side-chain had been defined as having just the last four atoms coplanar, but we found it necessary to set the last five atoms coplanar in order to prevent large, inexplicable shifts in the arginine side-chain atoms (possibly due to an undetected error in the program).

The initial model protein co-ordinates were those of Matthews *et al.* (1974). In all cycles of refinement, an inner resolution limit of 10·0 Å was used.

The computing time on a VAX 11/780 computer needed for a cycle of refinement for the thermolysin space group $P6_122$, with approximately 2500 atoms, was about 15 central processor unit hours for the 14,000 reflections in the 2·3 Å resolution data set and was 21 central processor unit hours with a floating point accelerator in place for the 35,000 reflections in the high-resolution data set. (The floating point accelerator halved the amount of central processor time needed by this particular program.) Over 90% of the computing time required for a cycle is spent calculating the structure factors and their derivatives.

(b) *Course and progress of refinement*

The progress of the refinement of the thermolysin structure is summarized graphically in Figure 1. The first 20 cycles of refinement were carried out with the 2·3 Å precession data set, the initial cycle consisting of stereochemical idealization only. Difference maps were calculated and water molecules added to the structure after cycles 7 and 15. Halfway through, after the tenth cycle, individual atomic temperature factors were introduced.

One set of weights for structure factors and stereochemistry was used throughout. The entire refinement proceeded very smoothly, with the only manual adjustment to the protein itself being a change in the temperature factor of a methionine sulfur atom. The results of this initial refinement proved to be very valuable in understanding, and later refining, the structures of several thermolysin–inhibitor complexes (Holmes & Matthews, 1981).

High-resolution refinement began at cycle 21 with the introduction of the 1·6 Å resolution merged oscillation–precession data set. All the data between 10·0 Å and 1·6 Å were used immediately, and an additional 37 cycles of refinement were performed. These cycles include various "experimentations" with the stereochemical idealization parameters. A difference map was calculated and water molecules were added to the structure after cycle 34. At various times, water

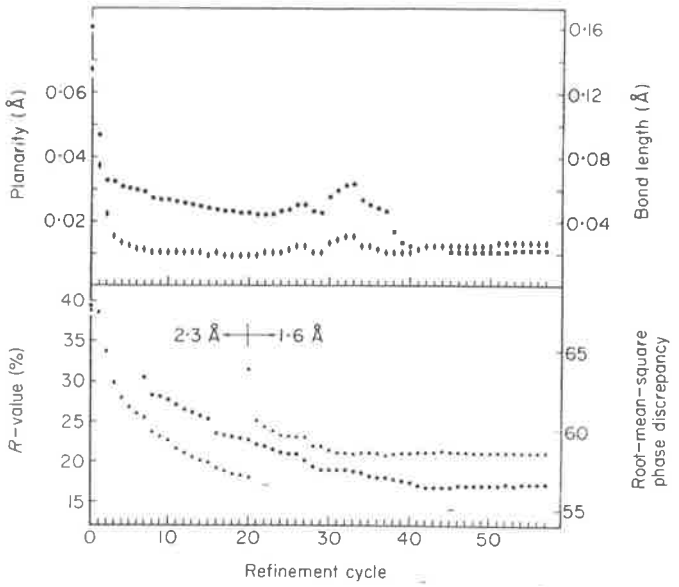


FIG. 1. Summary of the course of the thermolysin refinement showing the crystallographic R-value (▲), root-mean-square discrepancy between the phase angles determined by isomorphous replacement and those calculated from the model (●), and root-mean-square discrepancy of the model from ideal bond lengths (■) and planarity (◆).

molecules that had very high temperature factors or formed bad interatomic contacts were deleted from the co-ordinate list. During the high-resolution refinement, some changes were made with the structure factor weighting. In the latter cycles, the standard deviations used were either the value estimated from film processing (cf. Table 2) or a minimum threshold value of 30. This rather conservative approach was used, since the estimated values do not take into account systematic errors, for example due to absorption, and it was felt that they

TABLE 2

Structure factor statistics for the refined model

Resolution (Å)	No. of reflections	$\overline{\sigma(F_0)}$	$\overline{\lvert F_0 - F_c \rvert}$	R(%) All F_0	R(%) $F_0 > 3\sigma(F_0)$
10·0–5·0	1450	10·5	98	22·8	22·1
5·0–3·5	2880	8·5	76	15·2	15·0
3·5–2·9	3638	8·1	59	17·8	17·5
2·9–2·5	4223	10·8	46	19·5	18·9
2·5–2·2	4370	15·0	41	20·4	19·4
2·2–2·0	4173	19·1	37	21·3	20·0
2·0–1·9	4227	23·2	33	24·0	21·5
1·9–1·8	4114	29·1	31	28·2	23·7
1·8–1·7	3741	34·2	35	35·5	33·5
1·7–1·6	1855	41·1	41	42·6	
Overall	34,671			21·3	
	28,712				19·5

might be unrealistically low in some cases. Towards the end of the refinement, the bond length and bond angle weights were tightened to improve the stereochemistry of the model. Later, interatomic contact restraints were added and torsion angle and chiral center weights were tightened. The final crystallographic R-value for the 1·6 Å data is 21·3% for a model that includes 173 water molecules. The root-mean-square shift between the starting co-ordinates and the final refined values is 0·56 Å for all atoms in the structure and 0·48 Å for main chain atoms.

The structure factors and refined parameters for the thermolysin structure have been deposited in the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977).

The fact that no manual adjustment of the protein conformation was required during refinement is unusual. In retrospect, a number of factors contributed to the excellent set of starting co-ordinates. The original electron density map of thermolysin (Matthews *et al.*, 1972*a*; Colman *et al.*, 1972) was of high quality and was further improved by the inclusion of additional isomorphous derivatives (Matthews *et al.*, 1974). Also, the resolution (2·3 Å) was higher than most isomorphous replacement maps. Finally, the atomic co-ordinates were obtained by placing markers directly on the map sections rather than measuring them from a wire model (Matthews *et al.*, 1974).

(c) *Agreement of the refined model with the diffraction data*

During the course of the refinement of the thermolysin structure, the R-value decreased from 39·1% to 21·3% (Fig. 1; Table 2). The discrepancy between the phases calculated from the model and the multiple isomorphous replacement phases from 10·0 Å to 2·3 Å resolution decreased from a root-mean-square value of 68° to one of 57°. This agreement especially improved after cycles 7 and 15, when new water molecules were added to the model (Fig. 1).

If the initial isomorphous replacement phases are very poor, then the model phases may tend to move away from the isomorphous replacement phases as the refinement proceeds. On the other hand, if the initial isomorphous replacement phases are good, then these phases will be a better approximation to the correct values than those provided by the original model, which not only includes co-ordinate errors, but usually does not include solvent, and assumes an overall thermal factor. In such cases, of which thermolysin is an example, the discrepancy between the isomorphous replacement and the model phases will decrease as the refinement proceeds to a limiting value determined by errors in the isomorphous replacement phases and also, presumably, by limitations in the refined model.

At the completion of the high-resolution refinement, the mean value of the phase discrepancy $|\alpha_{MIR} - \alpha_{calc}|$ was 35°, bespeaking the accuracy of the isomorphous replacement phases. For these phases, the mean figure of merit was 0·79 (Matthews *et al.*, 1974), corresponding to a mean error in phase of 38° (Blow & Crick, 1959; Dickerson *et al.*, 1961), in good agreement with the value of 35°.

(d) *Agreement of the model with ideal stereochemistry*

The deviations of the refined model of thermolysin from ideal geometry are given in Table 3. The distribution of the deviations of the bond lengths and angles from

TABLE 3

Deviation from ideal geometry

Bond lengths (Å)	0·021
Bond angles (deg.)	2·9
Planarity (Å)	0·013
Chiral volume (Å^3)	0·20
Non-bonded contacts	
Separated by a single torsion angle (Å)	0.30
Separated by multiple torsion angles, including van der Waals' contact and hydrogen bonds (Å)	0·16
Torsion angles (deg.)	
Planar	1·5
Staggered	19·9
Orthonormal	20·2

The entries in the Table give the root-mean-square deviations of the listed parameters from "ideal" values. The chiral volume for a C^α, for example, is enclosed by C^α, C^β, C and N and ensures that the correct handedness is maintained; planar torsion angles occur in peptides and amides, staggered torsion angles occur between tetrahedral carbon atoms and orthonormal torsion angles occur between tetrahedral and trigonal carbons.

their ideal values corresponds to the range of values observed in "small molecule" structure determination. Thus the refinement program allows physically reasonable deviations from ideal geometry to occur. The average bond length and bond angle values for the peptide group are given in Table 4. In no case is there a significant difference between the average value of a parameter and the "optimum" restrained value, although it can be noted that the tetrahedral angles at the alpha-carbon atoms are more variable than the other angles. The average value of the angle at the alpha-carbon is greater than the tetrahedral value of 109·6°, but only marginally so.

The refinement program PROLSQ has the capability of restraining van der

TABLE 4

Average peptide geometry

Bonds	Restrained value (Å)	Average value (Å)	Root-mean-square deviation (Å)
C^α–C	1·53	1·52	0·020
C–O	1·24	1·25	0·016
C–N	1·32	1·32	0·016
N–C^α	1·47	1·47	0·019
Angles (deg.)			
C^α–C–O	121·0	120·3	2·3
C^α–C–N	114·0	116·4	2·2
O–C–N	124·9	123·2	2·3
C–N–C^α	123·0	122·3	2·8
N–C^α–C	109·6	111·0	3·5
ω†	180·0	178·9	1·1

† $\omega = 180°$ corresponds to a planar peptide bond.

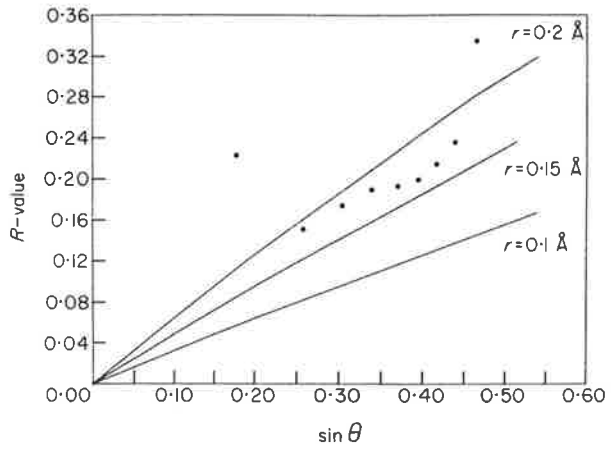


FIG. 2. Luzzati plot (Luzzati, 1952). The plotted R-values are for reflections with $F_0 > 3\sigma F_0$.

Waals' contacts. When a pair of atoms tend to move too close together, a term is added to the least-squares matrices to ensure that the approach is within acceptable limits. This type of restraint was not included in the first 49 cycles of refinement, because the program did not recognize the possibility of various hydrogen bonds involving water molecules. Neither were interactions involving the metal ions recognized. Since the contacts restraint was not used, the model was periodically screened for bad interatomic contacts. The worst contacts always involved water molecules, in which cases the offending molecules would be deleted from the co-ordinate list. This procedure gave quite acceptable root-mean-square deviations from ideality for interatomic contacts; namely, 0·34 Å for contacts involving atoms 1 and 4 of a single torsion angle and 0·19 Å for contacts between atoms further removed in the chemical structure of the protein (including water molecules). The last eight cycles of refinement included corrected contact restraints (although salt linkages were still not recognized), and resulted in only slightly improved deviations from ideality.

(e) *Error in the refined model*

Figure 2 is a Luzzati diagram (Luzzati, 1952) for the refined thermolysin structure. Except for the extreme low-angle and high-angle data, the points lie between the line for a mean co-ordinate error of 0·15 Å and that for an error of 0·20 Å. However, the Luzzati plot leads to an overestimation of the error in the model, as it assumes that error in the co-ordinates is the only cause of disagreement between F_{calc} and F_{obs}. Thus we conclude that the mean error in the refined thermolysin structure is not more than 0·15 Å.

4. Conformation of the Molecule

As discussed above, the refinement of thermolysin caused only small adjustments in the atomic co-ordinates. In no case was it necessary to invert a peptide bond or

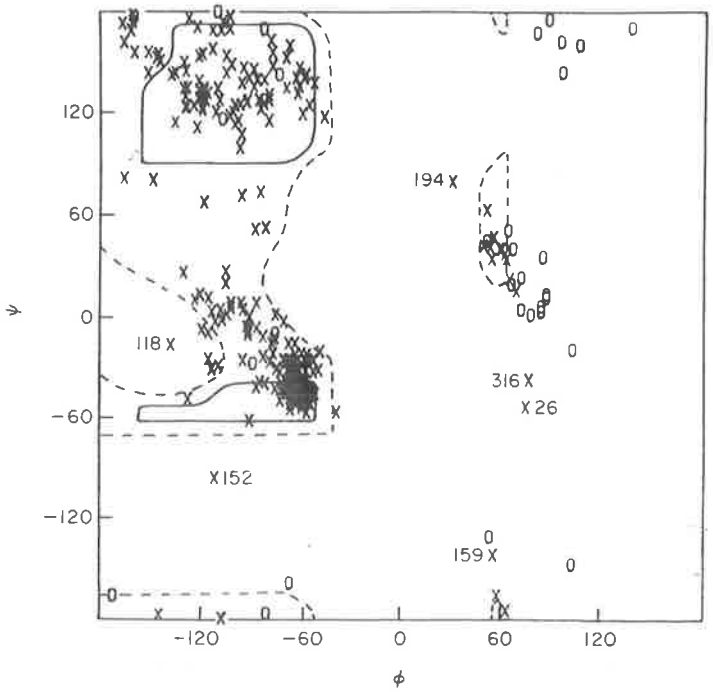


FIG. 3. Conformational diagram for the backbone of thermolysin. Residues that are outside the "allowed" regions for a hard-sphere model are numbered.

otherwise modify the backbone conformation. For this reason, the extensive discussion of the secondary structure of thermolysin given previously (Matthews *et al.*, 1974) applies without change, and need not be repeated here. Unusual features of the structure confirmed by the refinement include a *cis* peptide bond at Pro51, a γ-turn at Thr26 and a single turn of left-handed α-helix including Asp226-Asn227-Gly228-Gly229.

A plot of the backbone dihedral angles for refined thermolysin is shown in Figure 3. This plot is very similar to that one based on the initial model (Matthews *et al.*, 1974). Most of the non-glycine residues that were outside the "allowed" regions for an empirical hardsphere model in the initial structure remain outside in the refined structure. This result indicates that the "disallowed" combinations of dihedral angles seen in Figure 3 are not due to experimental error, since they have survived the refinement process.

The co-ordination of the active-site zinc ion is approximately tetrahedral (Table 5) with three protein ligands and a bound water molecule. Rees *et al.* (1981) have recently reported that for carboxypeptidase A the zinc co-ordination of Glu72 is bidentate, resulting in overall pentaco-ordination of the metal. However, for thermolysin the co-ordinating oxygen, OE1 of Glu166, is 2·1 Å from the zinc, whereas the second oxygen, OE2, is 2·7 Å away. Therefore, the refinement supports tetrahedral zinc co-ordination in native thermolysin, although pentaco-ordination is observed in some inhibitor complexes (e.g. see Holmes & Matthews, 1981).

TABLE 5

Zinc co-ordination

Ligand	Distance (Å)
His142 NE2	2·10
His146 NE2	2·08
Glu166 OE1	2·08
H_2O 0392	1·88
Ligands	**Angle (deg.)**
H_2O–Zn^{2+}–NE2 144	118·6
H_2O–Zn^{2+}–NE2 146	118·1
H_2O–Zn^{2+}–OE1 166	94·4
142NE2–Zn^{2+}–OE1 166	127·7
146NE2–Zn^{2+}–OE1 166	94·7
142NE2–Zn^{2+}–NE2 146	102·4

TABLE 6

Calcium ligand distances

Ligand	Distance (Å)		Ligand	Distance (Å)	Ligand	Distance (Å)
	Ca1	Ca2		Ca3		Ca4
Asp185 OD1		2·43	Asp57 OD1	2·23	Tyr193 O	2·43
Asp185 OD2	2·52		Asp57 OD2	2·71	Thr194 O	2·54
Glu177 OE1	2·53		Asp59 OD1	2·34	Thr194 OG1	2·48
Glu177 OE2		2·42	Glu61 O	2·18	Ile197 O	2·31
Glu190 OE1	2·40		H_2O O419	2·40	Asp200 OD1	2·24
Glu190 OE2	2·51	2·40	H_2O O482	2·36	H_2O O354	2·42
Glu187 O	2·29		H_2O O503	2·29	H_2O O480	2·35
Asp183 O		2·37				
Asp138 OD1	2·46					
H_2O O346	2·62					
H_2O O353		2·46				
H_2O O475		2·16				

The geometry at the four calcium-binding sites is summarized in Table 6. Apart from small adjustments in the individual ligand distances, the description of the calcium-binding sites given previously (Matthews *et al.*, 1974) still applies. The calcium to calcium distance at the double metal binding site is confirmed to be 3·83 Å.

5. Protein Dynamics

The individual atomic temperature factors obtained by refining a protein structure contain contributions from thermal vibrations, disorder of the crystal, and conformational variability among the molecules in the crystal (e.g. see Frauenfelder *et al.*, 1979; Artymiuk *et al.*, 1979; Sternberg *et al.*, 1979). Frauenfelder *et al.* analyzed atomic displacements in metmyoglobin, considering the sum of

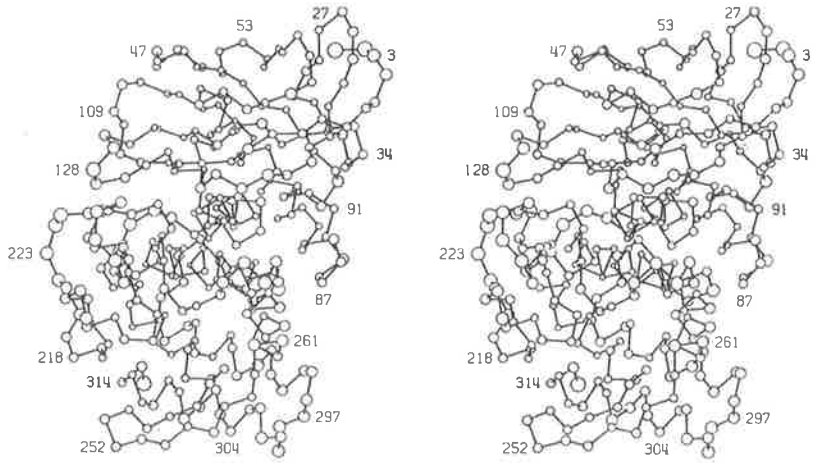


FIG. 4. Stereo diagram illustrating the apparent thermal motion of the thermolysin molecule. Larger circles correspond to residues with greater-apparent motion. The radius of each circle was obtained by taking the average B value for all atoms in that residue, subtracting a constant value of 4·0 Å^2 (in order to make differences in apparent motion more obvious) and drawing the circle at the 50% probability level (Johnson, 1965).

vibrational and conformational contributions and subtracting out crystal disorder contributions. They reasoned that the conformational contribution is large, since the atomic displacements reach large values relative to typical vibrational displacements. Artymiuk *et al.* carried out independent refinements of two homologous lysozymes that crystallize in unrelated space groups and found that the thermal displacements in the two enzymes agreed closely, suggesting that crystal packing effects are not serious and that displacements in the crystal are related to those that occur in solution. With these findings in mind, we have analyzed the temperature factors obtained during the thermolysin refinement.

Figure 4 is a stereo drawing of the alpha-carbon atoms of thermolysin, with the radius of each atom proportional to the displacement calculated from the average temperature factor for the residue. Figure 5 shows the average mean-square displacement plotted along the polypeptide chain. For the most part, the residues on the surface of the molecule have higher average temperature factors, while the internal residues have lower average temperature factors. Regions of highest mobility occur for exposed surface loops in the vicinity of residues 128, 180 and 225, as well as the amino and carboxyl termini. The two long α-helices that cross in the middle of the molecule (residues 137 to 151 and 160 to 180) have relatively low motion, as does the other long helix (residues 65 to 87), which is also largely buried.

Thermolysin consists of two distinct domains, joined by a single helix (residues 137 to 151). As can be seen in Figure 5, the thermal motion in the C-terminal domain, which is predominantly α-helical, is, on the average, slightly larger than in the N-terminal domain, which contains mostly β-sheet. Because of the distinctive domain structure, we have analyzed the average displacement as a function of distance both from the center of each domain and from the center of the whole

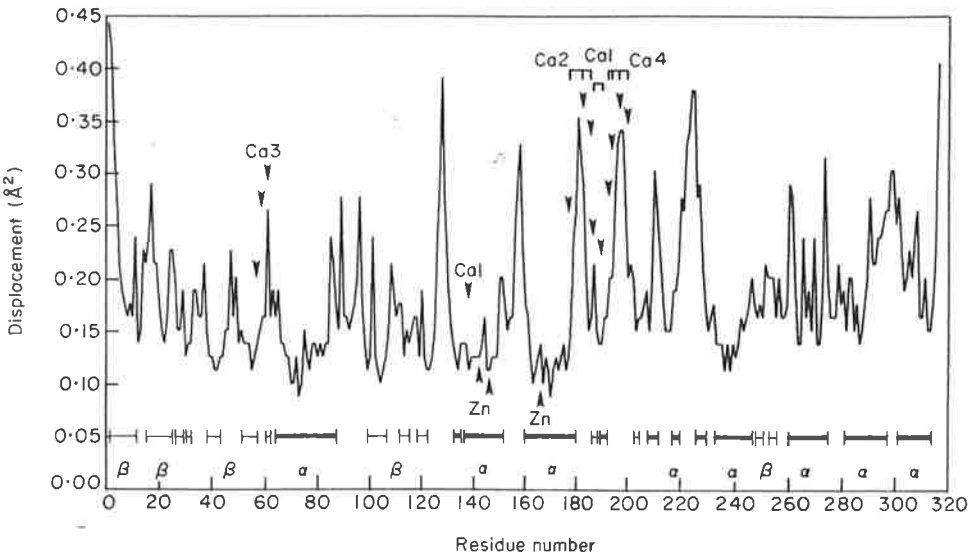


FIG. 5. Apparent thermal motion of thermolysin plotted as a function of position along the polypeptide chain. Values plotted are the average for all atoms in a given residue. Thin bars and thick bars show the locations of β-sheet strands and helices, respectively. Also shown are the locations of all residues that are ligands to the active-site zinc ion and the 4 bound calcium ions.

molecule (Fig. 6). The obvious result, as seen previously by others, is that atoms within the molecule (i.e. within about 10 Å of either the center of gravity of the individual domains (Fig. 6(a) and (c)) or the whole molecule (Fig. 6(b) and (d)) have relatively low and more-or-less equal motion, whereas residues outside 10 Å, which includes surface residues, begin showing a wide range of displacements. Careful inspection of Figure 6 suggests that the atoms with the least motion tend to be closer to the domain center (Fig. 6(a) and (c)) rather than the molecular center. In other words, the residues that are buried most deeply within the protein tend to be those with the least apparent motion. This finding tends to support the argument that apparent "thermal motion" seen in protein crystals is relevant to motion in solution (cf. Frauenfelder *et al.*, 1979; Artymiuk *et al.*, 1979). If the "*B* values" of a protein crystal were dominated by crystal disorder, then they would be expected to have a minimum value at the center of gravity of the whole protein. The fact that this is not the case for thermolysin suggests that crystal disorder is not the dominant factor.

There is no clear correlation between the locations of the bound calcium ions and regions of low or high motion (Fig. 5). However, the three zinc ligands, as well as all the residues in the active site region (Fig. 5) do have relatively low apparent displacements. This is in contrast to lysozyme where Artymiuk *et al.* (1979) found the active site is to be located in a region of high displacement. However, in contrast to hen lysozyme, bound competitive inhibitors are not observed to cause conformational changes in thermolysin (e.g. see Holmes & Matthews, 1981, and references therein).

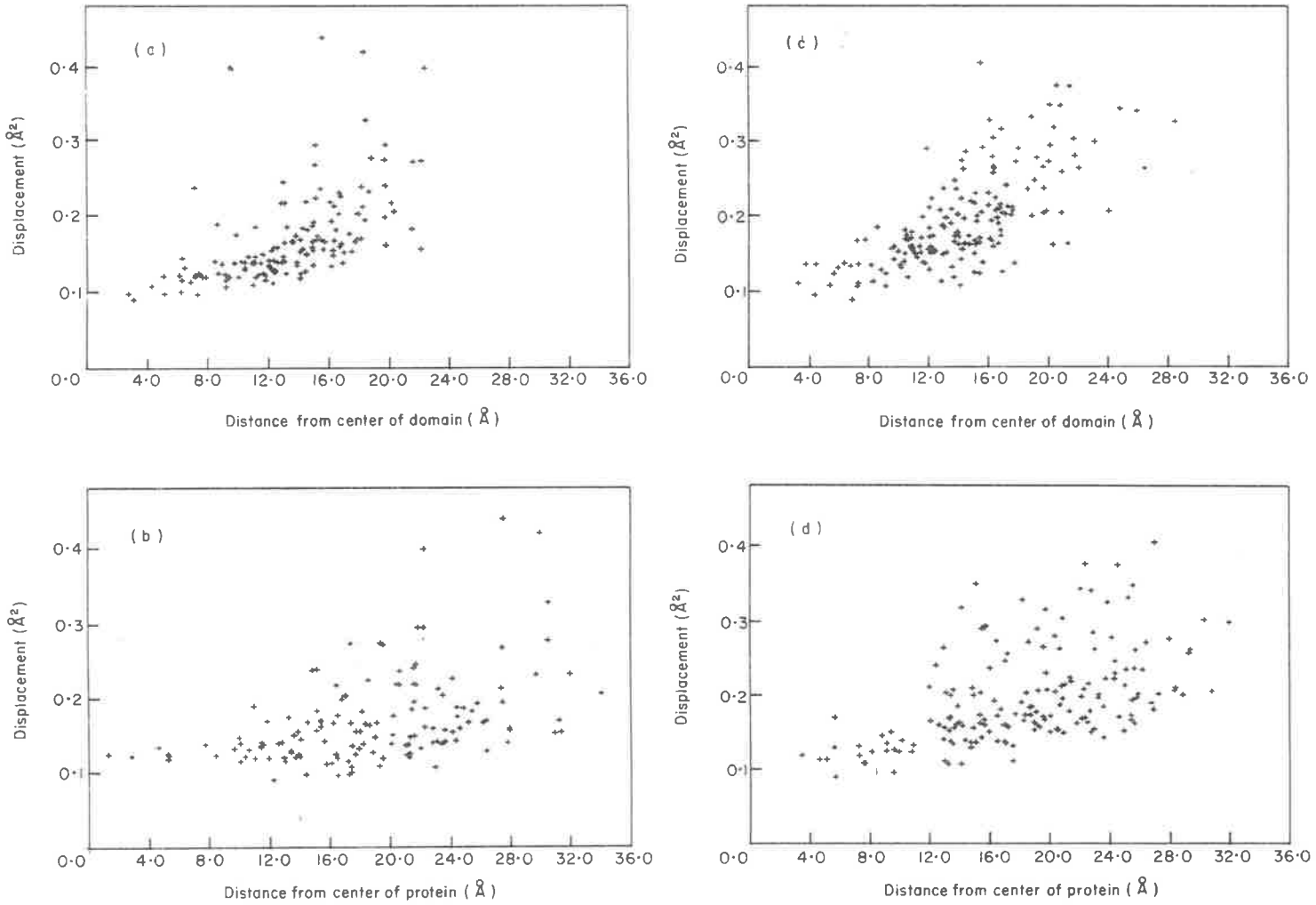


FIG. 6. Apparent thermal motion for each residue of thermolysin plotted as a function of the distance of the residue from the center of gravity of the whole molecule and the center of gravity of the domain within which the residue is located. (a) and (b) Residues 1 to 143, i.e. the N-terminal domain; (c) and (d) residues 144 to 316, which constitute the C-terminal domain. Displacements plotted are the average values for all atoms in the residue.

6. Solvent Structure

A total of 173 water molecules was included during refinement. For these, the refined individual temperature factors range from 9 $Å^2$ to 82 $Å^2$. Occupancies of the water molecules were not refined. Presumably, the overall R-value could have been reduced somewhat by allowing water molecules with partial occupancy, but this was not attempted.

The active site of the enzyme contains about a dozen water molecules, many of which are displaced when inhibitors and, presumably, substrate molecules bind to thermolysin. The active site Zn^{2+} has a water molecule as a ligand in addition to three side-chain ligands, and the four calcium ions are also co-ordinated by both protein and water ligands (Tables 5 and 6). Other water molecules bind to surface atoms of the protein or fill small hydrophilic crevices (cf. Birktoft & Blow, 1972; Bode & Schwager, 1975; Watenpaugh *et al.*, 1978). Some of the water molecules bound to surface protein atoms are shared between two neighboring protein molecules. There are seven water molecules that are "buried" within the protein, two of these occurring in the region of the Ca1–Ca2 binding site.

A summary of the environment of the water molecules is given in Table 7. To construct the Table, we listed all atoms within 3·2 Å of each water molecule (cf. Watenpaugh *et al.*, 1978) and then divided the water molecules into groups according to whether they had one, two, three. . .close contacts with surrounding atoms. These contacts can be regarded as potential hydrogen bonds, although we did not check the detailed geometry in each case. As can be seen from the Table, most of the water molecules make one, two or three hydrogen bonds with neighboring atoms (including other solvent molecules). There are four water molecules with six potential ligands within 3·2 Å. In each of these cases, the water molecule in question is a calcium ligand and is, in addition, hydrogen-bonded to one, two or three other water molecules. Thus, the calcium binding sites tend to be both hydrated and compact.

On average, the solvent atoms that are seen in the crystal structure make more contacts with the protein than with other solvent atoms. Solvent atoms with the maximal number of contacts are those most intimately associated with the protein. As seen in Table 7, these maximally bonded atoms are those with the lowest apparent motion, no doubt because their motion tends to reflect that of the protein to which they are hydrogen-bonded. Also, the variability in thermal motion of the multiliganded solvent atoms, as estimated by the standard deviations quoted in Table 7, is less than that for solvent atoms with only one or two contacts.

One striking result that can be seen in Table 7 is that many more water molecules form hydrogen bonds with oxygen atoms than with nitrogen atoms. Including other solvent atoms, the ratio is 5 to 1, and for hydrogen bonds to the protein it is 3 to 1. In part, this discrepancy is due to the fact than an amide nitrogen can donate only a single hydrogen bond, whereas a carbonyl oxygen can accept hydrogen bonds from more than one water molecule. Also, an oxygen atom can *accept* hydrogen bonds over a rather wide angular range (i.e. from water molecules that occupy a range of positions), whereas a water molecule *accepting* a hydrogen bond from a nitrogen must occupy a position more-or-less colinear with the N–H bond.

TABLE 7

Water molecule environment

Number of interactions per water molecule	Number of water molecules in class	Average mean-square displacement ($Å^2$)	Type of interacting atom						
			H_2O	O	OX	N	NX	Ca^{2+}	Zn^{2+}
1	54	0.51 ± 0.19	18	18	11	4	3		
2	52	0.43 ± 0.18	37	26	23	7	11		
3	36	0.29 ± 0.16	40	25	18	14	9	1	1
4	11	0.21 ± 0.12	14	7	14	4	2	3	
5	1	0.25	2	1	2				
6	4	0.21 ± 0.05	8	7	5			4	
Total	158		119	84	73	29	25	8	1

All atoms within 3.2 Å of a given solvent molecule are considered to be interacting atoms. The atom identification is as follows: H_2O, solvent atom; O, backbone carbonyl oxygen; OX, any other protein oxygen; N, backbone amide nitrogen; NX, any other nitrogen. There are 15 solvent molecules that have no other atoms within 3.2 Å. These molecules have contact distances just above 3.2 Å and/or make contact with a neighboring protein molecule in the crystal. The mean-square displacements were calculated from the relation $B = 8\pi^2\overline{x^2}$ and the quoted range of values corresponds to the standard deviation of the distribution.

In addition, side-chain hydroxyl and protonated carboxyl groups are versatile in that they can either donate or accept hydrogen bonds, while side-chain nitrogen atoms are capable only of donating hydrogen bonds.

It appears that every backbone nitrogen and carbonyl oxygen that is "inside" the protein is in a position to participate in hydrogen bonding of one sort or another. Considering the entire protein, 84 of the peptide carbonyl oxygen atoms are within hydrogen-bonding distance of a water molecule, whereas this occurs for only 29 peptide nitrogen atoms.

7. Thermostability

One of the initial motivations for determining the structure of thermolysin was to obtain insights into the mechanism of thermostability. Following the initial structure determination, we concluded that the enhanced thermostability of thermostable proteins cannot be attributed to a single determinant but, in general, arises from a combination of small differences in hydrophobic character, metal binding, hydrogen bonding, ionic interactions and so on (Matthews *et al.*, 1974; Weaver *et al.*, 1976). Subsequent studies have provided additional support for this rationalization.

In the case of the bacterial ferredoxins, Perutz & Raidt (1975) have argued that salt bridges are of prime importance whereas, in hemoglobin, salt bridges, hydrogen bonds and non-polar bonds may provide extra energy of stabilization. Walker *et al.* (1980) have also argued that ionic interactions are important. Argos *et al.* (1979) have carried out a statistical analysis of amino acid substitutions in thermophilic and mesophilic ferredoxins and dehydrogenases in an attempt to explain the differences in stability of these molecules. In agreement with Matthews *et al.* (1974), Argos *et al.* (1979) concluded that thermal stability is achieved by the addition of many small changes throughout a protein molecule. They also suggested that stabilization of thermostable proteins is enhanced by having a higher proportion of helix-stabilizing residues in helices, and similarly for sheets.

In an alternative approach to understanding the molecular basis of the stabilization of proteins, we have analyzed a series of temperature-sensitive mutants of phage lysozyme (e.g. see Grütter *et al.*, 1979; Schellman *et al.*, 1981). These studies show directly that substantial changes in the thermal stability of a protein can result from the replacement of a single amino acid residue, with practically no change in the backbone structure of the molecule. In addition, it has been found that the replacement of one amino acid residue with another can, in different cases, lead to very different changes in stability. For example, one temperature-sensitive mutant lysozyme has the substitution Thr → Ala, while another such mutant has exactly the opposite substitution, i.e. Ala → Thr (M. G. Grütter & B. W. Matthews, unpublished results). In this case, there is no net change of charge. Clearly, the consequences of substituting one amino acid residue for another depend on the environment of that particular residue. These findings are consistent with our previous inferences from the thermolysin structure concerning the thermostability of proteins. The important point is that the net free energy of stabilization of a protein is numerically small (e.g. 10 to 20 kcal), but this

number represents a difference between energy terms that are themselves large; the principal contribution to stabilization coming from hydrophobic interactions, and the principal destabilization term being due to the entropy cost of holding the protein in its defined tertiary conformation. Thermostable proteins, such as thermolysin, differ from their mesophilic counterparts in having a slightly higher net energy of stabilization but, as the thermolysin structure shows, this does not require special structural features. In the case of thermolysin, the bound calcium ions must contribute to the overall stability of the molecule and, as we have suggested previously (Matthews *et al.*, 1974), may help protect surface loops of the enzyme against autolysis. However, we believe that the use of bound metals is just one possible mode of stabilization. Similarly, ionic interactions may also play a significant role but we do not believe that increased thermostability *per se* is always due exclusively to an increased number of ionic interactions. Rather, we suggest that in a given instance a variety of stabilizing effects can contribute; these include hydrophobic interactions, ionic and hydrogen bonding, metal binding, disulfide bonds, stabilization of secondary structure, and so on.

Prior to the work described here, a partial refinement of the thermolysin structure at 2·3 Å was carried out by Dr Lynn F. Ten Eyck (unpublished result). We thank Dr Ten Eyck for several computer programs and for discussions on refinement techniques. We also thank our co-workers Drs R. Fisher, M. F. Schmid and L. H. Weaver as well as Drs P. Briley, W. A. Hendrickson and A. R. Sielecki for various help with data collection and with computational and refinement procedures.

This work was supported in part by grants from the National Science Foundation (PCM 8014311), the National Institutes of Health (GM 20066) and the M. J. Murdock Charitable Trust.

REFERENCES

Argos, P., Rossmann, M. G., Grau, U. M., Zuber, H., Frank, G. & Tratschin, J. D. (1979). *Biochemistry*, **18**, 5698–5703.

Artymiuk, P. J., Blake, C. C. F., Grace, D. E. P., Oatley, S. J., Phillips, D. C. & Sternberg, M. J. E. (1979). *Nature (London)*, **280**, 563–568.

Bernstein, F. C., Koetzle, I. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. C., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Birktoft, J. J. & Blow, D. M. (1972). *J. Mol. Biol.* **68**, 187–240.

Blow, D. M. & Crick, F. H. C. (1959). *Acta Crystallogr.* **12**, 794–802.

Bode, W. & Schwager, P. (1975). *J. Mol. Biol.* **98**, 693–717.

Colman, P. M., Jansonius, J. N. & Matthews, B. W. (1972). *J. Mol. Biol.* **70**, 701–724.

Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *Acta Crystallogr.* **14**, 1188–1195.

Frauenfelder, H., Petsko, G. A. & Tsernoglou, D. (1979). *Nature (London)*, **280**, 558–563.

Grütter, M. G., Hawkes, R. B. & Matthews, B. W. (1979). *Nature (London)*, **277**, 667–669.

Hendrickson, W. A. & Konnert, J. H. (1980). In *Computing in Crystallography* (Diamond, R., Ramaseshan, S. & Venkatesan, K., eds), pp. 13.01–13.23, Indian Academy of Sciences. Bangalore.

Holmes, M. A. & Matthews, B. W. (1981). *Biochemistry*, **20**, 6912–6920.

Johnson, C. K. (1965). Oak Ridge National Laboratory, Publication ORNL-3794.

Kester, W. R. & Matthews, B. W. (1977). *Biochemistry*, **16**, 2506–2516.

Luzzatti, P. V. (1952). *Acta Crystallogr.* **5**, 802–810.

Matthews, B. W., Jansonius, J. N., Colman, P. M., Shoenborn, B. P. & Dupourque, D. (1972*a*). *Nature New Biol.* **238**, 37–41.

Matthews, B. W., Klopfenstein, C. E. & Colman, P. M. (1972*b*). *J. Phys. E. Sci. Instrum.* **5**, 353–359.
Matthews, B. W., Weaver, L. H. & Kester, W. R. (1974). *J. Biol. Chem.* **249**, 8030–8044.
Perutz, M. F. & Raidt, H. (1975). *Nature (London)*, **255**, 256–259.
Rees, D. C., Lewis, M., Honzatko, R. B., Lipscomb, W. N. & Hardman, K. D. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 3408–3412.
Rossmann, M. G. (1979). *J. Appl. Crystallogr.* **12**, 225–238.
Schellman, J. A., Lindorfer, M., Hawkes, R. & Grütter, M. (1981). *Biopolymers*, **20**, 1989–1999.
Schmid, M. F., Weaver, L. H., Holmes, M. A., Grütter, M. G., Ohlendorf, D. H., Reynolds, R. A., Remington, S. J. & Matthews, B. W. (1981). *Acta crystallogr. sect. A*, **37**, 701–710.
Sielecki, A. R., Hendrickson, W. A., Broughton, C. G., Delbaere, L. T. J., Brayer, G. D. & James, M. N. G. (1979). *J. Mol. Biol.* **134**, 781–804.
Sternberg, M. J. E., Grace, D. E. P. & Phillips, D. C. (1979). *J. Mol. Biol.* **130**, 231–253.
Walker, J. E., Wonacott, A. J. & Harris, J. I. (1980). *Eur. J. Biochem.* **108**, 581–586.
Watenpaugh, K. D., Margulis, T. N., Sieker, L. C. & Jensen, L. H. (1978). *J. Mol. Biol.* **122**, 175–190.
Weaver, L. H., Kester, W. R., Ten Eyck, L. F. & Matthews, B. W. (1976). In *Symposium on Enzymes and Proteins from Thermophilic Microorganisms* (Zuber, H., ed.), pp. 31–39, Birkhauser-Verlag, Basel.
Weaver, L. H., Kester, W. R. & Matthews, B. W. (1977). *J. Mol. Biol.* **114**, 119–132.

Edited by R. Huber

SECTION B. BACTERIOPHAGE LYSOZYME

Reprinted from
Proc. Nat. Acad. Sci. USA
Vol. 71, No. 10, pp. 4178–4182, October 1974

The Three Dimensional Structure of the Lysozyme from Bacteriophage T4

(protein structure/x-ray diffraction/mutants)

B. W. MATTHEWS AND S. J. REMINGTON

Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oreg. 97403

Communicated by Virgil Boekelheide, August 1, 1974

ABSTRACT **The three dimensional structure of the lysozyme from bacteriophage T4 has been determined from a 2.5 Å resolution electron density map. About 60% of the molecule is in a helical conformation and there is one region consisting of antiparallel β-structure. The polypeptide backbone folds into two distinct lobes linked in part by a long helix. In the region between the two lobes, there is a cleft which deepens into a hole or cavity, about 6–8 Å in diameter, extending from one side of the molecule to the other. This opening is closed off by side chains which extend to within 3–5 Å of each other. A number of mutant lysozymes in which residues in the vicinity of the opening are modified have markedly reduced catalytic activity, suggesting that this region of the molecule may be catalytically important. The three dimensional structure of T4 phage lysozyme is quite different from that of hen egg-white lysozyme although it is not clear at this time whether or not the mechanisms of catalysis of the respective enzymes are related.**

T4 phage lysozyme is an enzyme produced in cells of *Escherichia coli* after infection with bacteriophage T4. The enzyme has similar catalytic activity to that of hen egg-white lysozyme, both being endoacetylmuramidases (1). The molecular weight of the enzyme is 18,700, and the amino-acid sequence has been determined (2). Furthermore, a number of lysozymes have been isolated from mutant strains carrying frame-shift or amber mutations in the phage genome and have been used to demonstrate *in vivo* certain features of the genetic code (3–5).

In this preliminary report, we describe the three dimensional structure of T4 phage lysozyme as determined by x-ray crystallography from a 2.5 Å resolution electron density map. Additional details will be given in a subsequent publication.

Crystallization

The protein was purified using essentially the method of Tsugita *et al.* (1) except that 1 mM mercaptoethanol was added to all buffers. Conditions for obtaining the crystals have been described previously (6). Before x-ray photography the crystals were equilibrated with a standard mother liquor consisting of 1.05 M K_2HPO_4, 1.26 M NaH_2PO_4, 0.23 M NaCl, 1.4 mM mercaptoethanol, pH 6.7. In this solution the crystals float.

The crystals have space group $P3_221$ with cell dimensions $a = b = 61.1$ Å, $c = 96.3$ Å and one molecule in each of the six asymmetric units.

Data collection

Diffraction data were recorded photographically using conventional Buerger precession cameras and integrated intensities measured with a computer controlled drum film scanner (7). Sets of 12 films sufficed to measure 92% of the data to 2.5 Å and 53% of the data between 2.4 Å and 2.5 Å. With the films used, the Bijvoet differences were obtained for about half the reflections. Systematic errors in the measurement of the Friedel pairs were reduced by a method of local scaling to be described elsewhere (ref 8, B. W. Matthews *et al.*, manuscript in preparation). Altogether, about 27,000 intensities were measured for the parent crystals and the two heavy atom derivatives, and were reduced to about 7600 unique reflections. Excluding the weak reflections, the agreement between structure amplitudes measured on different films was about 4%.

Heavy atom derivatives

Isomorphous heavy atom derivatives were sought in the usual way by soaking the crystals in a variety of reagents containing heavy atoms. In particular, efforts were made to exploit potentially available groups on the protein, and two such attempts, utilizing the cysteine and the methione residues, proved successful.

Although exposure of the crystals to mercurials caused cracking, it was determined that diffraction data could be collected from crystals soaked in solutions of mercuric ion. As a result of a series of experiments, the conditions finally adopted were to soak the crystals for 2 days in 2.5×10^{-4} M $HgCl_2$ in standard mother liquor, in the absence of mercaptoethanol. After this period, a solution 1.0×10^{-6} M in mercuric ion was substituted. Under these conditions, both cysteines appear to have been fully substituted with mercury.

In the case of methione, it was possible to obtain substitution of three of the five residues in the molecule by soaking the crystals for 5 days in 3.3×10^{-4} M K_2PtCl_4 in standard mother liquor, without mercaptoethanol.

The sites of heavy atom substitution were first determined from difference Patterson syntheses and confirmed by difference Fourier methods and parameter refinement (9) in the usual way. In these preliminary tests, the centrosymmetric (*h0l*) zone, which in space group $P3_221$ gives the x, y, and z coordinates of each atom, was used exclusively, including data, up to a resolution of 2.5 Å. Final refinement of the heavy atom parameters was by a "lack of closure" procedure (10) including all the data, using a program written by Dr. Lynn Ten Eyck. The refined parameters are given in Table 1.

The space group was shown by two methods similar to those used for thermolysin (11, 12) to be $P3_221$ rather than $P3_121$.

Phase angles were calculated using the method of Blow and Crick (13), including the anomalous scattering data (14). The overall figure of merit was 0.57 and the average value of

TABLE 1. *Refinement statistics for T4 lysozyme*

Derivative	Reaction site	Z	x	y	z	B	$\langle f_H \rangle$	E	R
$HgCl_2$	Cys 54	104	0.551	0.660	0.025	25	201	103	0.48
	Cys 97	88	0.572	0.965	0.083	38			
K_2PtCl_4	Met 106	127	0.680	0.340	0.034	26	267	126	0.44
	Met 120	139	0.127	0.453	0.114	84			
	Met 1	109	0.711	0.953	0.069	49			

Z is the occupancy of the heavy atom site; x, y, z the fractional coordinates; B, the isotropic thermal parameter in A^2; $\langle f_H \rangle$, the root mean square heavy atom scattering; E, the root mean square lack of closure error, and R, the reliability index for centrosymmetric reflections. Z, $\langle f_H \rangle$, and E are on the same arbitrary scale.

[lack of closure/E (or E')] at the most probable phase was 0.66, indicating that the errors in the data were somewhat less than had been estimated (15).

Electron density map

After tracing the electron density map onto sheets of transparent film, a number of helices were immediately apparent and the direction of the polypeptide chain within such helices could be ascertained from the characteristic forward thrust of the carbonyl oxygens and backward tilt of the side chains. On consideration of the two mercury binding sites as possible locations for the respective cysteines, it became apparent that a length of helix running from site Hg 1 must correspond to the sequence Cys 97, Ala 98, . . . Strong density, practically the highest in the map, could be seen at a position compatible with Met 102, apparently in an interior region, while further along the helix Met 106 was located at a position such that it accounted for the major site of chloroplatinite binding. From this initial interpretation, and by making use of the amino-acid sequence, it was not difficult to trace the entire course of the polypeptide chain. There are a few regions where a detailed interpretation is difficult, notably including residues 50–56, which are at an extremity of the molecule and include Cys 54, the site of reaction of Hg 2. Also, the density of the two residues at the carboxyl terminus is very weak, indicating free motion of these residues in the crystal, and consistent with the observation that they can be removed by carboxypeptidase A with no effect on the enzymatic activity (5). Nevertheless, the overall conformation is unambiguous and is confirmed by the coincidence of three of the five methionines with the chloroplatinite binding sites (the other two methionines are internal) and the two cysteines with the two mercury binding sites (Table 1). In addition, characteristic electron density for many of the side chains can be identified, and there is no obvious incompatibility between the electron density map and the amino acid sequence as determined by chemical methods (2).

In the initial interpretation of the electron density map, markers were placed at each alpha-carbon position. Subsequently, a Kendrew model was constructed using a modified Richards optical comparator (16), in which the electron density sections were mounted parallel to the mirror (12).

Description of the structure

The molecule of T4 phage lysozyme has overall dimensions about 50 × 30 × 30 Å and the general arrangement of the polypeptide backbone is illustrated in Figs. 1 and 2.

As is clear from the figures, the structure consists of two quite distinct lobes, with the carboxy terminal part of the polypeptide chain lying exclusively in the upper domain, while the lower domain contains most of the amino terminal portion of the molecule, although the amino terminal helix interacts extensively with the upper domain and appears to help link the two lobes together.

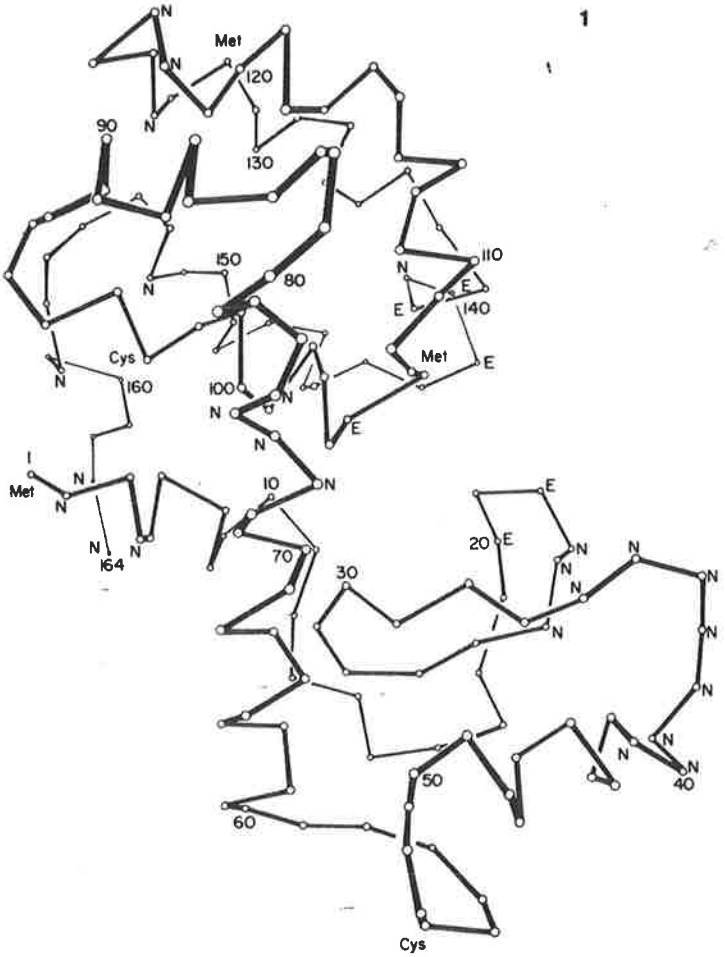
In the lower lobe residues, 18–34 form three strands of somewhat distorted antiparallel pleated sheet, this being the only extended β-structure in the molecule. On the other hand, the upper domain is markedly helical in character and appears to be based on a combination of five helices, each of about 10 residues, arranged so that their axes form the walls of a cylinder surrounding a pronounced hydrophobic core. This hydrophobic region connects through the waist region of the molecule with another hydrophobic region in the lower half of the molecule.

The most obvious connection between the two domains is a long helix of 20 residues which extends almost from one extremity of the molecule to the other. In addition, the two lobes are also connected through the amino terminal helix as mentioned above.

Altogether, about 60% of the molecule is comprised of helices, including residues 3–11, 39–49, 60–79, 82–90, 95–106, 115–123, 129–134, 137–141, and 143–155. Adding residues 18–34, which are in antiparallel β-structure, approximately 70% of the protein is involved in more-or-less regular secondary structure. Spectroscopic studies of the enzyme in solution also indicate a high helix content (M. A. Elwell and J. A. Schellman, personal communication).

Extending across the waist region, roughly from left front to right rear in Figs. 1 and 2, there is a depression in the surface of the molecule which, partway along its length, deepens into what can best be described as a hole, or opening, about 6–8 Å in diameter, extending from one side of the molecule through to the other. In Figs. 1 and 2, the direction of view is approximately normal to the "hole" which is located near the center of the figures. In the three dimensional structure, the polypeptide backbone in the vicinity of Thr 21–Glu 22 is about 8 Å from the backbone in the region of Glu 141–Thr 142 and the respective side chains of these residues extend to within 3–5 Å of each other so that the opening is more occluded than appears in Figs. 1 and 2. The guanido group of Arg 145 also extends across the mouth of the opening to within about 4 Å of the carboxyl of Glu 21, although the two groups do not obviously form a salt link.

This opening is perhaps the most unusual feature of the structure, and is without precedent among the proteins whose three dimensional structures have been determined to date. It remains to be determined whether or not this region of the molecule accommodates either the alternating *N*-acetyl-



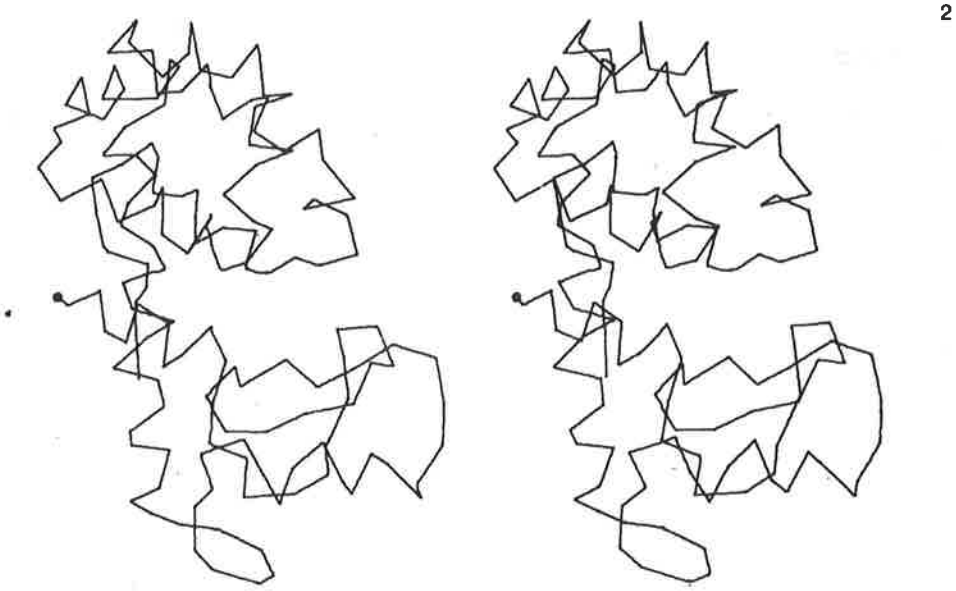


FIG. 1 (top). Perspective drawing illustrating the polypeptide backbone of bacteriophage T4 lysozyme. The approximate position of each alpha carbon atom is indicated by an open circle. Also shown are the methionine and cysteine residues involved in heavy-atom binding. Residues labeled E and N were respectively designated on the basis of genetic studies as "Essential" and "Non-essential" for full catalytic activity, as described in the *text*.

FIG. 2 (bottom). Stereo diagram illustrating the fold of the polypeptide chain in T4 phage lysozyme. The direction of view is close to that in Fig. 1 and the amino terminus is indicated by a small circle.

muramyl-*N*-acetylglucosaminyl portion of the cell wall or the interstrand polypeptide cross-link emanating from each *N*-acetylmuramyl unit (17), but if either of these should prove to be the case, then it seems certain that in order to allow the substrate to enter, the enzyme would have to undergo a fairly substantial conformational change.

Although there is no direct evidence to demonstrate that substrates might bind in the vicinity of the cleft and/or the opening described above, there is indirect evidence from genetic studies, described in the following section, which strongly suggests that this region of the molecule is catalytically important.

Also, there is some chemical evidence to support this assumption. Firstly, reaction of the two sulphydryls with *p*-chloromercuribenzoate (2) does not reduce the enzymatic activity to any appreciable extent, and in the three dimensional structure both cysteines are far from the region of the cleft (Fig. 1). Secondly, the chloroplatinite ion, used to obtain an isomorphous heavy atom derivative, reacts with three methionines, including Met 106 which is located toward one end of the cleft region, and the activity of the $PtCl_4^{2-}$-substituted enzyme is 30% of that of the unsubstituted enzyme (unpublished experiments performed in collaboration with F. W. Dahlquist and A. Y. Maynard), suggesting that the $PtCl_4^{2-}$ group either obstructs the substrate binding site or perturbs the conformation of the enzyme required for optimum catalytic activity.

Mutant lysozymes

The genetics of T4 phage lysozyme have been studied extensively over a number of years by Streisinger, Tsugita and colleagues, and a variety of mutant enzymes have been isolated and characterized (e.g., see refs. 3–5). From the analysis of a series of frame-shaft mutants, Tsugita (5) suggested that the residues Asp 20, Glu 22, Glu 105, Trp 138, Asn 140, and Glu 141 are all essential for full catalytic activity, on the grounds that changes in any of these amino acids drastically reduce the catalytic effectiveness of the enzyme. It is striking to find that each of these residues, indicated by the letter E in Fig. 1, is located in the vicinity of the "hole". In contrast, residues which Tsugita classed as non-essential, denoted N in Fig. 1, are scattered throughout the molecule.

It is also of interest to consider, in the light of the tertiary structure, those mutants in which additional amino acids have been inserted in the polypeptide chain. For example, it is hardly surprising that the replacement of Thr-Glu at positions 21,22 by Lys-Thr-Glu reduces the enzymatic activity almost to zero. One the other hand, it might be expected that the replacement of Ala 74, in the long helix connecting the two lobes of the molecule, by Asp-Val, would have drastic consequences; yet the activity observed for this mutant is almost the same as that for the wild-type enzyme. Obviously, a more detailed study of selected mutants of this type is in order.

Comparison with hen egg-white lysozyme

The lysozymes from T4 bacteriophage and from hen egg-white have similar catalytic activity but non-homologous amino acid sequences (2). It is, therefore, of interest to compare the tertiary structure of the phage enzyme with that of hen egg-white lysozyme determined by Phillips and collaborators (18–20).

Apart from the fact that the respective molecules are both folded into two lobes, there seems to be no similarity whatsoever between their respective three dimensional structures.

It has been suggested by Dunnill (21) that the amino-acid sequences of the phage and hen egg-white lysozymes are correlated and that the two enzymes have an evolutionary relationship, but the present structure determination does not support this hypothesis. For example, using the sequence alignment proposed by Dunnill (21), the respective helices of the two enzymes do not obviously correlate and, in addition, Asp 47 and Glu 64 in T4 phage lysozyme, postulated by Dunnill to be catalytically important, are located away from the presumed substrate binding site and are 20 Å apart.

Comparisons of the structure of α-chymotrypsin with that of subtilisin (22), and carboxypeptidase A with that of thermolysin (12), have shown that enzymes with similar catalytic activity may have quite different tertiary structures, yet have active sites with some elements in common. We have therefore examined the T4 phage lysozyme structure for possible analogues of Glu 35 and Asp 52, the two catalytically important residues in hen egg-white lysozyme (23). Two possible candidates might be Glu 11 and Asp 20, which are located on opposite sides of the opening described above, and have their carboxyl groups about 8 Å apart, comparable with 7 Å between the two carboxyls in the hen egg-white enzyme (20); however, there is an obvious difference in that neither acid group is located in a predominantly nonpolar region, as is Glu 35 in hen egg-white lysozyme. In the phage enzyme, Glu 11 is located in the inner wall of the opening through the molecule and forms a salt link with Arg 145, an interaction for which there is no obvious counterpart in hen egg-white lysozyme. Asp 20 is located in a "hairpin bend" which juts across the mouth of the molecular opening (Fig. 1) and is included in the residues designated as "essential" on the basis of the genetic studies (5).

It may be noted that residue 31, the only histidine in the phage lysozyme, is positioned in the depression between the two domains of the molecule, only a few angstroms from the mouth of the opening. Furthermore, the imidazole of His 31 participates in an ionic interaction with Asp 70 similar to that observed in the serine proteases (24, 25) and thermolysin (26). Whether this potential "charge relay system" plays a catalytic role in T4 lysozyme remains to be determined.

In summary, the three dimensional structures of hen egg-white lysozyme and T4 lysozyme are seen to be quite different but it is not clear at this time whether or not their respective mechanisms of catalysis may be related.

We are particularly grateful to Miss A. Yvonne Maynard for her excellent technical assistance in purifying the phage lysozyme and to Dr. Lynn Ten Eyck for his help in the refinement of the heavy atom parameters. We have benefited from discussions with F. W. Dahlquist, G. Streisinger, J. A. Schellman, and J. (Emrich) Owen, and would also like to thank H. F. Matthews and other colleagues in this Institute for their assistance. This work was supported in part by grants from the National Science Foundation (GB30823X), the National Institutes of Health (GM15423, GM20066, GM00715), and by the award to one of us (B.W.M.) of an Alfred P. Sloan Research Fellowship and a Public Health Service Career Development Award (GM 70585) from the Institute of General Medical Sciences.

1. Tsugita, A., Inouye, M., Terzaghi, E. & Streisinger, G. (1968) *J. Biol. Chem.* **243**, 391–397.
2. Tsugita, A. & Inouye, M. (1968) *J. Mol. Biol.* **37**, 201–212.

3. Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. & Inouye, M. (1966) *Cold Spring Harbor Symp. Quant. Biol.* **31**, 77–84.
4. Ocada, Y., Amagase, S. & Tsugita, A. (1970) *J. Mol. Biol.* **54**, 219–246.
5. Tsugita, A. (1971) in *The Enzymes*, ed. Boyer, P. D. (Academic Press, New York), Vol. 5, 3rd ed., pp. 343–411.
6. Matthews, B. W., Dahlquist, F. W. & Maynard, A. Y. (1973) *J. Mol. Biol.* **78**, 575–576.
7. Matthews, B. W., Klopfenstein, C. E. & Colman, P. M. (1972) *J. Phys. E: Sci. Instrum.* **5**, 353–359.
8. Matthews, B. W., Sigler, P. B., Henderson, R. & Blow, D. M. (1967) *Nature* **214**, 652–656.
9. Hart, R. G. (1961) *Acta Crystallogr.* **14**, 1194–1195.
10. Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961) in *Computing Methods and the Phase Problem in X-ray Crystal Analysis* (Pergamon Press, Oxford), pp. 236–251.
11. Matthews, B. W. (1966) *Acta Crystallogr.* **20**, 230–239.
12. Colman, P. M., Jansonius, J. N. & Matthews, B. W. (1972) *J. Mol. Biol.* **70**, 701–724.
13. Blow, D. M. & Crick, F. H. C. (1959) *Acta Crystallogr.* **12**, 794–802.
14. Matthews, B. W. (1966) *Acta Crystallogr.* **20**, 82–86.
15. Matthews, B. W. (1970) in *Crystallographic Computing*, eds. Ahmed, F. R., Hall, S. R. & Huber, C. P. (Munksgaard, Copenhagen), pp. 146–159.
16. Richards, F. M. (1968) *J. Mol. Biol.* **37**, 225.
17. Weidel, W. & Pelzer, H. (1964) *Advan. Enzymol.* **26**, 193–232.
18. Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1964) *Nature* **206**, 757–763.
19. Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967) *Proc. Roy. Soc. Sect. B* **167**, 365–377.
20. Imoto, T., Johnson, L. N., North, A. C. T., Phillips, D. C. & Rupley, J. A. (1972) in *The Enzymes*, ed. Boyer, P. D. (Academic Press, New York), Vol. 7, 3rd ed., pp. 665–868.
21. Dunnill, P. (1967) *Nature* **215**, 621–622.
22. Robertus, J. D., Alden, R. A., Birktoft, J. J., Kraut, J., Powers, J. C. & Wilcox, P. E. (1972) *Biochemistry* **11**, 2439–2449.
23. Blake, C. C. F., Johnson, L. N., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967) *Proc. Roy. Soc. Sect. B* **167**, 378–388.
24. Blow, D. M., Birktoft, J. J. & Hartley, B. S. (1969) *Nature* **221**, 337–340.
25. Wright, C. S., Alden, R. A. & Kraut, J. (1969) *Nature* **221**, 235–242.
26. Matthews, B. W., Colman, P. M., Jansonius, J. N., Titani, K., Walsh, K. A. & Neurath, H. (1972) *Nature New Biol.* **238**, 41–43.

J. Mol. Biol. (1978) **118**, 81–98

Structure of the Lysozyme from Bacteriophage T4: An Electron Density Map at 2·4 Å Resolution

S. J. Remington, W. F. Anderson, J. Owen, L. F. Ten Eyck
C. T. Grainger† and B. W. Matthews

Institute of Molecular Biology and Departments of Physics and Biology, University of Oregon, Eugene, Oreg. 97403, U.S.A.

(Received 29 March 1977)

The three-dimensional structure of the lysozyme from bacteriophage T4 has been determined from a 2·4 Å resolution electron density map. The map, determined using isomorphous replacement and anomalous scattering differences from platinum and mercury-substituted crystals, reveals the approximate conformation of the molecule unambiguously. Details of the structure determination are presented and the conformation of the molecule is described in detail. The observed properties of those mutant lysozymes for which the changes in the amino acid sequence have been determined are discussed in the light of the three-dimensional structure. In most cases the behavior of the mutant lysozymes can be easily rationalized, but there are a few instances where this is not so.

1. Introduction

Bacteriophage T4 lysozyme is an endoacetylmuramidase. Produced late in the infection of *Escherichia coli* by T4 bacteriophage, lysozyme cleaves the bacterial cell wall, allowing the release of the progeny phage particles (Streisinger *et al.*, 1966; Tsugita *et al.*, 1968; Tsugita, 1971). In a preliminary report (Matthews & Remington, 1974) we described the three-dimensional structure of T4 phage lysozyme as determined by X-ray crystallography from a 2·5 Å resolution electron density map. Subsequently, the interpretation of the original electron density map has been confirmed and the structure determined more accurately by the inclusion of additional data. In this paper we present details of the structure determination and also describe in more detail the conformation of the protein as it is currently known.

2. Crystallization

Crystals of phage T4 lysozyme suitable for X-ray diffraction are most readily obtained by batch crystallization with phosphate as the precipitant, in a manner similar to that described previously (Matthews *et al.*, 1973). Typically, to a portion of the protein solution (20 mg/ml in 0·55 M-NaCl, 14 mM-mercaptoethanol, 1 mM-$MgCl_2$, 0·01 M-sodium phosphate, pH 6·7) one slowly adds 0·8 to 1·1 portions of phosphate solution (2·2 M-NaH_2PO_4, 1·8 M-K_2HPO_4, pH 6·7). The solution is left at 4°C and the crystals grow over a period of one to two months. One problem which we have not been able to eliminate is the tendency of the larger crystals (greater than 0·5 mm^3) to develop irregular cracks. Before X-ray photography, and for storage,

† Permanent address: School of Physics, University of New South Wales, Kensington, N.S.W., Australia.

the crystals were equilibrated with a solution of 1·05 M-K_2HPO_4, 1·26 M-NaH_2PO_4, 0·23 M-NaCl, 1·4 mM-mercaptoethanol (pH 6·7). The phosphate-grown crystals have space group $P3_221$ with cell dimensions $a = b = 61{\cdot}2$ Å $\pm$ 0·1 Å, $c = 96{\cdot}8$ Å $\pm$ 0·2 Å.

The same crystal form has been obtained on occasion by using ammonium sulfate as the precipitant. In a typical procedure, 1·3 portions of saturated ammonium sulfate (pH 7·1), buffered with 0·05 M-phosphate, are added to one portion of the protein solution described above, and left to stand at 4°C. These crystals are well-formed rhombs, similar to those obtained from phosphate, although with somewhat different cell dimensions, namely $a = b = 61{\cdot}1$ Å, $c = 95{\cdot}3$ Å. Crystals were grown from ammonium sulfate on more than one occasion, but this proved difficult to reproduce routinely, and the crystals grew very slowly. For this reason the phosphate-grown crystals were used for the structure analysis.

One disadvantage of the phosphate-grown crystals is that they have more X-ray absorption than do those grown from sulfate. When crystals were transferred from phosphate to sulfate by gradually changing the composition of the mother liquor, the crystals disintegrated within a few hours, suggesting that one or more phosphate ions might participate in intermolecular contacts.

3. Heavy-atom Derivatives

Two isomorphous heavy-atom derivatives were employed for the structure determination, one using $HgCl_2$ which reacted with the two cysteines, and the second using K_2PtCl_4 which interacted with three of the five methionines. The mercury derivative was obtained by soaking the crystals for two days in $2{\cdot}5 \times 10^{-4}$ M-$HgCl_2$ in standard mother liquor, in the absence of mercaptoethanol, and then substituting a solution containing $1{\cdot}0 \times 10^{-6}$ M-$HgCl_2$ prior to data collection. The platinum derivative was obtained by soaking the crystals for five days in $3{\cdot}3 \times 10^{-4}$ M-K_2PtCl_4 in mercaptoethanol-free mother liquor.

For the original electron density map, sets of 12 precession films for each derivative were used to record 92% of the isomorphous differences to a resolution of 2·4 Å, and about half the Bijvoet differences. Prior to the calculation of the present map, the data set was improved by replacing a number of the original films and also expanded in the following manner. In space group $P3_221$ the planes hNl contain pairs of reflections which are related by the 2-fold symmetry axis but are not Bijvoet pairs, whereas the planes h, $h-N$, l contain pairs of Friedel-related reflections. Therefore, in the original data set, Bijvoet differences were obtained for all reflections recorded in the six h, $h-N$, l planes but this was not the case for reflections in the hNl planes. The reflections Bijvoet-related to those in a plane hNl occur in the plane $h\ \bar{N}\ l$. In space group $P3_221$ these complementary planes can be readily distinguished on the precession camera by inspection of the intensity distribution in the $h0l$ zone seen during alignment of the crystal. Thus, by collecting both the hNl and the $h\bar{N}l$ planes, in addition to the h, $h-N$, l planes, all the Bijvoet pairs were obtained. In practice, this meant that the data set for each derivative was expanded from the original 12 films to 17. Although the anomalous scattering differences obtained from the hNl and $h\ \bar{N}\ l$ films suffer in accuracy because they are not obtained from a single film, this is partly offset by the fact that F_+ and F_- are each obtained as the average of two intensity measurements.

TABLE 1

Data processing statistics

Derivative intensity statistics†	Native	$HgCl_2$	K_2PtCl_4
Films	12	17	17
Average R_{sym}	0·048‡	0·074§	0·076§
Average R_{sca}	0·051	0·049	0·051
R_{merge}	0·028	0·037	0·041
Cell dimensions			
a, b (Å)	61·2	61·2	61·1
c (Å)	96·8	96·6	97·1

†The respective R values are as defined previously (Matthews *et al.*, 1972). R_{sym} gives the agreement between symmetry-related reflections recorded on the same film, and R_{sca} the agreement between symmetry-averaged intensities recorded on the stronger and weaker film in a film pack. The averages quoted are for all films in which two symmetry-related reflections were averaged. R_{merge} gives the agreement between structure amplitudes measured on different films.

‡Average of two reflections, symmetry-related for h, N, l films or Bijovet-related for h, $h-N$, l films.

§Average of two symmetry-related reflections, h, N, l planes only.

The precession films were measured with a rotating drum scanner (Matthews *et al.*, 1972) and the data collection statistics are summarized in Table 1. Systematic errors in the measurement of the Bijvoet differences were reduced by the procedure of "local scaling" (Matthews & Czerwinski, 1975).

In space group $P3_221$ each precession film can be indexed absolutely but it is still desirable to check the assignment of the Friedel-related reflections by comparing measurements of anomalous differences which are duplicated on different h, $h-N$, l films, or on different pairs of h, $\pm N$, l films (cf. Colman *et al.*, 1972). For this purpose correlation coefficients

$$C_{ij} = \Sigma\Delta_i\Delta_j / \{\Sigma\Delta_i^2\ \Sigma\Delta_j^2\}^{\frac{1}{2}}$$

TABLE 2

Correlation matrix for isomorphous and anomalous differences: mercury derivative

Film		1	2	3	4	5	6	7	8	9	10	11	12
h, 0, l	1		97	95	96	94	96	93	97	94	94	89	93
h, $\pm$ 1, l	2	—		98	99	95	96	97	96	95	95	95	94
h, $\pm$ 2, l	3	—	45		98	93	96	97	96	92	97	90	91
h, $\pm$ 3, l	4	—	40	40		96	96	98	95	93	95	94	90
h, $\pm$ 4, l	5	—	38	49	54		93	91	91	92	86	88	85
h, $\pm$ 5, l	6	—	6	27	32	15		97	95	90	90	84	92
h, h, l	7	—	77	77	71	61	66		—	—	95	—	—
h, $h-1$, l	8	—	50	48	49	13	46	—		96	—	93	86
h, $h-2$, l	9	—	68	60	38	28	28	—	—		—	82	87
h, $h-3$, l	10	—	45	40	33	19	8	65	—	—		—	—
h, $h-4$, l	11	—	46	32	42	20	22	—	39	—	—		90
h, $h-5$, l	12	—	37	15	26	12	$\underline{-9}$	—	3	34	—	46	

The correlation coefficients for the isomorphous differences are given in the upper right triangle, and those for the anomalous differences in the lower left triangle. All correlations are multiplied by 100 and the one negative correlation is underlined.

TABLE 3

Correlation matrix for isomorphous and anomalous differences: platinum derivative

Film		1	2	3	4	5	6	7	8	9	10	11	12
$h, 0, l$	1		97	96	97	91	92	96	95	96	97	84	94
$h, \pm 1, l$	2	—		98	98	96	95	99	97	96	98	88	97
$h, \pm 2, l$	3	—	43		97	95	96	98	97	94	97	83	96
$h, \pm 3, l$	4	—	43	22		95	95	94	95	95	98	90	95
$h, \pm 4, l$	5	—	33	7	19		95	94	95	94	93	86	91
$h, \pm 5, l$	6	—	31	23	−5	33		98	94	92	95	83	93
h, h, l	7	—	72	35	34	35	37		—	—	97	—	—
$h, h-1, l$	8	—	56	33	23	46	22	—		94	—	87	94
$h, h-2, l$	9	—	34	38	11	23	−16	—	—		—	79	91
$h, h-3, l$	10	—	56	20	29	36	31	87	—	—		—	—
$h, h-4, l$	11	—	43	23	26	26	11	—	57	—	—		88
$h, h-5, l$	12	—	48	30	37	31	39	—	54	50	—	42	

were calculated, where $\Delta_i = (F_+ - F_-)_i$ is the anomalous difference measured on film i, and Δ_j the same quantity measured on film j. The summation is over all anomalous scattering measurements common to films i and j.

The results for the mercury derivative are given in Table 2 and for the platinum derivative in Table 3. It is clear that in each case the anomalous differences are measured reproducibly from film to film, with only 3 of 90 correlations negative, and overall correlation coefficients of 0·35 and 0·29. For comparison, the correlations between the isomorphous differences recorded on different pairs of films were also calculated and are included in Tables 2 and 3. In this case the agreement is striking, with no correlation less than 0·79 for either derivative, and overall coefficients of 0·93 in both cases.

While the correlation coefficients provide an excellent indication of the quality of the data and the reproducibility of the anomalous scattering and isomorphous differences, they do not, of themselves, guarantee the usefulness of a given derivative. In particular, it is easy to imagine a case where the isomorphous correlation could be high, but the derivative non-isomorphous. On the other hand a favorable correlation for the anomalous differences ought to be less ambiguous, since anomalous scattering differences can only be generated by anomalous scatterers in the crystal.

TABLE 4

Heavy-atom parameters

Derivative	Reaction site	Z	x	y	z	B
$HgCl_2$	Cys54	94·2	0·5511	0·6610	0·0246	26·9
	Cys97	87·8	0·5729	0·9661	0·0836	41·7
K_2PtCl_4	Met106	120·6	0·6796	0·3399	0·0344	24·9
	Met120	146·7	0·1268	0·4535	0·1140	94·7
	Met1	102·1	0·7113	0·9525	0·0691	50·3

Z is the occupancy, on an arbitrary scale, of the heavy-atom site; x, y, z the fractional co-ordinates; and B the isotropic thermal parameter in Å^2.

The heavy-atom binding sites were initially located from difference Patterson and difference Fourier projections and subsequently refined by minimizing the lack of closure of the phase triangles. In the refinement the anomalous scattering data were used to improve the phase angles but were not otherwise included in the refinement. The final heavy-atom parameters are given in Table 4 and the refinement statistics are summarized in Table 5 and Figure 1. The overall mean figure of merit was 0·70 and the average ratio of the lack of closure of the phase triangles to E

TABLE 5

Refinement statistics

Parameter	$HgCl_2$	K_2PtCl_4
Number of reflections in refinement	7242	7125
Mean amplitude	466	486
Mean isomorphous difference	151	189
Mean heavy-atom scattering	164	220
Lack of closure, E	110	137
Residual for centric reflections, R_c	0·46	0·44
Mean anomalous difference	42	49
Mean anomalous scattering	27	29
Lack of closure, E'	28	37

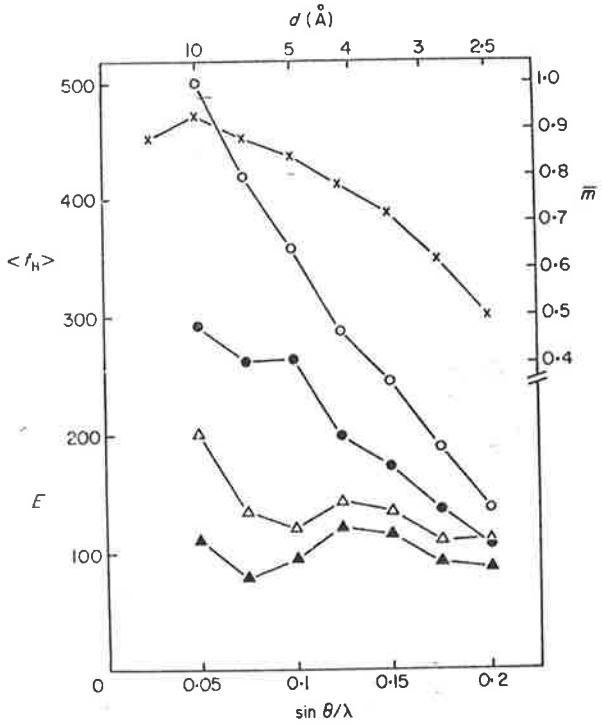


FIG. 1. Refinement statistics for centrosymmetric reflections. The circles give the r.m.s. heavy-atom scattering $\langle f_H \rangle$ and the triangles the r.m.s. lack of closure, E. Filled symbols: K_2PtCl_4; open symbols: $HgCl_2$. The crosses indicate the mean figure of merit, $\bar{m}$, for all reflections.

(or E') was 0·88 (Matthews, 1970). The r.m.s. isomorphous lack of closure error E was estimated from the centrosymmetric data and E' was put equal to the average lack of closure of the anomalous scattering phase triangles at the most probable phase.

The space group was shown to be $P3_221$ rather than its enantiomorph, $P3_121$, by the following test using the *hhl* data, for which the Bijvoet differences are relatively accurate (Tables 2 and 3). Phase angles were calculated for the combined mercury and platinum derivates, including isomorphous and anomalous differences, assuming the space group to be firstly $P3_221$ and secondly $P3_121$. The mean figure of merit was 0·634 in the first case and 0·609 in the second. A more striking indication that the former space group is the correct one is provided by the value of the average lack of closure of the phase triangles divided by E (or E'). This quantity increased from 0·73 to 0·84, the lower value indicating better closure of the phase circles in the correct space group. The choice of space group was independently confirmed by determining phases using the anomalous differences only, and then calculating difference Fourier maps as described in more detail in connection with the use of local scaling (Matthews & Czerwinski, 1975).

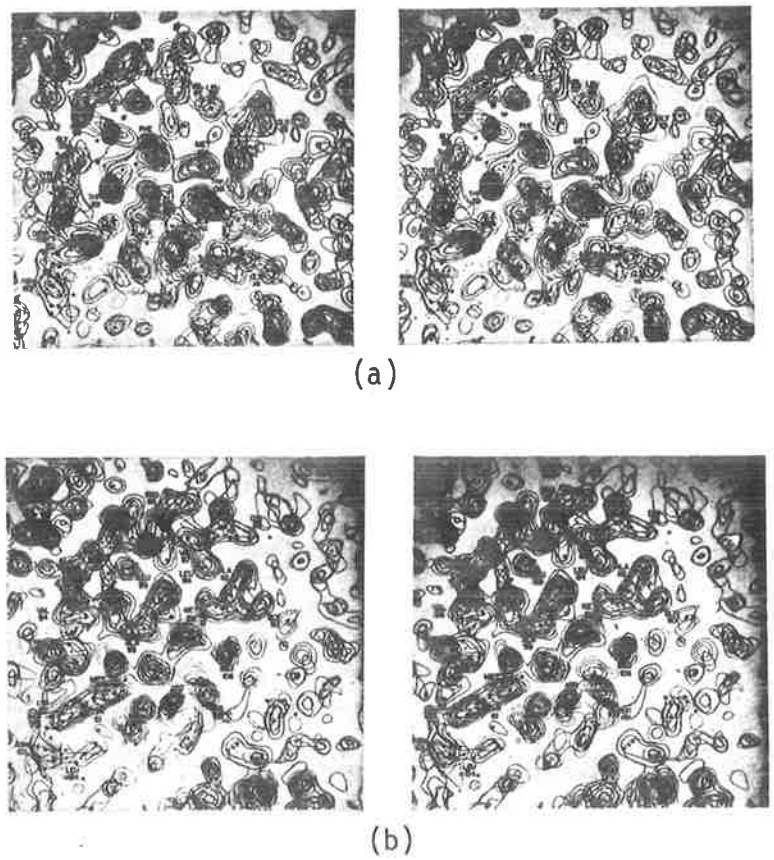


FIG. 2. Representative portion of the phage lysozyme electron density map. (a) Sections $z = -2/132$ to $4/132$. (b) Sections $z = 3/132$ to $9/132$.

4. Electron Density Map

The electron density was calculated and plotted onto transparent sheets in an optical comparator (Richards, 1968; Colman *et al.*, 1972) and a skeletal model built in the usual way using model parts from Cambridge Repetition Engineers.

As the amino acid sequence was known (Tsugita & Inouye, 1968), the electron density map was of sufficient quality that the approximate conformation of most of the molecule could be readily seen. The new map confirms the description of the molecule obtained from the preliminary electron density map (Matthews & Remington, 1974) and in addition shows the folding of the polypeptide chain in the vicinity of residues 50 to 56, which was initially uncertain, much more clearly. Representative portions of the electron density map are shown in Figure 2.

As portions of the model were completed, initial atomic co-ordinates were obtained by placing markers in the electron density map at the presumed atomic sites. The advantages of this procedure have been described previously (Matthews *et al.*, 1974). These "raw" co-ordinates were then checked and idealized by using a program written by L. F. Ten Eyck in which stereochemical bond length, bond angle and planarity constraints are enforced by constraining the interatomic distances between selected pairs of atoms (Ten Eyck *et al.*, 1976; Dodson *et al.*, 1976). The co-ordinates have been presented elsewhere (Remington *et al.*, 1977) and deposited at the Brookhaven Data Bank.

5. Description of the Molecule

The general description of the molecule given previously (Matthews & Remington, 1974) is confirmed by the improved electron density map. The molecule (Figs 3 and 4)

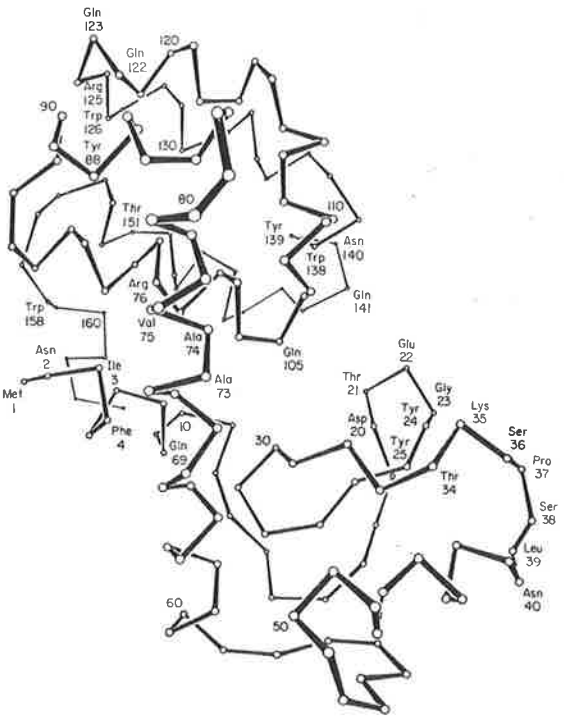


FIG. 3. Schematic diagram showing the polypeptide backbone of T4 phage lysozyme. The abeled residues are those which have been substituted in one or more mutant enzymes.

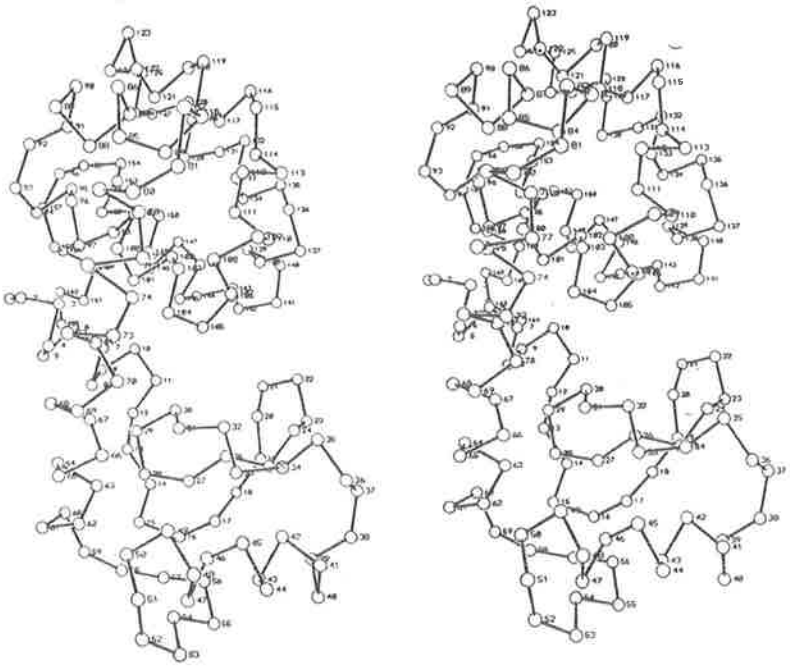


FIG. 4. Stereo diagram showing the conformation of the lysozyme molecule.

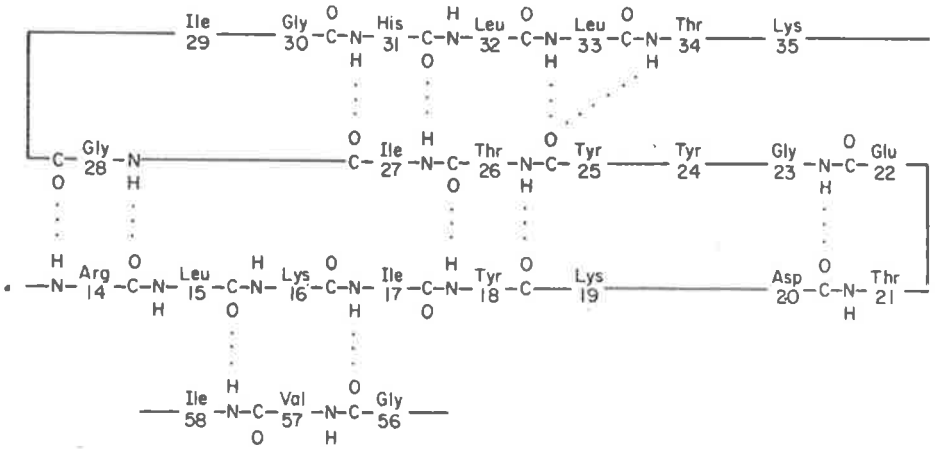


FIG. 5. Schematic illustration of the β-sheet region of T4 phage lysozyme.

is folded into two distinct lobes, joined by a long helix, with the amino terminal lobe incorporating an irregular region of extended antiparallel β-sheet and some helix, and the carboxyl terminal lobe consisting predominantly of helices surrounding a pronounced hydrophobic core. The hydrogen bonding observed in the region of β-sheet is illustrated in Figure 5. As is shown in the Figure, the extended β-structure consists primarily of three more-or-less extended strands of polypeptide chain, although an additional pair of hydrogen bonds is made to the backbone between Gly56 and Ile58 which can therefore be considered as a fourth strand.

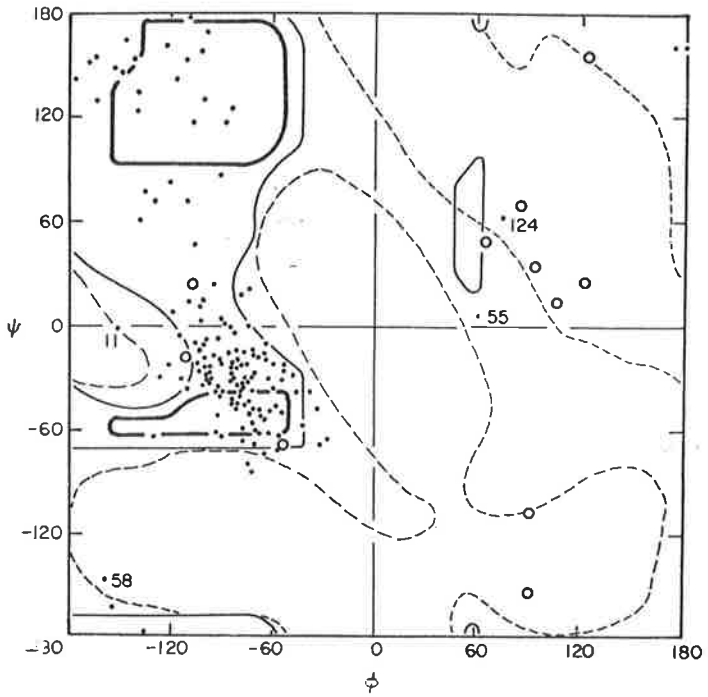


FIG. 6. Conformation angles for T4 phage lysozyme; glycine residues are indicated by open circles. The allowed regions for a hard-sphere model (Ramachandran & Sasisekharan, 1968) are indicated by solid lines, and those predicted by a quantum mechanical method (Pullman *et al.*, 1970) are indicated by a broken line.

The torsion angles (ϕ, ψ) of the polypeptide backbone are listed in Table 6 and plotted in Figure 6. These angles were calculated for the stereochemically idealized, but otherwise unrefined raw co-ordinates described above. As such, the torsion angles should be regarded as unrefined, with an estimated error of $\pm$ 30°. It is possible that one or two (ϕ, ψ) values could be radically in error, especially in the vicinity of the carboxyl terminal residues Lys162–Leu164 where the electron density is weak (Fig. 2(b)). Other than glycine residues, only Asn55 and Lys124 have conformations with ϕ near 60°. The distribution of (ϕ, ψ) values shown in Figure 6 illustrates rather strikingly the strongly helical character of T4 phage lysozyme.

The residues in α-helices in T4 phage lysozyme are listed in Table 7. The criteria for inclusion in an α-helix are that a residue has (ϕ, ψ) values approximating those of an α-helix (i.e. — 57, — 47), and also forms at least one α-helical-type hydrogen bond.

TABLE 6

Dihedral angles

1	MET	0.0	0.0	61	ASP	-69.4	-39.4	121	LEU	-104.8	-19.7
2	ASN	178.9	159.5	62	GLU	-91.2	-39.2	122	GLN	-74.6	-31.2
3	ILE	-71.1	-31.4	63	ALA	-60.1	-37.1	123	GLN	-103.5	12.5
4	PHE	-71.4	-55.7	64	GLU	-79.7	-23.0	124	LYS	74.8	63.4
5	GLU	-83.5	19.9	65	LYS	-73.5	-50.1	125	ARG	-121.2	84.6
6	MET	-113.5	-31.6	66	LEU	-81.4	-30.8	126	TRP	-72.4	-83.9
7	LEU	-56.6	-24.8	67	PHE	-67.3	-62.8	127	ASP	-39.1	-36.5
8	ARG	-77.5	-34.1	68	ASN	-65.1	-16.4	128	GLU	-115.3	-4.4
9	ILE	-105.9	-10.3	69	GLN	-82.0	-46.7	129	ALA	-104.7	-11.9
10	ASP	-103.0	-14.4	70	ASP	-93.4	-37.0	130	ALA	-74.9	1.7
11	GLU	-154.6	-1.9	71	VAL	-80.3	-19.7	131	VAL	-120.7	-22.8
12	GLY	52.5	-153.2	72	ASP	-94.6	-16.5	132	ASN	-93.6	-6.5
13	LEU	-163.8	129.0	73	ALA	-84.5	-44.0	133	LEU	-100.0	-23.5
14	ARG	-138.7	134.3	74	ALA	-71.4	-67.3	134	ALA	-72.6	-15.9
15	LEU	-109.8	13.8	75	VAL	-45.3	-39.1	135	LYS	-106.5	48.3
16	LYS	-153.5	147.6	76	ARG	-83.7	-28.1	136	SER	-138.1	-175.3
17	ILE	-80.2	136.3	77	GLY	-54.5	-68.6	137	ARG	-77.6	-34.7
18	TYR	-153.1	-163.6	78	ILE	-53.5	-50.1	138	TRP	-67.2	-50.8
19	LYS	-122.6	160.5	79	LEU	-53.9	-31.5	139	TYR	-80.0	-19.2
20	ASP	-102.7	157.2	80	ARG	-100.4	13.8	140	ASN	-79.1	-34.9
21	THR	-64.4	-19.6	81	ASN	-139.3	125.2	141	GLN	-103.5	-25.4
22	GLU	-118.4	6.9	82	ALA	-58.9	-21.6	142	THR	-130.3	73.4
23	GLY	106.7	14.6	83	LYS	-98.7	-31.6	143	PRO	-43.8	-55.5
24	TYR	-111.1	152.1	84	LEU	-93.4	-60.5	144	ASN	-75.3	-80.1
25	TYR	-100.3	129.5	85	LYS	-34.3	-45.5	145	ARG	-48.9	-26.0
26	THR	-141.1	152.9	86	PRO	-75.0	-30.6	146	ALA	-58.1	-61.4
27	ILE	-147.8	144.6	87	VAL	-70.3	-27.3	147	LYS	-88.6	3.3
28	GLY	92.3	-106.9	88	TYR	-75.8	-44.7	148	ARG	-78.5	-33.1
29	ILE	-135.4	76.2	89	ASP	-84.9	-13.4	149	VAL	-99.7	-8.6
30	GLY	66.6	48.8	90	SER	-83.6	-27.6	150	ILE	-101.5	-33.7
31	HIS	-106.8	117.2	91	LEU	-100.1	170.5	151	THR	-76.3	-30.1
32	LEU	-87.5	116.8	92	ASP	-109.3	177.1	152	THR	-63.1	-59.9
33	LEU	-89.8	-65.3	93	ALA	-77.5	-65.2	153	PHE	-53.4	-24.6
34	THR	-175.6	142.1	94	VAL	-54.0	-20.4	154	ARG	-74.9	-44.2
35	LYS	-82.6	-3.6	95	ARG	-85.4	-14.5	155	THR	-126.6	-28.7
36	SER	-84.6	125.2	96	ARG	-81.7	-23.5	156	GLY	93.3	35.6
37	PRO	-74.6	21.3	97	CYH	-85.0	-29.4	157	THR	-168.5	153.9
38	SER	-116.2	141.1	98	ALA	-64.9	-52.8	158	TRP	-111.4	72.2
39	LEU	-99.9	-18.9	99	LEU	-71.3	-36.6	159	ASP	-103.0	3.1
40	ASN	-55.6	-27.2	100	ILE	-62.4	-47.0	160	ALA	-92.8	-51.5
41	ALA	-55.3	-26.5	101	ASN	-70.4	-17.5	161	TYR	-98.5	25.0
42	ALA	-82.1	-14.0	102	MET	-72.1	-45.9	162	LYS	-31.6	-68.2
43	LYS	-90.1	-38.7	103	VAL	-76.9	-15.4	163	ASN	-130.9	-62.6
44	SER	-57.3	-46.1	104	PHE	-77.1	-57.8				
45	GLU	-74.8	-22.5	105	GLN	-81.5	-12.3				
46	LEU	-64.6	-73.8	106	MET	-110.6	-35.3				
47	ASP	-64.5	-10.8	107	GLY	123.2	156.4				
48	LYS	-92.9	-35.3	108	GLU	-68.9	-48.3				
49	ALA	-69.5	-61.0	109	THR	-90.6	5.8				
50	ILE	-100.6	-11.9	110	GLY	-111.0	-16.3				
51	GLY	84.4	70.4	111	VAL	-98.3	-24.6				
52	ARG	172.1	160.4	112	ALA	-79.0	19.6				
53	ASN	-91.6	86.9	113	GLY	-108.0	27.4				
54	CYH	-81.2	-45.5	114	PHE	-138.2	61.8				
55	ASN	61.1	5.5	115	THR	-63.2	-28.8				
56	GLY	123.2	26.5	116	ASN	-98.6	-29.3				
57	VAL	-163.1	152.8	117	SER	-81.6	-32.8				
58	ILE	-158.8	-147.1	118	LEU	-57.1	-70.0				
59	THR	-147.2	163.8	119	ARG	-27.9	-62.6				
60	LYS	-67.8	-17.7	120	MET	-84.1	0.0				

The residues shown in parentheses in Table 7 essentially continue the helix, but have conformations nearer those of the 3_{10} or α_{II} helices (i.e. -49, -26 or -93, -18) and do not appear to form the requisite hydrogen bonds. It should be emphasized that in no case does an "α-helix" appear to be an idealized Pauling–Corey α-helix. As in other proteins, helices are more-or-less distorted, and for many residues the carbonyl oxygen tilts away from the axis resulting in ϕ, ψ values distributed around (-80, -30) rather than (-57, -47). This tendency is readily seen in Figure 6. Counting the residues designated in Table 7 as being in α-helices, T4 phage lysozyme is 60% helical, which is slightly higher than the value of 51% estimated by circular dichroism for the enzyme in solution (Elwell & Schellman, 1975).

TABLE 7

Helices

Residues included	Type of helix
3–10–(11)	α, α_{II}
39–49–(50)	α, α_{II}
60–79–(80)	α, α_{II}
82–90	α
93–106	α
(108–113)	Approx. α_{II}
115–122–(123)	α, α_{II}
126–134	Approx. α
137–141	α
143–155	α

There are several "hairpin bends" in the phage lysozyme structure, the most obvious being at the ends of the extended β-sheet described above. These and other sets of three linked peptides in which a hydrogen bond appears to be formed between the first and third peptide unit are listed in Table 8. The different bends are also categorized according to the classification of Venkatachalam (1968). In the type I bends, there is no restriction on the residues in the bend, whereas in the type I′ bend the third residue is expected to be glycine, and this is in fact the case (Gly56).

TABLE 8

Hairpin bends

Residue number	Sequence	Type
20–23	Asp–Thr–Glu–Gly	I
28–29–30	Gly–Ile–Gly	Irregular
54–57	Cys–Asn–Gly–Val	I′
159–162	Asp–Ala–Tyr–Lys	I

6. Secondary Structure Prediction

As described in detail elsewhere (Matthews, 1975), T4 phage lysozyme was used as a test for the various methods which have been proposed to predict the secondary structure of proteins from their amino acid sequences. Since the locations of secondary structure described above differ slightly from those assumed previously, we have

recalculated the agreement between prediction and observation for the various methods (Table 9). The two measures of agreement, Q and C, are as used previously. Q_α, originally defined by Chou & Fasman (1974), is a quotient of agreement between the predicted locations of α-helical residues and the observed locations. Q_β is used for β-sheet predictions and Q_t for turns. Q has an expected value of 100% for a perfect prediction and 50% for a random prediction. The correlation coefficient C_α (Matthews, 1975) gives the correlation between residues predicted as being in helices and those actually observed in helices, and ranges from 1·0 for a perfect prediction through 0 for a random prediction to − 1·0 for a completely incorrect prediction. The correlation coefficients C_α and agreement quotients Q_α for helix prediction given in Table 9 are very similar to those quoted previously, and require no additional comment.

TABLE 9

Secondary structure predictions

Method†	Helix		Sheet		Turn	
	Q_α	C_α	Q_β	C_β	Q_t	C_t
Barry & Friedman	58	0·17	62	0·19	—	—
Burgess *et al.*	68	0·39	46	−0·10	74	0·32
Chou & Fasman	70	0·42	68	0·24	63	0·18
Finkelstein *et al.*	64	0·28	53	0·07	68	0·31
Guzzo	64	0·29	—	—	—	—
Leberman	63	0·28	—	—	—	—
Lim	60	0·20	53	0·07	—	—
Nagano & Hasegawa	62	0·24	60	0·18	59	0·11
Prothero	63	0·25	—	—	—	—
Ptitsyn & Finkelstein	63	0·26	48	−0·07	—	—
Schellman	71	0·42	76	0·40	—	—

†References are given by Matthews (1975).

In the case of the β-sheet it was previously required (Matthews, 1975) that a residue be approximately in the extended conformation and form at least one sheet–hydrogen bond to be designated as in β-sheet, but if one requires only the formation of the hydrogen bonds, then the following residues may be classed as belonging to β-sheet: 14, 16, 18 to 20, 23 to 28, 31 to 34, 57 to 58 (cf. Fig. 5). The agreement indices between prediction and observation for this assignment of β-structure (Table 9) are generally slightly higher than those quoted previously (Matthews, 1975).

The correlation coefficients for the turn predictions, using the assignments in Table 8, are also given in Table 9. In the original assignments (Matthews, 1975), two "bends" were included at 92 to 95 and 126 to 129 on the grounds that in each case a "1–3" hydrogen bond appeared to be formed at these loci, and the conformation was similar to that of a type III bend (i.e. a 3_{10} conformation). The dihedral angles obtained from the present model (Table 6) are consistent with the previous assignments, but it now appears that residues 92 to 95 and 126 to 129 are better considered as forming part of the (distorted) α-helices 93 to 106 and 126 to 134. This accounts for the differences between the values of C_t and Q_t in Table 9 and those reported previously (Matthews, 1975).

It should be emphasized that the reassignments made here in the location and extent of the secondary structure of T4 phage lysozyme do not arise from major

changes in the assumed conformation of the molecule, but rather are due to the subjective definition of secondary structure, and the fact that observed protein conformations often do not obviously belong in one or another structural category.

As was concluded previously (Matthews, 1975), and is apparent from Table 9, the use of T4 phage lysozyme as a test case indicates that current predictive methods can be quite misleading, and that all methods leave a great deal to be desired.

7. Mutant Lysozymes

A number of years ago Streisinger developed a technique for identifying and isolating phage particles with mutations in the lysozyme gene, and subsequently this capability has been used to demonstrate *in vivo* the nature of the genetic code and to study the molecular basis of mutation (e.g. see Streisinger *et al.*, 1966; Ocada *et al.*, 1970; Okada *et al.*, 1972; Tsugita, 1971).

During these studies, many mutations in the lysozyme gene have been identified, and in a number of cases the resultant changes in the amino acid sequence of the protein have been determined. Since any such change in the amino acid sequence of a protein might modify its stability and/or catalytic activity, it is of interest to examine the properties of the mutant lysozymes, in so far as they are known, in terms of the three-dimensional structure of the enzyme.

Figure 7 summarizes the majority of the known amino acid replacements in mutant phage lysozymes. This Figure is adapted from the previous summary of Tsugita (1971) and also includes unpublished observations of J. (Emrich) Owen and G. Streisinger. The number in parenthesis following the mutant name gives the approximate percentage activity of the mutant lysozyme relative to wild-type lysozyme. The activities have been estimated under variable conditions, in some cases for the purified protein and in others for the crude lysate, and should be taken as no more than a rough indication of activity (Tsugita, 1971). The letter (W) indicates a lysozyme with activity roughly comparable with wild-type lysozyme and (X) indicates very low activity, these estimates being based on halo size. In general "low activity" may be due either to low intrinsic activity or to instability of the lysozyme. Figure 7 also indicates those residues which occur in regular secondary structure, and shows whether a given residue is internal, external or partially exposed. For convenience the different mutant lysozymes will be discussed in groups according to their location in the amino acid sequence.

(a) *Residues 1 to 5*

Several mutant lysozymes with an insertion in the amino terminal region of the molecule have been found to have essentially normal activity, indicating that considerable flexibility is allowed in the makeup of this region of the structure. This is consistent with the fact that the amino terminal pentapeptide lies across the surface of the molecule largely exposed to the solvent. On the other hand, deletion of Phe4 substantially reduces the catalytic effectiveness of the enzyme. This mutant phage, when plated in the usual way, incubated at 37°C for several hours, and then exposed to chloroform vapors gives a very small halo. G. Streisinger has suggested to us that halo size reflects *stability* of the phage lysozyme rather than intrinsic *activity*, because the formation of a large (wild-type) halo necessitates that the enzyme remain active during several hours incubation at 37°C. Thus it can be argued that the result of deleting Phe4 is in part to destabilize the lysozyme molecule, which is

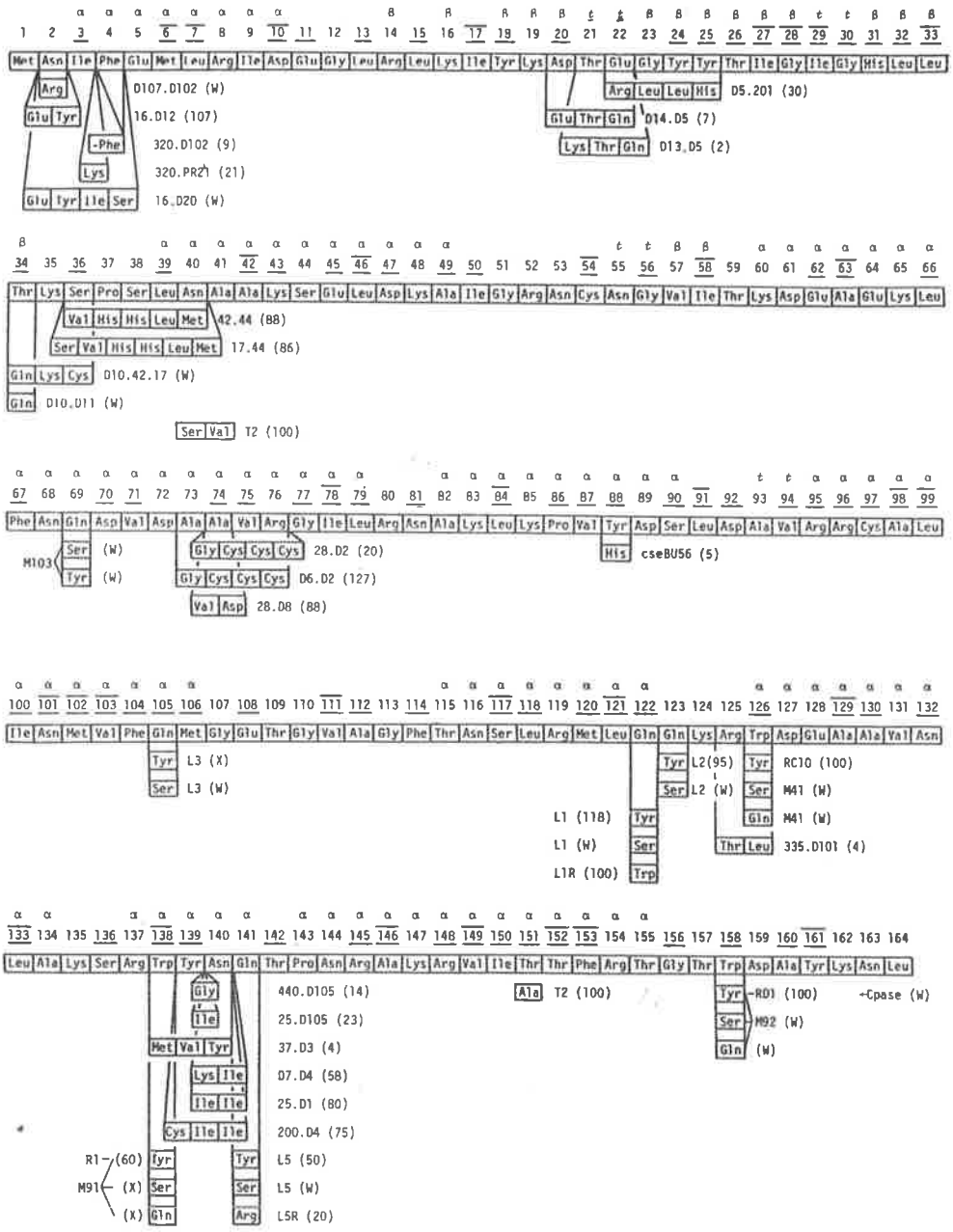


FIG. 7. Mutant lysozymes. The location and identity of the respective mutations are shown; L or C indicates an ochre mutant and M an amber mutant. Specific activities relative to wild-type lysozyme are shown in parentheses (see text); (W) indicates activity similar to wild-type and (X) very low activity. The residues located in helices, β-sheet or turns are labeled, respectively, α, β or t. Where the amino acid side-chain is partly exposed to solvent the residue number is underlined, while residues which are essentially internal are indicated with a line above and below the number. Residues with unmarked numbers are essentially fully exposed to the solvent.

consistent with the location of the amino terminus, far from the active site. Presumably the destabilization arises not so much from the loss of Phe4, which is fully exposed to solvent, as the consequential replacement of the partially buried Ile3, which contributes to the hydrophobic core of the molecule, by the hydrophilic side-chain of Asn2.

(b) *Residues 20 to 25*

Residues 20 to 25 comprise a hairpin bend within the β-sheet region described above. This loop extends across the mouth of a pronounced cleft in the molecule, forming a hole, or opening, about 6 to 8 Å in diameter, extending from one side of the molecule to the other (Matthews & Remington, 1974). In unpublished crystallographic experiments we have shown that *N*-acetyl glucosamine and tri-*N*-acetyl glucosamine bind in this cleft, confirming it to be the active site region.

The fact that the lysozyme from the mutant JD5J201: Glu-Gly-Tyr-Tyr → Arg-Leu-Leu-His(22 to 25) has moderately high activity suggests that none of these residues, in particular Glu22, is involved directly in catalysis. The conformation of Gly23 is such that a leucine molecule could be accommodated at this position without alteration of the protein structure. Ocada *et al.* (1970) reported that the lysozyme from the mutant JR14JD5: Asp-Thr-Glu → Glu-Thr-Gln(20 to 22) has low activity, and argued that since Glu22 was not involved in catalysis (for the reasons described above), Asp20 must have an important role in enzymic activity. We have confirmed that the purified lysozyme from the mutant JR14JD5 has low activity, and note that the halos formed by this mutant are moderately large, suggesting that the reduction in activity is not due to destabilization of the enzyme structure. These observations, together with the location of Asp20 in the active site cleft, provide compelling evidence that this residue is essential for catalysis.

There are a number of reasons why lysozyme from mutant JD13JD5: Thr-Glu → Lys-Thr-Glu(21 to 22) might have low activity, as reported by Ocada *et al.* (1970), the most obvious being that the insertion of the additional residue would cause significant conformational changes in the active site region.

(c) *Residues 31 to 41*

Residues 34 to 41 form an open irregular loop on the surface of the molecule not involved in any obvious structural or catalytic role. As can be seen in Figure 7, substantial changes can be made to the amino acid sequence in this region of the molecule, including, in one instance, an insertion of an extra residue, with little effect on the function of the enzyme. One of the differences in the amino acid sequences of the lysozymes from phage T4 and T2, namely Asn-Ala → Ser-Val(40, to 41) occurs in this region of the molecule (Inouye & Tsugita, 1968). The other altered residue, Thr → Ala(151), also occurs on the surface of the molecule.

(d) *Residues 69 to 76*

One of the most obvious structural elements in the phage lysozyme molecule is the long helix, including residues 60 to 79, which extends practically from one end of the molecule to the other (Fig. 3). Since it might be expected that this helix would be important in preserving the relative alignment of the two lobes of the molecule it was surprising to note that lysozymes from the mutants J28JD2:

Ala-Val-Arg → Gly-Cys-Cys-Cys(74 to 76) and J28JD8: Ala → Val-Asp(74), which resulted in the addition of an extra residue in the middle of this long helix, had been reported by Imada *et al.* (1970) to be moderately stable and quite active. In fact J28JD8 was reported to be almost as stable as the wild-type enzyme, and to have comparable activity. Because the reported properties of J28JD8 are difficult to reconcile with the observed three-dimensional structure we are currently reinvestigating the properties of this mutant lysozyme, and in our hands find that the halos obtained with this mutant phage are extremely small, suggesting that the mutant enzyme is very unstable. This behavior is what might be expected following an insertion in a region of essential secondary structure.

Gln69 is in the long helix described above, but is fully exposed to solvent, so that the substitution of Ser or Tyr at this locus would not be expected to significantly affect the properties of the lysozyme molecule.

(e) *Residue 88*

Liebscher and co-workers have recently isolated a phage T4 mutant with a cold-sensitive phenotype. The mutant phage grows normally at 37°C but at 20°C the efficiency of plating is reduced by about four orders of magnitude relative to wild-type. Liebscher *et al.* (1974*a*,*b*) showed that the lysozyme from the mutant strain cseBU56 was modified at position 88 where a tyrosine was replaced by a histidine, and attributed the cold-sensitivity to the increased activation energy of the modified enzyme. However, we have found that their strain contains two unlinked mutations, one in the lysozyme gene and another in an unidentified gene. We separated the mutations by a genetic cross. The recombinant carrying the latter mutation is cold-sensitive; it produces no plaques at room temperature and turbid plaques at 37°C. The recombinant carrying the lysozyme mutation produces lysozyme with moderate activity at room temperature, 37°C and 43°C as well. This is consistent with the three-dimensional structure of the lysozyme molecule, since there is no obvious reason why the substitution of a histidine for Tyr88 should significantly alter the activity of the molecule.

(f) *Residue 105*

Gln105 is located in the wall of the active site cleft and, as judged from the observed location of binding of *N*-acetyl glucosamine and tri-*N*-acetyl glucosamine (unpublished experiments), makes van der Waals' contact with a cell wall substrate when it binds to the enzyme. The fact that the mutant lysozyme with Gln105 replaced by a serine is fully active suggests that Gln105 is not essential for catalysis, but may be involved in substrate binding. On the other hand the mutant with Gln replaced by Tyr has almost zero activity, which is no doubt due to steric hindrance of the substrate by the bulky phenolic group.

(g) *Residues 122 to 127*

Residues 122 to 127 include the terminal residues of two helices, and the short connecting polypeptide between them. The segment is on the surface of the molecule and is remote from the active site. Therefore it is not surprising to find that point mutations at Gln122, Gln123 or Trp126 do not significantly alter the enzymatic activity. On the other hand it is not clear why mutant 335D101: Arg-Trp →

Thr-Leu(125 to 126) has rather low activity since the side-chain of Arg125 juts out from the surface of the molecule, and point mutations at Trp126 seem to have little effect.

(h) *Residues 138 to 141*

Residues 137 to 141 form a very short helix of about 1·5 turns which connects two longer helices. Trp138 is almost completely buried in the wall of the active site cleft, whereas Tyr139, Asn140 and Gln141 are more exposed to solvent. As can be seen in Figure 3, all four residues are located on the edge of the active site cleft, although all four are at least 5 Å from the presumed location of the saccharide portion of a bound substrate. Apparently the conformation of this short helix is not critical, since mutant lysozymes with a glycine or leucine inserted between Tyr139 and Asn140 retain some activity.

Elwell & Schellman (manuscript in preparation) have carried out a detailed thermodynamic study of the stability of mutant eRl in which Trp138 is replaced by a tyrosine, and also the triple mutant eRRR in which all three tryptophans, at positions 126, 138 and 158, are substituted by tyrosine. They found that Trp138 has a crucial role in maintaining T4 lysozyme in a stable, active form. Replacement of this tryptophan by tyrosine significantly decreases the thermal stability of the enzyme, but this is not the case for the tryptophans at positions 126 and 158. These observations are consistent with the fact that Trp138 is largely buried, whereas the other two tryptophans are largely exposed. Elwell and Schellman also studied the guanidine-denaturation of wild-type and tryptophan-substituted lysozymes, but in this case the results were less easy to interpret. It was found that wild-type and eRl lysozymes were almost equally stable in the presence of guanidine·hydrochloride, whereas the triple mutant eRRR was significantly less stable.

(i) *Residues 158 to 164*

Trp158 is located far from the active site, partly exposed to the solvent and partly contributing to the hydrophobic core in the amino terminal lobe of the molecule. It is not surprising that a tyrosine in this position could play a similar role, and Elwell & Schellman (manuscript in preparation) have shown that this substitution does not significantly affect the stability of the molecule. It might be anticipated that mutants with smaller side-chains at this position would be less stable; however, as shown in Figure 7, the substitution of serine or glutamine for Trp158 does not have a large effect on the enzymatic activity. The carboxy terminal residues 162 to 164 are fully exposed to solvent, and the last two residues, in particular, appear to be somewhat disordered in the crystal, so it is not surprising that the removal of the terminal dipeptide with carboxypeptidase A has little effect on the enzyme.

We are particularly grateful for the excellent technical assistance of A. Y. Maynard, J. Straub and K. Rine; also we thank R. Reynolds for help with photography, and our colleagues M. Elwell, F. W. Dahlquist, C. Schellman, J. Schellman and G. Streisinger for many helpful discussions.

Mutant strains eJR14eJD5 and eJ28eJD8 were kindly provided by Drs A. Tsugtia and M. Eda, and cseBU56 by Dr D.-H. Liebscher.

This study was supported in part by grants from the National Institutes of Health (GM21967, GM19795, GM20066), the National Science Foundation (PCM75–18407), and a Public Health Service Career Development Award to one of us (B.W.M.) (GM 70585) from the Institutes of General Medical Science.

REFERENCES

Chou, P. Y. & Fasman, G. D. (1974). *Biochemistry*, **13**, 222–245.
Colman, P. M., Jansonius, J. N. & Matthews, B. W. (1972). *J. Mol. Biol.* **70**, 701–724.
Dodson, E. J., Isaacs, N. W. & Rollett, J. S. (1976). *Acta Crystallogr.* **32**, 311–315.
Elwell, M. & Schellman, J. (1975). *Biochim. Biophys. Acta*, **386**, 309–323.
Imada, M., Inouye, M., Eda, M. & Tsugita, A. (1970). *J. Mol. Biol.* **54**, 199–217.
Inouye, M. & Tsugita, A. (1968). *J. Mol. Biol.* **37**, 213–223.
Liebscher, D.-H., Dauberschmidt, R., Rapoport, S. M., Heitmann, P. & Rosenthal, H. A. (1974*a*). *Mol. Gen. Genet.* **132**, 321–333.
Liebscher, D.-H., Dauberschmidt, R., Rapoport, S. M., Heitmann, P. & Rosenthal, H. A. (1974*b*). *FEBS Letters*, **44**, 173–177.
Matthews, B. W. (1970). In *Crystallographic Computing* (Ahmed, F. R., ed.), pp. 144–159, Munksgaard, Copenhagen.
Matthews, B. W. (1975). *Biochim. Biophys. Acta*, **405**, 442–451.
Matthews, B. W. & Czerwinski, E. W. (1975). *Acta Crystallogr. sect. A*, **31**, 480–487.
Matthews, B. W. & Remington, S. J. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 4178–4182.
Matthews, B. W., Klopfenstein, C. E. & Colman, P. M. (1972). *J. Phys. E: Sci. Instrum.* **5**, 353–359.
Matthews, B. W., Dahlquist, F. W. & Maynard, A. Y. (1973). *J. Mol. Biol.* **78**, 575–576.
Matthews, B. W., Weaver, L. H. & Kester, W. R. (1974). *J. Biol. Chem.* **249**, 8030–8044.
Ocada, Y., Amagase, S. & Tsugita, A. (1970). *J. Mol. Biol.* **54**, 219–246.
Okada, Y., Streisinger, G., Owen, J., Newton, J., Tsugita, A. & Inouye, M. (1972). *Nature (London)*, **236**, 338–341.
Pullman, B., Maigret, B. & Perahia, D. (1970). *Theor. Chim. Acta*, **18**, 44–56.
Ramachandran, G. N. & Sasisekharan, V. (1968). *Advan. Protein Chem.* **23**, 283–438.
Remington, S. J., Ten Eyck, L. F. & Matthews, B. W. (1977). *Biochem. Biophys. Res. Commun.* **75**, 265–270.
Richards, F. M. (1968). *J. Mol. Biol.* **37**, 225–230.
Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. & Inouye, M. (1966). *Cold Spring Harbor Symp. Quant. Biol.* **31**, 77–84
Ten Eyck, L. F., Weaver, L. H. & Matthews, B. W. (1976). *Acta Crystallogr. sect. A*, **32**, 349–350.
Tsugita, A. (1971). In *The Enzymes* (Boyer, P. D., ed.), vol. 5, 3rd edit., pp. 343–411, Academic Press, New York.
Tsugita, A. & Inouye, M. (1968). *J. Mol. Biol.* **37**, 201–212.
Tsugita, A., Inouye, M., Terzaghi, E. & Streisinger, G. (1968). *J. Biol. Chem.* **243**, 391–397.
Venkatachalam, C. M. (1968). *Biopolymers*, **6**, 1425–1436.

(Reprinted from Nature, Vol. 277, No. 5698, pp. 667-669, February 22 1979)
© Macmillan Journals Ltd., 1979

Molecular basis of thermostability in the lysozyme from bacteriophage T4

MOST proteins are denatured at temperatures above 50–60 °C, although some enzymes, especially those from thermophilic organisms, remain active at temperatures up to 80–90 °C. The determination of the three-dimensional structure of the thermostable protease thermolysin showed that heat-stable proteins do not contain unusual structural features absent from less stable proteins[1,2]. Furthermore, the amino acid sequences of similar proteins from both mesophilic and thermophilic sources have been shown to be homologous, suggesting that the respective structures are similar[3,4]. Nevertheless, such homologous amino acid sequences also include many differences which obscure those amino acid changes actually responsible for differences in thermostability. We report here the structure of a temperature sensitive (ts) mutant of T4 phage lysozyme. This permits the first direct comparison of two protein structures in which all differences are directly related to a change in thermal stability. It is shown that, except for the replacement of a partially exposed arginine by a histidine, the three-dimensional structure of the ts lysozyme is virtually identical with that of native lysozyme.

Bacteriophage T4 lysozyme, an endoacetylmuramidase, is produced late in the infection of *Escherichia coli* by T4 bacteriophage. It cleaves the bacterial cell wall, allowing the release of the progeny phage particles[5]. The amino acid sequence of the lysozyme has been determined[6], crystals suitable for detailed X-ray study have been obtained, the three-dimensional structure determined[7,8] to a resolution of 2.4 Å, and a thermodynamic analysis of the folding of the enzyme is underway (ref. 9 and unpublished results).

Since Streisinger developed a technique for identifying and isolating phage particles with mutations in the lysozyme gene[10], many different mutations in the lysozyme gene have been identified, the resultant changes in the amino acid sequence of the protein determined, and the properties of these mutant lysozymes discussed in the light of the three-dimensional structure of the wild-type enzyme[8,11].

Lysozyme mutant strains were generated by use of 2-aminopurine, which induces single base pair substitutions. Temperature-sensitive mutants were selected applying the recognition method for lysozyme gene mutations[10]. Plaques of wild-type phage on plates incubated at 37 °C become surrounded by large haloes after exposing the plate to chloroform vapour. The haloes result from phage lysozyme diffusing from the plaque into the surrounding area and lysing the chloroform-treated bacteria. Halo size is a direct measure of the stability of the lysozyme. For ts mutants, the halo size decreases with increasing temperature on plates incubated at 30, 37 and 41 °C whereas wild-type plaque haloes stay large. By testing the ability of the mutant phage to recombine with a number of precisely mapped deletion mutants in the lysozyme gene, the locus of the amino acid change was determined to lie between residues 65 and 105.

The ts mutant lysozyme was purified according to the standard procedure for wild-type lysozyme[12], and was found to have a melting temperature about 14° lower than that of the native enzyme (Fig. 1). Crystals isomorphous with the native enzyme were obtained by the standard procedure[8]. The largest cell dimension change, in c, was 0.4 Å. We therefore were able to analyse the mutant by the standard difference Fourier technique (for example see ref. 13). A data set to 2.4 Å resolution was collected and a difference electron density map calculated between the native and the mutant lysozyme. Full details of the structure determination will be published elsewhere.

The difference density map is essentially featureless except at the location of the side chan of Arg 96, where a strong negative feature is seen at the site of the guanidinium group, and a positive peak overlaps the positions of C^{γ} and C^{δ} (Fig. 2). Clearly, Arg 96 in wild-type lysozyme has been replaced in the ts mutant by an amino acid with a medium-size side chain. The location of the substitution is consistent with that determined genetically. Determination of the amino acid composition of the mutant lysozyme revealed that there were two histidines,

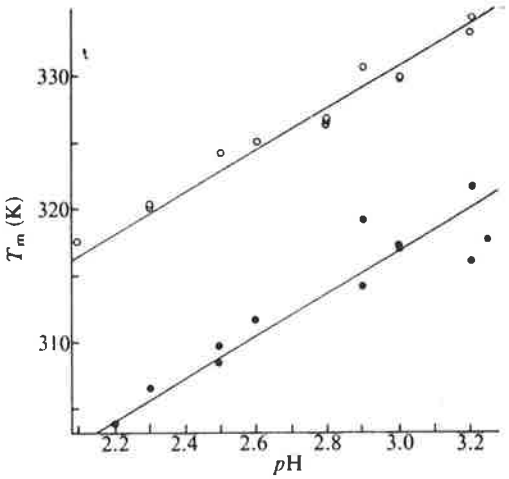


Fig. 1 The melting temperature (T_m) of wild-type (○) and temperature-sensitive (●) lysozymes as a function of *p*H. Denaturation was monitored by following changes in peptide circular dichroism at 223 nm. Details of the experimental method and of the analysis used to calculate T_m are given elsewhere[9]. Solvent conditions were as follows: 0.2 M NaCl, 10^{-4} M dithiothreitol, 10^{-6}–10^{-5} M lysozyme.

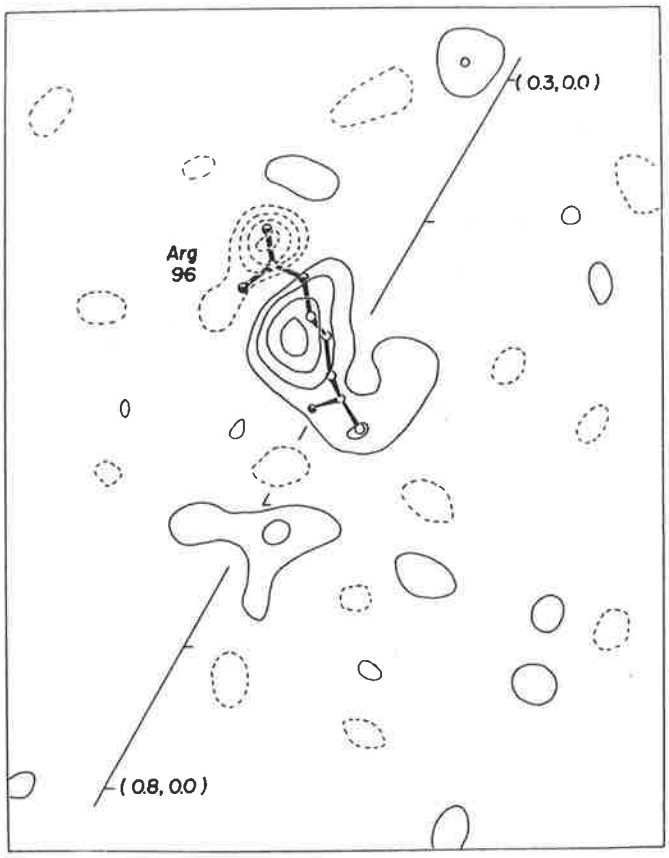


Fig. 2 Portion of the map showing the difference in electron density between native phage lysozyme and the temperature-sensitive mutant lysozyme with Arg 96 replaced by a histidine. The section shown, $z = 11/132$, approximately intersects the side chain of Arg 96. Contours are drawn at arbitrary equal intervals, with positive contours solid and negative broken. In the superimposed side chain of Arg 96, carbon atoms are represented by open circles and nitrogen by half solid circles.

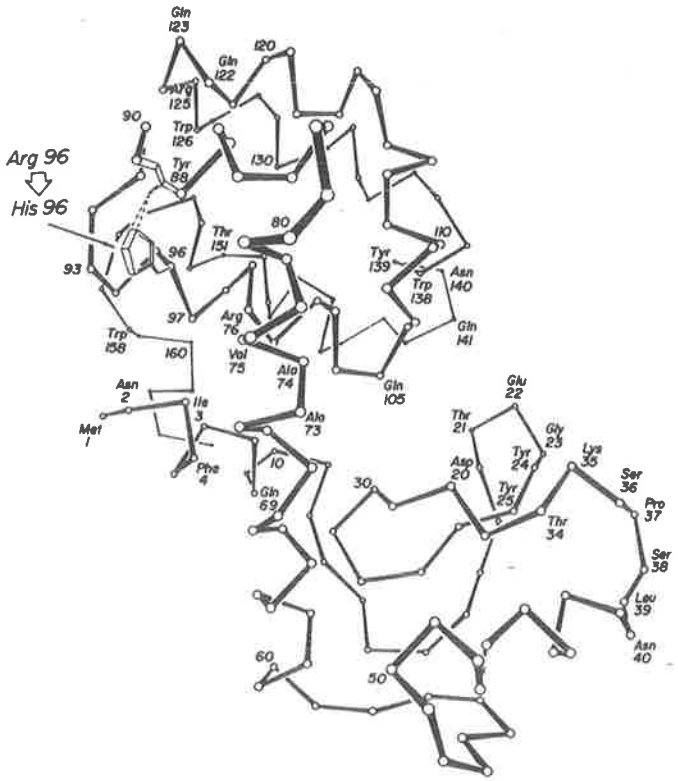


Fig. 3 Schematic drawing of the backbone of bacteriophage T4 lysozyme showing the location of His 96 in the temperature-sensitive mutant. (Adapted from ref. 8).

compared with one in the wild-type enzyme, indicating that Arg 96 has been replaced by a histidine. Furthermore, the electron density map (Fig. 2) also strongly suggests that the ts lysozyme has a histidine at position 96.

In the native enzyme, the side chain of Arg 96 is well-ordered, and lies on the protein surface between the ring of Tyr 88 and the main chain and C^{β} of Leu 91 (Fig. 3). The hydrocarbon portion of the arginine side chain contributes to the major hydrophobic core of the C-terminal lobe of the molecule[7,8]. Substitution of an imidazole at this position presumably destabilises the hydrophobic core somewhat, providing at least part of the molecular basis for the observed temperature sensitivity. The imidazole ring of His 96 lies parallel to the ring of Tyr 88, and appears to donate a hydrogen bond to the carbonyl oxygen of the same residue (Fig. 3), possibly destabilising the α-helix 82–90. Detailed analysis of these interactions will take some time.

The main finding, however, which appears clearcut, is that the three-dimensional structure of the ts lysozyme is virtually identical with that of native lysozyme, except for the replacement of a partially exposed arginine by a histidine. Excluding residue 96, there is no evidence for structural changes larger than a few tenths of an angstrom. This finding is consistent with our earlier suggestion[2,14], based on the structure of the thermostable protease thermolysin, that differences in the thermostability of proteins are, in general, due to subtle changes in hydrophobic interactions, hydrogen bonds, and so on, and not to a single determinant such as metal binding or changes in secondary structure. In agreement with Perutz[15], our study shows that extra energy of thermal stabilisation can be provided without disturbance of the tertiary structure of the protein, although we do not see any change in the number of salt bridges in the mutant lysozyme. Also, our results confirm general thermodynamic arguments which suggest that the net free energy of stabilisation of proteins is small, and derives from a delicate balance between large stabilising forces, principally due to hydrophobic interactions, and large destabilising ones, principally due to chain entropy[16–18].

In addition to the mutant described here, three other temperature sensitive mutant lysozymes, with amino acid substitutions in other parts of the molecule, have been crystallised isomorphously with the wild-type enzyme. Clearly, these mutant lysozymes also have structures similar to that of the native enzyme. We hope that structural comparisons of these and other modified lysozymes will eventually provide a detailed accounting of the contributions of individual amino acid residues to the stability of a protein. Related studies of mutant lysozymes with modified activity should help clarify the mechanism of action of the enzyme.

We thank Kerrie Rine for technical assistance; Drs G. Streisinger and J. Owen for advice on lysozyme genetics, and Dr J. Schellman for discussions on lysozyme stability. M.G.G. was supported by the Swiss NSF. This work was supported in part by NIH grants GM 21967, GM 20066 and GM 20195.

MARKUS G. GRÜTTER
RICHARD B. HAWKES
BRIAN W. MATTHEWS

Institute of Molecular Biology and Departments of Physics and Chemistry, University of Oregon, Eugene, Oregon 97403

Received 5 October; accepted 18 December 1978.

1. Matthews, B. M., Jansonius, J. N., Colman, P. M., Schoenborn, B. P. & Dupourque, D. *Nature new Biol.* **238**, 37–41 (1972).
2. Matthews, B. W., Weaver, L. H. & Kester, W. R. *J. biol. Chem.* **249**, 8030–8044 (1974).
3. Tanaka, M. *et al. J. biol. Chem.* **246**, 3953–3960 (1971).
4. Bridgen, J. & Harris, J. I. *Abs. Ninth Int. Congr. Biochem.*, Stockholm, p. 59 (1973).
5. Koch, G. & Dreyer, W. J. *Virology* **6**, 291–293 (1958).
6. Tsugita, A. & Inouye, M. *J. molec. Biol.* **37**, 201–212 (1968).
7. Matthews, B. W. & Remington, S. J. *Proc. natn. Acad. Sci. U.S.A.* **71**, 4178–4182 (1974).
8. Remington, S. J. *et al. J. molec. Biol.* **118**, 81–98 (1978).
9. Elwell, M. & Schellman, J. *Biochim. biophys. Acta* **386**, 309–323 (1975).
10. Streisinger, G., Mukai, F., Dreyer, W. J., Miller, B. & Horiuchi, S. *Cold Spring Harb. Symp. quant. Biol.* **26**, 25–30 (1961).
11. Tsugita, A. *Enzymes* **5**, 3rd edn, 343–411 (1971).
12. Tsugita, A. & Inouye, M. *J. biol. Chem.* **243**, 391–397 (1968).
13. Matthews, B. W. in *The Proteins* Vol. 3 (eds Neurath, H. & Hill, R. L.) 3rd edn, 403–590 (Academic, New York, 1977).
14. Weaver, L. H., Kester, W. R., Ten Eyck, L. F. & Matthews, B. W. *Experientia* suppl. **26**, 31–39 (1976).
15. Perutz, M. F. & Raidt, H. *Nature* **255**, 256–259 (1975).
16. Schellman, J. A. *Compt. rend. Trav. Lab. Carlsberg Ser. Chim.* **29**, 230–259 (1955).
17. Kauzmann, W. *Adv. Protein Chem.* **14**, 1–63 (1959).
18. Brandts, J. F. in *Thermobiology* (ed. Rose, A. H.) 25–72 (Academic, New York, 1967).

J. Mol. Biol. (1981) **147**, 523–543

Crystallographic Determination of the Mode of Binding of Oligosaccharides to T4 Bacteriophage Lysozyme: Implications for the Mechanism of Catalysis

W. F. Anderson†, M. G. Grütter, S. J. Remington‡, L. H. Weaver and B. W. Matthews§

Institute of Molecular Biology and Department of Physics
University of Oregon, Eugene, OR 97403, U.S.A.

(Received 16 September 1980)

Phage lysozyme has catalytic activity similar to that of hen egg white lysozyme, but the amino acid sequences of the two enzymes are completely different.

The binding to phage lysozyme of several saccharides including *N*-acetylglucosamine (GlcNAc), *N*-acetylmuramic acid (MurNAc) and $(GlcNAc)_3$ have been determined crystallographically and shown to occupy the pronounced active site cleft. GlcNAc binds at a single location analogous to the C site of hen egg white lysozyme. MurNAc binds at the same site. $(GlcNAc)_3$ clearly occupies sites B and C, but the binding in site A is ill-defined.

Model building suggests that, with the enzyme in the conformation seen in the crystal structure, a saccharide in the normal chair configuration cannot be placed in site D without incurring unacceptable steric interference between sugar and protein. However, as with hen egg white lysozyme, the bad contacts can be avoided by assuming the saccharide to be in the sofa conformation. Also Asp20 in T4 lysozyme is located 3 Å from carbon $C_{(1)}$ of saccharide D, and is in a position to stabilize the developing positive charge on a carbonium ion intermediate. Prior genetic evidence had indicated that Asp20 is critically important for catalysis. This suggests that in phage lysozyme catalysis is promoted by a combination of steric and electronic effects, acting in concert. The enzyme shape favors the binding in site D of a saccharide with the geometry of the transition state, while Asp20 stabilizes the positive charge on the oxocarbonium ion of this intermediate. In phage lysozyme, the identity of the proton donor is uncertain. In contrast to hen egg white lysozyme, where Glu35 is 3 Å from the glycosidic D–O–E bond, and is in a non-polar environment, phage lysozyme has an ion pair, Glu11 ... Arg145, 5 Å away from the glycosidic oxygen. Possibly Glu11 undergoes a conformational adjustment in the presence of bound substrate, and acts as the proton donor. Alternatively, the proton might come from a bound water molecule.

1. Introduction

The lysozyme from bacteriophage T4 is produced late in the infection of *Escherichia coli* by bacteriophage T4 and cleaves the bacterial cell wall, allowing

† Present address: Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2H7, Canada.

‡ Present address: Max-Planck-Institute für Biochemie, D-8033 Martinsried bei München, Germany.

§ Author to whom correspondence should be addressed.

0022–2836/81/120523–21 $02.00/0

© 1981 Academic Press Inc. (London) Ltd.

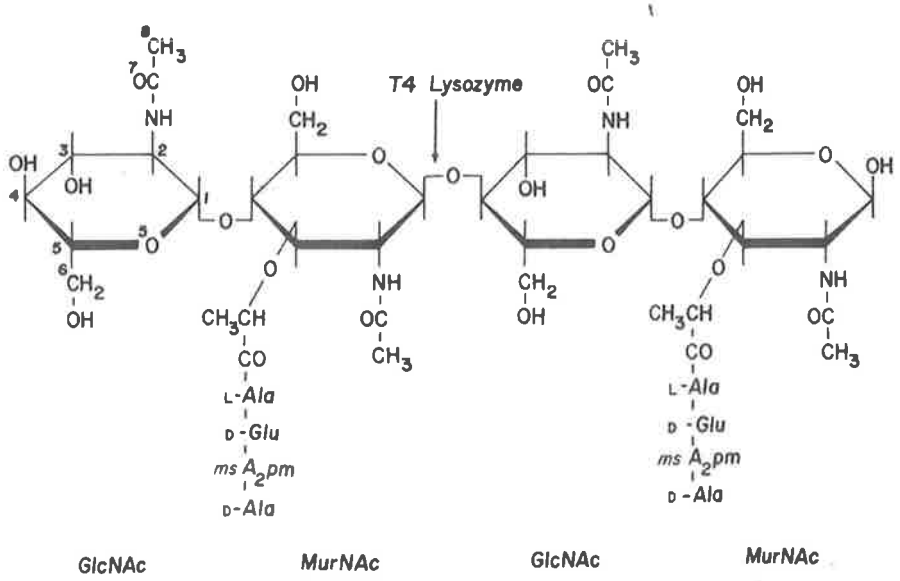


FIG. 1. Portion of the cell wall of *E. coli* cleaved by phage lysozyme.

the release of progeny phage particles (e.g. see Streisinger *et al.*, 1966; Tsugita *et al.*, 1968; Tsugita, 1971). The enzyme is an endoacetylmuramidase, cleaving the $\beta(1\text{-}4)$ glycosidic bond between *N*-acetylmuramic acid and *N*-acetylglucosamine, i.e. the same bond as is cleaved by hen egg white lysozyme (Tsugita *et al.*, 1968). Figure 1 shows a cell wall saccharide with the bond which is cleaved. The atom numbering we have used is the same as that for hen egg white lysozyme (e.g. see Imoto *et al.*, 1972). T4 lysozyme is approximately 250-fold more active than hen egg white lysozyme toward *E. coli* cell walls, but lacks the ability to cleave chitin (i.e. long polymers of GlcNAc†) (Jensen *et al.*, 1976).

The structure of hen egg white lysozyme is well known (Blake *et al.*, 1965; Imoto *et al.*, 1972), and, based on the X-ray structure, Phillips and his collaborators (Phillips, 1967; Blake *et al.*, 1967) have proposed a mechanism of action for the enzyme which is generally held to be correct, although there has been some dispute concerning the role played by "strain" in catalysis (Warshel & Levitt, 1976; Schindler *et al.*, 1977). The structure of human lysozyme, which is homologous with HEWL, has also been determined to high resolution and shown to be similar to that of HEWL (Artymiuk *et al.*, 1978).

In contrast, the amino acid sequence of phage lysozyme is not homologous with hen egg white lysozyme or with any of the "chicken" family of lysozymes. Notwithstanding this difference in amino acid sequence, there is a structural resemblance between parts of the three-dimensional structures of phage and hen egg white lysozyme (Rossmann & Argos, 1976; Levitt & Chothia, 1976; Remington & Matthews, 1978).

† Abbreviations used: GlcNAc, *N*-acetylglucosamine; MurNAc, *N*-acetylmuramic acid; HEWL, hen egg white lysozyme.

In this paper we describe the modes of binding of a series of oligosaccharides to phage lysozyme, determined by X-ray crystallography. These binding experiments have confirmed the location of the active site of phage lysozyme, and suggest a mechanism of action for the enzyme. There are seen to be similarities between the active sites of phage and hen egg white lysozymes which are discussed in detail in the following paper.

2. Experimental

The crystals of bacteriophage T4 lysozyme were obtained as described previously (Remington *et al.*, 1978). *N*-Acetylglucosamine and *N*-acetylmuramic acid were purchased from Sigma Chemical Co. The *N*-acetylglucosamine trisaccharide $(GlcNAc)_3$ and pentasaccharide were gifts from F. W. Dahlquist. $(GlcNAc)_4$ and $(GlcNAc)_6$ were supplied by J. Rupley. Crystals of T4 lysozyme were soaked for several days in 2·2 M-phosphate solutions containing substrate analogues to obtain the crystalline complexes.

The binding of the series of sugars GlcNAc through $(GlcNAc)_6$ was first examined in projection. In the (*h*0*l*) projection of space group $P3_221$, a bound sugar, is, in principle, seen at 3 different projected positions, permitting the binding site to be located in 3 dimensions. In practice, the electron density at each of these 3 positions varies because of the different direction of view of the sugar. At one position we always observed strong density (at least twice as high as any background peak) when saccharide site C (see below) was substantially occupied. The appearance of this peak in an (*h*0*l*) projection could therefore be used to monitor the binding of different saccharides to the crystalline enzyme.

As one proceeds from shorter to longer oligosaccharides, their solubility in the 2·2 M-phosphate solution required to stabilize the lysozyme crystals decreases markedly. GlcNAc itself is very soluble, and a concentration of 0·5 M was used for the crystallographic analysis. However, the solubility of the hexasaccharide $(GlcNAc)_6$ was less than 0·4 mM. By using saturated solutions of the longer oligosaccharides we were able to obtain detectable binding for GlcNAc through $(GlcNAc)_4$, but for $(GlcNAc)_5$ and $(GlcNAc)_6$ little binding could be detected. The lack of binding might be due simply to the low concentrations of these longer saccharides, or might also be due to some steric factor.

Following this preliminary survey, full 3-dimensional data sets were collected for GlcNAc, MurNAc and $(GlcNAc)_3$. Data collection (Table 1) was by standard precession techniques

TABLE 1

Data processing statistics

Statistic	Native	GlcNAc	MurNAc	$(GlcNAc)_3$
Sugar concentration M	—	0·5	0·7	~0·05
Films	12	12	12	12
Average R_{sym}(%)†	4·8	4·5	6·6	5·0
Average R_{sca}(%)	5·1	4·7	6·0	5·2
R_{merge}(%)	2·8	2·5	2·8	3·5
Isomorphous difference (%)	—	7·4	10·8	12·2
Cell dimensions				
a, *b* (Å)	61·2	61·3	61·1	61·5
c (Å)	96·8	96·9	97·2	96·3

† The respective R values are as defined previously (Matthews *et al.*, 1972). R_{sym} gives the agreement between symmetry-related reflections recorded on the same film, R_{sca} is for reflections on successive films in a film pack, and R_{merge} is for the merging of reflections measured on different films.

(Matthews *et al.*, 1972; Remington *et al.*, 1978). Difference electron density maps were calculated using isomorphous replacement phases (Remington *et al.*, 1978) to a nominal resolution of 2·4 Å. Models of the bound inhibitors were constructed in an optical comparator (Richards, 1968; Coleman *et al.*, 1972) and the new co-ordinates were measured from markers placed directly in either the difference map or on maps with coefficients of the form $[n F_{\text{Deriv.}} - (n-1) F_{\text{Nat.}}]$, where $n = 2$, 3 or 4. For each saccharide in turn, idealized co-ordinates were obtained by a least-squares fitting of the known structure of *N*-acetyl glucosamine (Johnson, 1966) to the raw co-ordinates. Where the electron density maps of the bound sugar indicated rotations in the hydroxymethyl group or the acetamido group, such rotations were allowed.

3. Results

(a) N-*Acetylglucosamine*

Although GlcNAc is not reported to be an inhibitor of T4 lysozyme and the enzyme will not hydrolyze GlcNAc oligomers or chitin, the complex of the protein and GlcNAc will form to a measurable extent in the crystal if the saccharide concentration is greater than about 0·1 M. Under these conditions GlcNAc binds to the enzyme at a site close to the opening or cleft in the protein structure that was predicted to be the active site (Matthews & Remington, 1974). The three-dimensional difference density map (Fig. 2) shows the location of binding clearly, with a maximum density 14σ, where σ is the root-mean-square value of the difference electron density. By analogy with hen egg white lysozyme, this subsite will be labeled the C or GlcNAc site.

As can be seen in Figure 2(a) and (b), the density of the hydroxymethyl group is weak, indicating that this group, which is almost fully exposed to the solvent, has some freedom to rotate. In contrast, on the opposite side of the inhibitor, the acetamido group has well defined high density, indicative of strong interaction with the protein. In fact, the amide nitrogen of this group donates a hydrogen bond to the main-chain carbonyl oxygen of Phe104, and the acetylcarbonyl oxygen accepts a hydrogen bond from the main-chain nitrogen of Leu32. The arrangement of hydrogen bonds essentially makes a very short parallel β-sheet, with the acetamido moiety of the sugar aligned between two peptides of the protein backbone. As will be discussed in the following paper, the geometry is strikingly similar to that which is observed at the C site of hen egg white lysozyme. The side-chains of Leu32 and Phe104 provide van der Walls' interactions with the sugar ring and the acetamido methyl group, respectively. Other than the hydrogen bonds described above, there are no other obvious hydrogen bonds between the sugar and the protein, so that the specificity of binding at the C site is directed towards the presence of the acetamido group.

N-Acetylglucosamine appears to be bound in the C site in the chair equatorial conformation. Only the β anomer of GlcNAc is bound in T4 lysozyme, in contrast to HEWL which binds both the α and β anomers in somewhat different, but overlapping, sites (Blake *et al.*, 1967). From the map showing the difference between GlcNAc-substituted and native lysozyme (Fig. 2), it is clear that GlcNAc occupies a single site (site C) with high occupancy and does not bind significantly in the crystal to any other site. There is a weak tongue of density extending

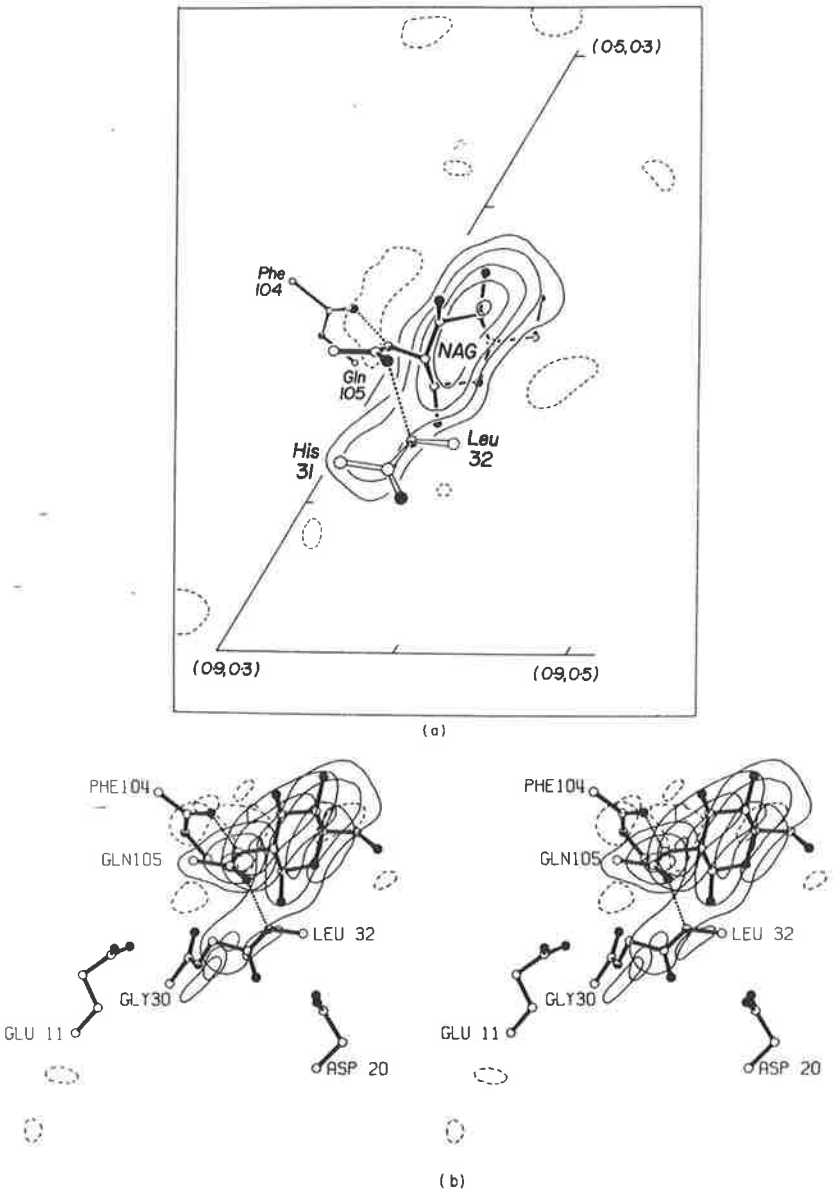


FIG. 2. (a) Section $z = 0{\cdot}129$ of the (GlcNAc − native) difference map. The bound saccharide is drawn with solid bonds, and the dotted lines show hydrogen bonds from the acetamido group to the protein. Contours at arbitrary equal intervals; negative contours broken. Oxygen atoms drawn solid, carbon open, and nitrogen half open.

(b) Stereo drawing corresponding to (a).

"downward" toward site D which could well be due to solvent being ordered when saccharide binds, but might also suggest that there is an alternative low-occupancy mode of binding. Other than this density feature, there is no significant difference density elsewhere in the active site cleft.

(b) N-*Acetylmuramic acid*

If phage lysozyme crystals are soaked in solutions containing high concentrations (above 0·1 M) of *N*-acetylmuramic acid, this sugar also binds to the protein at the C site. A three dimensional difference map for crystals soaked in 0·7 M-NAM is very similar to that obtained for GlcNAc, suggesting that the phage enzyme, MurNAc and GlcNAc both bind at the same site.

The near-identity of binding was confirmed by calculating a (MurNAc − GlcNAc) difference density map. The map is almost featureless, with the highest positive density (5σ) at the position corresponding to the lactyl group (Fig. 3). There is an equally strong negative feature adjacent to $C_{(3)}$ which suggests that the addition of the lactyl side-chain has caused a slight adjustment in the location of this part of the sugar, but the movement cannot exceed a few tenths of an ångstrom unit. Clearly, the same interactions which determine the binding of GlcNAc to phage lysozyme, in particular the hydrogen bonds to the acetamido group, also determine the binding of MurNAc to phage lysozyme. This is different from HEWL where MurNAc cannot bind at the C site because of steric interference between the lactyl group and the protein. Inspection of the model of MurNAc bound to phage lysozyme reveals no interactions between the lactyl group and the protein. Thus, in phage lysozyme, the differentiation between GlcNAc and MurNAc in a cell wall

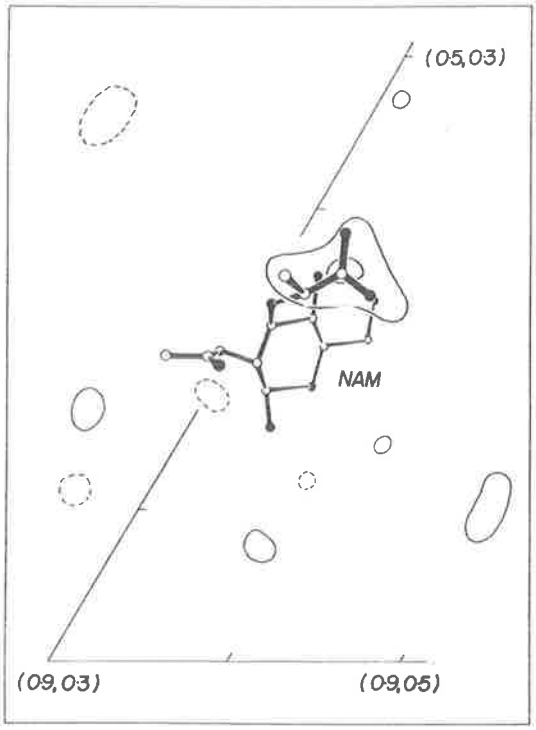


FIG. 3. Section $z = 0{\cdot}159$ of the (MurNAc − GlcNAc) difference map. The section intersects the lactyl group of the muramic acid.

saccharide (Fig. 1) does not occur in site C. Rather, the differentiation probably occurs in site E (see below).

(c) *Tri*-N-*acetylglucosamine*

The complex of phage lysozyme with the *N*-acetylglucosamine trimer was studied in order to examine the details of the protein–saccharide interactions at subsites other than C. The choice of $(GlcNAc)_3$ was a compromise between a desire to study the longest possible oligosaccharide, and the decrease in occupancy of binding of the longer saccharides due to their lower solubility.

The 3-dimensional difference map calculated with coefficients $[F_{GlcNAc_3} - F_{Nat}]$ has its highest positive density (11σ) in the C site (Fig. 4(a)). By comparison with the GlcNAc difference map, the occupancy of $(GlcNAc)_3$ is about 80% that of GlcNAc alone. The trisaccharide also clearly occupies a second site, analogous to site B of HEWL, but beyond this the difference density is very weak (Fig. 4(a)). In order to try to determine the location of the third saccharide unit we calculated a map with coefficients $[4F_{GlcNAc_3} - 3F_{Nat}]$. These coefficients are equal to $[F_{Nat} + 4(F_{GlcNAc_3} - F_{Nat})]$ and were intended to emphasize the weak density in the difference map. The resultant map, shown diagrammatically in Figure 4(b), has strong density for the two saccharide units bound in sites B and C, and has very weak density extending into both the A site and the D site, but the interpretation of this weak density is not obvious. The overall result of binding $(GlcNAc)_3$ to phage lysozyme is rather similar to the binding of the same trisaccharide to HEWL, namely, the strongest site of binding is subsite C, the next is subsite B, and the binding at subsite A is poorly defined, with the saccharide probably occupying a range of positions. In the case of phage lysozyme there is a suggestion that some of the bound trisaccharide molecules may occupy subsites B, C and D.

The location of the saccharide bound in site C is virtually identical with that of GlcNAc alone (Figs 4 and 5). In the B site the saccharide appears to be in the chair equatorial conformation, and is rotated 180° about the glycosidic bond so that the relative conformation of the linked saccharides is the same as for HEWL, and in the proposed structure of chitin (Carlstrom, 1962; Ramachandran *et al.*, 1963). There appears to be a hydrogen bond between the ring oxygen, $O_{(5)}$, of saccharide B, and the $O_{(3)}$ hydroxyl of saccharide A, as in HEWL.

The saccharide bound in site B (Fig. 5) is completely exposed on one side to the solvent, and makes less extensive and less specific interactions with the protein than does the sugar in site C. In site B, the saccharide lies against a loop of the polypeptide backbone from Gly107 to Thr109 which is at the beginning of the short irregular α-helix Glu108 to Gly113. There appear to be two hydrogen bonds, albeit irregular, between the protein backbone and the saccharide (Fig. 5), one between the amide nitrogen of Gly110 and the acetamido oxygen, and the other between the amide nitrogen of Thr109 and the $O_{(3)}$ hydroxyl. There also may be hydrogen bonding from the peptide nitrogen of Glu108, *via* a water molecule, to the methyl hydroxyl $O_{(6)}$. While the binding site for the saccharide in site B is formed by the amino-terminal turn of an α-helix, the binding site C is, on one side of the sugar,

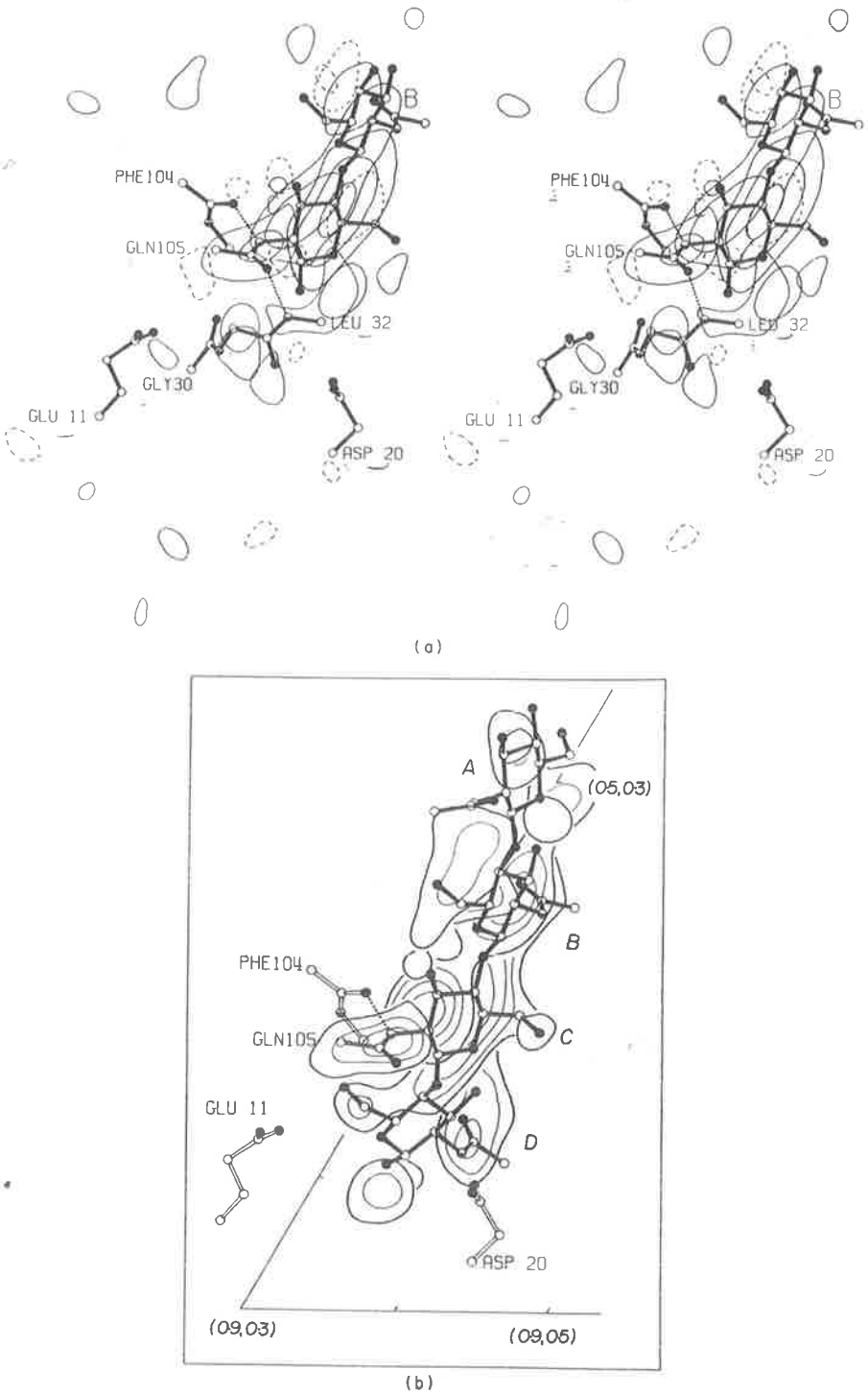


FIG. 4. (a) Stereo drawing showing the difference density ($GlcNAc_3$ – native). (b) Electron density map with coefficients $[4F_{GlcNAc_3} - 3F_{Nat.}]$.

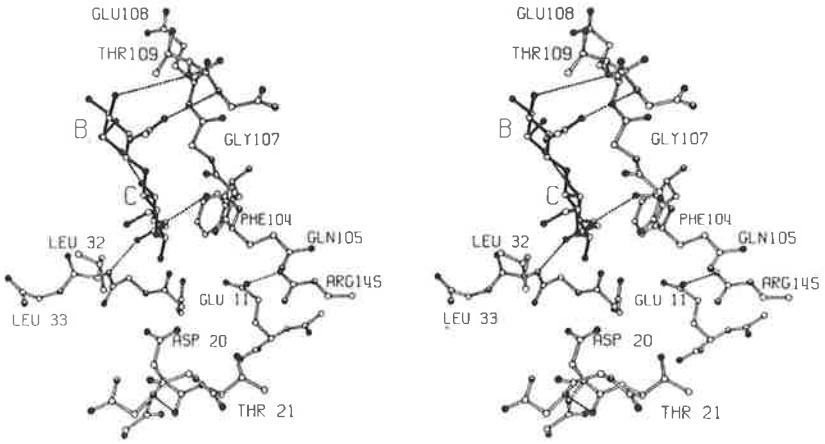


FIG. 5. The binding of a disaccharide in the B and C subsites of phage lysozyme as determined crystallographically. The direction of view is from the right of Fig. 4.

formed by the carboxyl terminus of helix Ala93 to Met106. In fact, the B–C binding site is formed largely by a region of the protein backbone in which helix 93 to 106 is terminated and the backbone turns sharply to immediately begin helix 108 to 113.

In the $(GlcNAc)_3$ minus native lysozyme difference electron density map there is weak positive and negative density suggesting that the hairpin loop Asp20 to Gly23, which folds across the mouth of the active site, may move slightly away from the saccharide bound in site C.

(d) *Model building*

Because the difference electron density maps show only the locations of saccharide binding in the B and C subsites, we have had to resort to model building to explore the other presumed subsites. In model building, we used, as a starting point, the extended structure of chitin. This conformation is similar to that observed in the binding of the tetrasaccharide lactone to hen egg white lysozyme (Ford *et al.*, 1974). Efforts were made to satisfy potentially favorable interactions between the saccharide and the protein, and to avoid prohibitively short van der Waals' contacts.

In the electron density maps showing $(GlcNAc)_3$ bound to phage lysozyme, there is weak density in both the A and D sites (Fig. 4(b)), but this density is so ill-defined that it shows no more than the approximate locations of binding. In the A site, the sugar is quite exposed to solvent, and could either lie against the surface of the protein, or move more-or-less freely in solution. There are several hydrophilic side-chains in this vicinity, including Arg80, Glu108 and Thr109, which could hydrogen bond to $O_{(3)}$, $O_{(4)}$ and $O_{(6)}$.

In contrast to the A site, which is quite exposed, the D and E sites of phage lysozyme are almost completely enclosed by protein (Fig. 6), which severely limits the possible location of the bound saccharide. In site D, if the normal chair

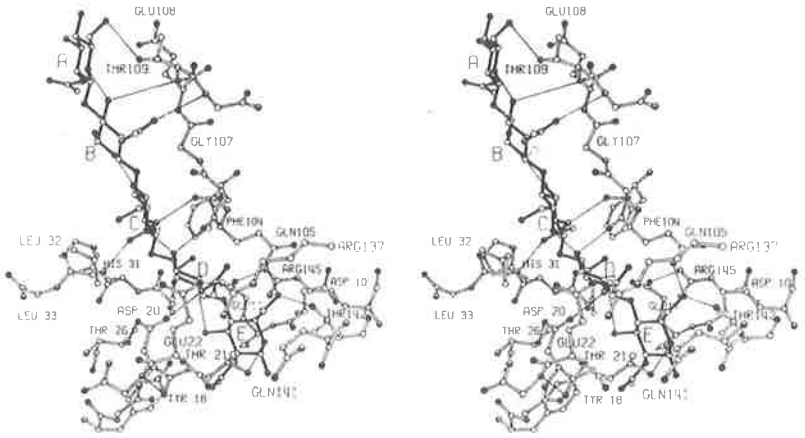


FIG. 6. View of the presumed binding of a pentasaccharide in the active site of phage lysozyme.

conformation is used, it is not possible to avoid unacceptably close contacts between the saccharide and the protein. In Table 2 we list all the contacts less than 3·5 Å between the saccharide and the protein. The Table includes the contacts for both site C, which is based on the experimental data described above, and for site D, which is based on model building. (Table 2 does not include a contact of 2 Å between $C_{(8)}$ of sugar D and the carboxyl of Glu22 since this side-chain is quite exposed to solvent, and is free to rotate away from the substrate.) As can be seen, the notable bad contact in site D is the distance of 2·3 Å between the hydroxymethyl carbon $C_{(6)}$ and the carbonyl oxygen of Gly30. Any attempt to relieve this close approach results in a shortening of the $O_{(3)}$–O = C(Gln105) hydrogen bond, which is already too short (2·6 Å), and a decrease in the close approach of 3·0 Å between $C_{(4)}$ and O = C(Gln105). The two protein atoms which give rise to these contacts are main-chain oxygens, located on opposite sides of the active site cleft, and in each case appear to be an integral part of the two lobes which together make up the phage lysozyme molecule. It therefore appears that the protein–saccharide contacts cannot be relaxed without a bodily "opening up" of the active site cleft. On the other hand it must be kept in mind that the present co-ordinates have an estimated uncertainty of about 0·5 Å, so that errors of 0·7 Å or so in interatomic distances could occur (see below).

The fact that there is steric interference in site D of phage lysozyme is of particular interest, because a similar situation occurs for hen egg white lysozyme. In the case of the HEW enzyme, Phillips and co-workers found impossibly close contacts in the D site between the enzyme and the $C_{(6)}$–$O_{(6)}$ atoms of a sugar bound in the normal unstrained chair conformation. By assuming the saccharide to be in a half-chair-like conformation with $C_{(2)}$, $C_{(1)}$, $O_{(5)}$ and $C_{(5)}$ coplanar, and the $C_{(5)}$–$O_{(6)}$ bond axial, the overcrowding could be relieved, leading to the proposal that strain plays an important role in the mechanism of catalysis (Phillips, 1967; Blake *et al.*, 1967). Subsequent to their original proposal for the HEWL mechanism, in which the site D geometry was based on model building, Phillips and co-workers have

obtained direct evidence for the site D geometry by determining the mode of binding of a tetrasaccharide δ-lactone derived from $(GlcNAc)_4$ (Secemski & Lienhard, 1971; Ford *et al.*, 1974). This inhibitor binds in sites A, B, C, D and can be regarded as a transition-state analogue. Based on careful analysis of a 2·5 Å resolution electron density map of the complex of the tetrasaccharide–lactone with HEWL at pH 2·6, Ford *et al.* (1974) concluded that the δ-lactone ring has a conformation close to a sofa or boat, such that the hydroxymethyl group $C_{(6)}$–$O_{(6)}$ is axial.

Inspection of the phage lysozyme model suggested that a saccharide with the

TABLE 2

Saccharide–protein contacts less than 3·5 Å

Site C Sugar	Distance (Å)	Protein	
C1	3·3	O=C	Gln105
C3	3·3	O=C	Phe104
N2	3·0 (H)	O=C	Phe104
C7	3·0	CB	Leu32
	3·1	O=C	Gly30
O7	2·7 (H)	N	Leu32
	3·2	O=C	Gly30
C8	2·9	CE1	Phe104
	3·0	O=C	Gly30
	3·1	CZ	Phe104
O1	3·1	CA	Gly104

Site D Sugar	Distance (Å) Chair	Distance (Å) Sofa	Protein	
C1	3·1	3·4	OD1	Asp20
C2	3·2	3·2	OD1	Asp20
	3·2	3·2	O=C	Gln105
C3	3·1	3·0	OD1	Asp20
	3·1	3·3	O=C	Gln105
C4	3·0	3·1	O=C	Gln105
O5	4·2	3·9	OD1	Asp20
	4·0	4·2	OE1	Glu11
	4·7	4·9	NEE1	Arg145
C6	2·3	3·0	O=C	Gly30
	3·2	4·3	C	Gly30
	4·3	3·4	CA	Gln105
O6	2·9 (H)	2·7 (H)	O=C	Gly30
	3·2	3·4	CD1	Phe104
	3·4	4·1	CA	Gln105
	3·5	2·7 (H)	OE1	Glu11
	3·8	2·9	OE2	Glu11
N2	2·8 (H)	2·7 (H)	OD1	Asp20
O7	3·6	3·4	O=C	Glu105
O3	2·6	3·0	O=C	Glu105
O1	4·1	3·6	OD1	Asp20
	3·7	4·1	NEE1	Arg145
	4·8	5·1	OE1	Glu11

TABLE 2 *(continued)*

Site E Sugar	Distance (Å)	Protein	
C1	3·2	OG1	Thr21
C2	3·2	OG1	Thr21
C4	3·4	CB	Thr21
C5	3·3	NEE1	Arg145
O5	2·7 (H)	OG1	Thr21
	3·3	CB	Thr21
	3·3	CG2	Thr21
C6	3·2	OG1	Thr142
	3·3	NOE	Gln105
O6	2·4 (H)	NOE	Gln105
	2·6 (H)	OG1	Thr142
C8	3·3	O=C	Glu11
O4	3·0 (H)	OG1	Thr21

Presumed hydrogen bonds are indicated (H).
Selected longer contacts in the vicinity of the D–O–E glycosidic bond are also included.

sofa conformation might also be accommodated in the D site, since the sofa conformation would increase the distance between $C_{(6)}$ and O = C(Gly30). To test the fit more carefully, we attempted to place in site D a sugar with the geometry for *N*-acetylgluconolactone determined by Ford *et al.* (1974), at the same time checking all distances between the sugar and the surrounding protein. The results are shown in Table 2. As can be seen, it is possible to place the gluconolactone so that all interatomic distances are within acceptable limits. Not only is the $C_{(6)} \cdots O$(Gly30) distance increased from 2·3 Å to the allowed value of 3·0 Å, but also a number of favourable interactions occur, including hydrogen bonds to $O_{(6)}$, $N_{(6)}$, $O_{(3)}$ and possibly $O_{(7)}$ (Table 2). In going from the chair to the sofa conformation, the saccharide as a whole remains essentially in the same position, with the hydroxymethyl group $C_{(6)}$–$O_{(6)}$ undergoing the greatest motion. If $O_{(6)}$ were kept in the same relative position in going from the chair to the sofa conformation, then impossibly short contacts would result, but by making an approximately 90° rotation about the $C_{(5)}$–$C_{(6)}$ bond, $O_{(6)}$ is brought snugly into a position where it can hydrogen bond to Glu11 and the carbonyl oxygen of Gly30 (Table 2). The differences in geometry of binding in site D for the chair and sofa conformations can be compared in Figure 7(a) and (b). However, it must be emphasized that the co-ordinates used to calculate the approach distances quoted in Table 2 are, apart from stereochemical idealization, unrefined, and have an estimated uncertainty of about 0·4 Å. During the progress of this work we have been refining the native co-ordinates and those of the complex with $(GlcNAc)_3$. The refinements are not complete, but, as a check we recalculated the approach distances using the best protein co-ordinates currently available, which have crystallographic residual of 21% for data to 1·9 Å resolution. Typically, the recalculated distances agree with those quoted in Table 2 within 0·3 Å, but there are some larger changes. As a result

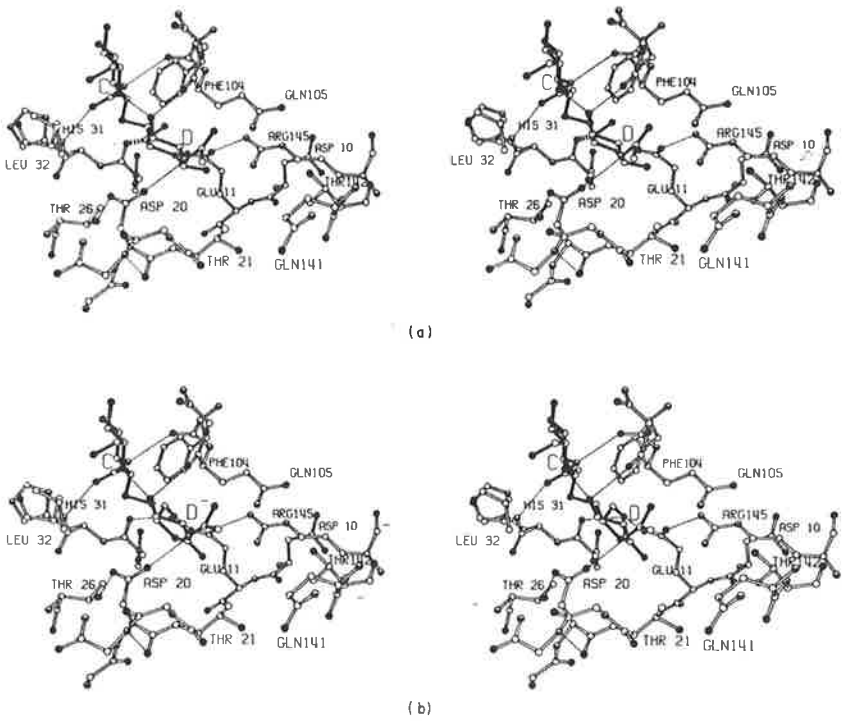


FIG. 7. Comparison of the binding in site D of the chair and sofa conformations. (a) Saccharides in sites C and D in the normal chair conformation. The unacceptable contact is indicated by the thick broken line. (b) Saccharide in site D in the sofa conformation.

of the refinement, some of the critical close approaches between saccharide D in the chair form and the protein are lengthened. Although these contacts still remain below the normally accepted limits, the preference for the sofa rather than the chair conformation is not as clear-cut as suggested by Table 2. Clearly, the energetics of binding will have to be reconsidered when the fully refined co-ordinates are available.

Having placed a saccharide in subsite D, assumed to be in the chair conformation, we then attempted to model-build the next sugar in subsite E. In this direction the protein has an opening or hole through which the saccharide must pass, and be essentially enclosed by protein. For this reason, the general course which the saccharide must follow is obvious, but it is difficult to place the saccharide so that bad contacts are avoided. We were able to find one stereochemically satisfactory location which is illustrated in Figure 6. The contacts between sugar and protein are listed in Table 2. In calculating all the distances in this Table, the protein co-ordinates are those reported previously for the native molecule (Remington *et al.*, 1977), except for the side-chain of Thr21 which was rotated 120° about its C^{α}–C^{β} bond to avoid bad contacts with saccharide E. As can be seen in Figure 6 and Table 2, we adopted the conformation for $O_{(6)}$ of saccharide E such that it makes hydrogen bonds to the amide of Gln105 and the hydroxyl

oxygen of Thr142. The calculated lengths of these presumed hydrogen bonds are a little short, but could easily be lengthened by slight rotations of the participating protein side-chains.

In the model for saccharide bound in site E it is difficult, if not impossible, to accommodate the lactyl group of MurNAc. Certainly, the cross-linking peptide which emanates from this group in the *E. coli* cell wall (Fig. 1) could not be located at this position. Therefore, in phage lysozyme the differentiation between the binding of GlcNAc and MurNAc occurs in site E rather than site C, as is the case for HEWL.

Beyond site E the extended saccharide chain extends into the solution, and makes little if any contact with the protein.

Co-ordinates for saccharides B and C, based on the electron density maps, and for saccharides A, D and E, based on model building, are given in Table 3.

4. Discussion

In comparison with hen egg white lysozyme, there is very little kinetic data available for T4 phage lysozyme. This is, no doubt, due to the fact that the enzyme is not so readily available as hen egg white lysozyme, and because no convenient small molecule substrate is available and the assay for enzymatic activity is to monitor the degradation of whole cell walls (Tsugita *et al.*, 1968). Recently, Bienkowska & Taylor (1979) have shown that phage lysozyme will hydrolyze the tetrasaccharide muropeptide consisting of the tetrasaccharide GlcNAc-MurNAc-GlcNAc-anMurNAc in which the lactyl group of MurNAc and anMurNAc are substituted by the tetrapeptide L-Ala-D-Glu-*ms*A_2pm-D-Ala (anMurNAc = 1,6-anhydro-*N*-acetylmuramic acid; *ms*A_2pm = *meso*-diaminopimelic acid). This muropeptide is a fragment of the *E. coli* cell wall obtained by digestion with λ-endolysin. Phage lysozyme cleaves the central glycosidic bond of the above muropeptide to produce two disaccharide-tetrapeptides, one of which, GlcNAc-MurNAc-L-Ala-D-Glu-*ms*A_2pm-D-Ala, had previously been shown by Jensen *et al.* (1976) to be a good competitive inhibitor (K_1 = 0·22 mM). The tetrasaccharide GlcNAc-MurNAc-GlcNAc-MurNAc is neither a substrate nor an inhibitor of phage lysozyme (Mirelman *et al.*, 1975). For all inhibitors tested, Jensen *et al.* (1976) found no indication of a conformational change in the protein, although Troitskii *et al.* (1979) have reported evidence for such a change. In contrast to hen egg white lysozyme, there is no evidence that phage lysozyme will catalyze transglycosylation.

The catalytic activity of phage lysozyme toward whole cell walls is maximal at pH 7·2 to 7·4 (Tsugita *et al.*, 1968). No chemical modification studies of the enzyme have been reported.

In the case of phage lysozyme, a great deal of information is potentially available from studies of mutants of the enzyme which have modified activity. Several years ago Tsugita (1971) examined all the known mutants of phage lysozyme and divided the sites at which amino acid substitutions occurred into two classes, "essential" and "non-essential". An essential amino acid was one for which any substitution substantially reduced the catalytic activity of the enzyme, while a non-essential

TABLE 3

Co-ordinates for GlcNAc in sites A to E

		X	*Y*	*Z*
Site A	$O_{(4)}$	23·95	1·83	15·35
	$C_{(1)}$	28·11	1·09	15·27
	$C_{(2)}$	27·33	0·85	16·59
	$C_{(3)}$	25·82	0·73	16·30
	$C_{(4)}$	25·34	2·01	15·65
	$C_{(5)}$	26·10	2·14	14·28
	$O_{(5)}$	27·52	2·20	14·55
	$C_{(6)}$	25·77	3·36	13·43
	$O_{(6)}$	24·90	3·03	12·40
	$N_{(2)}$	27·78	−0·42	17·17
	$C_{(7)}$	27·82	−0·51	18·56
	$O_{(7)}$	27·64	0·38	19·31
	$C_{(8)}$	28·13	−1·95	19·00
	$O_{(3)}$	25·13	0·67	17·61
	$O_{(1)}$	29·50	1·27	15·62
Site B	$C_{(1)}$	33·07	0·78	13·26
	$C_{(2)}$	31·76	1·28	12·65
	$C_{(3)}$	30·80	1·82	13·68
	$C_{(4)}$	30·49	0·65	14·66
	$C_{(5)}$	31·83	0·29	15·35
	$O_{(5)}$	32·74	−0·16	14·34
	$C_{(6)}$	31·78	−0·83	16·39
	$O_{(6)}$	31·03	−1·78	17·34
	$N_{(2)}$	32·11	2·39	11·74
	$C_{(7)}$	31·61	2·36	10·44
	$O_{(7)}$	30·94	1·51	9·98
	$C_{(8)}$	31·86	3·68	9·69
	$O_{(3)}$	29·55	2·10	13·02
	$O_{(1)}$	33·87	0·11	12·35
Site C	$C_{(1)}$	37·80	−1·61	12·21
	$C_{(2)}$	36·86	−2·00	13·36
	$C_{(3)}$	35·41	−1·69	13·02
	$C_{(4)}$	35·29	−0·21	12·76
	$C_{(5)}$	36·13	0·13	11·53
	$O_{(5)}$	37·54	−0·22	11·86
	$C_{(6)}$	36·19	1·57	11·07
	$O_{(6)}$	36·87	2·38	11·97
	$N_{(2)}$	37·01	−3·48	13·57
	$C_{(7)}$	37·57	−3·90	14·76
	$O_{(7)}$	38·14	−3·24	15·54
	$C_{(8)}$	37·36	−5·43	14·96
	$O_{(3)}$	34·61	−1·95	14·24
	$O_{(1)}$	39·01	−1·57	12·52
Site D (Chair)	$C_{(1)}$	41·63	−2·89	9·45
	$C_{(2)}$	40·57	−1·84	9·06
	$C_{(3)}$	40·03	−1·13	10·29
	$C_{(4)}$	39·41	−2·16	11·20
	$C_{(5)}$	40·51	−3·12	11·67
	$O_{(5)}$	41·06	−3·77	10·45
	$C_{(6)}$	40·13	−4·27	12·58

TABLE 3 *(continued)*

		X	*Y*	*Z*
	$C_{(6)}$	39·35	−5·21	11·93
	$N_{(2)}$	41·25	−0·84	8·18
	$C_{(7)}$	40·65	−0·54	6·97
	$O_{(7)}$	39·73	−1·09	6·49
	$C_{(8)}$	41·34	0·66	6·28
	$O_{(3)}$	38·92	−0·27	9·85
	$O_{(1)}$	41·88	−3·52	8·33
Site E	$C_{(1)}$	45·30	−5·29	6·69
	$C_{(2)}$	45·34	−4·75	8·13
	$C_{(3)}$	43·93	−4·53	8·67
	$C_{(4)}$	43·24	−3·53	7·78
	$C_{(5)}$	43·11	−4·14	6·38
	$O_{(5)}$	44·49	−4·39	5·89
	$C_{(6)}$	42·46	−3·31	5·29
	$O_{(6)}$	41·51	−4·11	4·69
	$N_{(2)}$	46·02	−5·79	8·97
	$C_{(7)}$	47·07	−5·38	9·77
	$O_{(7)}$	47·58	−4·33	9·76
	$C_{(8)}$	47·52	−6·51	10·73
	$O_{(3)}$	44·06	−3·88	9·98
	$O_{(4)}$	46·51	−5·27	6·15
Site D (Sofa)	$O_{(4)}$	39·02	−1·66	12·59
	$C_{(1)}$	41·54	−3·03	9·49
	$C_{(2)}$	40·59	−1·89	9·12
	$C_{(3)}$	40·15	−1·20	10·38
	$C_{(4)}$	39·45	−2·25	11·25
	$C_{(5)}$	40·39	−3·38	11·64
	$O_{(5)}$	41·41	−3·66	10·64
	$C_{(6)}$	39·55	−4·63	11·86
	$O_{(6)}$	40·15	−5·85	11·95
	$N_{(2)}$	41·28	−0·92	8·25
	$C_{(7)}$	40·70	−0·56	7·05
	$O_{(7)}$	39·64	−1·01	6·66
	$C_{(8)}$	40·86	0·47	5·66
	$O_{(3)}$	39·22	−0·13	10·06
	$O_{(1)}$	42·41	−3·33	8·72

The co-ordinates are in Å in an orthogonal reference frame parallel to a^*, b and c, as used for the lysozyme molecule (Remington *et al.*, 1977).

amino acid was one for which at least one substitution did not impair enzymatic activity. On this basis, Tsugita designated Asp20, Gln105 and Trp138 as essential, and probably Glu22, Asn140 and Glu141 as well. When the three-dimensional structure of the enzyme was first determined, it was striking to see that each of these residues was clustered in the vicinity of the presumed active site cleft (Matthews & Remington, 1974). Clearly, any proposal for the mechanism of catalysis of phage lysozyme ought to be consistent with the behavior of these mutants. Not only should the essential amino acids be involved in a direct way with substrate binding and/or catalysis, but also catalysis should not depend critically on a residue which is known to be non-essential.

There are three segments of the polypeptide chain which are in the vicinity of the active site cleft, and in which mutations have been observed. These are residues 20 to 25, 105 and 138 to 141. Residues 20 to 25 (Asp-Thr-Glu-Gly-Tyr-Tyr) form a sharp hairpin loop which extends across the mouth of the active site (Fig. 6). The mutant JD5J201 has Glu-Gly-Tyr-Tyr (22 to 25) replaced by Arg-Leu-Leu-His, and has moderately high activity, suggesting that these residues, in particular Glu22, are not involved directly in catalysis. However, mutant JR14JD5, in which Asp-Thr-Glu (20 to 22) is replaced by Glu-Thr-Gln, has low activity, suggesting that, since Glu22 is not involved, then Asp20 must have a vital role in catalysis (Ocada *et al.*, 1970). In the model for the saccharide bound in site D, one of the carboxyl oxygens of Asp20 approaches within 3·4 Å of $C_{(1)}$. Furthermore, as we discuss in more detail in the following paper, Asp20 is in a position corresponding to that of Asp52 in hen egg white lysozyme, and would therefore be expected to play an analogous role. This is also the strongest evidence that, in phage lysozyme, it is the glycosidic bond between saccharides D and E which is hydrolyzed. In contrast to HEWL, this has not been confirmed experimentally. The fact that the replacement of Asp20 by a chemically similar Glu residue reduces activity markedly indicates that the location of the carboxyl group is critical. In terms of the model, substitution of a glutamate at residue 20 places the carboxyl closer to $C_{(1)}$ of saccharide D, causing close contacts which would require the carboxyl to be rotated away from $C_{(1)}$. Also this amino acid substitution would disrupt the Asp20 . . . Thr26 hydrogen bond which may be necessary to hold the carboxyl of Asp20 in its correct position (Fig. 6).

The second segment of the polypeptide chain thought, from studies of mutants, to be important for catalysis includes Gln105. This glutamine can be replaced with a serine, with no loss of activity, but replacement with tyrosine results in almost zero activity. In terms of the model, the side-chain of Gln105 forms a short hydrogen bond to $O_{(6)}$ of saccharide E (Fig. 6). Possibly a serine, or serine plus a water molecule, could make the same hydrogen bond. Hydrogen bonds made by the backbone carbonyl oxygen of residue 105 (Fig. 6; Table 2) could be made whether the side-chain was serine or glutamine. On the other hand, substitution of a tyrosine at this location would completely block site E, preventing productive binding.

There are also a number of mutations known in the region 138 to 141 (Remington *et al.*, 1978). In general, these mutant lysozymes have moderately high activity, suggesting that this region is not directly involved with catalysis. Mutants in which Trp138 is replaced by amino acids with small side-chains tend to have lower activity, but this is probably due to the destabilization of the molecule which is expected to occur in such cases. The mutant in which Trp138 is replaced by a tyrosine is somewhat less stable than the native structure, and has about 60% its activity. Elwell & Schellman (1979) have suggested that in this instance the replacement of tryptophan by tyrosine may cause a local "unravelling" of the structure. In the model, Trp138 is more-or-less buried in the wall of the active site cleft, but neither this residue, nor Tyr139-Glu141 has any direct contact with the bound saccharide.

In summary, the following factors appear to be important in catalysis.

(1) Site C is the strongest site of binding for a single saccharide unit.

(2) In site D, the geometry of the enzyme is very restrictive, such that binding of a saccharide in the sofa conformation is favored relative to the normal chair conformation.

(3) Asp20 (analogous to Asp52 of HEWL) is directly involved in catalysis.

The similarity of these three factors with the situation which occurs in hen egg white lysozyme is striking. The immediate suggestion is that catalysis is promoted by a combination of steric and electronic effects which act in concert to favor the formation of a carbonium ion intermediate in which the positive charge on $C_{(1)}$ is stabilized by Asp20. However, the question is, what is the analogue of Glu35, the presumptive proton donor of hen egg white lysozyme? In phage lysozyme, Glu11 has *approximately* the same relative location as Glu35 (see the following paper, Matthews *et al.*, 1981), but is 5 Å away from the glycosidic oxygen, and in addition forms an ion pair with Arg145. In HEWL Glu35 is in a non-polar environment which is thought to raise its pK and help enhance the effectiveness of this group as a proton donor (Phillips, 1967; Blake *et al.* 1967; Dahlquist & Raftery, 1968).

In the model, with saccharide D in the sofa conformation, OE1 of Glu11 is 5·0 Å from the glycosidic oxygen, and NEE1 of Arg145 is only 4·1 Å away. (For the chair conformation the distances decrease to 4·8 Å and 3·7 Å, respectively.) The oxygen and nitrogen are at a hydrogen bonding distance from each other. With this geometry, there is almost enough space to interpose a water molecule between the ion pair and the glycosidic oxygen. On the other hand, there is no evidence in the current refined model of the *native* structure for a bound water molecule at this position.

Therefore, the identity of the proton donor, and the exact role played by Glu11 in phage lysozyme remain uncertain. It is possible that Glu11 might act as the proton donor, analogous to Glu35 in hen egg white lysozyme. This would require a conformational adjustment of the Glu35 side-chain to bring the carboxyl closer to the D–O–E glycosidic bond. It would also be necessary to move the side-chain of Arg145 away from Glu11 and away from the glycosidic oxygen to make room for Glu11. Certainly, some conformational change has to occur to allow the substrate to enter the active site (see below), and it is possible that an adjustment in the position of Glu11 and Arg145 might take place at the same time. It would also be necessary for Glu11 to remain protonated above pH 7·0, as required by the observation that the pH optimum for phage lysozyme is 7·3. (On the other hand it should be noted that, depending on the substrate and the ionic strength, hen egg white lysozyme is active from pH 4 to pH 10, with an optimum for cleavage of cell walls as high as pH 9·2 (Davies *et al.*, 1969).) The second alternative is that a water molecule might act as the proton donor, although there is no direct evidence for this. In any event, it appears that the proton donation, *per se*, is not a driving force in catalysis, as has been suggested by Vernon (1967) might be the case for hen egg white lysozyme. It is also worth noting that oxygen OE1 of Glu11 is 4·2 Å from $O_{(5)}$ of saccharide D, and is therefore well placed to assist in the stabilization of the oxocarbonium ion intermediate.

The original suggestion of Phillips and co-workers (Phillips, 1967; Blake *et al.*, 1967) that "strain" is an important factor in the mechanism of hen egg white

lysozyme has been contentious. Warshel & Levitt (1976) (also see Warshel, 1978) have suggested the electrostatic stabilization rather than strain of the substrate due to steric effects is important in promoting catalysis. Also Schindler *et al.* (1977) concluded from the observed binding constants of a series of oligosaccharides to HEWL that strain can only account for a small part of the catalytic effect of the enzyme. Recently, Kelly *et al.* (1979) have shown crystallographically that the trisaccharide MurNAc-GlcNAc-MurNAc binds to HEWL in subsites B, C and D. The MurNAc residue in subsite D is seen to be in the chair rather than the sofa conformation, but on the other hand this residue is not as deeply buried in the site as is observed for the tetrasaccharide lactone (Ford *et al.*, 1974). It is generally accepted that enzymes promote hydrolysis by favoring the formation of the transition state, thereby lowering the activation energy barrier of the rate-limiting step. In any given instance a number of factors may act in concert to help stabilize the transition state. Our preliminary results suggest that, in the case of phage lysozyme, both electrostatic and steric factors are of importance in catalysis.

Unfortunately, the crystallographic results obtained to date do not show the location of the crosslinking cell wall peptide which emanates from the muramic acid in site D, and is necessary for catalysis. Presumably this peptide is required for optimal binding of the substrate, and might also act as a trigger to promote opening of the active site cleft to allow the substrate to enter.

Jensen and co-workers examined the specificity requirements of phage lysozyme by determining the effectiveness of *Escherichia coli* B and *Micrococcus luteus* cell wall hydrolysis products as inhibitors (Mirelman *et al.*, 1975; Jensen *et al.*, 1976). They showed that the enzyme was inhibited by a fragment as small as MurNAc-L-Ala-D-Glu(A_2pm) (K_I = 0·32 mM). Interestingly, only a single saccharide is required for inhibition. From the series of compounds studied it was inferred that the one modification which abolishes inhibition is the addition of glycine at the α-carboxyl of the D-glutamic acid. Evidently, the presence of this free carboxyl is necessary for inhibition, but on the other hand the compound with the glycine present is a hydrolysis product of phage lysozyme, so that the free carboxyl is, apparently, not essential for hydrolysis.

If one attempts to model-build the L-Ala-D-Glu(A_2pm) . . . peptide, it can be placed across the surface of the upper lobe of the molecule (to the right in Fig. 6), in the vicinity of Met106, Phe114 and Asn132-Arg137. At this stage it does not seem worthwhile to postulate a detailed model for the geometry of binding; however, it is interesting to note that the α-carboxyl of the glutamate is directed inwards toward the active site cleft, and is also close to the guanidinium of Arg137. Addition of a glycine at the α-carboxyl would result in a serious steric interference with the protein. Presumably the crosslinking peptide is involved in some way with the entry of the substrate into the active site. In the native enzyme structure, as seen in the crystals, the active site cleft is occluded, so that some movement of the protein must take place in order for the substrate to enter (Figs 6 and 7). This movement could be of two types: either a local hinge motion of the Lys19 to Tyr24 hairpin loop away from the mouth of the active site cleft, or a large-scale separation of the two lobes of the molecule. The latter separation might disrupt the Glu11 . . . Arg145 interaction and remove Arg145 from the immediate vicinity of the D–O–E

glycosidic bond. Hopefully, cocrystallization of the enzyme complexed with a sizeable peptidoglycan fragment will eventually provide answers to some of these questions.

We thank Drs F. W. Dahlquist, H. B. Jensen, D. Mirelman, J. A. Rupley and N. Sharon for gifts of saccharides. Also we thank Dr Dahlquist for helpful discussions on lysozyme catalysis.

This work was supported in part by a National Institutes of Health Postdoctoral Fellowship (GM 05972) (to W.F.A.), a Swiss National Science Foundation Fellowship (to M.G.G.), National Institutes of Health research grants GM 20066 and GM 21967, National Science Foundation grant PCM 8014311, and a grant from M. J. Murdock Charitable Trust.

REFERENCES

Artymiuk, P. J., Blake, C. C. F. & Oatley, S. J. (1978). *Acta Crystallogr. sect. A*, **34**, S63.
Bienkowska, K. & Taylor, A. (1979) *Eur. J. Biochem.* **96**, 581–584.
Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Nature (London)*, **206**, 757–761.
Blake, C. C. F., Johnson, L. N., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967). *Proc. Roy. Soc. ser. B*, **167**, 378–388.
Carlstrom, D. (1962). *Biochim. Biophys. Acta*, **59**, 361–364.
Colman, P. M., Jansonius, J. N. & Matthews, B. W. (1972). *J. Mol. Biol.* **70**, 701–724.
Dahlquist, F. W. & Raftery, M. A. (1968). *Biochemistry*, **7**, 3277–3280.
Davies, R. C., Neuberger, A. & Wilson, B. M. (1969). *Biochim. Biophys. Acta*, **178**, 294–305.
Elwell, M. L. & Schellman, J. A. (1979). *Biochim. Biophys. Acta*, **580**, 327–338.
Ford, L. O., Johnson, L. N., Machin, P. A., Phillips, D. C. & Tjian, R. (1974). *J. Mol. Biol.* **88**, 349–371.
Imoto, I., Johnson, L. N., North, A. C. T., Phillips, D. C. & Rupley, J. A. (1972). In *The Enzymes* (Boyer, P., ed.), vol. 7, 3rd edit., pp. 665–868, Academic Press, New York.
Jensen, H. B., Kleppe, G., Schindler, M. & Mirelman, D. (1976). *Eur. J. Biochem.* **66**, 319–325.
Johnson, L. N. (1966). *Acta Crystallogr.* **21**, 885–891.
Kelly, J. A., Seilecki, A. R., Sykes, B. D., James, M. N. G. & Phillips, D. C. (1979). *Nature (London)*, **282**, 875–878.
Levitt, M. & Chothia, C. (1976). *Nature (London)*, **261**, 552–558.
Matthews, B. W. & Remington, S. J. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 4178–4182.
Matthews, B. W., Klopfenstein, C. E. & Colman, P. M. (1972). *J. Phys. E: Sci. Instrum.* **5**, 353–359.
Matthews, B. W., Remington, S. J., Grütter, M. G. & Anderson, W. F. (1981) *J. Mol. Biol.* **147**, 545–548.
Mirelman, D., Kleppe, G. & Jensen, H. B. (1975). *Eur. J. Biochem.* **55**, 369–373.
Ocada, Y., Amagase, S. & Tsugita, A. (1970). *J. Mol. Biol.* **54**, 219–246.
Phillips, D. C. (1967). *Proc. Nat. Acad. Sci., U.S.A.* **57**, 484–495.
Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). In *Aspects of Protein Structure* (Ramachandran, G. N., ed.), pp. 121–135, Academic Press, London.
Remington, S. J. & Matthews, B. W. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 2180–2184.
Remington, S. J., Ten Eyck, L. F. & Matthews, B. W. (1977). *Biochem. Biophys. Res. Commun.* **75**, 265–269.
Remington, S. J., Anderson, W. F., Owen, J., Ten Eyck, L. F., Grainger, C. T. & Matthews, B. W. (1978). *J. Mol. Biol.* **118**, 81–98.
Richards, F. M. (1968). *J. Mol. Biol.* **37**, 225–230.
Rossmann, M. G. & Argos, P. (1976). *J. Biol.* **105**, 75–96.
Schindler, M., Assaf, Y., Sharon, N. & Chipman, D. M. (1977). *Biochemistry*, **16**, 423–431.
Secemski, I. I. & Lienhard, G. E. (1971). *J. Amer. Chem. Soc.* **93**, 3549–3550.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzahi, E. & Inouye, M. (1966). *Cold Spring Habor Symp. Quant. Biol.* **31**, 77–84.
Troitskii, A. V., Chirgadze, Y. N. & Brazhnikov, E. G. Doklady (1979). *Biokhimiya (Moscow)*, **44**, 1864–1876.
Tsugita, A. (1971). In *The Enzymes* (Boyer, P. D., ed.). vol. 5 3rd edit., pp. 343–411. Academic Press, New York.
Tsugita, A., Inouye, M., Terzagi, E. & Streisinger, G. (1968). *J. Biol. Chem.* **243**, 391–397.
Vernon, C. A. (1967). *Proc. Roy. Soc. ser. B*, **167**, 389–401.
Warshel, A. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 5250–5254.
Warshel, A. & Levitt, M. (1976). *J. Mol. Biol.* **103**, 227–249.

J. Mol. Biol. (1982) **154**, 525–535

Amino Acid Substitutions Far From the Active Site of Bacteriophage T4 Lysozyme Reduce Catalytic Activity and Suggest that the C-terminal Lobe of the Enzyme Participates in Substrate Binding

M. G. Grütter† and B. W. Matthews‡

Institute of Molecular Biology and Department of Physics
University of Oregon, Eugene, Ore. 97403, U.S.A.

(Received 19 August 1981)

As part of a systematic study of the effect of single amino acid substitutions on the structure, function and stability of the lysozyme from bacteriophage T4, we have characterized a mutant lysozyme for which the catalytic activity is reduced to 4% that of the wild-type enzyme, yet the location of the amino acid substitution is on the far side of the molecule, 25 Å from the active site.

Genetic and chemical analysis show that the mutant lysozyme differs from wild-type only in the replacement of Glu128 by a lysine residue. Crystallographic analysis confirms the location of the amino acid substitution, and also shows that the replacement of Glu128, which is on the surface of the protein, causes very little change in the three-dimensional structure of the lysozyme molecule. Thermodynamic analysis shows that the stability of the mutant lysozyme is comparable with that of native lysozyme.

These results suggest that Glu128 participates directly in substrate binding and/or catalysis, and argue against the possibility that the low activity of the mutant enzyme is due to an indirect effect such as destabilization of the lysozyme molecule, or modification of its three-dimensional structure.

Glu128 is located in the part of the carboxy-terminal domain of phage lysozyme that has no counterpart in the structure of hen egg-white lysozyme. We suggest that the role of this C-terminal domain is to bind the peptide cross-link that connects neighboring saccharide strands within the cell walls of *Escherichia coli*. This postulate is consistent with the known differences in specificity between phage lysozyme and hen egg-white lysozyme, and is also compatible with the activity of native and mutant phage lysozymes toward different bacterial cell walls.

1. Introduction

The lysozyme from bacteriophage T4 hydrolyzes the $\beta(1\rightarrow4)$ glycosidic linkage between the alternating units of *N*-acetylmuramic acid and *N*-acetylglucosamine in bacterial cell walls (Fig. 1). This is the same bond as that cleaved by hen egg-

† Present address: Abt. Biophysikalische Chemie, Biozentrum der Universitat Basel, CH-4056 Basel, Switzerland.

‡ Author to whom all correspondence should be addressed.

0022–2836/82/030525–11 $02.00/0 © 1982 Academic Press Inc. (London) Ltd.

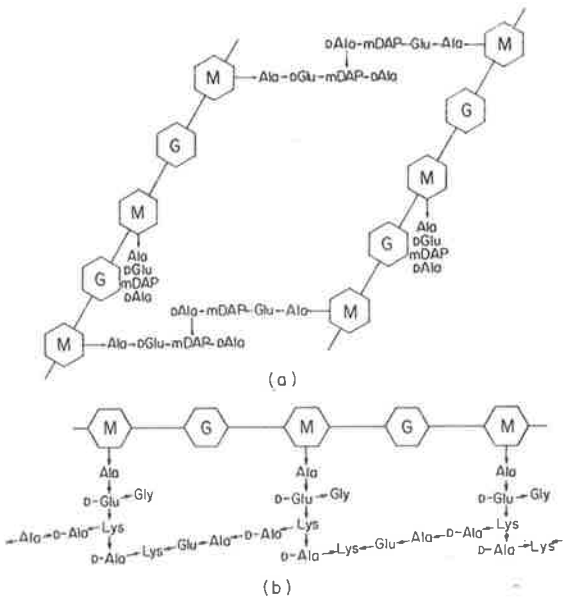


FIG. 1. Schematic representation of the peptidoglycan from (a) *E. coli*, and (b) *M. lysodeikticus*. From Ghuysen (1968) and Tsugita (1971). In (b) an intra-strand connection is shown, although both intra- and inter-strand bridges occur (e.g. see Schleifer & Kandler (1972) and references therein).

white lysozyme. However, for cell walls of *Escherichia coli*, the natural substrate of T4L†, the phage enzyme is 250-fold more active than that from hen egg-white. The specificities of the respective enzymes are different in that HEWL will cleave polymers of GlcNAc, or alternating polymers of GlcNAc-MurNAc, whereas T4L requires the presence of at least a part of the peptide which, in *E. coli* cell walls, links neighboring oligosaccharides in the bacterial cell wall (Fig. 1). (Tsugita *et al.*, 1968; Tsugita & Inouye, 1968; Mirelman *et al.*, 1975; Jensen *et al.*, 1976; Bienkowska & Taylor, 1979.)

Although the amino acid sequences of HEWL and TL4 are not homologous, the respective three-dimensional backbone conformations of the two lysozymes do have some elements that are similar (Rossman & Argos, 1976; Levitt & Chothia, 1976; Remington & Matthews, 1978). We have recently shown that there are also similarities between the two lysozymes in the way in which they bind oligosaccharides, and probably in their respective mechanisms of catalysis (Anderson *et al.*, 1981; Matthews *et al.*, 1981*b*). Taken together, these similarities suggest that the lysozymes from hen egg-white and bacteriophage T4 may have evolved from a common precursor (Matthews *et al.*, 1981*a*).

The part of T4L that corresponds with the structure of HEWL includes mainly the first half of the molecule. Most of the C-terminal half of phage lysozyme, which forms the upper domain in Figure 2, has no counterpart in HEWL (e.g. see Figs 1 and 3 of Matthews *et al.*, 1981*b*). In this paper we present evidence suggesting that

† Abbreviations used: T4L, bacteriophage T4 lysozyme; HEWL, hen egg-white lysozyme; GlcNAc, *N*-acetylglucosamine; MurNAc, *N*-acetylmuramic acid.

the role of the C-terminal domain, which is present only in the phage enzyme, is to bind the cross-linking peptide of *E. coli* cell walls.

2. Experimental

(a) *Phage strains*

Bacteriophage T4 strain cseBU56 was originally obtained as a generous gift from Dr D.-H. Liebscher (Liebscher *et al.*, 1974*a*,*b*) and was found to have 2 unlinked mutations, one in the lysozyme gene and another in an unidentified gene. The mutations were separated by a genetic cross (cf. Streisinger *et al.*, 1966) and the recombinant strain (hereinafter CSE) carrying the lysozyme mutation was crossed with known deletion mutants to map the region of the mutation within the lysozyme gene.

(b) *Purification; enzyme assay; crystallization*

The mutant lysozyme was purified as described (Tsugita & Inouye, 1968; Tsugita *et al.*, 1968; Matthews *et al.*, 1973). Enzymatic assays with *E. coli* cell walls and *Micrococcus lysodeikticus* cell walls were performed according to Tsugita (1971).

Crystals isomorphous with the native enzyme crystals were obtained from the mutant enzyme under the same, or closely similar, conditions as used for the native enzyme (Remington *et al.*, 1978).

(c) *Chemical characterization of amino acid substitution*

Following fragmentation of the mutant lysozyme with cyanogen bromide we isolated a peptide running from Asp127 to the C terminus (Leu164) (Tsugita & Inouye, 1968). The peptide was sequenced by Dr J. Becker (Oregon State University) by standard Edman degradation. In contrast to the native enzyme, where cleavage occurs at Met120, the mutant is cleaved at the Trp126-Asp127 linkage (presumably in addition to Met120).

(d) *Crystallographic data collection*

A full 3-dimensional data set was measured to 2·4 Å resolution by standard precession photography (Matthews *et al.*, 1972; Remington *et al.*, 1978). Data collection statistics are given in Table 1.

TABLE 1

Data processing statistics

	Native	Mutant CSE
Films	12	12
Average R_{sym}† (%)	4·8	6·7
Average R_{sca} (%)	5·1	7·1
R_{merge} (%)	2·8	3·8
Average isomorphous difference (%)	—	10·6
Cell dimensions		
a, *b* (Å)	61·2	61·4
c (Å)	96·8	96·8

† The respective *R*-values are as defined previously (Matthews *et al.*, 1972). R_{sym} gives the agreement between symmetry-averaged intensities on the same film; R_{sca} gives the agreement between symmetry-averaged intensities recorded on stronger and weaker films in the same film pack; and R_{merge} gives the agreement between structure amplitudes measured on different films.

Difference electron density maps were calculated using the isomorphous replacement phases, and a model of the substituted amino acid side-chain was built in an optical comparator.

3. Results

(a) *Amino acid substitution*

From genetic crosses with bacteriophage containing known mutations in the lysozyme gene, it was determined that the amino acid replacement was located between residue 120 and the carboxy terminus (residue 164). Edman degradation of the cyanogen bromide fragment containing the C-terminal residues showed its sequence to be Asp-Lys-Ala-Ala-Ala-Val-Asn-Leu-Ala-Lys-Ser-Arg- . . . This is identical with the amino acid sequence of the native enzyme, beginning at Asp127, except that Glu128 is replaced by a lysine residue (Tsugita & Inouye, 1968). The lysozyme from mutant CSE therefore has the substitution Glu128 → Lys. This is also consistent with the restrictions placed on possible mutations by the nature of the genetic code. When bromouracil is used as a mutagen, as was done in this case (Liebscher *et al.*, 1974*a*), the resultant mutations are expected to generate changes in the DNA of the form A·T ↔ G·C (Freese, 1963). Taking into account all possible codons for glutamic acid, and considering all possible new codons that might be generated by base changes of the above type, it can easily be shown that any such new codon will code either for lysine (as is observed) or glycine. As discussed below, the crystallographic results also show clearly that the substitution is at Glu128. In the three-dimensional structure of phage lysozyme (Fig. 2), Glu128 is located at the "back" of the C-terminal domain, 25 Å from the active site cleft.

Our results disagree with the findings reported by Liebscher *et al.* (1974*a*,*b*). These workers reported that in the lysozyme from the mutant cseBU56, Tyr88 is replaced by a histidine residue. In addition, we do not find that the lysozyme from the mutant strain cseBU56 is unusually cold-sensitive, as was reported by Liebscher *et al.*

(b) *Crystallographic results*

The 2·4 Å resolution electron density map showing the difference between the mutant and wild-type lysozyme is virtually featureless, except in the vicinity of the side-chain of Glu128. Here there is a pronounced negative peak of magnitude 7·1σ and a nearby positive peak (5·8σ), where σ is the r.m.s. value of the difference density through the unit cell. The difference density map in the vicinity of Glu128 is shown in Figure 3(a) and (b). The major negative peak is larger by a factor of two than any other feature in the difference electron density map throughout the unit cell.

Clearly, Glu128 has been replaced by a less electron-dense side-chain. The other positive and negative features suggest that the side-chain of Arg125 moves toward the position previously occupied by the carboxyl of Glu128. There are also indications (Fig. 3(b)) that some of the Asp127 side-chains in the crystal rotate so as to interact with Arg125 and with the new lysine side-chain at position 128. There

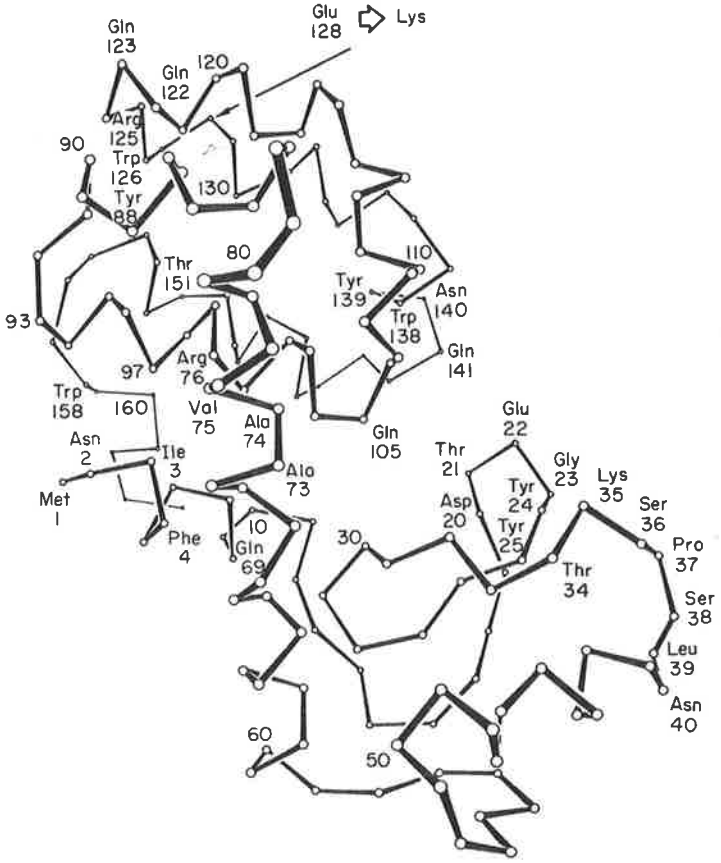


FIG. 2. Backbone conformation of bacteriophage T4 lysozyme showing the location of Glu128 relative to the active site cleft.

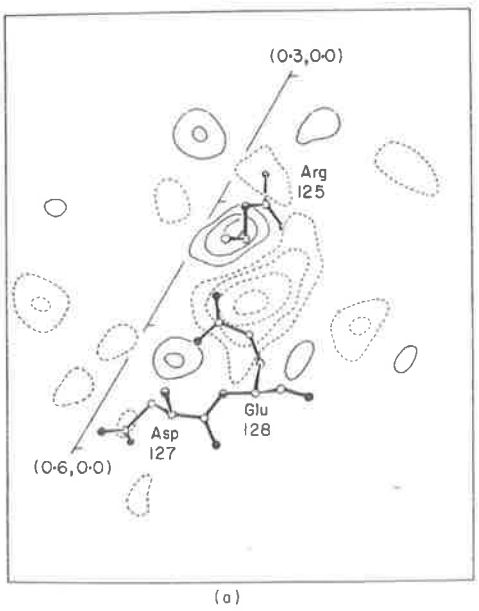
is no indication of an adjustment in the position of the polypeptide backbone of the protein in the vicinity of either Glu128 or Arg125.

The fact that the electron density map in parts of the lysozyme molecule away from Glu128 is featureless indicates that if any structural changes do occur elsewhere in the molecule they must be very small, probably not exceeding 0·1 Å. In the vicinity of the active site, the difference electron density map has the featureless appearance characteristic of most of the unit cell. There is no indication of a significant structural alteration being induced in this part of the molecule.

(c) *Activity and stability*

The measured activity of the purified mutant lysozyme toward *E. coli* cell walls is 4% that of the wild-type enzyme, the measurements being made under identical conditions.

Measurements of the thermal denaturation of wild-type phage lysozyme and a



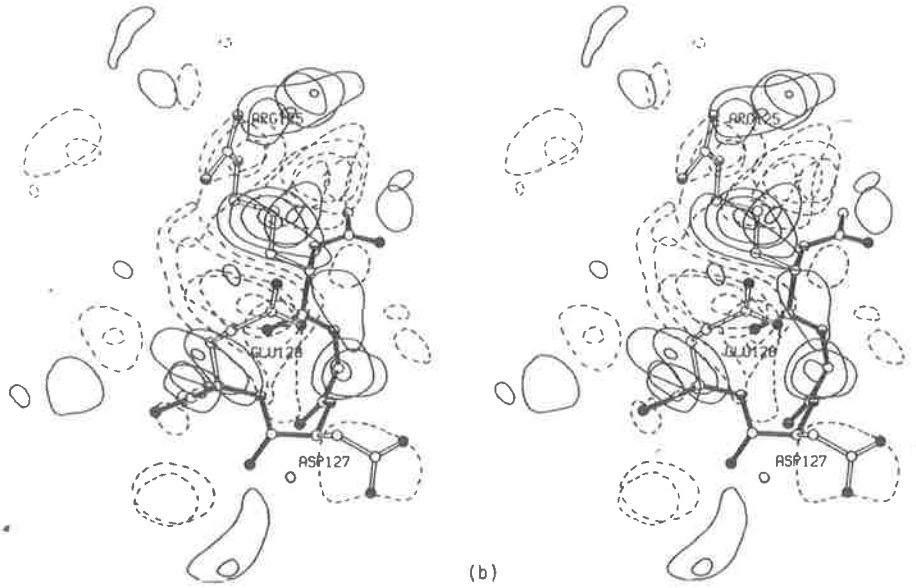


FIG. 3. Difference electron density between CSE mutant lysozyme and native lysozyme (coefficients: $F_{CSE} - F_{nat}$). Positive contours are shown as unbroken lines, negative contours as broken lines; contours are drawn at arbitrary equal intervals. Carbon atoms are drawn as open circles, oxygen as filled circles, and nitrogen as half-filled circles. The backbone bonds are filled, side-chain bonds are open (a) Mono drawing showing part of section $z = 0{\cdot}129$. (b) Stereo drawing showing the difference density in the vicinity of Glu128. In the stereo drawing, the direction of view has been reversed so that the reader is looking from the outside toward the surface of the molecule.

series of mutant lysozymes, including the CSE lysozyme, have been carried out (Schellman & Hawkes, 1980; Schellman *et al.*, 1981; Grütter *et al.*, unpublished results). Over a range of pH values, the melting temperature of the CSE mutant lysozyme is about 5 deg. C lower than that of wild-type lysozyme, whereas temperature-sensitive mutant lysozymes typically unfold at temperatures ten to 20 deg. C lower than wild-type lysozyme. Clearly, the stability of the CSE lysozyme is closer to that of the native enzyme than it is to any of the temperature-sensitive mutants that we have characterized. Also, under the standard assay conditions (20°C), the measured activity of these temperature-sensitive mutant lysozymes is comparable with that of wild-type lysozyme, even though their stability is lower.

4. Discussion

The above results, taken together, provide compelling evidence that the substitution of Glu128 by a lysine residue is directly responsible for a 25-fold reduction in the catalytic activity of the enzyme toward *E. coli* cell walls. The experimental data rule out the possibility that the reduction in activity is due to an indirect effect, such as destabilization of the molecule, or to large conformational changes in the enzyme induced by the amino acid substitution. The question, then, is how the replacement of a surface amino acid residue, 25 Å from the active site, can cause a substantial reduction in catalytic power.

There are a number of known mutations in the lysozyme genome that cause a substantial reduction in the activity of the enzyme (Tsugita, 1971). However, if one ignores those cases where the loss of activity is due to destabilization of the lysozyme molecule, in almost every case the location of the amino acid substitution that results in low catalytic activity is within the active-site cleft (Matthews & Remington, 1974; Remington *et al.*, 1978). The CSE mutant described here is one exception to this rule. The second known exception is the lysozyme from mutant strain 335D101, in which Arg125-Trp126 is substituted by Thr-Leu (G. Streisinger & J. Emrich Owen, personal communication; Remington *et al.*, 1978). In this case, the activity of the altered lysozyme is 4% that of wild-type; i.e. the same reduction in activity as occurs with the CSE lysozyme. In the case of the Arg-Trp → Thr-Leu substitution, it appears that the loss of activity can be attributed primarily to the replacement of the arginine rather than the tryptophan residue, since replacement of the tryptophan by a tyrosine, serine or glutamine residue leaves the activity of the lysozyme virtually unaffected (Remington *et al.*, 1978).

The part of the lysozyme structure that includes Arg125, Trp126 and Glu128 is shown in Figure 4. As can be seen, Arg125 and Glu128 occur within the first turn of an α-helix. Both the arginine and glutamate side-chains extend into the solvent. In the native crystal structure, for which the mother liquor is high salt (2·3 M-phosphate, 0·23 M-NaCl), the respective side-chains do not appear to form an ionic interaction, although this could well occur in a low-salt environment.

In order to rationalize the above results, we suggest that the role of Arg125 and Glu128 is to participate in the recognition and binding of the cross-linking peptide that links adjacent oligosaccharide strands within the cell wall of *E. coli* (Fig. 1).

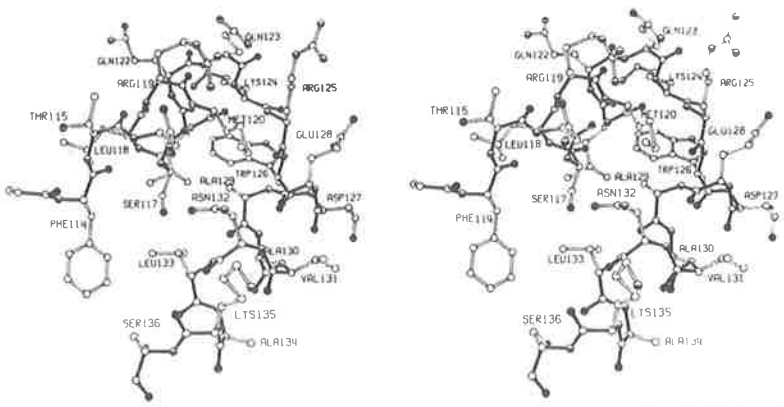


FIG. 4. Stereo view of the part of the C-terminal lobe of phage lysozyme that includes Arg125 and Glu128. The view is from the outside of the molecule, looking toward its surface.

In previous experiments (Anderson *et al.*, 1981) we have determined the modes of binding of a series of mono- and oligosaccharides to phage lysozyme. By analogy with hen egg-white lysozyme, we have labeled the respective sugar binding sites as A, B, C, D and E, where hydrolysis occurs between saccharides D and E. From the known orientation of the saccharide in site D, it can be inferred that the cross-linking peptide emanating from the 3-hydroxyl must extend across the face of the upper, C-terminal lobe of the lysozyme molecule (Anderson *et al.*, 1981). If one assumes that the cross-linking peptide continues to make contact with the surface of the lysozyme molecule, then the peptide occupies the general location illustrated in Figure 5, and passes in the immediate vicinity of Arg125 and Glu128. We postulate that this does in fact occur, and that Arg125 and Glu128 participate in the binding and alignment of the peptide. These interactions, in turn, help to position the saccharide part of the cell wall within the active-site cleft, in the optimum alignment for hydrolysis.

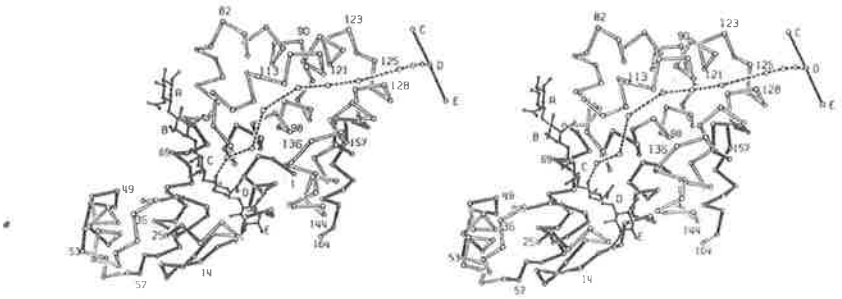


FIG. 5. Stereo diagram showing the proposed position (broken line) occupied by the peptide portion of the *E. coli* peptidoglycan. The Figure shows the backbone of phage lysozyme together with a bound oligosaccharide, as inferred from previous crystallographic studies. The filled bonds connect the parts of the phage lysozyme backbone that are "equivalent" to hen egg-white lysozyme; the open-bonded parts of the backbone have no counterpart in hen lysozyme. The direction of view is essentially the same as in Fig. 4, so that this Figure can be used to visualize the location of the active site relative to the portion of the molecule shown in Fig. 4.

The tertiary structure of the peptidoglycan in a typical bacterial cell wall is uncertain. There is a general agreement that successive saccharides in the glycan chain are rotated by approximately 180°, as we also propose (Fig. 5), but the geometry of the cross-linking peptide is in dispute. Keleman & Rogers (1971) propose an extended β-sheet conformation, whereas Formanek *et al.* (1974) favor a more compact $2{\cdot}2_7$ helical arrangement. X-ray diffraction patterns of dried cell walls show rather diffuse maxima, suggesting that the peptidoglycan is not very well-ordered, and should not be considered as crystal-like (Formanek *et al.*, 1974). It seems likely that peptidoglycan conformation is somewhat irregular, with many "lattice defects", and it is at these sites that lysozyme may well act.

It is already known that at least part of the cross-linking peptide must be present for hydrolysis by phage lysozyme to occur (Bienkowska & Taylor, 1979). We here extend this requirement by suggesting that the rate of catalysis can be influenced by interactions with the enzyme extending to the sixth or seventh peptide linkage, as illustrated in Figure 5.

A prediction of this hypothesis is that changes in the peptide linkage would modify the activity of the enzyme. That this is the case, is shown in Table 2. Cell walls of *Micrococcus lysodeikticus* are hydrolyzed by phage lysozyme at 1/2400th the rate of *E. coli* cell walls. As shown in Figure 1, the peptide linkage for *M. lysodeikticus* is different from that for *E. coli* (Ghuysen, 1968; Schleifer & Kandler, 1972). Furthermore, the activity of the CSE mutant lysozyme toward *M. lysodeikticus* cell walls is essentially identical with that of the wild-type enzyme, providing further evidence that Glu128 is involved in the binding of *E. coli* cell walls. For *M. lysodeikticus* cell walls, a similar interaction with Glu128 is apparently not possible, and the substitution of Glu128 by a lysine residue has no effect on the rate of catalysis.

These findings are also consistent with the structural relation between T4 phage and hen egg-white lysozymes. As discussed in the Introduction, the structures of T4L and HEWL have some similarity, but most of the C-terminal lobe of T4L has no counterpart in HEWL. This is illustrated in Figure 5, where the solid connections are used to designate the α-carbons of T4L that are structurally equivalent to HEWL. As can be seen, most of the upper, C-terminal lobe in the

TABLE 2

Lysozyme activity

	E. coli cell walls	*M. lysodeikticus* cell walls
Wild-type T4 phage lysozyme	100	0·041
CSE mutant lysozyme	4·1	0·036
Hen egg-white lysozyme†	0·4	0·007

Lysozyme activity is measured as cell walls hydrolyzed per gram of lysozyme per second, and is here given in arbitrary units normalized to wild-type activity toward *E. coli* cell walls equal to 100 units.

† Data from Tsugita (1971).

Figure consists of open bonds, indicating that HEWL has no counterpart to the region of T4L that we postulate to interact with the peptide cross-linkage. It is easy, then, to see how the function of the additional C-terminal residues in T4L is to increase the specificity and enhance the activity of this enzyme toward *E. coli* cell walls. Conversely, the phage enzyme needs the peptide cross-links to be present in order for hydrolysis to occur. In contrast, HEWL is seen to be an "all purpose" muramidase of broad specificity but lower efficiency.

We thank Drs G. Streisinger and J. (Emrich) Owen for advice and guidance in lysozyme genetics, K. Rine and E. G. Kottas for excellent technical assistance and Dr J. Becker of Oregon State University for amino acid sequence determinations. Mutant strain cseBU56 was kindly provided by Dr D.-H. Liebscher.

This study was supported in part by grants from the National Institutes of Health (GM21967; GM20066), the National Science Foundation (PCM-8014311), and a grant from the M. J. Murdock Charitable Trust.

REFERENCES

Anderson, W. F., Grütter, M. G., Remington, S. J. & Matthews, B. W. (1981). *J. Mol. Biol.* **147**, 523–543.

Bienkowska, K. & Taylor, A. (1979). *Eur. J. Biochem.* **96**, 581–584.

Formanek, H., Formanek, S. & Wawra, H. (1974). *Eur. J. Biochem.* **46**, 279–294.

Freese, E. (1963). In *Molecular Genetics* (Taylor, J. J., ed.), vol. 1, pp. 207–269, Academic Press, New York and London.

Ghuysen, J.-M. (1968). *Bacteriol. Rev.* **32**, 425–464.

Jensen, H. B., Kleppe, G., Schindler, M. & Mirelman, D. (1976). *Eur. J. Biochem.* **66**, 319–325.

Keleman, M. V. & Rogers, H. J. (1971). *Proc. Nat. Acad. Sci., U.S.A.* **68**, 992–996.

Levitt, M. & Chothia, C. (1976). *Nature (London)*, **261**, 552–558.

Liebscher, D. H., Dauberschmidt, R., Rapoport, S. M., Heitmann, P. & Rosenthal, H. A. (1974*a*). *FEBS Letters*, **44**, 173–177.

Liebscher, D.-H., Dauberschmidt, R., Rapoport, S. M., Heitmann, P. & Rosenthal, H. A. (1974*b*). *Mol. Gen. Genet.* **132**, 321–333.

Matthews, B. W. & Remington, S. J. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 4178–4182.

Matthews, B. W., Klopfenstein, C. E. & Colman, P. M. (1972). *J. Phys. E. Sci. Instrum.* **5**, 353–359.

Matthews, B. W., Dahlquist, F. W. & Maynard, A. Y. (1973). *J. Mol. Biol.* **78**, 575–576.

Matthews, B. W., Grütter, M. G., Anderson, W. F. & Remington, S. J. (1981*a*). *Nature (London)*, **290**, 334–335.

Matthews, B. W., Remington, S. J., Grütter, M. G. & Anderson, W. F. (1981*b*). *J. Mol. Biol.* **147**, 545–558.

Mirelman, D., Kleppe, G. & Jensen, B. B. ((1975). *Eur. J. Biochem.* **55**, 369–373.

Remington, S. J. & Matthews, B. W. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 2180–2184.

Remington, S. J., Anderson, W. F., Owen, J., Ten Eyck, L. F., Grainger, C. T. & Matthews, B. W. (1978). *J. Mol. Biol.* **118**, 81–98.

Rossmann, M. G. & Argos, P. (1976). *J. Mol. Biol.* **105**, 75–96.

Schellman, J. A. & Hawkes, R. B. (1980). In *Protein Folding* (Jaenicke, R., ed.), pp. 331–334, Elsevier/North-Holland Biomedical Press, Amsterdam.

Schellman, J. A., Lindorfer, M., Hawkes, R. B. & Grütter, M. G. (1981). *Biopolymers*, **20**, 1989–1999.

Schleifer, K. H. & Kandler, O. (1972). *Bacteriol. Rev.* **36**, 407–477.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzahi, E. & Inouye, M. (1966). *Cold Spring Harbor Symp. Quant. Biol.* **31**, 77–84.

Tsugita, A. (1971). In *The Enzymes* (Boyer, P. D., ed.), vol. 5, 3rd edit., pp. 343–411, Academic Press, New York.
Tsugita, A. & Inouye, M. (1968). *J. Mol. Biol.* **37**, 201–212.
Tsugita, A., Inouye, M., Terzaghi, E. & Streisinger, G. (1968). *J. Biol. Chem.* **243**, 391–397.

Edited by R. Huber

SECTION C. BACTERIOCHLOROPHYLL PROTEIN

(Reprinted from Nature, Vol. 258, No. 5536, pp. 573–577, December 18, 1975)

Chlorophyll arrangement in a bacteriochlorophyll protein from *Chlorobium limicola*

R. E. Fenna & B. W. Matthews

Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oregon 97403

The three-dimensional structure of a chlorophyll-containing protein has been determined by X-ray crystallography and shown to consist of three identical subunits, each containing a core of seven bacteriochlorophylls arranged in an irregular fashion and enclosed within an envelope of protein.

THE three-dimensional structure of a bacteriochlorophyll protein has been determined by X-ray crystallography. This is the first chlorophyll-containing protein for which the structure is known, and provides new information on the arrangement of chlorophyll *in vivo*. The bacteriochlorophyll protein is seen to consist of three identical subunits, each containing a core of seven bacteriochlorophyll molecules arranged in an irregular fashion and enclosed within an envelope of protein.

Green photosynthetic bacteria contain small amounts of a water-soluble bacteriochlorophyll protein which can be extracted in aqueous salt solutions without the use of detergents[1,2]. This complex has been shown to mediate in the transfer of excitation energy from the antenna pigment chlorobium chlorophyll to the photochemical reaction centres of the organism[2–4]. The complex from *Chlorobium limicola* has been shown by chemical studies to have a molecular weight of about 150,000 and to contain 21 ± 2 molecules of bacteriochlorophyll *a* (Bchl *a*), tightly, though not covalently associated with the protein[5]. Analyses of the absorption and circular dichroism spectra, recorded at 77 K, gave evidence for interaction between at least five molecules of Bchl *a* (ref. 6). Preliminary X-ray studies of two different crystal forms of the Bchl protein have shown it to be a trimer consisting of three identical subunits[7]. The form in space group $P6_3$ with unit cell dimensions: $a = b = 112.4$ Å and $c = 98.4$ Å, has one subunit per asymmetric unit and was used for the X-ray structural studies reported here. In this paper we describe the arrangement of the Bchl *a* molecules within the complex as deduced from a 2.8-Å resolution electron density map. Description of the X-ray structure determination will be confined to essential features and a comprehensive analysis will be presented elsewhere.

Structure determination

Details of the procedure used for growing large crystals of the Bchl protein have previously been described[7,8]. Diffraction data from the native crystals and isomorphous heavy atom derivatives were recorded photographically using an Elliott rotating anode X-ray generator and Enraf–Nonius precession cameras. Integrated intensities were measured from conventional precession photographs by means of a computer controlled Optronics rotating drum microdensitometer[9]. After symmetry averaging and merging the intensities from 23 different film packs, each set of data reduced to about 16,500 unique reflections, excluding the Friedel related measurements. This total includes 90% of the data to Bragg spacings of 2.8 Å together with 42% between 2.8 Å and 2.75 Å spacings. Excluding the weakest reflections, the mean error in amplitude on scaling together planes within the native and derivative data sets ranged from 2.8 to 3.4%. The crystals contain about 70% by volume solvent, and presumably for this reason do not diffract strongly, particularly at high angles. The resolution of 2.8 Å adopted for this study is the highest which could reasonably be attained using Buerger precession photographs.

Isomorphous heavy atom derivatives were prepared by soaking crystals of the native protein in solutions of heavy atom compounds in 0.01 M Tris-HCl buffer containing 1 M NaCl and 10% w/v $(NH_4)_2SO_4$ at *p*H 7.8. Two suitable derivatives were obtained by soaking in the presence of 5×10^{-4} M potassium chloroplatinite and saturating concentrations of methyl mercuric iodide, each for 7 d. A third, double derivative, was made by soaking in a solution containing both 5×10^{-4} M potassium chloroplatinite and an excess of methyl mercuric iodide, again for a period of 7 d. One heavy atom site in each of the two single derivatives was initially located from a difference Patterson synthesis of the hk0 zone. Additional sites in the mercury and double derivatives were then found from projection and three-dimensional difference Fourier maps in the usual way. Further refinement of the heavy atom parameters was carried out using a lack-of-closure least squares refinement procedure applied to a subset of the three-dimensional data, selected to include all of the centric data together with 25% of the acentric reflections[10]. The heavy atom parameters used in the final phase calculation are listed in Table 1.

"Best" phases were calculated by the multiple isomorphous replacement method including contributions from measurements of anomalous scattering of the heavy atoms[11–13]. The mean figure of merit of 0.57 is probably a conservative estimate since the value of 0.66 for the average r.m.s. lack of closure over E indicates that the E values used had been overestimated. A comparison of the mean heavy atom scattering with the E values (Table 1) shows that the single site platinum derivative contributes substantially less to the phasing than do the mercury and double derivatives. This is not surprising in view of the difficulty in measuring the comparatively small changes in intensities produced by the introduction of a single platinum atom into an asymmetric unit of this size. The Fourier calculation was performed using a fast Fourier transform

Table 1 Heavy atom parameters

Derivative	Z	x	y	z	B	N	$\langle f_h \rangle$	E	R (%)
Potassium chloroplatinite	28.2	0.113	0.443	−0.251	27.5	5,445	44	35	52
Methyl mercuric iodide	40.3	0.379	0.582	−0.007	25.1	5,444	75	38	47
	16.7	0.301	0.439	−0.104	34.6				
	9.3	0.279	0.434	−0.143	9.9				
Double derivative	31.4	0.113	0.443	−0.250	35.0	5,460	85	43	52
	38.7	0.380	0.582	−0.006	25.9				
	14.1	0.301	0.437	−0.104	32.8				
	10.0	0.278	0.434	−0.142	16.4				
	9.6	0.510	0.204	−0.397	23.8				

Z, relative occupancy of heavy atom site in arbitrary units; x, y, z, fractional coordinates; B, isotropic temperature factor (Å²); N, number of reflections included in the refinement. $\langle f_h \rangle$, E and R are respectively the mean heavy atom scattering, the lack of closure error and the residual for the centric reflections to a resolution of 2.8 Å.

program provided by L. F. Ten Eyck[14]. The electron density distribution was sampled at intervals of 0.73 Å ($c/135$) normal to the c axis and the contour levels traced on to Dayco plastic sheets stretched on to aluminium frames, suitable for observation in an optical comparator[15,16].

Description of the molecule

We have suggested previously that in the $P6_3$ crystal form the trimers are positioned on the threefold rather than the 6_3 axes of the unit cell[7]. The electron density map clearly confirms this conclusion. The packing arrangement gives rise to extensive channels of solvent 112.4 Å (the a cell dimension) apart and approximately 62 Å in diameter centred on the 6_3 axes and running parallel to the c axis throughout the crystal. Each trimer extends 57 Å along the threefold axis and has a maximum diameter of 83 Å at right angles to it. Thus, the overall dimensions of the trimer correspond closely to those deduced earlier from considerations of symmetry and molecular packing. It was possible to define the inter-subunit boundary by making only one cut through contoured electron density. The subunits are, however, packed very closely together and extensive interactions between them undoubtedly occur at the interfaces. In contrast, the contacts between one trimer and its neighbours in this crystal form are very tenuous.

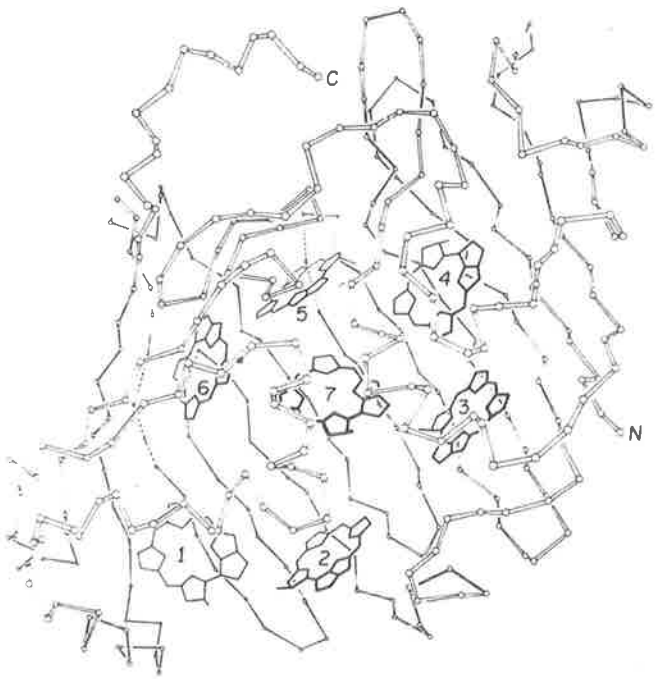


Fig. 1 Schematic diagram showing the arrangement of the polypeptide backbone and chlorophyll core of one subunit of the bacteriochlorophyll protein. Presumed α-carbon positions are indicated by circles. The connectivity of the polypeptide chain is in some instances uncertain (see text), and particularly ambiguous regions are indicated by broken lines. For clarity, the magnesium atoms, the chlorophyll ring substituents and the phytyl chains, except for the first bond, are omitted. The direction of view is from the threefold axis, which is horizontal, towards the exterior of the molecule.

The amino acid sequence of the Bchl protein has not yet been determined; it was therefore necessary to rely solely on the continuity of the electron density in attempting to determine the course of the polypeptide chain. By placing markers in the electron density map at probable α-carbon positions it was possible to follow extensive sections of the protein backbone. In some regions, however, particularly in loops on the surface of the molecule, breaks in the continuity do occur and at this stage we are unable to define a unique course for the whole of the polypeptide chain. But, apart from a few ambiguities, the continuity is generally quite good, and is sufficiently defined to give an overall view of the protein conformation. The total number of amino acid residues in each subunit, based on the number of marked α-carbon positions is 327, although it is probable that some residues have been missed in regions where the electron density is not well defined. Assuming a mean residue weight of 110, calculated from the amino acid composition[5], the minimum molecular weight of each subunit, including Bchl, becomes 42,300, and for the trimer 127,000.

The conformation of the polypeptide chain is illustrated in Figs 1 and 4, and can be seen to form a distorted hollow cylinder. In our present interpretation of the electron density map, one end of the cylinder seems to be "open", but is occluded by the carboxyl-terminal helix of an adjacent subunit of the trimer. It is also possible to interpret the electron density map in the vicinity of this inter-subunit contact in such a way that a single subunit would form a "closed" hollow cylinder. At present we cannot distinguish with certainty between these two possibilities. The hollow cavity within the cylinder contains the Bchl aggregate. The area of the cylindrical wall which is exposed to the solvent in the trimer is composed almost entirely of 15 strands of β sheet which seem to be mainly anti-parallel. The side of the cylinder in contact with an adjacent subunit has a more complex structure, consisting of four short lengths of α helix interspersed with regions of irregular conformation.

Bacteriochlorophyll core

Within the core of each subunit the electron density clearly corresponds to an array of chlorophyll molecules. Disks of density indicate the positions of the porphine rings, whereas the phytyl side chains can be seen as resolved ribbons of density. Seven Bchl molecules pack within one subunit, corresponding to a total of twenty-one for the trimer, in agreement with the chemical studies[5]. The chemical structure of Bchl a is known (Fig. 2), and differs from chlorophyll a in having two extra hydrogen atoms in ring II and an acetyl group in place of the vinyl group on ring I. The absolute configurations at the

asymmetric carbon atoms in rings II, IV and V have been established[17,18]. In addition the X-ray crystal structure of ethyl chlorophyllide *a* (a *trans*-esterified derivative of chlorophyll *a* in which the phytyl group is replaced by ethyl) has been determined[19–21].

The seven Bchl *a* molecules are confined within an ellipsoid of axial dimensions: 45×35×15 Å (Fig. 3). The average centre–centre nearest neighbour distance between porphine rings is 12 Å, which may be compared with 24 Å, the closest distance between chromophores in adjacent subunits. The orientations of the porphine rings do not conform to any repetitive pattern, although their planes do, to a first approximation, lie parallel with one another. This may be demonstrated by taking the mean of the direction cosines of the perpendiculars to each of the rings to define a plane which is most nearly parallel with the planes of the seven porphine rings. The angles made by the individual porphine planes with this calculated plane lie between 10° and 40°. In contrast, the azimuthal orientations of the near coplanar rings vary considerably. The normal to the calculated plane makes an angle of 62° with the threefold axis of the trimer and the centre of the aggregate in each subunit is at a distance of about 20 Å from the threefold axis. Thus, in the complete trimer the Bchl *a* aggregates of the individual subunits combine to form a triangular funnel about the threefold axis.

In each subunit the phytyl chains lie close together, sandwiched between five of the porphine rings on one side and the other two rings and the β sheet wall of protein on the other. The hydrocarbon chains are in most cases in an extended conformation, although in one instance (Bchl 1) the chain is bent into a complete U-shaped loop. The phytyl chains of chlorophylls 4, 5 and 6 lie parallel, forming a planar structure in close contact with the β-sheet wall of the protein. This hydrophobic core of phytyl chains is completely buried in the protein, a thermodynamically favourable arrangement analogous to the interior of other globular proteins in which most of the hydrophobic residues are found buried out of contact with the solvent. A characteristic of β structure is the alternation of the directions of the amino acid side chains relative to the plane of the sheet. It seems likely that in this case the inwardly directed side chains are predominantly hydrophobic, forming close interactions with the phytol aggregate, whereas the residues directed towards the exterior are largely hydrophilic, forming favourable interactions with the solvent. The β sheet thus provides a suitable amphipathic layer, shielding the hydrophobic Bchl aggregate from the aqueous surroundings. Since the direction of the side chains alternates along the polypeptide chain, it will be interesting to see whether the sequence of these strands conforms to a pattern of alternating hydrophobic and hydrophilic residues.

The electron density map gives no evidence for chemical interactions between neighbouring chlorophyll molecules, either direct or through bridging water molecules, other than the hydrophobic associations which undoubtedly occur between the phytyl chains. There is, however, good evidence for extensive interactions between the chlorophyll molecules and the surrounding protein shell. The presence of a fifth ligand to the Mg atom of each of the Bchl *a* molecules is indicated by electron density protruding from the centres of each of the porphine rings. In five cases this density extends to form a continuous bridge between the Mg atom and the polypeptide chain. Without knowledge of the amino acid sequence it is not possible at this stage to characterise these ligands. Peptide oxygens or amino acid side chains such as Asp, Glu and their amides, Ser, Thr, His or Met could interact with the Mg directly or by hydrogen bonding to a coordinated water molecule. In addition, the Bchl *a* ring substituents include a number of potential hydrogen bond acceptors such as the oxygen-containing groups mentioned earlier. Most of these groups are close to the periphery of the aggregate and are sufficiently close to the protein electron density to be expected to form hydrogen bonds with suitable donors in the polypeptide chain. Thus, it seems that the Bchl *a* molecules are anchored to the protein through extensive hydrogen bonding and liganding to the Mg atom, in addition to hydrophobic interactions. The distribution and orientations of the chlorophyll molecules are therefore governed largely

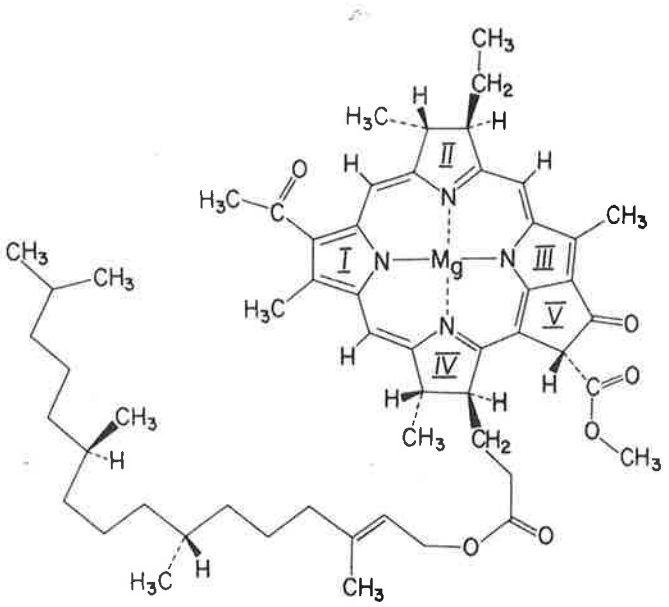


Fig. 2 Structure of bacteriochlorophyll *a*, indicating the absolute configurations of the asymmetric carbons atoms.

by specific interactions with the protein rather than with one another. The Bchl molecules do not occur in regular one- or two-dimensional arrays as has been observed in crystal structures of chlorophyll derivatives, and postulated as models for chlorophyll arrangement *in vivo*[10–22].

Energy transfer

Fluorescence studies suggest that in the green photosynthetic bacteria this Bchl *a* protein functions as an intermediary in the transfer of excitation energy from the antenna pigment, chlorobium chlorophyll 660, to the photochemical reaction centre, P840. Mechanisms for energy transfer between the components of the photosynthetic apparatus of both bacteria and higher plants have recently been discussed by several authors[23–25]. Energy transfer between chromophores separated by relatively short distances (10–20 Å) may be described in terms of an exciton model[26–28]. For a given electronic transition, coulombic interactions between the transition dipole moments of neighbouring identical, or near identical, molecules in an aggregate give rise to a set of N discrete exciton energy levels, where N is the minimum number of interacting species. Splitting of the main 809-nm absorption band and the presence of a complex CD spectrum at liquid nitrogen temperatures have been observed for this Bchl *a* protein. These observations were interpreted in terms of an exciton interaction involving a minimum of five or possibly six Bchl *a* molecules with estimated nearest neighbour distances of 12–15 Å (ref. 6 and J. M. Olson, unpublished). The work described here shows that there are in fact seven Bchl *a* molecules separated by average nearest neighbour distances of 12 Å in each subunit of the trimer. The interactions between Bchl *a* molecules within each subunit

thus appear to be appropriately described in terms of an exciton model. The magnitudes of the exciton interactions are critically dependent upon the relative orientations of the transition dipoles and on the inverse cube of the distances between the chromophores. We have observed a high degree of non-uniformity in the azimuthal orientations of neighbouring Bchl molecules in this aggregate (Fig. 3). This observation is in good agreement with an interpretation of the complex low temperature absorption and CD spectra in terms of the exciton theory. Furthermore, since the positions and orientations of the Bchl *a* molecules are now known, it should be possible to calculate the expected exciton interactions, and hence estimate the numbers and magnitudes of the absorption splittings and components of the CD spectrum. Such calculations for model systems using experimentally determined dipole moments or calculated point monopoles have been described[29,30]. It is hoped that this Bchl *a* protein will provide a useful model for testing current theories of exciton interactions between chlorophyll molecules; so enabling more reliable structural evaluations to be made for other chlorophyll containing systems for which multicomponent CD spectra have been observed[31-32].

In comparison with the Bchl protein described here, very little is known of the environments of the light-harvesting chlorobium chlorophyll 660 and the reaction centre, P840; and virtually nothing about the relative separation and spatial distributions of all three components[33]. Any discussion of the mechanism of energy transfer to and from the Bchl *a* protein remains speculative in the absence of such information. At chromophore separations of 50–100 Å, energy transfer is thought to occur by resonance, which, unlike exciton transfer, has an inverse sixth power dependence on separation distances[34]. If, however, the other chlorophyll components occur in proteins with molecular dimensions comparable with the Bchl *a* protein, and these proteins are closely associated in the membrane, then energy transfer would be required to occur over distances in the range of 20–50 Å (as for example between subunits of the Bchl-protein trimer) intermediary between the exciton theory and Forster's resonance transfer.

Chlorophyll proteins

The results described here represent the first structure determination of a chlorophyll-containing protein. In common with other globular proteins it contains the usual types of secondary structure such as α helices and β sheet. Unlike any other protein for which the X-ray crystal structure has been determined, however, it contains a large non-protein core as was originally proposed by J. M. Olson[35]. Presumably both the protein and the Bchl are necessary to maintain the integrity of the complex. The molecule is also unusual in being trimeric, having a cyclic arrangement of subunits involving extensive heterologous interactions. The functional significance of the trimeric nature of the molecule is obscure, though if the protein were extrinsically associated with the photosynthetic membrane it could indicate the presence of an intrinsic membrane component also possessing a threefold axis of symmetry.

Recent progress in the isolation and partial characterisation of a number of reaction centre and antenna chlorophyll proteins from photosynthetic membranes suggests that most, if not all, of the chlorophyll in both green plants and bacteria occurs in close association with relatively low molecular weight proteins[36,37]. The Bchl *a* protein described here is somewhat atypical in that it is highly water-soluble, whereas most other chrorophyll-proteins require the use of detergents to facilitate their solubilisation from membranes. There is however no evidence at present to suggest that the organisation of chlorophyll in these hydrophobic proteins should be grossly different from the model we have described. All of the current models describing the state of chlorophyll *in vivo* place emphasis on either direct

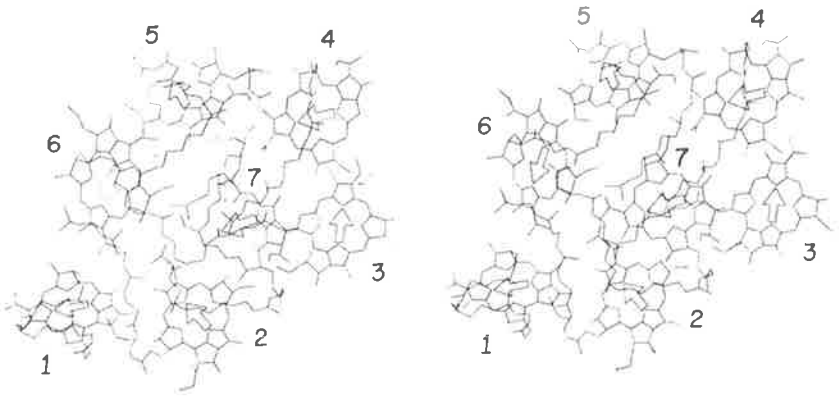


Fig. 3 Stereo diagram illustrating the chlorophyll arrangement within one subunit of the bacteriochlorophyll protein. The direction of view is approximately normal to the "average" chlorophyll plane (see text). The arrows, directed from ring III to ring I, are parallel to the Q_y transition moments of the respective bacteriochlorophylls. The magnesium atoms are omitted.



Fig. 4 Stereo drawing showing the arrangement of the chlorophyll core within the polypeptide envelope of the bacteriochlorophyll protein. The direction of view is the same as in Fig. 1, and the location of the threefold axis is indicated by the two horizontal lines.

interaction involving Mg liganding or hydrogen bonding via bridging water molecules as the major factors determining the relative orientations of the porphine rings[19,38,39]. Our new findings for this particular Bchl-protein indicate that interactions between chlorophyll and protein including liganding to the magnesium atom, hydrogen bonding and hydrophobic interactions are of major importance in determining the arrangement of the chlorophyll molecules. Furthermore we suggest that these types of interaction could be of universal significance in determining the state of chlorophyll *in vivo*.

Our results may be summarised as follows: the three-dimensional structure of a chlorophyll-containing protein has been determined by X-ray crystallography and shown to consist of three identical subunits, each containing a core of seven bacteriochlorophylls enclosed within an envelope of protein. The bacteriochlorophyll molecules are confined within a flattened disk-shaped region with their porphine rings lying roughly parallel to the disk. In contrast to current models for chlorophyll arrangement *in vivo*, the chlorophyll packing is dominated by interactions between chlorophyll and protein rather than between chlorophyll and lipid or between adjacent chlorophyll molecules. Furthermore the chlorophylls are arranged in an irregular fashion rather than in strictly ordered one-dimensional or two-dimensional arrays. It is suggested that the arrangement of chlorophyll seen here, in close association with protein, typifies the usual arrangement of chlorophyll *in vivo*.

We thank Dr J. M. Olson for supplying the bacteriochlorophyll protein and for his help, Dr S. Perez for help with data collection, Dr L. F. Ten Eyck for providing the Fourier transform and parameter refinement programs, S. J. Remington for the stereo-plotting program, and Dr L. H. Weaver, Mr W. R. Kester, Mrs H. F. Matthews and Ms J. Stephens for their assistance. This work was supported by grants from the US National Science Foundation and the National Institutes of Health. B. W. M. is an Alfred P. Sloan Research Fellow.

Received October 7; accepted October 22, 1975.

[1] Olson, J. M., and Romano, C. A., *Biochim. biophys. Acta*, **59**, 726–728 (1962).
[2] Sybesma, C., and Olson, J. M., *Proc. natn. Acad. Sci. U.S.A.*, **49**, 248–253 (1963).
[3] Sybesma, C., and Vredenberg, W. J., *Biochim. biophys. Acta*, **75**, 439–441 (1963).
[4] Sybesma, C., and Vredenberg, W. J., *Biochim. biophys. Acta*, **88**, 205–207 (1964).
[5] Thornber, J. P., and Olson, J. M., *Biochemistry*, **7**, 2242–2249 (1968).
[6] Philipson, K. D., and Sauer, K., *Biochemistry*, **11**, 1880–1885 (1972).
[7] Fenna, R. E., Matthews, B. W., Olson, J. M., and Shaw, E. K., *J. molec. Biol.*, 231–240 (1974).
[8] Olson, J. M., Koenig, D. F., and Ledbetter, M. C., *Archs Biochem. Biophys.*, **129**, 42–48 (1969).
[9] Matthews, B. W., Klopfenstein, C. E., and Colman, P. M., *J. Phys. E.*, **5**, 353–359 (1972).
[10] Dickerson, R. E., Kendrew, J. C., and Strandberg, B. E., *Acta Crystallogr.*, **14**, 1188–1195 (1961).
[11] Blow, D. M., and Crick, F. H. C., *Acta Crystallogr.*, **12**, 794–802 (1959).
[12] North, A. C. T., *Acta Crystallogr.*, **18**, 212–216 (1965).
[13] Matthews, B. W., *Acta Crystallogr.*, **20**, 82–86 (1966).
[14] Ten Eyck, L. F., *Acta Crystallogr.*, **A29**, 183–191 (1973).
[15] Richards, F. M., *J. molec. Biol.*, **37**, 225–230 (1968).
[16] Colman, P. M., Jansonius, J. N., and Matthews, B. W., *J. molec. Biol.*, **70**, 701–724 (1972).
[17] Fleming, I., *Nature*, **216**, 151–152 (1967).
[18] Brockman, J., Jr, *Angew. Chem. int. Edit.*, **7**, 221 (1968).
[19] Strouse, C. E., *Proc. natn. Acad. Sci. U.S.A.*, **71**, 325–328 (1973).
[20] Chow, H. C., Serlin, R., and Strouse, C. E., *J. Am. chem. Soc.* (in the press).
[21] Kratky, C., and Dunitz, J. D., *Acta Crystallogr.*, **B31**, 1586–1589 (1975).
[22] Fisher, M. S., Templeton, D. H., Zalkin, A., and Calvin, M., *J. Am. chem. Soc.*, **94**, 3613–3619 (1972).
[23] Robinson, G. W., *Brookhaven Symp. Biol.*, **19**, 16–48 (1966).
[24] Seely, G. R., *J. theor. Biol.*, **40**, 173–187 (1973).
[25] Knox, R. S., *Bioenergetics of Photosynthesis*, ch. IV (Academic, New York, 1975).
[26] Kasha, M., *Radiat. Res.*, **20**, 55–71 (1963).
[27] Kasha, M., Rawls, H. R., and El-Bayoumi, M. A., *Pure appl. Chem.*, **11**, 371–392 (1965).
[28] Tinoco, I., Jr, *Radiat. Res.*, **20**, 133–139 (1963).
[29] Houssier, C., and Sauer, K., *J. Am. chem. Soc.*, **92**, 779–791 (1970).
[30] Philipson, K. D., Tsai, S. C., and Sauer, K., *J. phys. Chem.*, **75**, 1440–1445 (1970).
[31] Sauer, K., Dratz, E. A., and Coyne, L., *Proc. natn. Acad. Sci. U.S.A.*, **61**, 17–24 (1968).
[32] Reed, D. W., and Ke, B., *J. biol. Chem.*, **248**, 3041–3045 (1973).
[33] Fowler, C. F., Nugent, N. A., and Fuller, R. C., *Proc. natn. Acad. Sci. U.S.A.*, **68**, 2278–2282 (1971).
[34] Forster, T., *Discuss. Faraday Soc.*, **27**, 7–17 (1959).
[35] Olson, J. M., in *The Chlorophylls* (edit. by Vernon, L. P., and Seely, G. R.), 413–425 (Academic, New York, 1966).
[36] Feher, G., *Photochem. Photobiol.*, **14**, 373–387 (1971).
[37] Thornber, J. P., *A. Rev. Physiol.*, 128–158 (1975).
[38] Cotton, T. M., Trifunac, A. D., Ballschmiter, K., and Katz, J. J., *Biochim. biophys. Acta*, **368**, 181–198 (1974).
[39] Fong, F. K., *Proc. natn. Acad. Sci. U.S.A.*, **71**, 3692–3695 (1974).

J. Mol. Biol. (1979) **131**, 259–285

Structure of a Bacteriochlorophyll *a*-Protein from the Green Photosynthetic Bacterium *Prosthecochloris aestuarii*

B. W. Matthews, R. E. Fenna†, M. C. Bolognesi‡, M. F. Schmid

Institute of Molecular Biology and Department of Physics
University of Oregon, Eugene, Ore. 97403, U.S.A.

AND

J. M. Olson

Biology Department, Brookhaven National Laboratory
Upton, N.Y. 11973, U.S.A.

(Received 9 November 1978)

The three-dimensional structure of a water-soluble bacteriochlorophyll *a*-containing protein from the green photosynthetic bacterium *Prosthecochloris aestuarii* has been determined by X-ray crystallography from a 2·8 Å resolution electron density map based on four isomorphous derivatives. Details of the crystallographic procedures used to obtain the map are presented.

The bacteriochlorophyll *a*-protein is shown to consist of three identical subunits, tightly packed around a 3-fold symmetry axis. Each subunit consists of a core of seven bacteriochlorophyll *a* molecules enclosed within a "bag" of protein. The polypeptide chain forms an extensive 15-strand β-sheet, which is almost planar in its central region, and twisted at its extremities, and wraps around the chlorophyll core to form an efficient amphipathic layer between the chlorophylls and the aqueous environment. There are extensive contacts between the phytyl chains of the seven bacteriochlorophylls within each subunit. These hydrocarbon chains constitute an inner hydrophobic core of the molecule which may be important in forming the complex. There are also extensive contacts between the protein and both the bacteriochlorophyll head groups and tails, but relatively few contacts between the respective head groups. The seven magnesiums all appear to be five co-ordinated. In five cases the presumed ligand is a histidine side-chain, in one case a polypeptide carbonyl oxygen, and in the other case a water molecule.

At low temperature, both the absorption and circular dichroism spectra of the bacteriochlorophyll *a*-protein show splitting which can be interpreted in general terms as due to exciton interactions between the seven chromophores, but calculations of the expected splitting based on the bacteriochlorophyll co-ordinates determined crystallographically are in poor agreement with the observed spectra. Furthermore, the observed red shift of the Q_y absorption band of bacteriochlorophyll *a*, from about 770 nm in organic solvents to 809 nm in the bacteriochlorophyll *a*-protein, is not explained by the exciton calculations. It seems likely that the red shift is due to perturbations of the spectra of the individual bacteriochlorophylls by the protein environment, but, pending the determination of the amino acid sequence, it is not possible at this time to define in detail all the protein–chlorophyll interactions. It is suggested that the bacteriochlorophyll *a*-protein

† Present address: Department of Biochemistry, University of Miami School of Medicine, Miami, Flor. 33152, U.S.A.

‡ Permanent address: Istituto de Cristallografia Università, via Bassi 4-27100, Pavia, Italy.

0022–2836/79/180259–27 $02.00/0

© 1979 Academic Press Inc. (London) Ltd.

serves as a good model for the organization of chlorophyll *in vivo*, and that the types of interaction seen here between chlorophyll and protein are likely to be found in other chlorophyll proteins.

1. Introduction

In the green photosynthetic bacteria, a water-soluble bacteriochlorophyll *a*-protein is believed to function as an intermediary in the transfer of excitation energy from the light-harvesting *Chlorobium* chlorophyll (Bchl† *c*, *d*, *e* or *f*) to the photochemical reaction centers (Olson & Romano, 1962; Sybesma & Olson, 1963; Sybesma & Vredenberg, 1963,1964). We describe here the three-dimensional structure of the Bchl *a*-protein from *Prosthecochloris aestuarii*, strain 2K, as determined by X-ray crystallography from a 2·8 Å resolution electron density map. (The strain of bacteria was previously identified incorrectly as *Chloropseudomonas ethylica* or *Chlorobium limicola* (Olson, 1978*a*).) Preliminary crystallographic results (Fenna *et al.*, 1974) showed that the molecule is a trimer of molecular weight 150,000 and that each of the three identical subunits contains seven molecules of Bchl *a* (Fenna &Matthews, 1975,1976).

The initial electron density map was based on three isomorphous heavy-atom derivatives. In this paper we describe the X-ray crystallographic procedures used to obtain a better electron density map, improved by the inclusion of a fourth derivative and by remeasurement of some of the data. A more detailed analysis of the structure is presented, and an attempt is made to rationalize the optical properties of the Bchl *a*-protein in the light of the three-dimensional structure.

2. Crystallization

Bacteriochlorophyll *a*-protein was prepared from mixed cultures of *P. aestuarii* and *Desulfuromonas acetoxidans* (Olson, 1971,1978*b*). Crystals suitable for X-ray diffraction analysis were obtained by a method based on that of J. M. Olson *et al.* (1969). Two droplets, one containing protein and the other precipitant were placed about 1 cm apart on a plastic or glass plate, joined by a narrow V-shaped liquid junction, sealed in a Petri dish and allowed to equilibrate, presumably by a combination of vapor diffusion and salt diffusion *via* the junction. The protein droplet contained 50 μl of the Bchl protein, ~14 mg ml^{-1}, in 10 mM-Tris·HCl buffer (pH 7·8), 1 M-NaCl, the precipitant droplet 50 μl of 10% (w/v) ammonium sulfate in the same buffer/NaCl mixture. The crystallization dishes were kept at 5°C and the crystals grew as hexagonal rods up to 0·5 mm diam. and 1 mm long, over several weeks. In order to minimize excessive nucleation, it was found helpful to centrifuge the protein solution twice for at least 40 min at 27,000 *g*, carefully transferring the supernatant between runs, before setting up the crystallization experiments. Prior to X-ray photography, and for storage, the crystals were equilibrated with a "standard mother liquor", the same as that in the precipitant droplet described above.

The crystals have space group $P6_3$ with cell dimensions $a = b = 112{\cdot}4$ Å, $c = 98{\cdot}4$ Å, with 1 subunit/asymmetric unit, implying that the Bchl protein consists of 3 identical subunits arranged with 3-fold symmetry (Fenna *et al.*, 1974). The packing in the crystals was very open, with only about 30% of the crystal occupied by Bchl protein, and the remainder by solvent, and was characterized by large channels, about 62 Å in diameter, parallel to the hexagonal axis, extending through the crystals. These channels can be readily seen in electron micrographs of stained crystals (J. M. Olson *et al.*, 1969; R. A. Olson *et al.*, 1969; Labaw & Olson, 1970). The X-ray structure determination has confirmed the overall molecular packing of the Bchl protein proposed from the electron microscope analyses (Matthews *et al.*, 1977), although the molecules are now known to be trimers rather than tetramers, as assumed in the earlier studies.

† Abbreviations used: Bchl, bacteriochlorophyll.

3. Data Collection and Processing

Diffraction data from the native crystals and heavy-atom derivatives were recorded photographically on Kodak no-screen X-ray film using an Elliott GX6 rotating anode X-ray generator operating at 1600 W and Enraf–Nonius precession cameras. A 0·4 mm standard collimator was used and the copper Kβ radiation reduced by means of a nickel foil filter. The crystal-to-film distance was 75 mm. The exposure time for each photograph was about 40 hours, dependent upon the size of the crystal, which was normally 0·45 mm in the shortest dimension.

A complete three-dimensional data set to 2·8 Å resolution required 23 screened precession photographs including the reciprocal lattice planes: *hkn* ($n = 0,3$); *hnl* ($n = 0,8$); $h(h-n)l$ ($n = 0,9$). This scheme, illustrated in Figure 1, includes 90% of

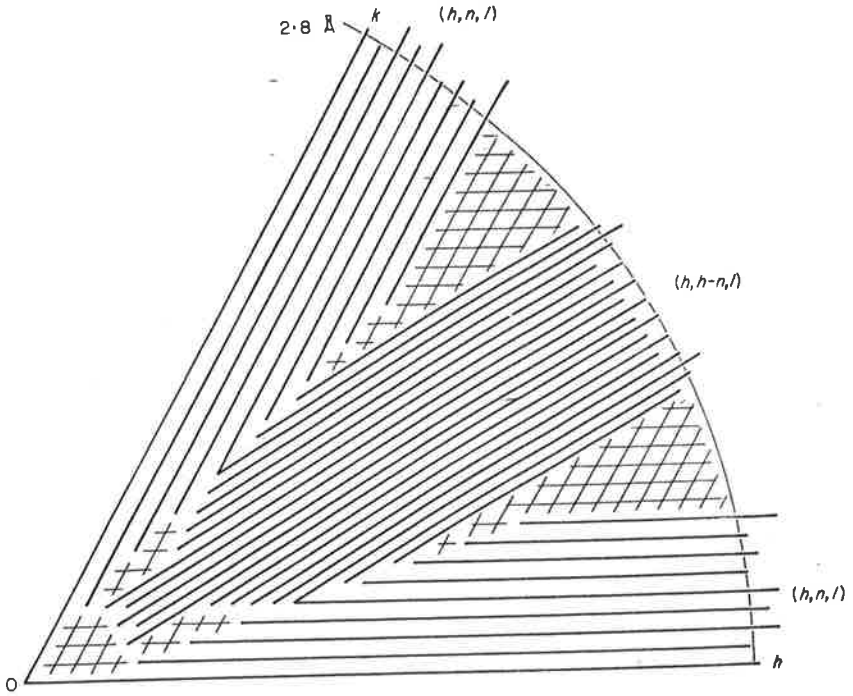


Fig. 1. Data collection scheme.

the reflections to 2·8 Å, 43% between 2·8 Å and 2·7 Å, and allows ample overlap between levels for scaling purposes. The direction of positive l in the reciprocal lattice was chosen arbitrarily and then individual films of the types *hnl* or $h(h-n)l$ were indexed in a manner consistent with this choice. For these planes, the Friedel pairs *hkl* and *hkl̄* can be measured from the same film. One-half of each film includes reflections indexed $h,n,\pm l$ or $h,(h-n),\pm l$ while the other half includes those with indices $n,k,\pm l$ or $k,(k-n),\pm l$. Depending upon the orientation of the crystal with respect to the direction of the X-ray beam, photographs of the type *hkn* have reflections with either positive l or negative l for the whole film, so that no anomalous scattering differences are obtained from these films. The choice of the sign of l is readily made; for example by inspection of the *hk*0 zone on a setting photograph.

Film optical densities at points on a two-dimensional grid of raster size 100 μm were measured with an Optronics rotating drum microdensitometer operated on line to a Varian 620/i computer (Matthews *et al.*, 1972). Individual intensities were taken as the sums of O.D. measurements within a rectangular box (typically 11×9 raster points) centered at the refined reciprocal lattice points. Background measurements were made vertically above and below reflections at points centered midway between reciprocal lattice rows. The number of point background measurements generally exceeded the number of point measurements for the reflections. A pair of films was used for each exposure, and comparison of front and back films enabled an estimate of the film non-linearity correction factor to be calculated for each exposure (Matthews *et al.*, 1972). Measurements which included point density readings exceeding 2·5 O.D. units were discarded.

Before applying appropriate Lorentz and polarization corrections, the intensities of symmetry-related reflections recorded from *hkn*, *h0l* and *hhl* films were averaged. Symmetry R factors were computed for these measurements together with the Friedel-related pairs of reflections on films of the types *hnl* or $h(h-n)l$ with $n \geq 1$. Each isomorphous derivative data film was scaled directly to the corresponding native film for each of the 23 different reciprocal lattice planes in the three-dimensional data set. In an attempt to maximize the quality of the data, poor films, as judged by the values of the R factors for symmetry and scaling, were discarded and repeat measurements made on new crystals. In a few cases where duplicate films of similar quality had been obtained these were averaged before inclusion in the data set. In all, duplicates for three native, five platinum, two mercury and four uranyl lattice planes were included. Local scaling to reduce systematic errors in the measurement of Friedel differences was tested according to the criteria described by Matthews & Czerwinski (1975). This procedure was found to produce significant improvement in 20% of the derivative films and was applied only to these films. Scale factors between films comprising the native three-dimensional data set were computed by the method of Hamilton *et al.* (1965). The same scale factors were then applied to each of the individually scaled derivative data films. Typically each set of 23 precession photographs gave 78,000 intensity measurements, which reduced to about 17,000 unique amplitudes after symmetry averaging and merging. In determining the scale factors between planes, the weaker amplitudes corresponding to 25% of the total of 5500 overlaps were omitted. The merging R factors quoted in Table 1 are, however, calculated from all overlapping measurements.

Correlation coefficients $C_{ij} = \Sigma \Delta_i \Delta_j / (\Sigma \Delta_i^2 \Sigma \Delta_j^2)^{\frac{1}{2}}$ between both isomorphous and anomalous intensity differences recorded on more than one precession photograph were computed as described by Colman *et al.* (1972) and Remington *et al.* (1978). This procedure served as an independent check on the self consistency of indexing within each data set. The overall correlation coefficients are listed in Table 1 together with the other statistics compiled during the data processing.

4. Heavy-atom Derivatives and Phasing

Four isomorphous heavy-atom derivatives were used in the Bchl protein structure analysis. A platinum derivative was made by soaking native crystals for seven days in 0·5 mM-potassium chloroplatinite in standard mother liquor consisting of 10% (w/v) ammonium sulfate, 1 M-sodium chloride and 0·01 M-Tris·HCl (pH 7·8). The

TABLE 1

Data processing statistics

Data set	Native	K_2PtCl_4	MeHgI	Pt + Hg	$K_5(UO_2)_2F_9$
Unique reflections	16,608	16,995	17,004	16,948	16,746
Average $R_{sym}(1)$†	0·058	0·063	0·062	0·058	0·069
Average $R_{sym}(2)$	0·107	0·100	0·103	0·099	0·101
Average R_{scal}	0·072	0·082	0·072	0·064	0·064
R_{merge}	0·046	0·051	0·047	0·049	0·043
Overall isomorphous correlation‡	—	0·38	0·68	0·71	0·79
Overall anomalous correlation	—	0·02	0·05	0·08	0·34

† The R values are as defined previously (Matthews *et al.*, 1972). R_{sym} gives the agreement between symmetry-related reflections on the same film. $R_{sym}(1)$ refers to hnl and $h(h-n)l$ planes having 2 symmetry-related reflections or 4 for $h0l$ and hhl. $R_{sym}(2)$ refers to the hkn planes having 6 symmetry-related reflections. R_{scal} gives the agreement between symmetry-averaged intensities recorded on the stronger and weaker film in a film pack. R_{merge} gives the agreement between structure amplitudes measured on different films.

‡ The correlation coefficients are overall values as described in the text.

second, mercury, derivative was obtained by soaking the crystals for seven days in a saturated solution of methyl mercuric iodide in the same mother liquor. A third, double derivative, was made by soaking crystals for seven days in the presence of both 0·5 mM-potassium chloroplatinite and saturated methyl mercuric iodide. The fourth derivative was formed by soaking crystals for two days in 20 mM-$K_5(UO_2)_2F_9$ in 10% (w/v) ammonium sulfate and 1 M-sodium chloride buffered with 0·05 M-PIPES at pH 6·5. In order to avoid crystal cracking, the reduction in pH from 7·8 to 6·5 was carried out in steps of 0·2 of a pH unit, accompanied by a change in buffer from Tris to PIPES at pH 7·0. Comparison of the diffraction patterns recorded from native crystals at pH 7·8 and pH 6·5 showed no significant changes in intensities or in the unit cell dimensions.

Single sites in each of the platinum and mercury derivatives were initially located from difference Patterson syntheses of the centro-symmetric $hk0$ zone using data to 5 Å resolution. In space group $P6_3$, self vectors occur at (x,y), $(2x,2y)$ and $(2x-y, x+y)$. In the platinum derivative difference Patterson (Fig. 2(a)) one of the double weight vectors occurs as the highest peak in the map. In the mercury derivative difference Patterson (Fig. 2(b)) a possible first site appeared as a set of three superimposing vectors, and hence occurring with multiple weight. This entirely fortuitous superposition of the three self vectors greatly facilitated the location of the first mercury site in this derivative. Two further mercury sites were found in a difference Fourier synthesis of the $hk0$ zone using phases calculated from the platinum and first mercury sites. Subsequently, the relative z co-ordinates of the platinum and mercury sites were determined from three-dimensional difference Patterson and Fourier syntheses. Similarly, sites in the platinum plus mercury double derivative and the uranyl derivative were first located in projection difference Fourier syntheses and the z co-ordinates later determined from three-dimensional Fourier syntheses phased on the platinum and mercury single derivatives.

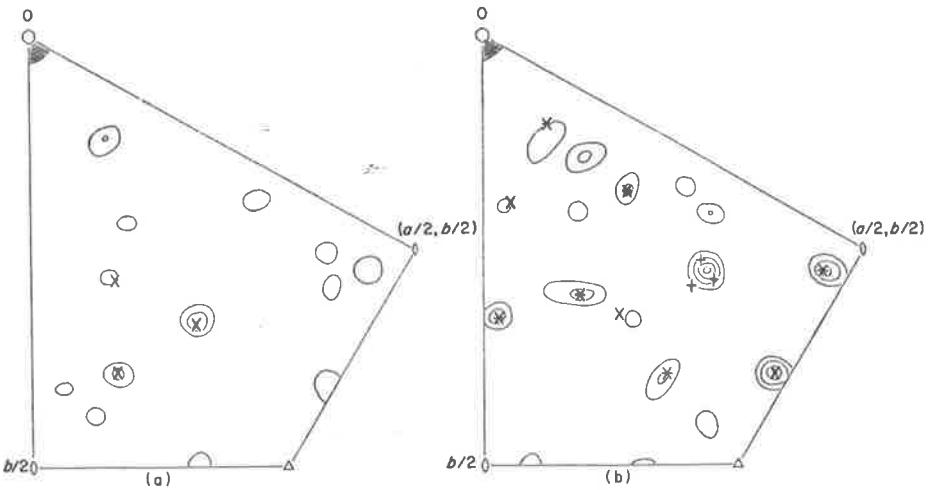


FIG. 2. Difference Patterson projections at 5 Å resolution for the *hk*0 zone. Self vectors (×) and (+), cross vectors (✱); symbols for double-weight peaks are drawn heavily. (a) Platinum derivative. (b) Mercury derivative showing two major sites.

The heavy-atom parameters for all four derivatives were then refined by using a lack-of-closure least-squares refinement procedure applied to a subset of the three-dimensional data, selected to include all the centric data together with 25% of the acentric reflections. Choice of the correct enantiomorph for the heavy-atom distribution was made by including phasing contributions from the anomalous components of the heavy-atom scattering in two separate refinement cycles, in one of which the heavy-atom co-ordinates were the mirror image of those used in the other cycle (cf. Matthews, 1966,1970). One of the choices was found to give significantly lower values (4 to 10%) for the root-mean-square lack of isomorphism, E', for the anomalous contributions of each of the derivatives. The low values found for the correlation of anomalous differences measured on different films (Table 2) indicate that, for the platinum derivative in particular, these differences are barely significant, and, for this reason, the anomalous contributions of the K_2PtCl_4 derivative were not included in the final phase calculation (North, 1965; Matthews, 1966). The heavy-atom parameters used in the four-derivative "best" phase calculation (Blow & Crick, 1959) are listed in Table 2. The program for the refinement was developed by Dr L. F. Ten Eyck and that for the Fourier synthesis by G. N. Reeke. A summary of the heavy-atom refinement statistics is given in Table 3 and the ratio of the root-mean-square heavy-atom scattering to the lack of closure, E, for each derivative and the figure of merit for the phasing are given as a function of the resolution in Figure 3. The overall mean figure of merit was 0·74 and the average ratio of the lack of closure of the phase triangles to E was 0·93 (Matthews, 1970).

Normally, three isomorphous heavy-atom derivatives might be considered sufficient for reasonably accurate phase determination, but this was not the case for the Bchl *a*-protein for the following reasons. (1) The introduction of a single platinum atom into an asymmetric unit of 45,000 molecular weight produced only small changes in the diffracted structure amplitudes (about 10%). Considerable efforts were made to minimize errors in the intensity measurement, and the correlation coefficient of 0·38 (Table 1) indicates that the platinum isomorphous differences are significant, but this derivative nevertheless contributes weakly to the overall phase determination.

TABLE 2

Heavy-atom parameters

Derivative	Z	x	y	z	B
K_2PtCl_4 (1 site)	27·9	0·113	0·443	−0·251	38·1
MeHgI (3 sites)	38·5	0·380	0·583	0·007	26·3
	15·6	0·301	0·439	−0·103	40·0
	8·2	0·278	0·434	−0·141	10·7
K_2PtCl_4 + MeHgI (5 sites)	32·2	0·113	0·443	−0·25	43·3
	37·4	0·380	0·582	0·007	27·8
	13·5	0·301	0·438	−0·104	36·5
	8·5	0·278	0·434	−0·142	12·7
	8·6	0·511	0·204	−0·396	29·6
$K_5(UO_2)_2F_9$ (11 sites)	34·5	0·437	0·502	0·069	33·5
	19·1	0·504	0·495	0·099	27·0
	19·0	0·324	0·022	0·254	43·2
	12·8	0·515	0·512	0·109	34·7
	11·4	0·472	0·415	0·132	32·2
	11·4	0·359	0·059	0·060	58·3
	9·9	0·400	0·292	0·408	29·9
	8·1	0·656	0·327	0·336	33·6
	7·9	0·457	0·462	0·050	41·0
	7·6	0·342	0·655	0·069	23·0
	7·2	0·451	0·370	0·182	22·2

Z, relative occupancies of heavy-atom sites in arbitrary units; x,y,z, fractional co-ordinates; B, isotropic temperature factor (Å^2).

TABLE 3

Refinement statistics

Parameter	K_2PtCl_4	MeHgI	Pt + Hg	$K_5(UO_2)_2F_9$
No. of reflections used in the refinement	5471	5461	5478	5459
Mean amplitude	364	365	364	367
r.m.s.† isomorphous difference	39	54	61	62
r.m.s. heavy-atom scattering	40	70	81	75
Lack of closure, E	34	38	43	37
$R_c(1)$‡	0·78	0·53	0·52	0·50
$R_c(2)$	0·51	0·43	0·47	0·45
r.m.s. anomalous difference	25	26	26	28
r.m.s. anomalous scattering	6	10	12	10
Lack of closure, E'	25	25	24	26

† r.m.s., root-mean-square.

‡ Values of $R_c(1)$ are the residuals for the centric reflections to 2·8 Å resolution phased on all 4 derivatives. Values of $R_c(2)$ are the residuals for the centric reflections to 2·8 Å resolution using the single derivative isomorphous replacement phases for each derivative. $R_c = \Sigma \mid (F_{PH} - F_P) - F_H \mid / \Sigma \mid F_{PH} - F_P \mid$ (Cullis *et al.*, 1961).

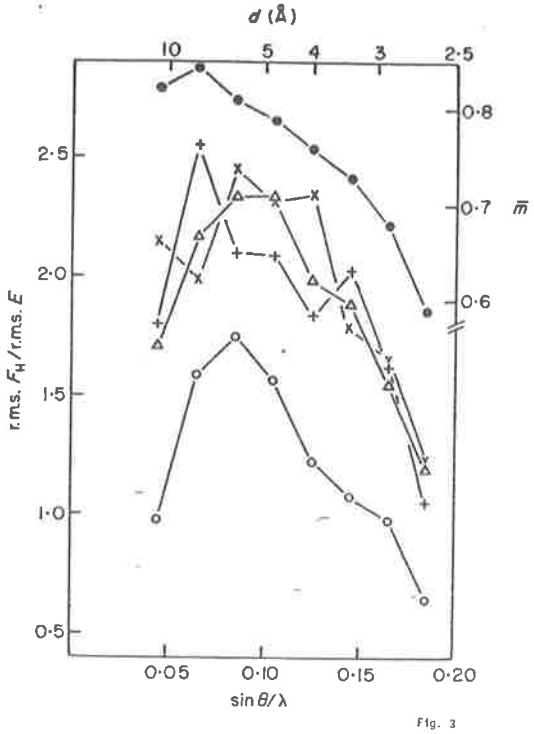


FIG. 3. Root-mean-square (r.m.s.) heavy-atom scattering divided by the lack of closure of the centrosymmetric reflections for 4 isomorphous derivatives: —○—○—, platinum; —+—+— mercury; —△—△— platinum plus mercury; —×—×— uranyl. —●—●—, mean figure of merit $\bar{m}$.

(2) The double derivative, although giving reproducibly measurable intensity differences, is inherently a poor substitute for a totally independent derivative because it duplicates to some extent the phase information provided by the mercury derivative. (3) The separation in z between the platinum atom and the major mercury site is only 0·7 Å from a quarter of a unit cell translation. As a consequence, the phase ambiguity for reflections with l even will be poorly resolved by the phasing contributions of the three derivatives, especially for the lower resolution data. For these reasons, the additional phasing information provided by the multi-site uranyl derivative is particularly significant in improving the overall quality of the phase determination. This improvement is reflected in a large increase in the figure of merit from 0·57 for the three-derivative phases (Fenna & Matthews, 1975) to 0·74 when the additional uranyl derivative is included in the phase determination.

5. Interpretation of the Electron Density Map

The incorporation of an additional heavy-atom derivative, combined with remeasurement of selected films in the original data set (Fenna & Matthews, 1975) has resulted in a substantial improvement in the quality of the electron density map. In common with the original map, the quality of the electron density distribution is

best in the centre of each subunit, and in the vicinity of the 3-fold axis (i.e. the centre of the Bchl trimer). Toward the periphery of the molecule, especially at the extremities of the surface loops, the electron density is weaker, and, in a few cases, very poorly defined.

Starting from the previously defined amino terminus, it was straightforward to follow the apparent course of the backbone from one end of the molecule to the other. The present electron density map has resolved the ambiguities in the interpretation of the original map, and has shown the probable course of the polypeptide backbone. Nevertheless, it has to be noted that in several places the presumed course of the backbone passes through weak density, which, in the absence of the amino acid sequence, could result in errors in chain tracing. There are no alternative interpretations of the chain folding which readily suggest themselves, but we cannot rule out this possibility.

The approximate course of the backbone was indicated by placing markers in the isomorphous replacement map (scale 2 cm $\equiv$ 1 Å) and a model was constructed in a modified optical comparator (Richards, 1968; Colman *et al.*, 1972) using Kendrew parts from Cambridge Repetition Engineers. Raw co-ordinates were obtained by placing markers in the electron density map to correspond both with the model and the density (cf. Matthews *et al.*, 1974). These co-ordinates were then adjusted to have acceptable stereochemistry using the procedure of Ten Eyck *et al.* (1976) and Dodson *et al.* (1976).

The co-ordinates of the 358 α-carbon atoms included in the present model are listed in Table 4 and have been deposited with the Protein Data Bank (Bernstein *et al.*, 1977). The co-ordinates are in ångström units along the orthogonal crystallographic axes a^*, b and c. It has to be emphasized that model building, in the absence of a known amino acid sequence, is subject to substantial errors. In particular, in regions of weak electron density we can do no more than "guess" the course of the polypeptide backbone, and it is likely that addition or subtraction of one or more residues will be necessary in these regions when the amino acid sequence is known. Particularly doubtful regions include residues 51 to 56, 166 to 168, 206 to 209, 274 to 275 and 334 to 336 and the quoted co-ordinates for these α-carbons should be treated with the greatest caution.

We have used the electron density map to determine a preliminary "X-ray" sequence which is included in Table 4. As is well known, such sequences should be taken as an indication of the size and shape of each residue, rather than a chemically correct identification of each side-chain (e.g. see Matthews, 1977). In Table 4 the amino acid name is followed by a number from 5 to 0 which is intended as a rough indication of the confidence of our assignment. In those surface loops described above where the electron density is so weak that no assignment was possible, the amino acid is listed XXX and the confidence level zero. In some instances we have listed in in Table 4 a second choice for the amino acid. This was, in general, not done for structurally similar alternatives such as valine and threonine, but for cases such as a surface amino acid with very weak but extended side-chain density which might suggest, for example, either a serine or a mobile lysine. The identifications given in Table 4 have been tailored, to some extent, to reflect the known amino acid composition of the protein (Thornber & Olson, 1968; Olson *et al.*, 1976*d*). In the present electron density map there is no obvious indication of a disulphide bridge. If past

experience is any indication, only about 50% of the identifications listed in Table 4 will be chemically correct. Nevertheless, such a tabulation can be a useful aid to chemical sequence determination, and will serve as a starting point for refinement of the structure.

In order to confirm the tracing of polypeptide backbone, we attempted to label the amino terminus with the water-soluble iodo-analogs of the Edman reagent, 4-iodo-, and 6-iodo-3-isothiocyanato-benzene sulfonic acid (unpublished method of T. Weiland & G. Petsko). From pH 7 to 8 these compounds are expected to react specifically with the α-amino group of the polypeptide chain. In the case of the bacteriochlorophyll protein, the attempted labeling was not successful, but, because of the potential use of these reagents by other groups, we will report here the results of our experiments.

Crystals of the Bchl protein were soaked for one week in solutions containing 1 M-NaCl, 15% (w/v) ammonium sulfate, 0·01 M-phosphate buffer (pH 7·5), and, respectively, 3 mM of the 6-iodo compound, and 3·4 mM of the 4-iodo compound. Phosphate buffer was used in place of the usual Tris to minimize side reactions. Diffraction data were measured photographically for the (*hk*0) and (*h*0*l*) zones, and projection difference Fourier maps calculated in the usual way.

The two projections for the 6-iodo compound are shown in Figure 4. As can be seen, the (*hk*0) projection suggests two sites of binding labeled A and B. The (*h*0*l*) projection clearly confirms site A, which has approximate co-ordinates (0·36, 0·49, 0·07). For site B, the interpretation is ambiguous, with possible sites indicated at (0·29, 0·40, 0·52) (labelled B1) and/or (0·29, 0·40, 0·97) (labelled B2), although neither solution is convincing.

The clear-cut binding site (site A) is on the surface of the protein, near residue 6, and about 20 Å from the presumed amino terminus. Residue 6 is located in the β-sheet wall, and the electron density appears quite continuous at this point, with no indication of a potential amino terminus. It is conceivable that the amino terminus of the protein, which appears to be somewhat disordered, extends beyond "residue 1", and folds back on itself to reach the observed binding site, but this seems unlikely. The possible binding site B1 is located in a region of intermolecular contact, close to residue 204 of one molecule and the presumed carboxyl terminus of another molecule. This might indicate that we have confused the amino and carboxy termini, although we think it unlikely. The electron density map is, in many places, of sufficient quality to indicate the polarity of the polypeptide backbone, and is consistent with our present chain tracing. Potential site B2 is located on the surface of the β-sheet, near residue 261, and, again, is not close to a potential terminus.

In the case of the 4-iodo compound, the results were quite different from the 6-iodo compound in that no binding was observed at sites A, B1, B2 or at any other site. This suggests that the binding observed in the 6-iodo case may be due to hydrophobic interactions involving the 6-iodo moiety rather than *via* the isothiocyanato group. In any event, there is no clear evidence that either reagent has reacted with the amino terminus of the protein. This could be due to an excessively slow rate of reaction, or to the amino terminus being blocked. Another possibility is that the reaction does occur, but that the terminus is so disordered in the crystal that the iodine cannot be located.

TABLE 4

Backbone co-ordinates

1	ALA	2	72.7	6.2	62.6	121	ALA	4	37.0	7.0	20.8	241	SER	3	49.6	36.8	33.2
2	GLY	2	70.3	8.7	61.2	122	THR LYS	2	36.4	9.4	22.2	242	LYS	4	48.8	38.6	34.3
3	VAL	2	69.2	9.1	57.7	123	ALA	3	33.4	6.3	24.2	243	ALA	3	52.4	37.7	35.1
4	THR HIS	2	66.6	11.9	57.6	124	SER	3	36.3	6.3	26.6	244	ALA	4	51.0	34.1	35.7
5	ALA	4	64.7	14.0	55.3	125	SER	2	35.6	9.9	27.4	245	SER	3	50.3	32.9	37.3
6	GLU HIS	3	62.8	17.1	56.3	126	ARG LYS	3	32.0	9.0	28.2	246	ILE	4	47.9	30.3	39.4
7	SER	3	60.6	19.0	53.9	127	ASN PRO	2	32.4	6.3	30.9	247	GLU	2	61.4	29.7	40.9
8	GLU ASP	4	57.8	21.6	54.0	128	VAL	3	33.8	7.5	32.0	248	GLY	5	63.5	27.0	42.7
9	TYR	4	55.0	21.8	51.3	129	ALA SER	2	35.8	10.3	34.5	249	GLY	4	67.0	26.1	43.8
10	GLU	3	52.0	23.9	50.7	130	ASP	3	39.5	11.0	34.3	250	PHE	4	68.9	22.9	43.9
11	VAL	5	49.5	25.4	48.4	131	PRO	4	41.6	9.5	31.5	251	LYS	3	72.2	22.3	45.7
12	THR ASP	3	47.5	28.6	48.3	132	ILE VAL	3	43.2	9.3	32.7	252	SER	3	73.7	19.0	46.8
13	ALA	4	44.7	27.9	45.7	133	TYR	4	46.8	7.3	29.9	253	ASP	4	77.3	17.9	47.4
14	GLU	3	42.7	30.6	43.9	134	MET	4	46.9	8.9	26.4	254	ALA	3	78.4	14.5	48.8
15	GLY	1	39.0	30.8	44.3	135	HIS	3	49.1	7.8	23.6	255	GLY	3	76.6	14.9	52.1
16	GLY ALA	2	36.7	31.0	41.4	136	GLY	5	50.1	9.4	20.3	256	GLY	4	73.0	15.5	51.1
17	ALA	2	37.5	30.5	37.7	137	GLU PHE	3	53.1	10.9	18.4	257	ILE	3	70.6	17.8	49.1
18	SER	3	41.2	31.2	38.4	138	LEU GLN	4	54.4	14.5	18.5	258	LYS	3	68.8	21.1	49.7
19	VAL THR	3	43.3	28.1	37.9	139	PHE	5	56.0	15.9	15.3	259	VAL	4	65.8	21.8	47.4
20	PHE	5	46.0	29.9	39.8	140	HIS	5	58.4	18.8	15.5	260	ASP	2	64.1	25.2	47.0
21	ALA	4	47.8	29.4	43.1	141	ARG	3	60.2	21.2	13.0	261	VAL	4	60.9	25.9	45.0
22	GLU LYS	3	51.0	29.8	45.1	142	ILE VAL	3	62.9	23.6	14.0	262	ARG	3	58.1	28.5	44.9
23	VAL	4	52.8	26.5	45.5	143	ASP	4	64.9	26.4	12.5	263	GLY	5	55.7	29.9	42.3
24	LYS ARG	3	56.0	26.2	47.6	144	LEU MET	4	68.0	27.5	14.4	264	ASP	4	52.2	31.2	41.3
25	GLY ALA	5	58.6	23.5	48.0	145	ALA	3	69.6	30.7	13.6	265	HIS VAL	3	50.2	30.0	38.3
26	SER ASP	2	61.5	23.1	50.3	146	ILE	3	73.1	30.9	15.1	266	PRO	5	47.4	28.1	36.5
27	ALA	5	63.3	19.8	50.7	147	SER	2	76.1	33.1	15.1	267	GLY	4	47.5	24.3	36.3
28	SER LYS	2	66.6	18.5	52.0	148	LEU ASP	3	79.4	31.7	14.0	268	SER	3	47.1	23.3	40.0
29	VAL	3	68.5	15.2	52.2	149	ASP GLU	3	81.6	34.0	15.9	269	THR	2	43.4	23.5	40.8
30	THR	2	71.7	14.2	53.9	150	ALA	4	83.8	31.5	17.8	270	ARG	5	41.6	21.8	37.9
31	SER VAL	3	72.0	11.5	51.2	151	ALA	3	86.3	29.0	16.4	271	GLU	4	37.9	20.9	37.9
32	PRO	5	74.7	11.3	48.4	152	ASP LEU	3	84.4	26.6	18.6	272	GLY ALA	2	36.7	17.4	39.0
33	ALA	4	73.5	12.0	44.9	153	LEU	4	81.1	27.1	16.8	273	GLY	1	33.3	18.0	37.3
34	SER THR	3	74.6	11.1	41.4	154	THR	2	82.7	27.2	13.4	274	XXX	0	29.8	18.8	38.2
35	LEU	5	75.9	12.8	38.4	155	LYS	5	84.9	24.2	14.1	275	XXX	0	30.3	21.8	40.5
36	PRO	4	79.4	11.3	38.2	156	VAL	3	82.3	22.0	15.7	276	GLU	2	33.6	20.4	41.9
37	LEU	4	80.5	14.8	36.1	157	TRP	5	80.3	23.0	12.6	277	PHE	3	37.2	21.5	41.3
38	GLN LEU	3	79.9	16.0	42.7	158	SER	2	82.5	21.7	9.8	278	GLY ALA	3	40.4	19.7	41.9
39	ASN ASP	5	79.4	19.5	43.9	159	GLY LYS	2	82.6	18.3	11.6	279	VAL LEU	3	44.2	19.7	42.2
40	VAL THR	4	76.8	20.9	41.6	160	ALA	2	78.8	18.4	11.5	280	LEU	5	45.6	19.0	38.8
41	ASP ASN	5	74.3	23.7	42.3	161	ALA	3	78.8	18.5	7.7	281	HIS	4	46.8	15.5	38.7
42	CYS HIS	4	71.5	25.0	39.9	162	GLY	3	81.1	15.5	7.1	282	HIS	5	50.5	15.1	39.7
43	ASP	4	69.1	27.9	40.0	163	ALA	2	78.8	13.6	9.5	283	VAL SER	2	52.6	14.7	36.5
44	LEU	4	65.8	27.8	37.9	164	ASP	3	75.8	14.6	7.4	284	GLU GLN	3	54.5	11.4	36.2
45	LYS	5	63.9	31.0	36.8	165	ASP	1	77.2	13.4	4.1	285	ILE	4	57.5	11.9	33.9
46	LEU	4	60.8	30.3	34.8	166	XXX	0	77.4	9.8	5.2	286	SER	4	58.8	13.7	37.0
47	ASP	3	58.7	33.4	34.0	167	XXX	0	74.1	8.2	5.2	287	LEU ASN	3	57.9	11.0	39.2
48	ALA	5	55.9	33.8	31.5	168	XXX	0	71.0	7.6	7.3	288	THR VAL	2	59.4	8.0	37.4
49	LYS SER	2	54.3	37.2	30.5	169	ASN	1	73.2	6.8	10.2	289	HIS	4	62.5	9.8	36.4
50	PRO	3	51.2	38.0	28.3	170	PHE	4	73.6	10.3	11.8	290	HIS	5	63.3	11.2	39.8
51	XXX	0	51.7	39.3	24.8	171	GLY	4	69.8	10.7	12.0	291	GLY PRO	4	62.5	7.9	41.4
52	XXX	0	49.3	42.2	25.4	172	ASP	4	70.0	7.6	14.2	292	MET	4	64.8	6.5	38.8
53	XXX	0	50.1	43.4	21.9	173	TRP	5	72.9	7.2	16.5	293	THR VAL	2	67.6	7.9	40.9
54	XXX	0	49.0	39.9	20.8	174	LEU	5	72.7	10.9	16.9	294	GLY	4	65.7	8.0	44.2
55	XXX	0	45.8	38.0	21.6	175	ARG	5	70.2	10.1	19.6	295	LYS GLN	3	65.7	4.4	45.0
56	XXX	0	46.6	34.4	22.1	176	GLU	4	73.2	8.7	21.5	296	PHE	4	69.4	4.4	45.5
57	SER ALA	3	50.8	34.7	22.0	177	PHE TYR	4	74.9	12.1	22.1	297	LEU	3	69.3	5.2	49.2
58	VAL	4	52.8	34.2	25.2	178	TRP	5	71.5	13.9	22.2	298	ASP THR	2	67.1	4.7	52.3
59	LYS	4	56.1	35.6	26.3	179	TRP	5	70.9	12.7	25.8	299	PHE	4	65.5	7.8	53.9
60	PHE	5	58.3	32.8	27.8	180	ILE	3	74.2	12.0	27.5	300	ASP	3	63.3	5.0	56.7
61	SER	3	61.7	33.3	29.6	181	GLY	4	74.2	8.7	29.5	301	VAL	3	60.2	11.0	56.1
62	VAL LEU	4	64.1	31.2	31.7	182	PRO	3	71.6	8.6	32.3	302	GLU	2	59.1	14.0	58.1
63	GLY ALA	4	67.3	31.5	33.7	183	ALA PRO	2	69.2	10.2	29.9	303	SER	4	57.2	16.2	55.6
64	LEU	5	69.5	28.7	34.8	184	PHE	5	68.9	7.2	27.5	304	ALA GLY	3	54.2	18.4	55.6
65	ALA LYS	2	72.8	29.0	36.8	185	GLY	3	67.3	5.1	30.3	305	LEU	3	51.2	19.5	53.5
66	SER ALA	3	74.7	25.7	37.1	186	GLY	3	64.7	7.8	30.8	306	LYS	3	49.8	23.0	53.4
67	SER ALA	2	77.8	24.7	38.9	187	VAL	4	63.7	7.8	27.1	307	LEU	4	46.6	23.5	51.6
68	ILE	5	79.6	21.4	38.4	188	ASN	3	62.4	4.3	27.6	308	SER	3	44.8	26.9	51.5
69	GLU	2	83.3	21.0	39.0	189	GLU	4	61.6	4.5	31.4	309	VAL	3	41.4	26.6	49.8
70	ALA PRO	1	83.6	24.7	39.8	190	GLY	4	58.3	6.0	31.0	310	PRO	5	39.6	29.9	49.2
71	ASP	4	82.3	24.9	36.3	191	GLY	4	57.3	4.7	27.6	311	LYS	4	37.4	31.1	52.0
72	LYS	3	79.6	27.5	35.8	192	HIS	5	59.1	7.4	25.7	312	GLY	3	34.1	29.5	51.9
73	LEU HIS	4	77.1	26.9	33.1	193	ARG	5	59.9	6.9	22.0	313	TYR ARG	3	35.4	26.7	49.7
74	SER	2	74.7	29.5	31.9	194	ILE	4	62.1	9.2	19.9	314	GLU	2	36.0	23.5	51.6
75	LEU ASN	3	71.2	29.2	30.3	195	ALA	4	62.0	9.3	16.1	315	ILE LEU	4	37.4	20.3	50.6
76	ALA	4	69.0	32.0	29.1	196	PRO	4	65.1	9.9	14.0	316	GLN TYR	2	35.0	17.6	49.5
77	VAL	4	66.2	31.8	26.5	197	VAL	4	66.7	13.2	12.9	317	TRP	4	37.1	14.5	49.2
78	GLU	4	63.6	34.3	25.7	198	THR	4	65.8	14.6	9.5	318	GLY	4	40.5	13.8	50.7
79	ALA	4	60.6	33.0	24.0	199	THR	3	68.0	17.2	7.9	319	GLY	3	42.5	10.7	50.0
80	GLU	4	57.2	33.9	22.5	200	SER LYS	2	65.9	19.3	5.5	320	PRO	3	45.1	10.4	51.6
81	LEU	4	54.7	31.2	21.6	201	SER LYS	2	68.6	22.0	5.0	321	LYS	3	43.1	12.4	54.2
82	ALA	3	51.2	30.1	21.0	202	ALA GLN	3	72.1	22.9	5.9	322	LEU	3	43.6	15.7	55.9
83	ALA	2	48.8	27.1	21.1	203	SER	2	73.3	26.4	5.1	323	SER	2	44.8	15.1	59.4
84	GLY	2	47.0	29.9	19.5	204	SER	3	75.5	29.3	6.1	324	ASN	1	44.1	18.6	60.6
85	ASP	4	46.7	28.2	16.1	205	GLU	3	74.0	32.7	5.3	325	GLU	3	42.3	21.5	59.0
86	SER	3	45.1	24.7	16.5	206	XXX	0	75.4	36.3	5.3	326	LYS	2	41.1	25.0	60.3
87	ALA LYS	3	44.5	22.6	19.5	207	XXX	0	75.7	37.7	8.9	327	THR ASP	1	38.5	27.0	56.5
88	ASP	4	46.9	20.2	17.8	208	XXX	0	79.3	36.4	5.3	328	GLU	2	40.4	27.3	55.3
89	LYS	5	49.7	22.6	16.7	209	XXX	0	79.1	32.7	9.7	329	ASP	4	44.0	26.3	55.6
90	ARG	4	51.9	24.9	18.7	210	ASN THR	2	75.5	31.8	10.4	330	TYR	4	43.7	22.5	55.2
91	SER	3	54.2	27.7	17.4	211	GLY	4	74.5	28.2	10.9	331	GLN LYS	1	46.7	20.5	56.3
92	GLY	5	57.3	29.3	18.7	212	THR	3	70.9	28.0	9.6	332	TRP	5	48.5	17.0	56.1
93	THR	3	59.9	32.0	18.5	213	SER VAL	3	69.8	24.4	10.1	333	ALA	1	51.1	15.3	58.0
94	GLY	5	63.1	32.7	20.4	214	SER	3	66.2	23.1	9.6	334	XXX	0	50.9	11.6	57.4
95	SER LYS	2	66.7	33.8	20.8	215	TRP	5	65.2	19.8	10.8	335	XXX	0	51.0	8.0	57.9
96	GLY	4	68.9	32.6	23.6	216	ALA	4	62.4	17.6	11.8	336	XXX	0	48.7	8.1	55.0
97	ALA	4	72.5	32.2	24.8	217	PHE	4	61.7	15.0	14.5	337	TYR	5	50.5	11.1	53.3
98	VAL ILE	4	74.5	29.5	26.7	218	LYS	3	58.8	13.3	16.3	338	ALA	2	53.5	10.2	54.5
99	THR	3	77.9	29.4	28.4	219	ILE	3	58.3	12.1	19.9	339	LYS	4	53.6	6.9	52.8
100	VAL ASN	3	80.9	27.8	30.2	220	GLY	5	56.6	8.8	20.8	340	TRP	5	52.5	8.3	49.5
101	GLY	2	83.8	29.6	31.9	221	GLY	5	54.7	7.9	24.0	341	VAL	4	55.4	10.7	49.6
102	GLU	2	84.7	31.2	28.6	222	SER	3	52.1	5.6	25.4	342	GLU	4	58.2	8.0	49.7
103	PHE	5	82.6	29.9	25.7	223	GLY	5	50.8	7.4	28.4	343	MET HIS	4	56.4	5.9	47.2
104	SER	3	79.2	30.6	24.3	224	VAL ASP	3	53.4	8.7	30.7	344	VAL	4	56.5	8.8	44.7
105	HIS	5	77.0	31.5	21.4	225	VAL	4	54.6	11.6	28.3	345	CYS	5	59.5	9.9	45.8
106	ALA LYS	2	73.8	33.4	20.6	226	ASN CYS	3	53.1	14.3	30.3	346	LYS	4	61.0	6.5	44.8
107	PHE	5	71.0	31.5	18.8	227	SER PHE	3	53.7	16.3	27.1	347	GLY PRO	4	59.3	6.3	41.4
108	THR	3	67.5	32.2	17.5	228	LEU	4	51.1	14.4	25.1	348	GLY ALA	4	55.8	5.1	41.9
109	PHE	5	64.7	30.2	16.3	229	ALA	3	48.6	13.0	27.6	349	VAL	3	52.6	6.9	40.6
110	GLN GLU	3	61.3	29.5	15.0	230	ARG	4	48.2	16.5	29.0	350	GLY	3	49.7	6.1	42.8
111	GLY	5	59.5	26.3	15.8	231	TRP	5	48.5	18.5	25.8	351	GLU	3	46.8	8.6	42.6
112	SER ASP	2	56.4	44.4	14.9	232	ASN THR	2	46.7	21.7	26.3	352	PHE	4	46.1	11.3	45.2
113	ALA VAL	3	54.6	21.3	16.5	233	GLN GLU	3	46.8	21.5	30.1	353	GLU ARG	1	42.9	13.1	44.4
114	THR	3	51.8	19.0	15.8	234	VAL	3	50.6	22.3	30.0	354	SER ALA	2	41.9	16.2	46.3
115	ILE	4	50.7	16.2	17.9	235	PHE	4	49.9	25.5	28.0	355	ASN	3	38.3	17.2	45.7
116	PHE ASP	3	49.0	13.0	16.8	236	PRO	5	47.2	27.4	29.9	356	TYR PHE	4	37.1	20.5	47.0
117	ARG	4	46.5	11.1	18.9	237	SER	3	47.7	30.3	27.5	357	GLY ALA	2	34.1	22.7	46.5
118	TYR	5	45.9	7.5	17.8	238	ASP	3	43.9	30.5	27.9	358	LYS	1	31.6	25.2	47.8
119	PHE	5	42.2	6.9	18.7	239	LYS	3	43.9	31.1	31.6						
120	GLY ALA	3	39.4	9.4	19.2	240	LEU	4	46.9	33.5	31.7						

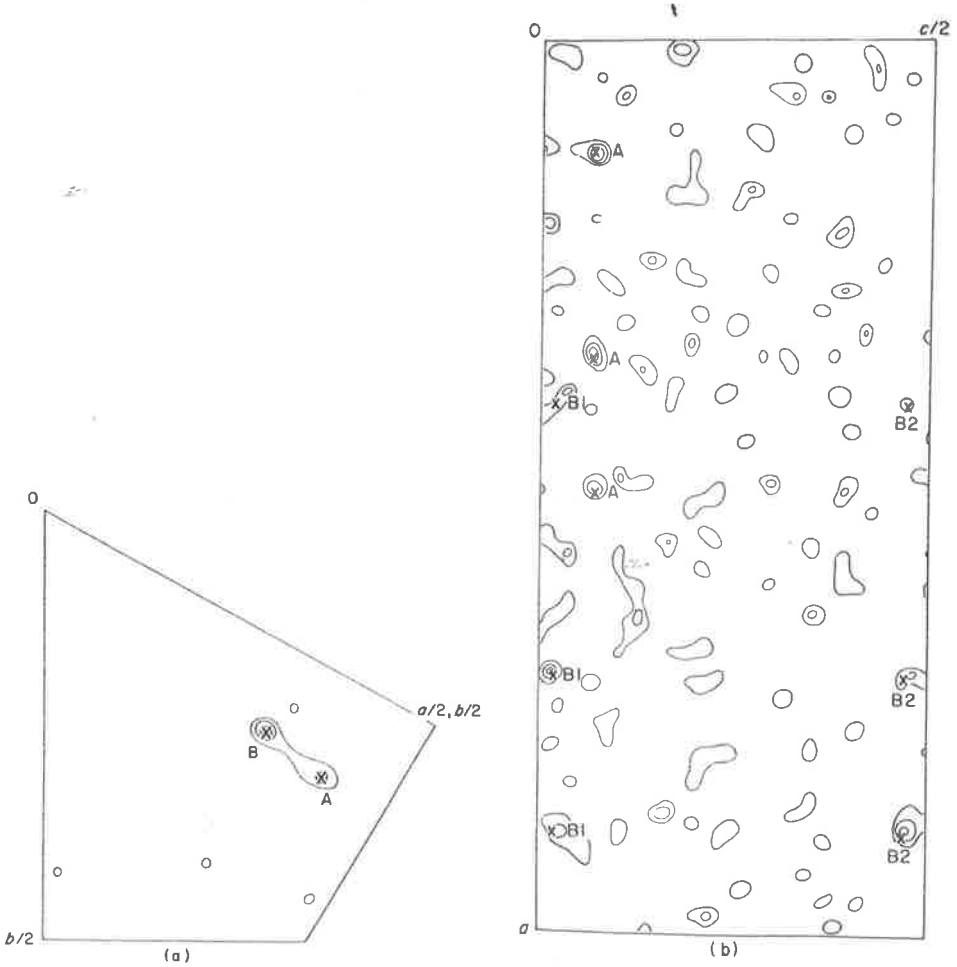


FIG. 4. Difference Fourier projections for 6-iodo-3-isothiocyanato-benzene sulfonic acid bound to the bacteriochlorophyll *a*-protein. A, major binding site; B1 and B2, possible secondary binding sites. (a) A 3·5 Å resolution *hk*0 projection. (b) A 2·8 Å resolution *h*0*l* projection.

6. Overall Conformation

The bacteriochlorophyll protein is a trimer, consisting of three identical subunits, arranged about a 3-fold crystallographic symmetry axis (Fig. 5).

In each subunit, the polypeptide chain folds to form a shell or bag, within which are enclosed seven bacteriochlorophyll molecules. As will be discussed below, the packing of the chlorophylls within the protein is space filling, analogous to the packing of hydrophobic residues in the core of a typical globular protein. There are few, if any, solvent molecules or unoccupied regions within each subunit. The overall arrangement of the seven bacteriochlorophylls within one subunit is shown in Figure 6.

The folding of the polypeptide chain is illustrated in Figures 6 and 7. As can be seen, the folding is dominated by a large wall of β-sheet which comprises much of the surface of the trimer exposed to solvent. The intersubunit contact region, which is quite extensive, is made up in part of helices and irregular folds. The "opening"

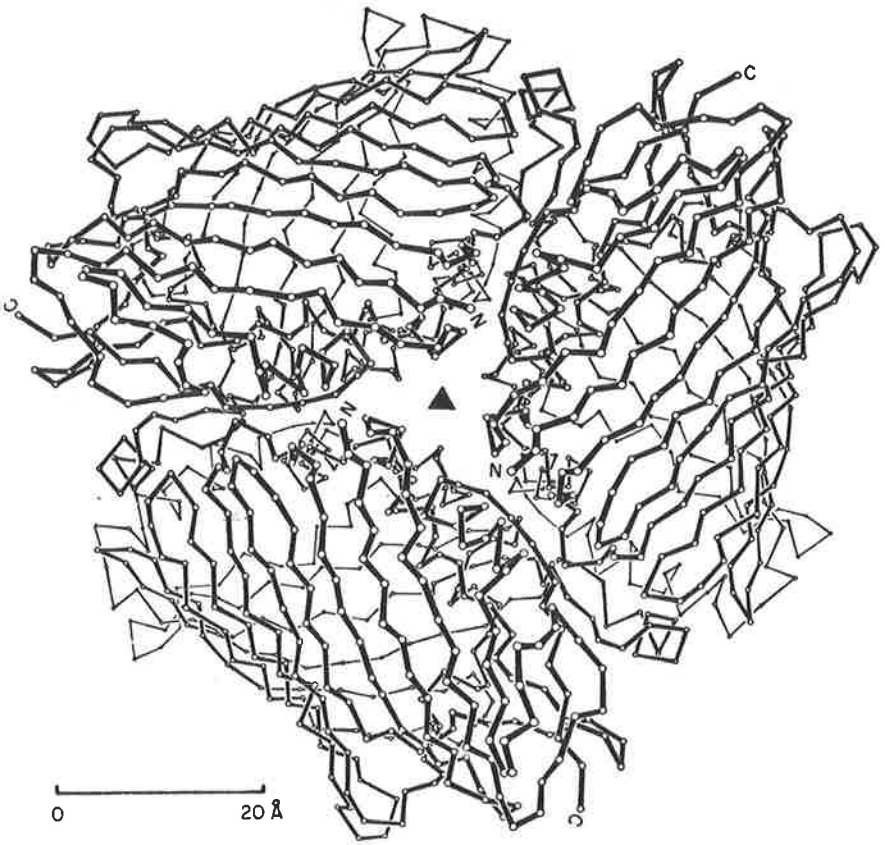


FIG. 5. Polypeptide backbone of the bacteriochlorophyll *a*-protein projected down the trimer axis.

in the protein shell, seen in Figures 6 and 7, is occluded by a neighboring subunit, so that in the trimer all of the bacteriochlorophyll molecules are shielded from contact with solvent. It seems clear that the β-sheet wall is a simple and economical fold by which the protein forms an amphipathic layer shielding the non-polar chlorophyll core from its aqueous surroundings. In fact, considering the ratio of chlorophyll to protein in the complex (about 1:6, w/w), it is difficult to imagine any way in which the available protein could enclose the chlorophyll, other than an arrangement including extensive β-sheet, as is observed.

The topology of the β-sheet is illustrated diagramatically in Figure 8. The arrows show the direction, and the thick lines indicate the approximate length, of each strand. The zigzag lines indicate α-helices. There are, in all, 15 strands in the β-sheet which are, with the exception of strands 1 and 13, exclusively antiparallel.

The β-sheet, taken as a whole, has the left-handed twist (viewed perpendicular to the chain direction) characteristic of other sheets (Chothia, 1973). The overall twist angle between the extreme strands (the angle between strands 9 and 14 in Fig. 8) is 132°. However, most of the twist occurs in the shorter chain segments at the beginning and end of the sheet. In fact, seven of the inner strands, between 1 and 5 in Figure 8, form a sheet which is practically flat, with an overall twist of only 6°.

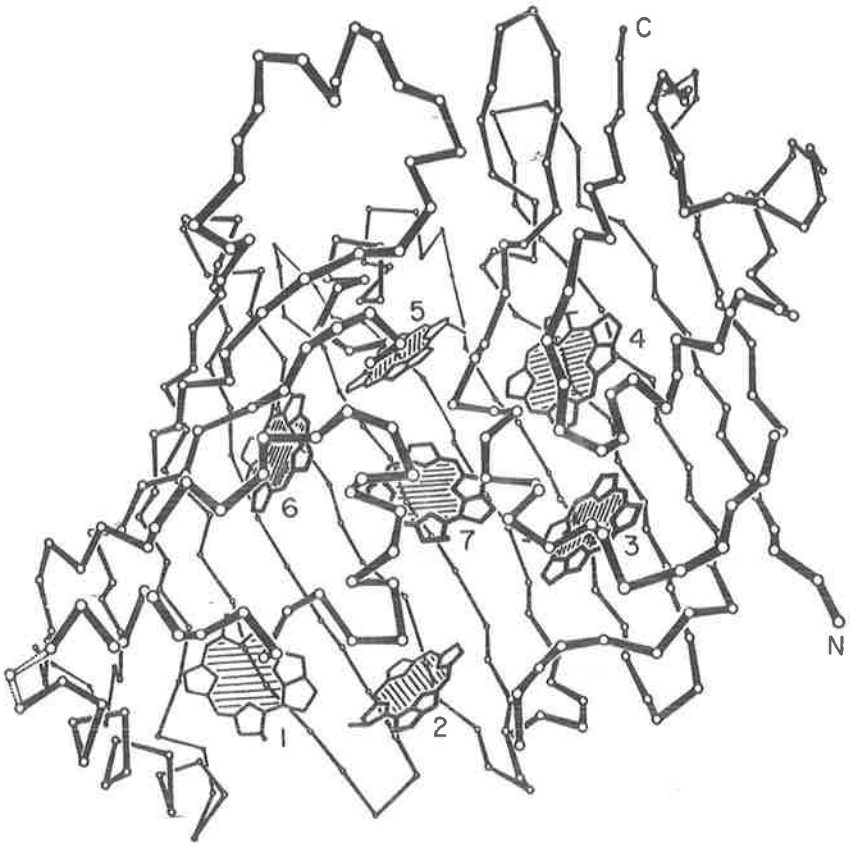


FIG. 6. View of the bacteriochlorophyll *a*-protein subunit from the 3-fold axis, which is horizontal, toward the exterior of the protein. For clarity the phytyl chains and other bacteriochlorophyll ring substituents have been omitted.

This flat region is shown in Figure 9, and is bounded by α-carbons 7 through 13 in strand 1 and 73 through 79 in strand 5. In this central region of the β-sheet, the individual polypeptide strands lie essentially in straight lines. This is an example of the general prediction that a large twist angle cannot occur between long, fully extended (i.e. straight), hydrogen-bonded, polypeptide strands. A large twist angle requires that the strands become further apart as one proceeds outwards, until hydrogen bonds can no longer be formed.

A somewhat different situation occurs in the long strands, 7 and 8 (residues 102 to 117 and 133 to 150 in Fig. 9). Here, hydrogen-bonding is maintained by a combination of twisting, plus a gradual curve in the strands. In fact, each of these two strands changes direction by about 90° in going from one end to the other.

7. Bacteriochlorophyll Core

The electron density for each of the seven chlorophyll molecules constituting the core of each subunit is well defined, allowing their positions to be determined reasonably accurately, and preliminary co-ordinates have been presented elsewhere (Fenna

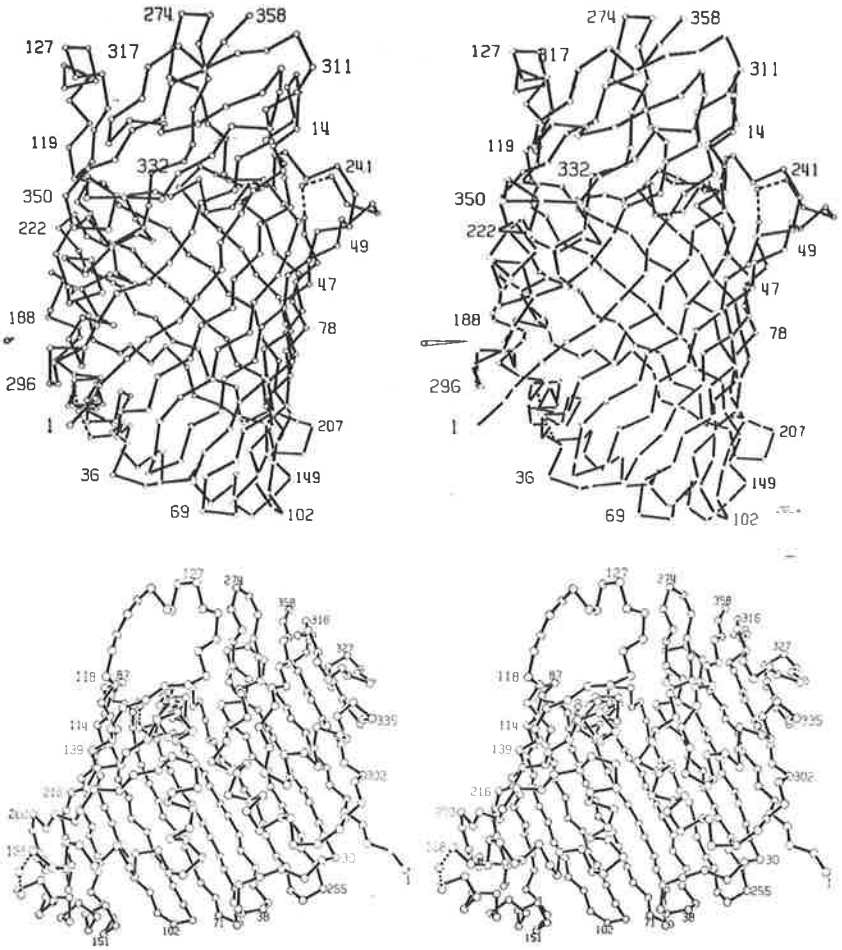


FIG. 7. Stereoscopic views of the backbone of the bacteriochlorophyll *a*-protein. (a) View down the 3-fold axis. (b) View at right angles to the 3-fold axis (cf. Fig. 6).

et al., 1977). Representative examples of the fit between the electron density and the assumed co-ordinates are shown in Figure 10. Consistent with the spectroscopic data (Olson *et al.*, 1963), the electron density for each of the seven chlorophylls appears to correspond to bacteriochlorophyll *a* rather than any other chlorophyll species. Also the total of 21 bacteriochlorophyll *a* molecules per Bchl protein trimer is in agreement with prior chemical studies (Olson, 1966).

As can be seen in Figure 10, the electron density for both the porphine rings and the phytyl side-chains is well defined, indicating that these groups are not disordered in the crystal. In fact there is every indication that the Bchl molecules, including the phytyl chains, are as well ordered as the amino acid side-chains within a typical globular protein.

In five cases, for Bchl molecules 1, 3, 4, 6 and 7, the central magnesium atom appears to be liganded by a histidine side-chain (residues 105, 290, 282, 140 and 289, respectively, in our present numbering scheme). The density for two of the presumed

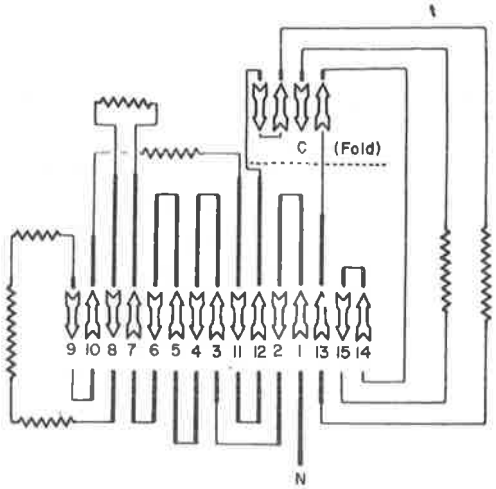


FIG. 8. Schematic illustration of the folding of the polypeptide in the bacteriochlorophyll *a*-protein. The arrows indicate the directions of the strands in the extended sheet region, and the zig-zag lines indicate helices.

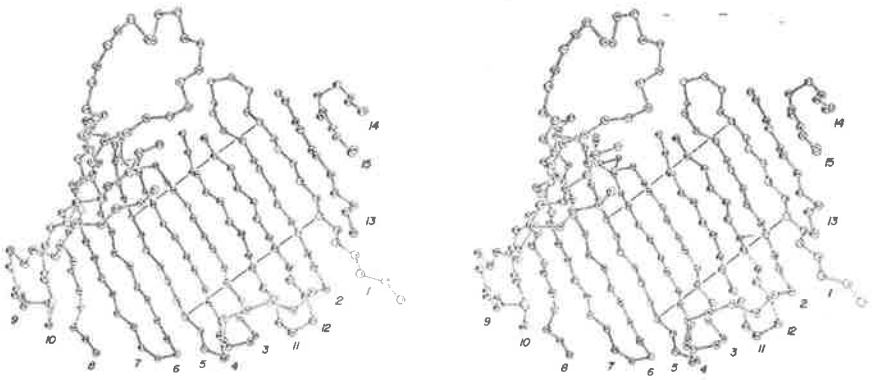


FIG. 9. Stereo view of the extended β-sheet. The straight lines indicate the boundaries of that part of the sheet which is practically flat (see text).

histidines can be seen in Figure 10(b) and (c). For Bchl molecule 5, the magnesium is obviously liganded by the protein, but the electron density map suggests that the interaction may be with a peptide oxygen of the backbone, at residue 234, rather than with an amino acid side-chain. In the case of Bchl molecule 2 (Fig. 10(a)), there is no obvious electron density axial to the magnesium indicating a protein ligand. The closest density, that of the side-chain of residue 73, extends to within about 3·5 Å of the magnesium and might be a potential ligand, although the electron density map suggests another possibility. As can be seen in Figure 10(a), there is a weak tongue of density extending from the magnesium toward the carbomethoxy ester C=0 group

FIG. 10. Stereo views of the 3 bacteriochlorophyll *a*-molecules with electron density superimposed. (●) Oxygen atoms; (◐) nitrogen atoms; (○) carbon atoms. (a) Bacteriochlorophyll molecule 2. (b) Bacteriochlorophyll molecule 4 with presumed imidazole as magnesium ligand. (c) Bacteriochlorophyll molecule 7 with presumed imidazole as magnesium ligand.


(a)
(b)
(c)

(0-2). This suggests that the metal and the oxygen could be bridged by two water molecules in a manner strikingly similar to that observed by Strouse and co-workers in the crystal structure of ethyl chlorophyllide $a \cdot 2H_2O$ (Strouse, 1974; Chow *et al.*, 1975).

For each bacteriochlorophyll, the electron density map suggests that the magnesium is displaced from the chlorin plane toward the metal ligand. The resolution of the electron density map is not sufficient to determine the out-of-plane distance accurately, but in each case the map is consistent with a displacement of about 0·4 Å, as observed in the structure of ethyl chlorophyllide *a* (Strouse, 1974). For Bchl molecules 1, 2, 4, 5 and 6, the magnesium is on the side of the chlorin ring opposite to C-30, the first carbon in the phytyl chain; i.e. on the same side as the carbomethoxy ester, as is the case for ethyl chlorophillide *a*. Bchl molecules 3 and 7 have their magnesiums displaced on the opposite side of the ring.

In no case is there an indication that any of the seven magnesium atoms is six-co-ordinated. For each Bchl, the side of the magnesium distal to the fifth ligand is screened from protein and from solvent by a phytyl chain, either from the same Bchl or from one of the others.

The overall arrangement of the seven Bchl molecules is shown in Figure 11. There are many contacts less than 4 Å between the ring substituents of adjacent rings, and

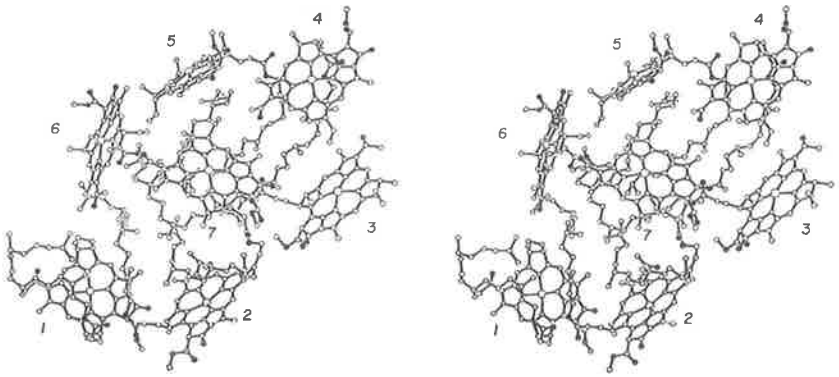


FIG. 11. The 7-chlorophyll core of the Bchl *a*-protein.

between rings and phytyl tails, but in no case does any pair of rings overlap. There is only one instance in which an atom in the π-electron system of one macrocycle approaches within 4 Å of another, this being the keto oxygen 0-1 of Bchl molecule 1, which is calculated to be 3·5 Å and 3·2 Å, respectively, from C-19 and C-20 of Bchl molecule 2. (All such calculated distances have an estimated uncertainty of about 0·4 Å for conjugated ring atoms, and about 0·7 Å for ring substituents.)

The packing of the phytyl chains in the Bchl protein can be considered as a model for lipid–lipid interactions. As can be seen in Figure 12, there are extensive contacts between the atoms in the phytyl chains. Approach distances less than 4 Å are marked. The arrows show the position at which each phytyl chain emanates from its chlorin ring. It is striking to note that essentially in every case the chains begin at the outside of the cluster and run toward the middle. This indicates that the hydrophobic self-aggregation of the phytyl chains must provide a significant part of the driving force

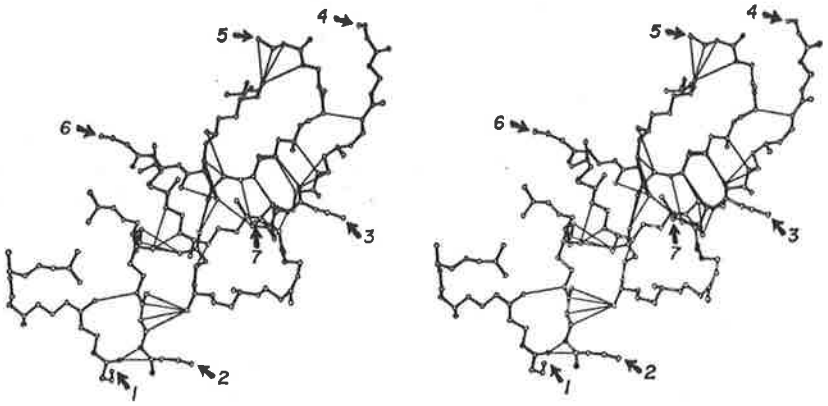


FIG. 12. View of the 7 phytyl chains with bacteriochlorophyll head groups removed. The sites of attachment are indicated by arrows. Approach distances of 4 Å or less are indicated by connecting single lines.

for the association of the seven Bchl molecules. The cluster of the seven phytyl chains can be considered as the inner core of the Bchl protein. Around this inner core are arranged the seven head groups, with the protein shell surrounding the whole (Figs 6 and 13).

Inspection of the Bchl protein model suggests that the removal of one or more of the Bchl molecules would drastically destabilize the protein. Also there is no evidence that the chlorophyll can be reversibly removed from the protein. It seems clear that the folding of the Bchl protein must be a co-operative process involving both the polypeptide and the bacteriochlorophyll. As can be seen from Figures 6 and 8, the protein can be visualized as being derived from a large, essentially planar, structure, folded around a core of bacteriochlorophyll. The actual folding pathway is open to conjecture, but is interesting to note that strands 3 through 8 of the β-sheet wall (Figs 8 and 9) form a simple, compact, contiguous structural element which might provide the nucleus for the folding process.

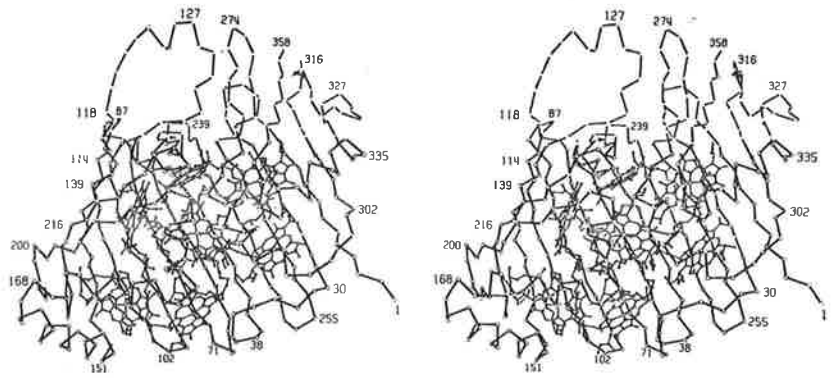


FIG. 13. Stereo view of the bacteriochlorophyll *a*-protein (cf. Fig. 6).

8. Spectral Properties

The spectrum of any array of chromophores is determined not only by the characteristic spectrum of the isolated chromophores, but by the state of aggregation and by the molecular environment of the members of the array. Obviously, one would like to use the observed spectral characteristics of components of photosynthetic units to determine the nature and arrangement of the chromophores within such units. To date, rather limited progress has been made in this area, largely because of the paucity of well-characterized model systems.

The structure determination of the bacteriochlorophyll protein provides such a model system. It is not only the first determination of the structure of a chlorophyll array, but is the first known structure of any chlorophyll-containing entity. There is, therefore, considerable interest in correlating the three-dimensional structure of the chlorophyll protein with its observed absorption, fluorescence and circular dichroism spectra.

Bacteriochlorophyll *a* in non-polar organic solvents has an absorption maximum at about 770 nm for the Q_y absorption band, whereas the maximum for the bacteriochlorophyll *a* protein at room temperature is at 809 nm (Olson *et al.*, 1963). At 77 K both the absorption and circular dichroism spectra begin to show considerable structure, with the absorption band splitting into four maxima between 789 and 825 nm, and the circular dichroism Q_y band showing evidence for at least five components between 786 and 823 nm (Olson, 1978*b*; Philipson & Sauer, 1972; Olson *et al.*, 1976*b*). Philipson and Sauer interpreted the low-temperature spectra in terms of exciton splittings arising from interactions among the bacteriochlorophyll *a* molecules. They were able to simultaneously fit the observed 809 nm absorption and circular dichroism spectral bands with five skewed Gaussian components centered at 787 to 792, 800 to 804, 812 to 813, 814 to 817, and 823 to 825 nm. The five components were interpreted to indicate an interaction between at least five bacteriochlorophyll *a* molecules in the Bchl protein. Also, from calculations of the expected exciton band splitting for hypothetical chlorophyll arrays, Philipson & Sauer estimated the center-to-center separation of the bacteriochlorophylls in the Bchl protein to be 12 to 15 Å. Both these predictions are confirmed by the structure as it is now known. More recently, Olson *et al.* (1976*b*) have shown that the observed spectra can be better fitted by adding a sixth component, providing evidence for an exciton interaction between at least six chromophores.

Now that the arrangement of bacteriochlorophylls within the Bchl protein is known, one can apply exciton theory (Kasha, 1963) to calculate the expected absorption and circular dichroism spectra of the complex. The calculation requires not only knowledge of the position and orientations of the Bchl *a* molecules, but also the transition moments of the individual chromophores. Because the transition moment of each bacteriochlorophyll molecule will be influenced by its environment, which, in the case of the Bchl protein includes specific interactions between the magnesiums, the chlorin substituents, and protein side-chains, the exciton calculation is not straightforward.

As a first step, we assumed the transition moments of the seven bacteriochlorophylls within one subunit to be identical and equal to that of monomeric Bchl *a* (which is a gross over-simplification), and calculated the transition interaction energies between all possible pairs of Bchl *a* molecules (Fenna & Matthews, 1976). The results, shown in Figure 14, indicate that the seven chromophores act as a single unit in terms of

the delocalization of excitation energy. Similar calculations involving all 21 chromophores in the Bchl protein trimer (Figs 14 and 15) suggest that there are significant interactions between the three number 7 Bchl molecules, and between Bchl molecule 5 and Bchl molecules 2 and 3 of a neighboring subunit. This adds a further complication to the exciton calculations.

Calculations of the energies and oscillator strengths of the exciton components of the absorption band, based on the simplified model described above, with seven equivalent Bchl molecules, do not agree particularly well with the observed spectra.

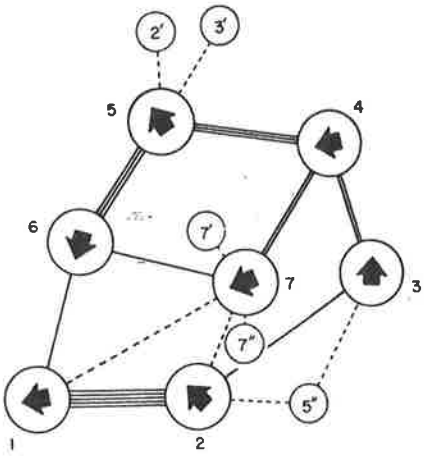


FIG. 14. Schematic illustration of the calculated interaction energies between the 7 bacteriochlorophylls within one subunit, and between subunits of the Bchl *a*-protein. Each solid connecting line indicates an interaction of about 40 wave numbers, and the broken connecting lines about 20 wave numbers.

In particular, it is not possible to account for the observed red shift of the band from 770 nm for monomeric Bchl *a* to about 809 nm for the Bchl protein. Other investigators (Pearlstein & Hemenger, 1978; L. L. Shipman & J. R. Norris, personal communication) have also found that first-order exciton calculations are unable to account for the observed red shift. It seems unlikely that this discrepancy is due to inaccuracy of the Bchl *a* co-ordinates as determined from the present electron density map. Shipman (1977) recently carried out extensive test calculations on different arrays of 40 chlorophyll *a* molecules, absorbing at 670 nm, plus a single chromophore absorbing at 700 nm, and found that transition density coupling always led to a broadening of the absorption bands, but never produced a red shift of more than about 15 nm. Furthermore, large exciton shifts require the chromophores to be parallel and close together, as, for example, in crystals of ethyl chlorophyllide $a \cdot 2H_2O$, where the estimated exciton shift is about 30 nm (Strouse, 1978). This is not the case for the Bchl protein.

We have considered the possibility that the red shift observed in the Bchl protein, relative to monomeric Bchl *a*, might be due to the presence of some species other than bacteriochlorophyll *a*. There is no evidence for this in the spectrum of the chlorophyll dissociated from the Bchl protein, which is identical with that of bacteriochlorophyll *a* (Olson *et al.*, 1963). Also, the structural differences between Bchl *a* and

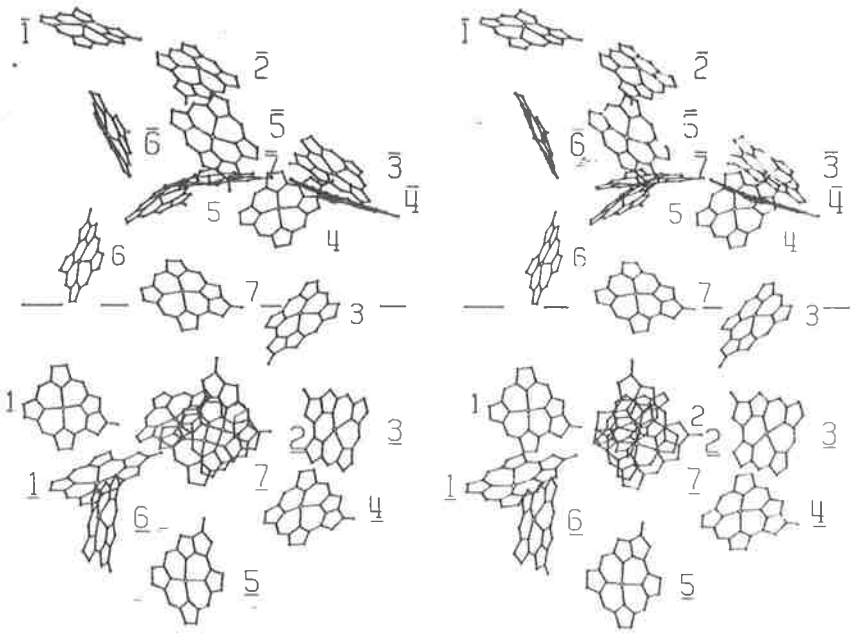


FIG. 15. Stereo view of the head groups of 21 bacteriochlorophylls within the Bchl *a*-protein trimer, with the ring substituents, excluding the keto oxygen, removed. The bacteriochlorophylls within one subunit are numbered 1 to 7, those within the second subunit $\bar{1}$ to $\bar{7}$, and those within the third subunit $\underline{1}$ to $\underline{7}$. The 3-fold axis relating the subunits is indicated by the horizontal broken line.

the other chlorophyll species ought to cause recognizable changes in the electron density map. For example, ring II is saturated in Bchl *a*, but not in chlorophyll *a* or *b* nor in the *Chlorobium* chlorophylls. This would cause a substantial change in the positions of the ring II methyl and ethyl substituents. For all seven rings, the electron density map appears to correspond to Bchl *a* rather than any of the other alternatives. We are therefore forced to conclude that the Bchl protein contains only bacteriochlorophyll *a*.

It seems likely, then, that the red shift observed for the Bchl protein must be due to perturbations of the spectra of the individual bacteriochlorophylls by the protein environment. Whether this is due to some generalized environment provided by the protein, or to specific protein–bacteriochlorophyll interactions, is not clear at this point. In the absence of the amino acid sequence, it is not possible to define in detail all the protein–chlorophyll interactions. We can do no more than inspect the environment of each bacteriochlorophyll and tentatively identify potential ligands, in particular to the magnesiums, to the ring V keto oxygens, and to the ring I acetyl oxygens, which might perturb the spectra. The magnesium ligands have already been described. In the case of the oxygens, there is often an amino acid residue close enough to be a potential hydrogen-bond doner, although this is not always the case. The keto oxygen of Bchl molecule 1 is close to the macrocycle carbons C-19 and C-20 of Bchl molecule 2 (Fig. 11) and does not obviously participate in any hydrogen-bonding. The keto oxygen of Bchl molecule 2 appears to form a hydrogen bond with

the protein backbone of a threefold-related subunit, although the subunit–subunit contact is such that in the trimer Bchl molecule 2 is not accessible to the external solvent. One suggestive interaction involving a keto oxygen is for Bchl molecule 6, where the apparent ligand appears to be an arginine side-chain partially buried in the surface of the protein. Further identification of the interactions involving the chlorophylls will have to await determination of the amino acid sequence.

Recently, Pearlstein & Hemenger (1978) have made the interesting observation that the agreement between the observed and calculated absorption and circular dichroism spectra for the Bchl protein is substantially improved by rotating the transition direction of each Bchl by 90°, i.e. by assuming that the lowest singlet transition of bacteriochlorophyll *a* is *x*-polarized rather than *y*-polarized. At present there is little theoretical justification for such an assumption, but the fact remains that it does lead to spectra which agree reasonably well with those observed. Clearly, further work in this area is necessary. In the meantime, it seems that considerable caution is necessary in deducing structural features solely by analyzing optical properties such as circular dichroism, fluorescence and absorption.

9. Organization of Chlorophyll *in Vivo*

Although membrane-bound *in vivo* (Olson *et al.*, 1976*a*) the bacteriochlorophyll protein is atypical in that it is water-soluble *in vitro*, whereas virtually all chlorophyll-containing components of photosynthetic systems are insoluble in aqueous solvents. Nevertheless we believe the Bchl protein provides a good model for the organization of chlorophyll *in vivo*, because the Bchl protein has the chlorophyll on the inside and the protein on the outside.

There is now substantial evidence that chlorophyll *in vivo* is usually associated with protein. For example, many components of photosynthetic membranes have been shown to have complex absorption and fluorescence spectra which revert to the simplified spectra of chlorophyll *a*, chlorophyll *b*, Bchl *a*, etc., on extraction with organic solvents. In many instances the wavelength maxima are appreciably red-shifted from those in solution, although not so much as for crystals. Thus there is little evidence supporting the earlier notion of crystalline arrays of chlorophyll in photosynthetic membranes, but increasing evidence favoring the occurrence of chlorophyll–protein complexes (e.g. see Sauer, 1975). Recently, a number of such complexes have been isolated. These include not only bacterial reaction centre particles (e.g. see Clayton & Wang, 1971; Okamura *et al.*, 1974), but also components of the antenna system (Sauer & Austin, 1978). Thornber and colleagues (e.g. see Thornber, 1975) have isolated a light-harvesting chlorophyll *a*/*b* protein, which is a major component of higher plants and green algae and appears to consist of three moles of chlorophyll *a* and three moles of chlorophyll *b* associated with a single polypeptide of 25,000 molecular weight.

These antenna chlorophyll–protein complexes all have spectra which can be interpreted in terms of exciton coupling between some of the chromophores. Furthermore, the complexes contain about 1 to 2 moles of chlorophyll per 10,000 molecular weight of protein, which is very comparable with the ratio of 1·6 moles of Bchl *a* to 10,000 molecular weight of protein in the water-soluble Bchl protein. For these reasons, it is reasonable to expect that the membrane-bound chlorophyll proteins have

structures of the same type as the water-soluble Bchl protein, i.e. with the chromophores forming the core of the complex, and surrounded by a shell of protein. Also, particularly for the chlorophyll *a*/*b* protein from green plants, in order to enclose the chromophores within the available length of polypeptide chain, it is likely that the protein coat will include a large amount of β-sheet. The main difference, then, between the water-soluble and membrane-bound chlorophyll proteins is expected to be in the surface residues of the respective proteins, which include more hydrophobic and less hydrophilic residues for the membrane-bound species. By enclosing the chlorophyll within an envelope of protein, the spectral properties of each chlorophyll–protein complex will be determined by the geometrical arrangement of the chromophores, in combination with the immediate environment provided by the protein. Within a photosynthetic membrane, the protein envelope would act as a shield, preventing perturbation of the chlorophyll spectrum by non-specific chlorophyll–lipid interactions. In such a lipid environment, it is possible that the shielding of the chromophores could be provided in part by the phytyl chains, which themselves might be partly exposed to the external lipid environment, but it seems unlikely that the chlorin moieties would be exposed.

It is not to be expected that the geometrical arrangement of chlorophylls within other chlorophyll-proteins will be similar to that in the water-soluble bacteriochlorophyll protein, although it can be anticipated that the arrangements will tend to be irregular, and not highly symmetric or repetitive. Notwithstanding predictions based on crystal structures of chlorophyll derivatives (Fischer *et al.*, 1972; Strouse, 1976), it seems unlikely that chlorophyll will be arranged in simple repetitive linear polymers, either in association with a membrane, or within chlorophyll-containing proteins.

10. Organization of the Bacteriochlorophyll Protein *in Vivo*

In the green photosynthetic bacteria, the bulk light-harvesting pigment, *Chlorobium* chlorophyll, is located in discrete chlorophyll bodies measuring 300 to 500 Å in diameter and 1000 to 1600 Å long (Cohen-Bazire *et al.*, 1964; Cruden & Stainer, 1970; Fowler *et al.*, 1971). The reaction center, on the other hand, is incorporated within the cytoplasmic membrane which is adjacent to, but not contiguous with, the underlying chlorophyll bodies (Fowler *et al.*, 1971; Olson *et al.*, 1976*c*; Boyce *et al.*, 1976). The bacteriochlorophyll protein is known to be closely associated with the membrane containing the reaction centers, and is thought to form a bridge between the chlorophyll body and the reaction center (Olson *et al.*, 1976*c*). This model is consistent with two of the atypical features of the Bchl protein, one being the fact that the Bchl protein occurs in the green bacteria, but seems to have no counterpart in the green plants. Because the antenna chlorophyll and the reaction center in the green bacteria are in separate structures, an intermediary, i.e. the Bchl protein, is required to facilitate efficient transfer of excitation energy from one to the other. On the other hand, in the other photosynthetic prokaryotes and in the green plants, both the antenna chlorophyll and the reaction center are in the same membraneous structure, and are presumably close together, so that efficient energy transfer from one to the other can occur directly. The second unusual property of the Bchl protein is its water solubility. This follows as a natural consequence of the Bchl protein being

attached to, rather than enclosed within, the cytoplasmic membrane, and presumably located in an aqueous environment. When the antenna system and the reaction centers are located in the same membrane, as is usually the case, there is no need for any water-soluble component.

Another unusual aspect of the Bchl protein is its 3-fold symmetry. As is well known, most oligomeric proteins have symmetry based on 2-fold axes, for example cyclic symmetry C_2 and dihedral symmetry D_2 (Matthews & Bernhard, 1973). Other than the Bchl protein only one other protein, a gluconate aldolase from *Pseudomonas putida* (Vandlen *et al.*, 1973) has been shown crystallographically to have cyclic symmetry C_3. Threefold symmetry may be unusual for water-soluble proteins, but it is the expected symmetry for close-packed planar arrays. In the purple membrane, for example, this symmetry is observed (Henderson & Unwin, 1975). Therefore the 3-fold symmetry of the Bchl protein could indicate close association of the protein in a planar array adjacent to the cytoplasmic membrane containing the reaction center. Such association of the protein in close-packed layers is in fact observed in a second crystal form of the protein which has a lower solvent content than the crystal form used for the structure determination (Fenna *et al.*, 1974; Matthews *et al.*, 1977). Finally, it will be noted in Figure 15 that, as a consequence of the 3-fold symmetry, the 21 chromophores within each Bchl protein trimer have an almost continuous range of azimuthal angles about the 3-fold axis, consistent with an ability to collect excitation energy from a wide range of angles. Figure 15 allows one to visualize the arrangement of the chromophores in the Bchl protein relative to the cytoplasmic membrane. Presumably this membrane is normal to the 3-fold axis of the Bchl trimer, although it is not known which "end" of the protein binds to the membrane. The amino acid sequence, when known, will reveal the nature of the residues on the surface of the protein which may indicate the orientation of protein relative to the membrane.

We thank Dr Lynn Ten Eyck for help with running his heavy-atom refinement and structure idealization programs. One of us (M.C.B.) thanks the Commonwealth Fund for a Harkness Fellowship. Work at the University of Oregon was supported in part by research grant PCM 77-19310 from the National Science Foundation. Work at Brookhaven National Laboratory was done under the auspices of the United States Department of Energy.

REFERENCES

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Crystallogr.* **12**, 794–802.
Boyce, C. O. L., Oyewole, S. H. & Fuller, R. C. (1976). *Brookhaven Symp. Biol.* **28**, 365.
Chothia, C. (1973). *J. Mol. Biol.* **75**, 295–302.
Chow, H.-C., Serlin, R. & Strouse, C. E. (1975). *J. Amer. Chem. Soc.* **97**, 7230–7237.
Clayton, R. K. & Wang, R. T. (1971). *Methods Enzymol.* **23**, 696–704.
Cohen-Bazire, G., Pfennig, N. & Kunisawa, R. C. (1964). *Cell Biol.* **22**, 207–225.
Colman, P. M., Jansonius, J. N. & Matthews, B. W. (1972). *J. Mol. Biol.* **70**, 701–724.
Cruden, D. L. & Stainer, R. Y. (1970). *Arch. Mikrobiol.* **72**, 115–134.
Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G. & North, A. C. T. (1961). *Proc. Roy. Soc. ser. A*, **265**, 15–38.
Dodson, E. J., Isaacs, N. W. & Rollett, J. S. (1976). *Acta Crystallogr. sect. A*, **32**, 311–315.
Fenna, R. E. & Matthews, B. W. (1975). *Nature (London)*, **258**, 573–577.

Fenna, R. E. & Matthews, B. W. (1976). *Brookhaven Symp. Biol.* **28**, 170–182.
Fenna, R. E., Matthews, B. W., Olson, J. M. & Shaw, E. K. (1974). *J. Mol. Biol.* **84**, 231–240.
Fenna, R. E., Ten Eyck. L. F. & Matthews, B. W. (1977). *Biochem. Biophys. Res. Commun.* **75**, 751–756.
Fischer, M. S., Templeton, D. H., Zalkin, A. & Calvin, M. (1972). *J. Amer. Chem. Soc.* **94**, 3613–3619.
Fowler, C. F., Nugent, N. A. & Fuller, R. (1971). *Proc. Nat. Acad. Sci., U.S.A.* **68**, 2278–2282.
Hamilton, W. C., Rollett, J. S. & Sparks, R. A. (1965). *Acta Crystallogr.* **18**, 129–130.
Henderson, R. & Unwin, P. N. T. (1975). *Nature (London)*, **257**, 28–32.
Kasha, M. (1963). *Radiat. Res.* **20**, 55–71.
Labaw, L. W. & Olson, R. A. (1970). *J. Ultrastruct. Res.* **31**, 456–464.
Matthews, B. W. (1966). *Acta Crystallogr.* **20**, 82–86.
Matthews, B. W. (1970). In *Crystallographic Computing* (Ahmed, F. R., ed.), pp. 144–159, Munksgaard, Copenhagen.
Matthews, B. W. (1977). In *The Proteins* (Neurath, H. & Hill, R. L., eds), 3rd edit., vol. 3, pp. 403–590, Academic Press, New York.
Matthews, B. W. & Bernhard, S. A. (1973). *Annu. Rev. Biophys. Bioeng.* **2**, 257–317.
Matthews, B. W. & Czerwinski, E. W. (1975). *Acta Crystallogr. sect. A*, **31**, 480–487.
Matthews, B. W., Klopfenstein, C. E. & Colman, P. M. (1972). *J. Phys. E.* **5**, 353–359.
Matthews, B. W., Fenna, R. E. & Remington, S. J. (1977). *J. Ultrastruct. Res.* **58**, 316–330.
Matthews, B. W., Weaver, L. H. & Kester, W. R. (1974). *J. Biol. Chem.* **249**, 8030–8044.
North, A. C. T. (1965). *Acta Crystallogr.* **18**, 212–216.
Okamura, M. Y., Steiner, L. A. & Feher, G. (1974). *Biochemistry*, **13**, 1394–1410.
Olson, J. M. (1966). In *The Chlorophylls* (Vernon, L. P. & Seely, G. R., eds), pp. 413–425, Academic Press, New York.
Olson, J. M. (1971). In *Photosynthesis Part A* (San Pietro, A., ed.) (vol. 23 of *Methods of Enzymology*), pp. 636–644, Academic Press, New York.
Olson, J. M. (1978*a*). *Int. J. Syst. Bacteriol.* **28**, 129–139.
Olson, J. M. (1978*b*). In *The Photosynthetic Bacteria* (Clayton, R. K. & Sistrom, W. R., eds), pp. 161–197, Plenum Press, New York.
Olson, J. M. & Romano, C. A. (1962). *Biochim. Biophys. Acta*, **59**, 726–728.
Olson, J. M., Filmer, D., Radloff, R., Romano, C. A. & Sybesma, C. (1963). In *Bacterial Photosynthesis* (Gest, H., San Pietro, A. & Vernon, L. P., eds), pp. 423–431, Antioch Press, Yellow Springs, Ohio.
Olson, J. M., Koenig, D. F. & Ledbetter, M. C. (1969). *Arch. Biochem. Biophys.* **129**, 42 48.
Olson, J. M., Giddings, T. H. & Shaw, E. K. (1976*a*). *Biochim. Biophys. Acta*, **449**, 197–208.
Olson, J. M., Ke, B. & Thompson, K. H. (1976*b*). *Biochim. Biophys. Acta*, **430**, 524–537.
Olson, J. M., Prince, R. C. & Brune, D. C. (1976*c*). *Brookhaven Symp. Biol.* **28**, 238–246.
Olson, J. M., Shaw, E. K. & Engleberger, F. M. (1976*d*). *Biochem. J.* **159**, 769–774.
Olson, R. A., Jennings, W. H. & Hanna, C. H. (1969). *Arch. Biochem. Biophys.* **130**, 140–147.
Pearlstein, R. M. & Hemenger, R. P. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 4920–4924.
Philipson, K. D. & Sauer, K. (1972). *Biochemistry*, **11**, 1880–1885.
Remington, S. J., Anderson, W. F., Owen, J., Ten Eyck, L. F., Grainger, C. T. & Matthews, B. W. (1978). *J. Mol. Biol.* **118**, 81–98.
Richards, F. M. (1968). *J. Mol. Biol.* **37**, 225–230.
Sauer, K. (1975). In *Bioenergetics of Photosynthesis* (Govindgee, ed.), pp. 115–181, Academic Press, New York.
Sauer, K. & Austin, L. A. (1978). *Biochemistry* **17**, 2011–2019.
Shipman, L. L. (1977). *J. Phys. Chem.* **81**, 2180–2184.
Strouse, C. E. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 325–328.
Strouse, C. E. (1976). In *Chemistry* (Lippard, D., ed.), vol. 21, pp. 159–177, John Wiley, New York.
Sybesma, C. & Olson, J. M. (1963). *Proc. Nat. Acad. Sci., U.S.A.* **49**, 248–253.
Sybesma, C. & Vredenberg, W. J. (1963). *Biochim. Biophys. Acta*, **75**, 439–441.

Sybesma, C. & Vredenberg, W. J. (1964). *Biochim. Biophys. Acta*, **88**, 205–207.
Ten Eyck, L. F., Weaver, L. H. & Matthews, B. W. (1976). *Acta Crystallogr. sect. A*, **32**, 349–350.
Thornber, J. P. (1975). *Annu. Rev. Plant Physiol.* **26**, 127–158.
Thornber, J. P. & Olson, J. M. (1968). *Biochemistry*, **7**, 2242–2249.
Vandlen, R. L., Ersfeld, D. L., Tulinsky, A. & Wood, W. A. (1973). *J. Biol. Chem.* **248**, 2251–2253.

OFFPRINTS FROM LIPID-PROTEIN INTERACTIONS, VOL. 1
EDITED BY PATRICIA JOST & O. HAYES GRIFFITH
COPYRIGHT © 1982 BY JOHN WILEY & SONS, INC.

One

Lipid-Protein Interactions in a Bacteriochlorophyll-Containing Protein

BRIAN W. MATTHEWS

Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oregon U.S.A.

I. INTRODUCTION 1
II. THE BACTERIOCHLOROPHYLL PROTEIN 2
A. Chlorophyll Core 5
B. Phytol Conformation 5
1. Bacteriochlorophyll 1
2. Bacteriochlorophyll 2
3. Bacteriochlorophyll 3
4. Bacteriochlorophyll 4
5. Bacteriochlorophyll 5
6. Bacteriochlorophyll 6
7. Bacteriochlorophyll 7
C. Inferences for Lipid-Protein Interactions 17
1. Lipid Conformations Are Irregular
2. Lipid Tails Prefer Parallel Interactions
3. Bends Occur at Substituted Positions
4. Lipid Tails Have Irregular Conformations But Are Not Crystallographically Disordered
5. Lipid-Lipid Interactions Do Not Dominate Over Lipid-Protein Interactions
6. Role of Protein Secondary Structure
III. SUMMARY 21
ACKNOWLEDGMENTS 22
REFERENCES 22

I. INTRODUCTION

Although more than 100 protein structures have now been determined at essentially atomic resolution, all of these proteins are more or less water soluble, and none

are intrinsic membrane proteins (e.g., see Matthews, 1977). To date, very limited progress has been made in obtaining good three-dimensional crystals of any integral membrane protein, although the use of octyl glucoside (Michel & Oesterhelt, 1980; Garavito & Rosenbusch, 1980) and Triton X-100 (Henderson & Shotton, 1980) to crystallize bacteriorhodopsin and the so-called matrix protein has recently shown promise. Once good crystals are obtained, there should be no difficulty in performing structure determination by the usual method of isomorphous replacement.

A striking advance has been made by Henderson and Unwin (1975) in using low-dose electron microscopy to determine the structure of bacteriorhodopsin (purple membrane protein). In this case the structure analysis was possible because the protein, within its natural membrane, can be induced to form very regular two-dimensional arrays. The reported structure determination, to a resolution of 7 Å, does not reveal anything of the lipid arrangement in the membrane, or of lipid-protein interactions, although it is to be hoped that such information will be forthcoming from a higher resolution analysis, which is now in progress (R. Henderson, personal communication).

In the meantime, a structural precedent for lipid-protein interactions has been established by the determination of the bacteriochlorophyll *a*-containing protein from *Prosthecochloris aestuarii* (Fenna & Matthews, 1975; Matthews et al., 1979). In this case the protein is peripherally associated with the cytoplasmic membrane of the green photosynthetic bacterium and can be readily dissociated from the membrane into aqueous solution. However, although the protein itself is water soluble, it encloses a cluster of seven bacteriochlorophyll *a* molecules that are intimately associated both with themselves and with the surrounding protein. The structure of this protein-chlorophyll complex therefore provides an example of the sorts of structural features that might be anticipated in other complexes between lipids and protein. The lipid conformation, lipid packing, and lipid-protein interactions within the bacteriochlorophyll protein constitute the focus of this chapter.

II. THE BACTERIOCHLOROPHYLL PROTEIN

The light-gathering apparatus of the green photosynthetic bacteria consists of three chlorophyll-containing entities. First, the "chlorophyll bodies", which reside beneath the cytoplasmic membrane, incorporate 95% of the total chlorophyll of the organism and constitute the principal light-gathering pigment. Second, the "bacteriochlorophyll-protein", which includes most of the remaining chlorophyll, resides between the chlorophyll bodies and the cytoplasmic membrane. Third, the reaction center itself, which resides in the cytoplasmic membrane. The bacteriochlorophyll-protein is thought to transmit excitation energy to a special pair of chlorophylls in the reaction center where the first step in energy transduction, namely a charge separation, occurs.

The bacteriochlorophyll-protein (Bchl-protein), first isolated and crystallized by John Olson (Olson & Romano, 1962; Olson et al., 1969), has been subject to crystallographic analysis and the structure determined to a nominal resolution of

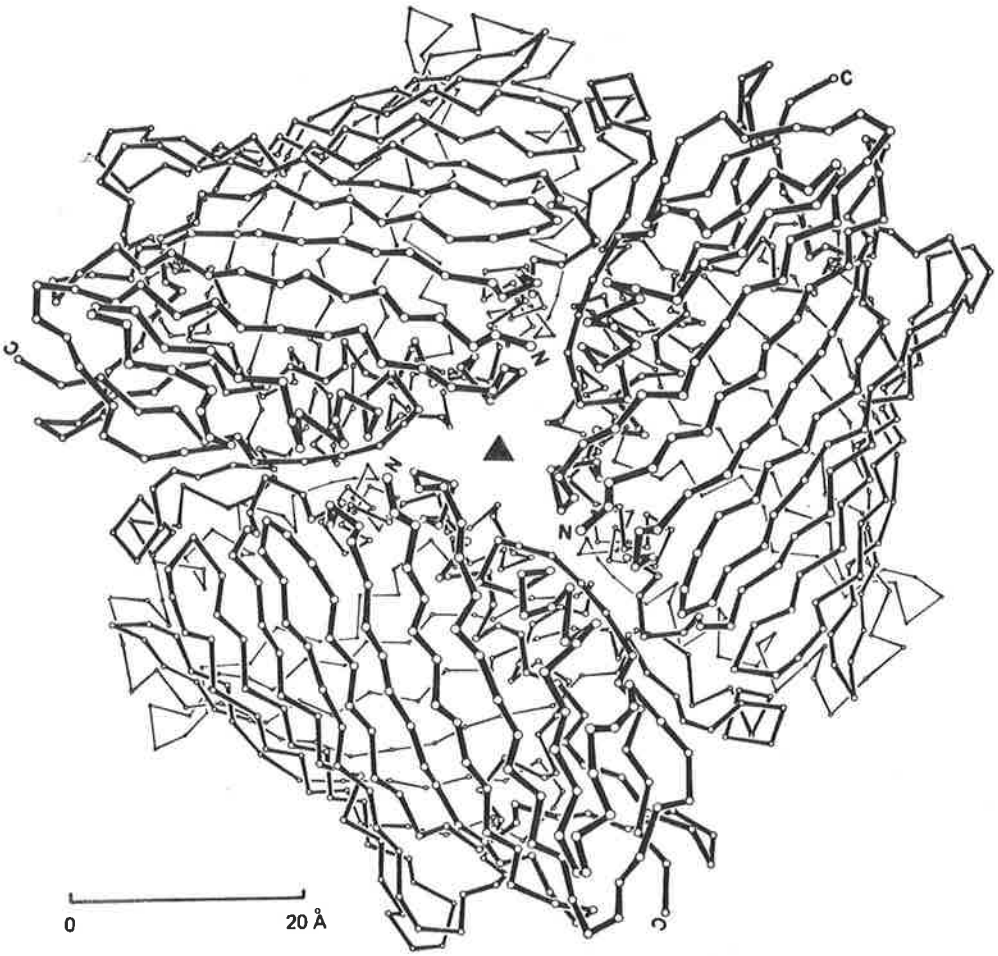


Figure 1. The protein backbone of the Bchl-protein seen projected down the 3-fold symmetry axis. The connected circles show the positions of successive α-carbon atoms. The amino (N) and carboxyl (C) termini of the polypeptide backbone are indicated. The seven bacteriochlorophylls enclosed within each of the three subunits are not shown. Matthews et al. (1979). Reprinted with permission from *The Proteins*. Copyright by Academic Press Inc. (London) Ltd.

2.7 Å (Fenna & Matthews, 1975; Matthews et al., 1979). The molecule (MW 150,000 daltons) consists of three identical subunits related by 120° rotations about a threefold axis of symmetry (Fig. 1). In each subunit the polypeptide backbone forms a large, twisted β-sheet of 15 strands, which forms the "outside" of the protein molecule (i.e., the part exposed to solvent) and encloses a central core of 7 bacteriochlorophyll *a* molecules. The part of the protein that forms the subunit-subunit contacts includes several α-helices (Fig. 2).

All the available evidence suggests that the trimeric structure seen in the crystals occurs *in vivo* and is not an artifact of crystallization. The subunit-subunit contacts

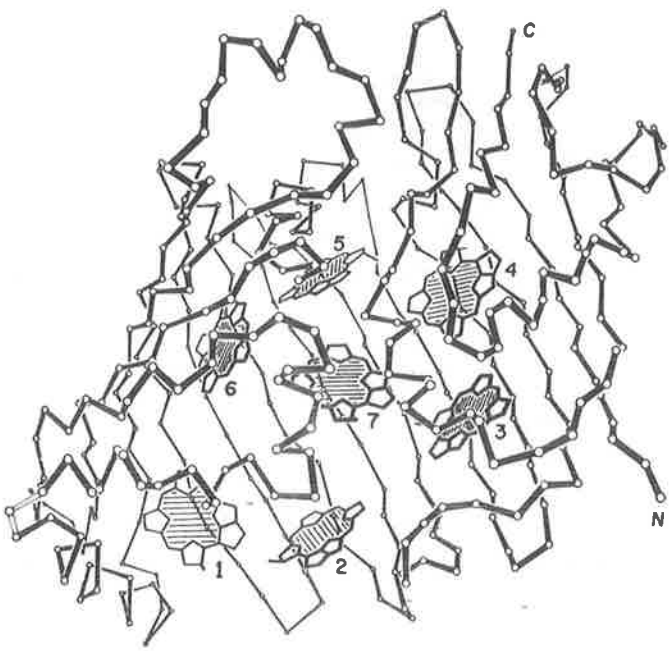


Figure 2. A single subunit of the Bchl-protein, viewed from the center of the trimer, with the 3-fold axis horizontal. For simplicity only the seven chlorin rings of the bacteriochlorophylls are shown, with the phytyl chains omitted. The numbering of the bacteriochlorophylls is consistent with that used in Figs. 3 and 5, and in the text. Reprinted with permission from Fenna et al. (1977).

seen in the crystal are very extensive, indicating strong intersubunit interaction. The three subunits cannot be dissociated from each other without also separating the chlorophyll from the protein. Therefore it is reasonable to assume that the structure as determined crystallographically is biologically relevant.

In contrast to many membrane proteins where the protein is enclosed within a lipid environment, the Bchl-protein has the lipid enclosed within the protein. Also, in contrast to a normal membrane bilayer, the lipids are not constrained to have polar head groups at the membrane surface, exposed to solvent, and hydrocarbon tails away from the surface. Thus, the constraints imposed by enclosing the bacteriochlorophylls within the protein envelope of the bacteriochlorophyll-protein may not be strictly comparable to the constraints that occur when a lipid interacts with a protein, and, at the same time, remains part of a membrane bilayer. Nevertheless, the Bchl-protein does provide an example of the interaction of hydrocarbon chains in the presence of protein, and it is reasonable to expect that the lipid-lipid and lipid-protein interactions seen here will illustrate the main features that are anticipated in other complexes where lipids are intimately associated with protein. Lipids with branched hydrocarbon tails, as in the phytyl chains, are found in naturally occurring membranes. For example, the predominant lipids of the purple membrane from *H. halobium* have ether-linked dihydrophytyl tails (Kushwaha et al., 1975).

A. Chlorophyll Core

The central core of each subunit of the Bchl-protein consists of seven bacteriochlorophyll molecules clustered together within an approximate ellipsoid measuring 45 × 35 × 15 Å (Fig. 3). As discussed in more detail below, there are extensive interactions both between the different Bchls and between the Bchls and the surrounding protein. To a rough approximation, the seven chlorin rings are parallel with each other (within about 40°), but they are not coplanar. As can be seen in Fig. 3, four of the chlorin rings (4,5,6,7) form a boundary on one side of the Bchl core, while chlorin rings 1 and 2 form a partial boundary on the other side of the core. Each of these six porphine rings interacts mainly with phytyl chains on one side, and protein on the other. The remaining porphine ring (3) is at the edge of the core and interacts mainly with protein on both sides of the ring.

Since the amino acid sequence of the Bchl-protein is not known, specific interactions between individual amino acid side chains and the protein are not known with certainty. Nevertheless, inspection of the electron density map suggests that for Bchl molecules 1, 3, 4, 6, and 7, the central magnesium ion is liganded by a histidine side chain. Bchl-5 appears to be liganded by a peptide oxygen of the protein backbone, and Bchl-2 appears to be liganded by a water molecule trapped within the protein.

B. Phytol Conformation

The phytyl chains pack together in an irregular way to form the inner hydrophobic core of the Bchl-protein subunit (Figs. 2 and 3). It is important to note that, although the phytols pack in an irregular manner, they are not disordered. The electron density map for the chlorophyll protein shows that each bacteriochlorophyll molecule, including both the head group and the phytyl tail, has a distinct, well-defined conformation that is essentially the same from one Bchl-protein molecule to the next. This is not to imply that the bacteriochlorophylls are held completely rigid. Rather, there must be some vibration and other thermal motion, but such excursions are typically on the order of a few tenths of an angstrom unit, as occur within the interior of a typical globular protein. In no case is there an indication that a phytyl chain occupies radically different conformations in different molecules.

The resolution of the current electron density map, nominally 2.7 Å, limits the precision with which the coordinates of the bacteriochlorophylls are known. The estimated uncertainty in the preliminary coordinates (Table 1) is about 0.4 Å for the conjugated chlorin ring system and twice this for the ring substituents, including the phytyl chains (Fenna et al., 1977). Data have been collected to 1.9 Å resolution, and a refinement of the structure, now in progress (M. F. Schmid & B. W. Matthews, unpublished results), will improve the accuracy of these coordinates. Also it should be possible to determine if there are marked differences in thermal motion as one proceeds along the individual phytyl chains. At present there is no obvious indication of such variations.

In the section that follows we briefly describe the conformation and environment

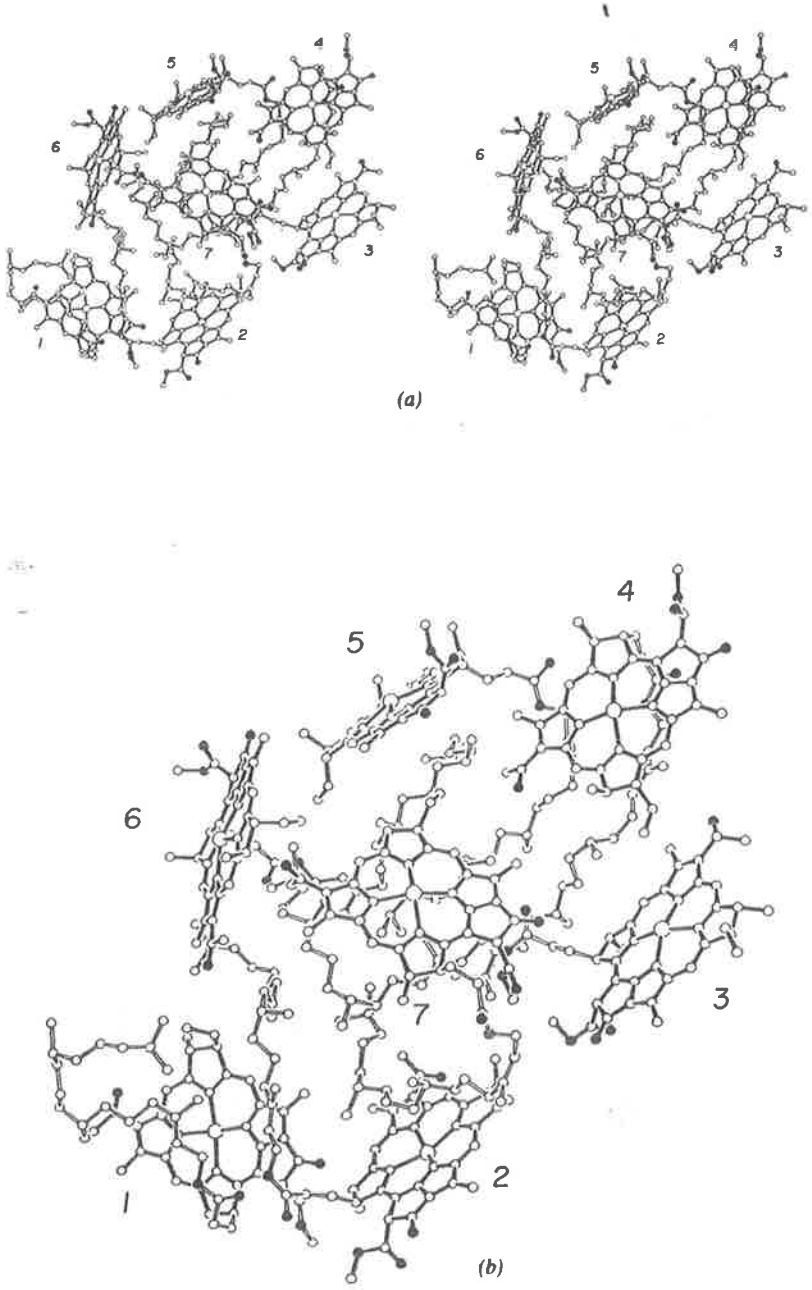


Figure 3. (a) Stereo drawing of the seven bacteriochlorophylls which constitute the core of the Bchl-protein. In this and subsequent figures, oxygen atoms are solid, nitrogen half-solid, and carbon atoms open circles. (*b*) Mono drawing of the seven bacteriochlorophylls: the chlorin head group bonds are solid and the phytol chains are drawn with open bonds. Matthews et al. (1979). Reprinted with permission from *The Proteins*. Copyright by Academic Press Inc. (London) Ltd.

of each of the seven bacteriochlorophylls in the Bchl-protein. For reference, the structure of bacteriochlorophyll *a* and the atomic numbering used are shown in Fig. 4. The conformations of the seven individual Bchl molecules are illustrated in the series of seven stereo drawings (Figs. 5*a*–5*g*). In each case the direction of view is the same, namely parallel to the crystallographic *z*-axis. The numbering of the bacteriochlorophylls is given in Figs. 2 and 3. In three cases (Bchl's 2, 4, and 7) the stereo drawings include the electron density contours from which the coordinates of the bacteriochlorophylls were derived. These figures are included to help give the reader an indication of the quality, but at the same time the limitations, of the present electron density map from which the structure of the Bchl-protein was obtained. In no case can individual atoms be resolved at a resolution of 2.7 Å, but in each case the electron density for the planar chlorin ring and for the long phytyl chain can be seen quite unambiguously. Furthermore, careful model building shows that the electron density map usually includes appropriate features for both the substituents of the chlorin rings and the methyl substituents of the phytyl tails. On the other hand, as can be appreciated from Figs. 5*b*, 5*d*, and 5*g*, the precise positions of these substituents and the exact conformations of the phytyl chains cannot be reliably determined at this time. These limitations must be borne in mind in the following discussion.

1. *Bacteriochlorophyll 1*

Bchl 1 is the most compactly folded of all the Bchls (Figs. 3 and 5*a*). The phytol lies, in large part, against its own chlorin head group, and has sharp bends at the three substituted positions C_{38}, C_{42}, and C_{46} (see Fig. 4), to form a complete hairpin bend. The hairpin loop interacts with an α-helix on one side and β-sheet on the other (Fig. 6). A short segment of the phytyl chain of Bchl 1, including the three carbons adjacent to the ester linkage, has hydrophobic interactions with the phytol of Bchl 2 (Fig. 7), but other than this, phytol 1 does not interact with any other phytol. The phytol-phytol interactions for Bchl 1 are far less than for any other Bchl (Fig. 7). This indicates that when lipids interact with a protein they need not retain the extended side-by-side conformation they normally adopt in a membrane bilayer. So long as the protein provides an appropriate hydrophobic environment, lipids interacting with the protein can be expected to adopt irregular conformations complementary to the protein surface. Also, hydrophobic lipid-lipid interactions do not necessarily predominate over hydrophobic lipid-protein interactions.

2. *Bacteriochlorophyll 2*

The phytyl tail of Bchl 2 (Figs. 3 and 5*b*) extends away from the head group and changes direction at C_{35}, C_{42}, and C_{47} (i.e., as with phytol 1, at or near substituted positions along the hydrocarbon chain). At various points along its length, phytol 2 interacts hydrophobically with phytols 1, 3, and 7 (Figs. 3 and 7). Within the regions of the phytols that participate in these interactions, the respective hydrocarbon chains are more or less parallel to each other, and in each case the chains

TABLE 1 Coordinates for the Bacteriochlorophyll Core

The coordinates of the atoms in the seven bacteriochlorophyll *a* molecules are in angstroms relative to orthogonal axes parallel to the crystallographic axes *a**, *b*, and *c*, with the origin at the origin of the unit cell. Reproduced with permission from Fenna et al. (1977).

BCHL	1	MG	75.7	24.6	20.8		2	C40	71.7	19.6	24.7		4	C24	58.8	13.7	41.6		6	C8	59.7	16.7	24.4
	1	N1	75.9	24.7	18.7		2	C41	70.6	19.7	23.6		4	C25	59.8	15.0	44.6		6	C9	59.3	18.1	23.9
	1	N2	73.6	24.4	20.5		2	C42	69.2	19.3	24.1		4	C26	60.9	14.1	44.1		6	C10	58.1	18.7	24.1
	1	N3	75.5	23.7	22.5		2	C43	68.3	20.5	24.1		4	C27	55.3	15.0	48.2		6	C11	57.7	20.0	23.8
	1	N4	77.8	24.0	23.8		2	C44	69.0	21.9	24.4		4	C28	52.4	16.9	47.5		6	C12	56.5	20.7	24.0
	1	O1	77.4	22.4	26.6		2	C45	67.9	23.0	24.6		4	C29	51.6	17.7	46.4		6	C13	56.6	22.0	23.5
	1	O2	79.6	24.9	25.1		2	C46	68.3	24.2	23.8		4	C30	50.9	20.4	44.2		6	C14	57.9	22.0	22.9
	1	O3	80.7	22.9	25.4		2	C47	67.4	25.4	24.2		4	C31	51.8	21.4	44.5		6	C15	58.3	23.3	22.3
	1	O4	78.7	20.1	22.5		2	C48	67.1	26.3	22.9		4	C32	50.2	18.3	41.2		6	C16	59.5	23.5	21.7
	1	O5	78.1	20.4	20.3		2	C49	66.1	25.5	22.1		4	C33	50.2	17.2	46.6		6	C17	59.8	24.9	21.1
	1	O6	73.6	26.0	15.3		2	C50	66.0	26.0	20.6		4	C34	47.9	17.8	46.4		6	C18	61.2	24.5	20.4
	1	C1	77.1	24.7	18.0		2	C51	64.7	25.5	20.0		4	C35	53.1	20.9	45.3		6	C19	61.5	23.2	21.0
	1	C2	76.7	25.0	16.6		2	C52	73.0	17.2	24.9		4	C36	55.3	20.9	45.5		6	C20	62.8	22.6	20.9
	1	C3	75.4	25.1	16.4		2	C53	69.4	18.8	25.6		4	C37	56.7	21.0	44.9		6	C21	65.7	21.4	20.5
	1	C4	74.8	24.9	17.8		2	C54	69.8	24.6	24.0		4	C38	57.6	20.0	44.7		6	C22	65.7	18.6	21.8
	1	C5	73.5	24.9	18.1		2	C55	67.2	25.7	19.8		4	C39	58.8	20.3	43.8		6	C23	65.5	17.3	22.5
	1	C6	72.9	24.7	19.4	BCHL	3	MG	65.6	14.0	44.9		4	C40	60.0	19.3	43.8		6	C24	61.4	15.3	23.1
	1	C7	71.4	24.6	19.5		3	N1	64.3	15.3	45.8		4	C41	61.1	19.8	42.9		6	C25	59.7	16.6	25.9
	1	C8	71.2	24.3	21.1		3	N2	65.6	12.8	46.6		4	C42	61.2	19.0	41.5		6	C26	59.6	18.0	26.6
	1	C9	72.7	24.1	21.5		3	N3	67.4	13.1	44.5		4	C43	62.0	19.7	40.4		6	C27	55.3	20.1	24.7
	1	C10	73.0	23.8	22.8		3	N4	66.2	15.6	43.5		4	C44	63.3	20.4	40.8		6	C28	56.0	23.2	23.3
	1	C11	74.3	23.6	23.4		3	O1	70.8	12.4	41.3		4	C45	63.4	21.8	40.0		6	C29	57.0	24.1	22.5
	1	C12	74.7	23.2	24.7		3	O2	70.5	15.9	42.1		4	C46	64.8	22.1	39.6		6	C30	60.0	26.0	22.0
	1	C13	76.1	23.1	24.6		3	O3	71.0	15.5	40.1		4	C47	65.0	22.1	38.1		6	C31	60.6	25.7	23.4
	1	C14	76.5	23.4	23.3		3	O4	64.5	17.7	38.0		4	C48	66.4	21.7	37.4		6	C32	61.1	24.4	18.9
	1	C15	77.9	23.4	23.1		3	O5	66.5	17.6	37.0		4	C49	66.6	21.4	36.3		6	C33	56.2	24.8	21.4
	1	C16	78.5	23.7	21.9		3	O6	60.5	16.8	47.8		4	C50	68.0	21.6	35.7		6	C34	56.5	25.7	19.3
	1	C17	80.0	23.6	21.7		3	C1	63.8	16.4	45.2		4	C51	67.8	21.7	34.2		6	C35	61.6	26.4	24.3
	1	C18	80.1	24.3	20.3		3	C2	62.6	16.8	45.9		4	C52	57.6	18.7	45.4		6	C36	62.7	26.7	26.5
	1	C19	78.7	24.3	19.8		3	C3	62.4	15.9	46.9		4	C53	61.9	17.6	41.9		6	C37	62.5	26.5	28.0
	1	C20	78.3	24.6	18.5		3	C4	63.5	14.9	46.8		4	C54	65.3	23.4	40.2		6	C38	63.3	25.7	28.8
	1	C21	77.8	25.1	15.5		3	C5	63.6	13.8	47.6		4	C55	68.8	20.4	36.0		6	C39	63.0	25.2	30.2
	1	C22	74.6	25.3	15.2		3	C6	64.6	12.8	47.5	BCHL	5	MG	53.3	23.9	31.0		6	C40	62.1	23.9	30.2
	1	C23	75.6	25.3	14.0		3	C7	64.7	11.5	48.4		5	N1	53.1	26.0	31.3		6	C41	60.6	24.3	30.4
	1	C24	70.8	26.0	19.2		3	C8	66.0	10.8	48.0		5	N2	54.7	24.3	29.5		6	C42	59.9	23.2	31.2
	1	C25	70.5	22.9	21.1		3	C9	66.4	11.6	46.8		5	N3	54.3	22.1	31.1		6	C43	58.5	23.8	31.5
	1	C26	70.1	22.5	19.7		3	C10	67.5	11.2	46.1		5	N4	52.6	23.7	33.1		6	C44	58.3	23.9	33.1
	1	C27	73.8	23.0	25.8		3	C11	68.0	11.9	44.9		5	O1	54.5	17.9	33.1		6	C45	56.9	23.9	33.5
	1	C28	77.3	22.7	25.4		3	C12	69.1	11.5	44.1		5	O2	51.8	18.1	34.7		6	C46	56.6	24.1	35.0
	1	C29	78.5	22.9	24.5		3	C13	69.1	12.5	43.1		5	O3	51.2	19.8	33.5		6	C47	56.0	25.5	35.2
	1	C30	80.6	22.2	21.7		3	C14	68.1	13.5	43.4		5	O4	51.8	23.1	39.7		6	C48	56.0	26.4	34.0
	1	C31	80.3	21.0	20.9		3	C15	68.0	14.5	42.5		5	O5	54.0	23.4	39.0		6	C49	55.3	27.8	34.3
	1	C32	80.7	25.7	20.3		3	C16	67.2	15.6	42.5		5	O6	53.9	29.9	29.4		6	C50	56.1	29.0	34.1
	1	C33	79.5	23.7	25.3		3	C17	67.1	16.8	41.6		5	C1	52.4	26.6	32.3		6	C51	55.5	30.2	34.8
	1	C34	81.7	23.7	25.9		3	C18	66.0	17.7	42.3		5	C2	52.3	28.0	32.0		6	C52	64.4	24.9	28.1
	1	C35	78.9	20.5	21.3		3	C19	65.5	16.8	43.4		5	C3	52.9	28.3	30.9		6	C53	59.7	21.9	30.4
	1	C36	76.7	20.5	20.6		3	C20	64.3	17.1	44.1		5	C4	53.5	26.9	30.4		6	C54	55.6	23.0	35.4
	1	C37	75.9	20.6	19.2		3	C21	61.8	18.0	45.6		5	C5	54.2	26.7	29.3		6	C55	56.1	29.3	32.6
	1	C38	74.6	20.7	19.2		3	C22	61.3	15.9	47.9		5	C6	54.8	25.5	28.8	BCHL	7	MG	63.4	14.0	32.5
	1	C39	73.9	21.2	17.9		3	C23	61.5	15.2	49.2		5	C7	55.7	25.4	27.5		7	N1	63.8	14.0	30.4
	1	C40	74.8	21.1	16.7		3	C24	64.9	11.8	49.9		5	C8	56.2	23.9	27.5		7	N2	62.0	15.5	32.2
	1	C41	74.0	21.4	15.4		3	C25	65.6	9.3	47.7		5	C9	55.4	23.4	28.9		7	N3	63.5	14.7	34.4
	1	C42	74.9	21.2	14.2		3	C26	66.7	8.3	48.1		5	C10	55.6	22.1	29.1		7	N4	65.4	13.1	32.7
	1	C43	73.9	20.9	13.0		3	C27	69.9	10.3	44.3		5	C11	55.1	21.4	30.2		7	O1	65.0	13.9	38.7
	1	C44	73.2	19.6	13.4		3	C28	69.8	12.9	41.9		5	C12	55.3	20.1	30.6		7	O2	68.2	12.7	37.4

| | | |

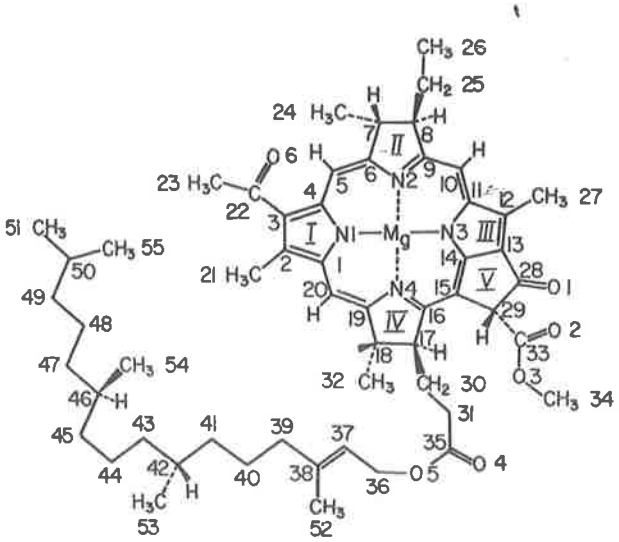
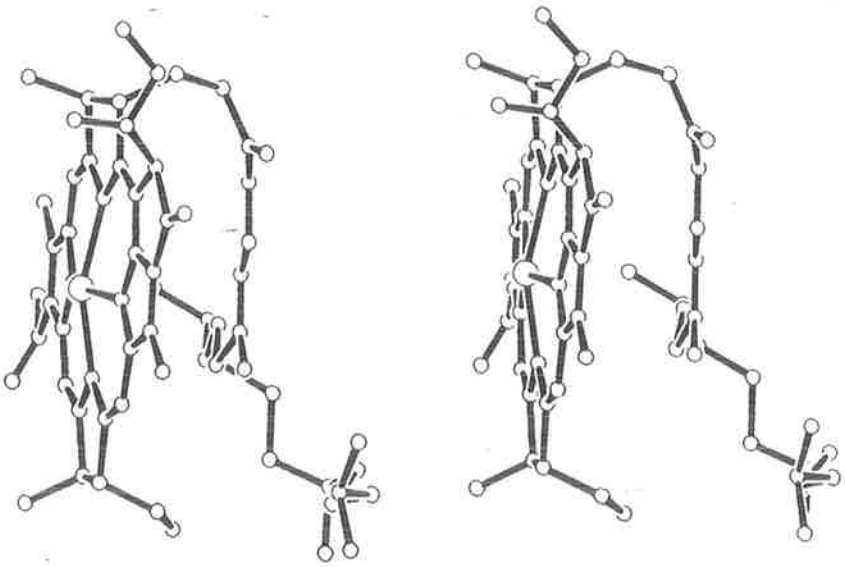


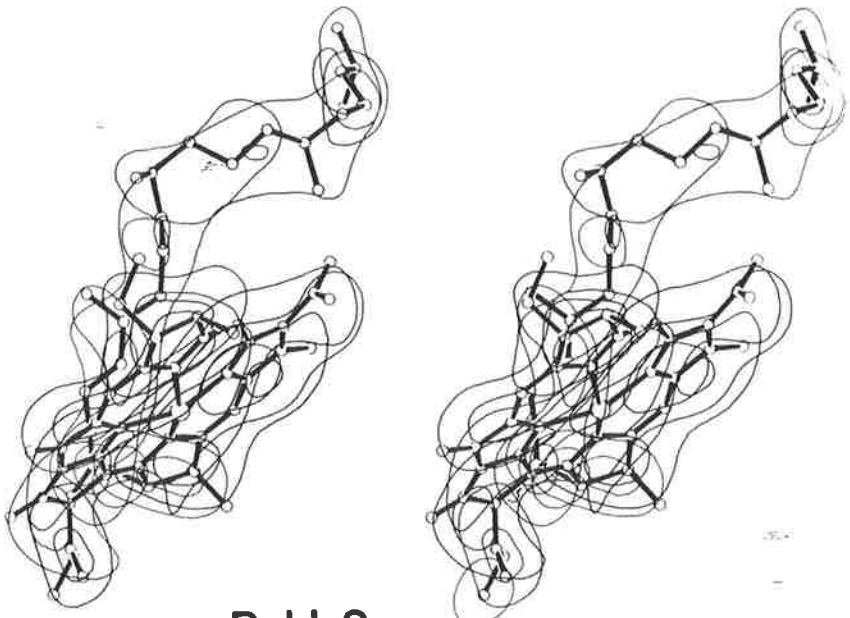
Figure 4. Schematic illustration of bacteriochlorophyll a showing the atomic numbering that is used. Reprinted with permission from Fenna et al. (1977).



Bchl 1

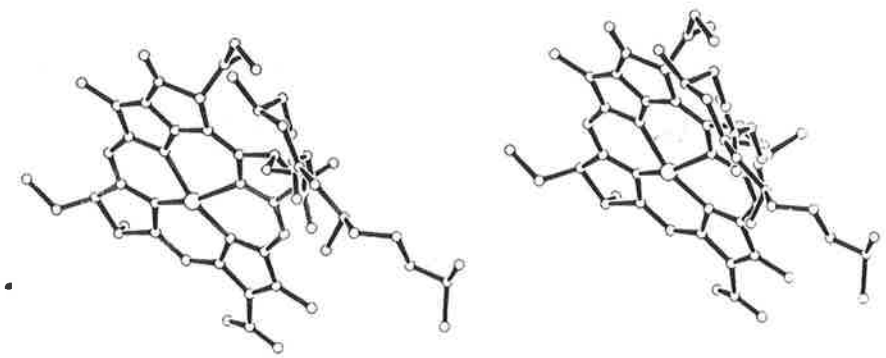
(a)

Figure 5. Stereo drawings of the seven bacteriochlorophylls viewed parallel to the c crystallographic axis (i.e., as in Fig. 1). (*a*) Bchl 1. (*b*) Bchl 2 with electron density superimposed. (*c*) Bchl 3. (*d*) Bchl 4 with electron density superimposed; the imidazole that acts as a ligand to the magnesium is also shown. (*e*) Bchl 5. (*f*) Bchl 6. (*g*) Bchl 7, including the electron density and liganding imidazole.



Bchl 2

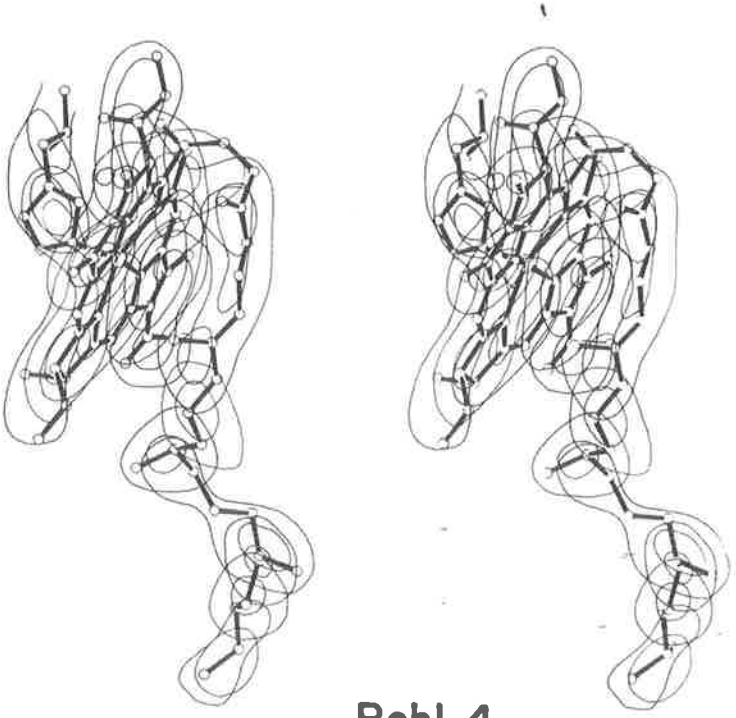
(b)



Bchl 3

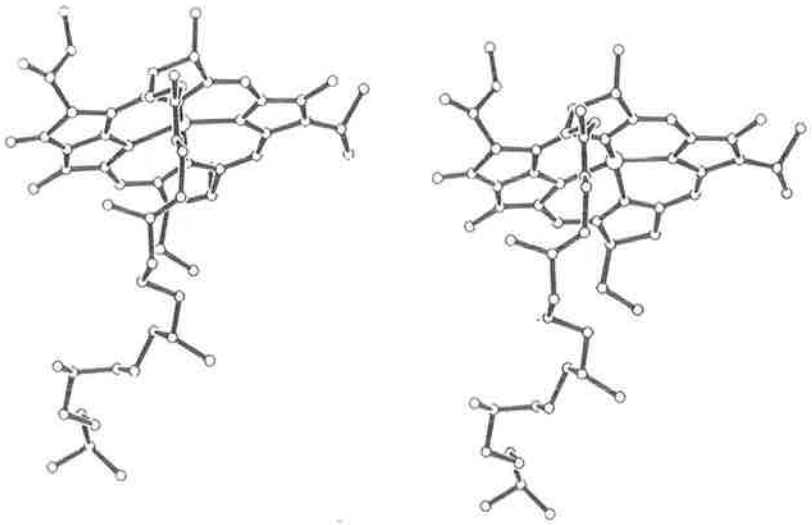
(c)

Figure 5. *(continued)*



Bchl 4

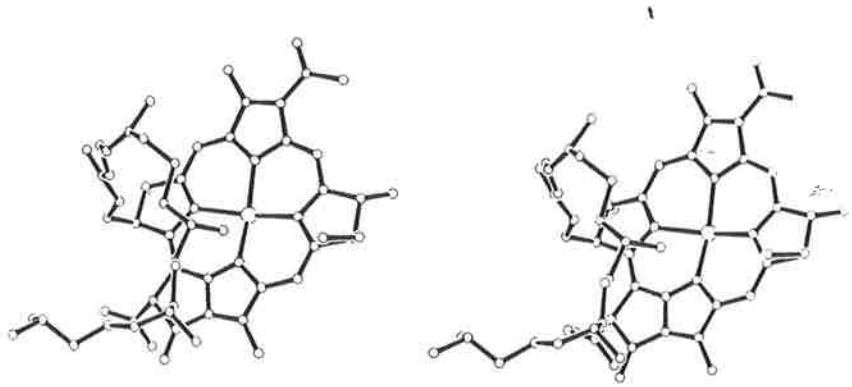
(d)



Bchl 5

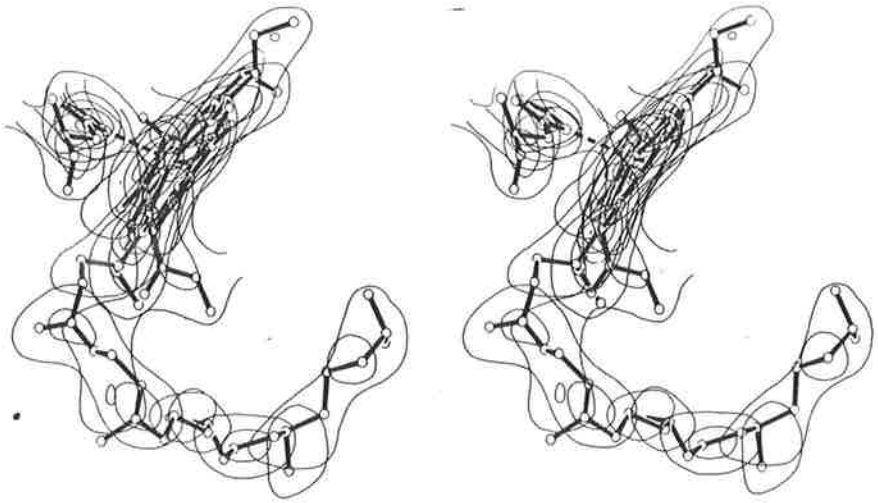
(e)

Figure 5. *(continued)*



Bchl 6

(f)



Bchl 7

(g)

Figure 5. *(continued)*

run in the same sense. In other words there seems to be a tendency to preserve (or at least to utilize) the parallel, side-by-side interactions that occur between lipids in a membrane bilayer. However, as we shall dicuss below, the preference for parallel packing is not always followed. Part of phytol 2 interacts with an α-helix and part with β-sheet (Fig. 6). As can be seen in Figs. 2 and 6, Bchl 2, and, to some extent, Bchl 1, do not appear to be enclosed by protein, and hence appear to be exposed to the solvent. In fact, these "exposed" regions are completely covered by the adjacent subunit of the trimer (Fig. 1). The current model of the Bchl-protein suggests that the protein surrounds the Bchl core in such a manner that little, if any, contact occurs between the solvent and any one of the Bchls. However, it is interesting to note that Bchl 2 is the only case in which the magnesium ligand appears to be a water molecule rather than part of the protein.

3. *Bacteriochlorophyll 3*

Bchl 3 adopts a very extended conformation and extends almost from one side of the Bchl core to the other (Figs. 3 and 5*c*). There is only one place at which the phytyl chain sharply changes direction, this being at C_{46}, again a substituted position. In addition to the interaction with phytol 2 mentioned above, phytol 3 has hydrophobic interactions with phytols 4 and 5 (Figs. 3 and 7).

In each case, in the region of contact the hydrocarbon chains lie side by side and are parallel (as distinct from antiparallel). Phytol 3 has practically no contact with the protein at all. Its predominant interactions are with the other phytols listed above, and with the chlorin ring of Bchl 7.

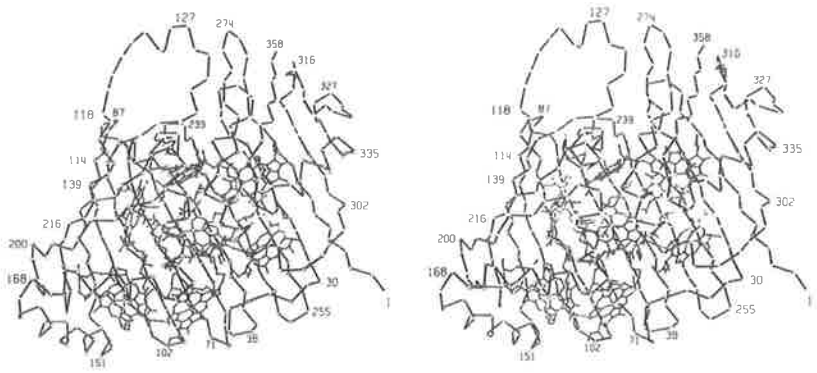


Figure 6. Stereo view of one subunit of the bacteriochlorophyll-protein showing the seven bacteriochlorophylls enclosed within an envelope of protein. The view direction is the same as in Fig. 2, and the 3-fold symmetry axis runs from left to right and is "in front" of the subunit. Only the protein backbone, without side chains, is shown. Matthews et al. (1979). Reprinted with permission from *The Proteins*. Copyright by Academic Press Inc. (London) Ltd.

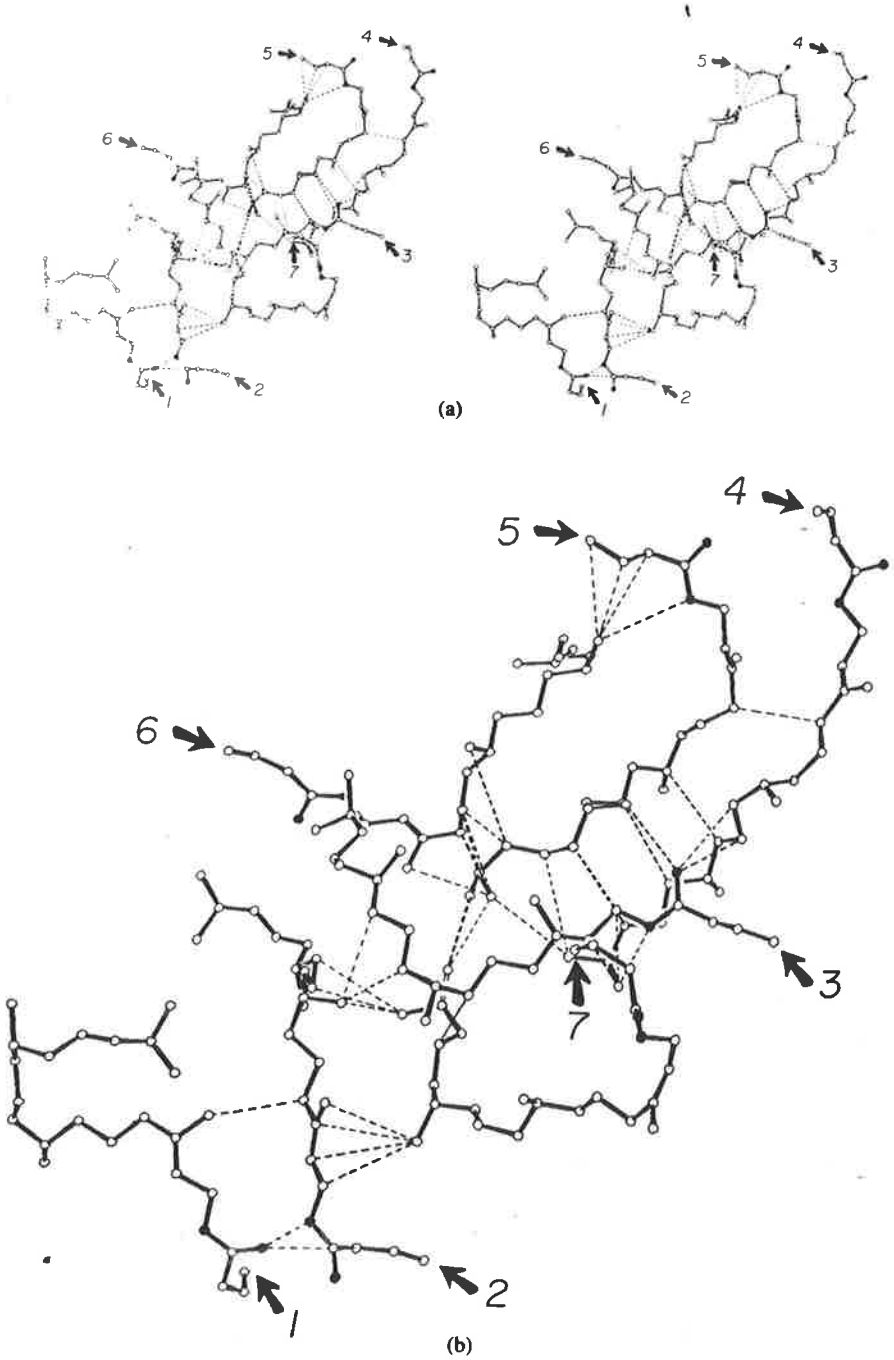


Figure 7. (a) Stereo view and (*b*) mono view of the seven phytyl chains, with chlorin rings removed, showing the interactions between the hydrocarbon chains. Approach distances of 4 Å or less between the chains are indicated by broken lines. The arrows show the point of attachment of the chlorin head groups.

4. *Bacteriochlorophyll 4*

The phytyl tail of Bchl 4 folds back across the face of the chlorin ring and then adopts an almost fully extended conformation, with slight changes of direction at C_{43} and C_{46} (Figs. 3 and 5*d*). Bchl 4 provides the one example of a lipid-protein interaction where the protein does not greatly alter the extended *trans* configuration of the lipid chain. For much of its length, phytol 4 lies parallel and adjacent to phytol 5 to produce the only example in the Bchl-protein of an extended lipid-lipid interaction as envisaged in a protein-free bilayer. As can be seen in Fig. 7, there are many favorable contacts between phytols 4 and 5 that extend along much of their respective lengths. In addition, the two chains interact extensively with the extended β-sheet of the protein that encloses the Bchl core (Fig. 6). The proximal, more polar, sections of the two phytols lie parallel to the β-sheet strands, but for the most part the hydrocarbon chains are at right angles to the β-sheet.

5. *Bacteriochlorophyll 5*

As mentioned above, the phytyl chain of Bchl 5 is largely parallel to that of Bchl 4. As with Bchl 4, the phytyl chain runs back across the face of the chlorin ring, but in this case part of another Bchl (phytol 6) is interposed between chain 5 and its own head group (Figs. 3 and 5*e*). The phytyl chain is more or less extended (at least it does not make sharp bends), but, as can be seen in Fig. 5*e*, there are a number of local irregularities and departures from the fully extended conformation. The observation that the phytols of Bchl 4 and Bchl 5 are largely extended illustrates that although bends tend to occur at substituted positions, the presence of a methyl-substituted carbon does not necessarily imply that a sharp bend will occur at that locus.

6. *Bacteriochlorophyll 6*

Bchl 6 is of interest because it provides the only example in the Bchl-protein of an antiparallel interaction between the phytyl chains. The phytyl tail has distinct bends at C_{39} and C_{46} (Fig. 5*f*), and from C_{39} to C_{46} it lies adjacent to the phytol of Bchl 5 (Figs. 3 and 7). However, in this instance, the respective hydrocarbon chains are antiparallel. There is another contact between phytol 6 and the end of phytol 2 (Figs. 3 and 7) where the chains also are run in opposite directions, but here the region of contact is shorter (excluding the hairpin bend in phytol 1). These two instances, both involving Bchl 6, are the only cases where antiparallel contacts between adjacent phytyl chains are observed.

7. *Bacteriochlorophyll 7*

Possibly the most unusual feature about Bchl 7 is that it provides the only instance of the phytyl chain emanating from one side of the chlorin ring, then folding back to occupy a position on the opposite side of the head group (Figs. 3 and 5*g*). The phytol conformation is somewhat irregular, with changes of direction at the substituted positions C_{35}, C_{38}, C_{42}, and C_{46}, but also departures from the all-*trans*

extended conformation at other places as well. Phytol 7 interacts most extensively with the chlorin ring and the phytyl tail of Bchl 2, but its terminal methyl groups are buried within the interior of the Bchl cluster and also interact with the phytols of Bchls 3 and 5.

C. Inferences for Lipid-Protein Interactions

As discussed above, the chlorophyll protein is not an intrinsic membrane protein, and the arrangement and interactions of the bacteriochlorophylls should be used with caution as a model for lipid arrangement and lipid-protein interactions in general. Nevertheless, until a better structural model for the interactions between proteins and lipids in a bilayer is obtained, it is of interest to ask what inferences can be drawn from the structure of the Bchl-protein. Drawing on the foregoing discussion of the individual Bchls, the following generalizations are suggested for the interaction between lipids and protein.

1. Lipid Conformations Are Irregular

In every case in the Bchl-protein the phytyl chains adopt more or less irregular conformations. In some cases the hydrocarbon chains are quite extended, but they are by no means in an idealized all-*trans* conformation. In one case the hydrocarbon chain folds completely back on itself. In this respect the Bchl-protein differs markedly from crystal structures of isolated lipids such as the double chain glycolipid cerebroside (Pascher & Sondell, 1977), dimyristoyl lecithin (Pearson & Pascher, 1979), and other single chain lipids (Pascher, 1976). In such crystal structures, the lipids adopt very regular, extended conformations. The Bchl-protein structure suggests that in the presence of protein, lipids may adopt quite irregular conformations. The single bonds present in all lipid tails provide many degrees of freedom, and rotations about such bonds can occur with little, if any, energy cost. There is no apparent reason why the surface of a protein should prefer to bind lipids in a fully extended conformation. Rather, the only requirement is that the lipid should adopt a conformation that is complementary to that of the protein, and not make prohibitively short interatomic contacts, or leave unfilled spaces. The free lipid can adopt many different conformations of essentially equal energy. In the presence of the protein, the favorable free energy to be gained from lipid-protein interactions will overwhelm the weak preference of the lipid to adopt the fully extended conformation. As judged by the Bchl-protein, the conformations lipids may adopt when they are in contact with protein can be very irregular.

2. Lipid Tails Prefer Parallel Interactions

In a membrane bilayer, the amphipathic nature of membrane lipids results in the well-known bilayer structure with polar head groups at the surface and nonpolar tails below the surface, interacting with each other. Because of these constraints, the lipid tails are forced to lie more or less parallel to each other, at least for the part of the tail close to the head group. In the Bchl-protein, the constraints on the

TABLE 2 Phytol-Phytol Interactions[a]

Phytol	1	2	3	4	5	6	7
1	−6	+5	·	·	·	·	·
2	+5	·	+6	·	·	−5	+6
3	·	+6	·	+5	+6	×	×
4	·	·	+5	·	+16	·	×
5	·	·	+6	+16	·	−10	×
6	·	−5	×	·	−10	·	·
7	·	+6	×	×	×	·	·

[a]Each entry gives the length of the phytyl chain that interacts for each pair of Bchls (i.e., the number of pairs of contiguous carbons along the respective chains that remain within approximately 5 Å of each other). A positive entry indicates a parallel interaction, a negative entry indicates an antiparallel interaction; × indicates that the chains cross and make contact for only one or two carbon atoms; · indicates that no contacts are made.

Bchl packing are very different from those in a membrane bilayer, but where the hydrocarbon tails do interact, they clearly prefer to do so in a parallel manner. This is illustrated in Table 2, which summarizes the contacts between the various phytyl tails. The number +5, for example, in the first row of the table, means that five contiguous carbons of Bchl 1 remain within approximately 5 Å of five contiguous carbons of Bchl 2. The plus sign denotes that the phytyl tails run parallel, a minus sign indicates antiparallel, and an × indicates that the chains make contact but are neither parallel nor antiparallel. The numbers in the table give a rough indication of the total amount of hydrophobic phytol-phytol contact which occurs within the Bchl-protein. Table 2 suggests that, when contacts do occur between the phytyl chains, they usually extend over at least five carbons. Furthermore, all such extended contacts are parallel, except those involving phytol 6 and the hairpin loop of phytol 1. It is hard to know whether this observed preponderance of parallel alignment, as opposed to antiparallel alignment, indicates an intrinsic preference for the parallel arrangement or is merely a coincidence. If one takes the chlorin head group and the ester linkage into account, then the phytyl chains do have a distinct polarity. In addition, because of the asymmetric centers at C_{42} and C_{46}, the phytols have an intrinsic polarity within the hydrocarbon chain itself. Possibly this inherent asymmetry favors parallel rather than antiparallel alignment of the hydrocarbon chains.

Insofar as protein-lipid interactions in general are concerned, the Bchl-protein suggests that, when several lipids bind to a protein, they may do so with irregular conformations, but at the same time the neighboring lipid-lipid tails will tend to adopt a conformation such that the tails lie parallel.

3. *Bends Occur at Substituted Positions*

As discussed above for the individual Bchls, the phytyl chains often contain sharp bends, and when such bends occur, they are invariably at, or adjacent to, substituted

positions (including C_{30} and C_{31}, between the chlorin ring and the phytol ester linkage). This is not to suggest that the hydrocarbon chains are fully extended except at or near the substituted positions. The current electron density map suggests that many local departures from the all-*trans* configuration do occur, but that gross changes in direction of the phytyl chains occur near the sites of substitution. In no case is there evidence for a sharp bend at, for example, C_{40} or C_{44}.

4. *Lipid Tails Have Irregular Conformations But Are Not Crystallographically Disordered*

In the well-known model for biological membranes proposed by Singer (1974), functional globular proteins are embedded in a fluid mosaic lipid bilayer. Lipids remote from the protein undergo relatively free fluid-like motion, whereas the motion of lipid in direct contact with the protein surface ("boundary lipid") is considerably restricted (Jost et al., 1973). Lipids in this boundary region, however, undergo some motion (Brotherus et al., 1980).

In the case of the Bchl-protein the phytyl chains have irregular conformations, but these conformations are quite distinct and are repeated from one Bchl-protein molecule to the next. The electron density map shows quite clearly that each of the seven phytyl chains occupies a well-defined position (see Figs. 5*b*, 5*d*, and 5*g*). At the present stage of the X-ray analysis, it is impossible to rule out slight variations in conformations at the termini of the phytyl tails, but large-scale disorder clearly does not occur. The amount of motion indicated for the Bchls is the same as that for buried side chains within a globular protein (i.e., a few tenths of an angstrom). Although recent nuclear magnetic resonance (NMR) studies show that such internal side chains do rotate from time to time, there is clearly a dominant low-energy conformation, which is the one seen in the crystal structure (Gelin & Karplus, 1975).

Therefore, the Bchl-protein structure supports the idea that when lipids interact with a protein, much of the motion that they exhibit in a membrane bilayer is "frozen out". The lipids bound to the protein are not to be envisaged as completely static and unable to move. Rather, lipids bound to the protein will exchange with free lipids in the membrane, but, when bound to the protein, the lipids will adopt well-defined sites.

5. *Lipid-Lipid Interactions Do Not Dominate Over Lipid-Protein Interactions*

Notwithstanding the foregoing comments concerning the preference for parallel interactions between the phytyl tails of the Bchls, the Bchl-protein structure does not suggest that hydrophobic interactions between the hydrocarbon chains dominate over interactions between the Bchls and the protein. If the interactions between the phytyl chains were overwhelmingly strong, one would expect a structure in which these chains were clustered together to maximize the interaction between them. This is clearly not the case. Rather, there are some phytol-phytol interactions (Fig.

7 and Table 1), but there are also phytol-protein and phytol-chlorin interactions. We have not attempted to estimate the contributions of these three types of interaction (such a calculation would preferably utilize refined coordinates and knowledge of the amino acid sequence), but inspection of the model suggests that phytol-phytol, phytol-chlorin, and phytol-protein interactions occur in approximately equal proportions. (In addition, there are extensive chlorin-protein interactions, with each chlorin ring having at least one face fully in contact with the surrounding protein, as shown in Figs. 3 and 6. On the other hand, the chlorin-chlorin interactions are weak, and involve only edge-to-edge contacts between the ring substituents.)

Therefore, it is to be expected that in lipid-protein interactions in general, the final structure adopted in the lipid-protein complex need not be dominated by lipid-lipid interactions, nor by lipid-protein interactions. Since protein structures are relatively rigid (at least compared to lipids in a bilayer), the protein will be less able to modify its structure to accommodate the lipid than the lipid is free to adjust its structure to conform to that of the protein. Therefore, while the protein structure may be dominant in determining the structure of the final complex, the lipids will contribute to the overall stability of the complex not only through their favorable interactions with the protein, but also by favorable interactions among themselves. The interactions between the lipid tails are of the type hydrocarbon-hydrocarbon, and the lipid tail-protein interactions are also hydrocarbon-hydrocarbon. Obviously, the energetics are similar whether the lipid is surrounded by protein or by other lipid, just so long as the lipid environment is not polar.

6. *Role of Protein Secondary Structure*

Water-soluble proteins are generally categorized as having three types of secondary structure: helix, sheet, and coil (i.e., irregular). On the average, these structures occur with approximately equal frequency, but in the case of individual proteins, the fraction of a given type of secondary structure may vary from essentially zero to 100%. The purple membrane protein has been shown to consist almost exclusively of α-helices (Henderson & Unwin, 1975), and helical conformations may turn out to be a common feature in many other membrane-bound proteins, although this is by no means certain.

Since the Bchl-protein contains a number of examples of an interaction between protein and lipid, it is of interest to examine the secondary structure of the protein in such regions to see if any general principles are suggested. The most obvious example occurs between the extended β-sheet that encloses one side of the Bchl core and the phytyl chains of Bchls 4, 5, and 6 (Figs. 3 and 6). In the region where the three hydrocarbon tails lie next to each other, they all lie against the β-sheet wall and are perpendicular to the polypeptide strands of the β-sheet. This mode of interaction could clearly be extended to allow a series of more or less fully extended lipids to interact with the surface of a protein. Such a mode of interaction would not appear to be of high specificity and might be a generalized mode of interaction between an intrinsic membrane protein and its lipid milieu. Clearly, this is not the

same basis for interaction that must occur between the purple membrane protein and its surroundings. In this case, the lipid chains must lie more or less parallel to the α-helical rods of the protein. There is an example in the Bchl-protein of an α-helix interacting with lipid tails, although the contacts are not extensive. The α-helix begins at residue 168 (Fig. 6). It interacts with the phytol of Bchl 1, which runs antiparallel to the helix, makes a hairpin turn, and ends up running parallel to the helix. The phytol of Bchl 2 also interacts with the same helix, but, in contrast, crosses at right angles to the helix axis. Most of the other helices in the Bchl-protein interact with the chlorin rings rather than the phytyl tails, often lying across the faces of the rings (Fig. 6).

Until the amino acid sequence of the protein has been determined, it is not possible to describe in detail the interactions between the lipid chains and the individual amino acid side chains. Where the extended β-sheet wall of the protein encloses the Bchl core, the amino acids that face inward, toward the core, appear to be predominantly valine, leucine, isoleucine, alanine, and glycine. These residues alternate with side chains that are directed outward toward the solvent, and presumably include the polar and charged amino acids.

III. SUMMARY

The Bchl-protein provides the only high-resolution structure of a complex of lipids with protein. In this case the lipids are completely enclosed within an envelope of protein. The lipids occupy well-defined but quite irregular conformations, and the same is to be expected for lipid-protein complexes in general. There is no evidence for crystallographic disorder in the lipid conformations. In a number of instances the hydrocarbon chains have sharp bends, and such bends tend to occur at or near the methyl-substituted positions. On the other hand, methyl substitution does not necessarily require the occurrence of a bend. There are instances where the hydrocarbon chains lie side by side, and in such cases parallel rather than antiparallel arrangements occur much more often. Nevertheless, phytol-phytol interactions do not dominate over phytol-protein, phytol-chlorin, or chlorin-protein interactions. Each of these types of interactions occurs with roughly equal frequency in the Bchl-protein structure.

On the other hand, there are relatively few chlorin-chlorin contacts. Different types of protein secondary structure such as α-helices and β-sheets are seen to interact with the phytyl chains. In the Bchl-protein, an extended region of interaction occurs between the β-sheet of the protein and three extended phytyl chains. However, interactions between the lipids and other, less regular, protein conformations, are also observed. In the bacteriohodopsin structure, and in other helical membrane proteins, it is to be expected that interactions between lipids and helices will predominate. In general, it can be anticipated that any type of secondary structure, including helices, sheets, and irregular structure, may participate in lipid-protein interactions.

ACKNOWLEDGMENTS

The contributions of Drs. R. E. Fenna, J. M. Olson, M. F. Schmid, and M. C. Bolognesi in determining the bacteriochlorophyll-protein structure are gratefully acknowledged. Also, I thank Drs. Pat Jost and O. Hayes Griffith for helpful discussions concerning this chapter and lipid-protein interactions in general.

Preparation of this chapter was supported in part by a grant from the National Science Foundation (PCM-8014311) and by a grant from the M. J. Murdock Charitable Trust.

REFERENCES

Brotherus, J. R., Jost, P. C., Griffith, O. H., Keana, J. F. W., & Hokin, L. E. (1980). Charge Selectivity at the Lipid-Protein Interface of Membranous Na, K-ATPase. *Proc. Natl. Acad. Sci. U.S.A.* **77,** 272–276.

Fenna, R., & Matthews, B. W. (1975). Chlorophyll Arrangement in a Bacteriochlorophyll Protein from *Chlorobium limicola*. *Nature (London)* **258,** 573–577.

Fenna, R. E., Ten Eyck, L. F., & Matthews, B. W. (1977). Atomic Coordinates for the Chlorophyll Core of a Bacteriochlorophyll *a*-Protein from Green Photosynthetic Bacteria. *Biochem. Biophys. Res. Commun.* **75,** 751–756.

Garavito, R. M., & Rosenbusch, J. P. (1980). Three-Dimensional Crystals of an Integral Protein: An Initial Analysis. *J. Cell Biol.* **86,** 327–329.

Gelin, B. R., & Karplus, M. (1975). Sidechain Torsional Potentials and Motion of Amino Acids in Proteins: Bovine Pancreatic Trypsin Inhibitor. *Proc. Natl. Acad. Sci. U.S.A.* **72,** 2002–2006.

Henderson, R., & Unwin, P. N. T. (1975). Three-Dimensional Model of Purple Membrane Obtained by Electron Microscopy. *Nature (London)* **257,** 28–32.

Henderson, R., & Shotton, D. (1980). Crystallization of Purple Membrane in Three Dimensions. *J. Mol. Biol.* **139,** 99–109.

Jost, P. C., Griffith, O. H., Capaldi, R. A., & Vanderkooi, G. (1973). Evidence for Boundary Lipid in Membranes. *Proc. Natl. Acad. Sci. U.S.A.* **70,** 480–484.

Kushwaha, S. C., Kates, M., & Martin, W. G. (1975). Characterization and Composition of the Purple and Red Membrane *Halobacterium cutirubrum*. *Can. J. Biochem.* **53,** 284–292.

Matthews, B. W. (1977). In *The Proteins* (H. Neurath & R. L. Hill, Eds.), Vol. 3, 3rd ed., pp. 403–590, Academic Press, New York.

Matthews, B. W., Fenna, R. E., Bolognesi, M. C., Schmid, M. F., & Olson, J. M. (1979). Structure of a Bacteriochlorophyll *a*-Protein from the Green Photosynthetic Bacterium *Prosthecochloris aestuarii*. *J. Mol. Biol.* **131,** 259–285.

Michel, H., & Oesterhelt, D. (1980). Three-Dimensional Crystals of Membrane Proteins: Bacteriorhodopsin. *Proc. Natl. Acad. Sci. U.S.A.* **77,** 1283–1285.

Olson, J. M., & Romano, C. A. (1962). A New Chlorophyll from Green Bacteria. *Biochim. Biophys. Acta* **58,** 726–728.

Olson, J. M., Koenig, D. F., & Ledbetter, M. C. (1969). A Model of the Bacteriochlorophyll-Protein from Green Photosynthetic Bacteria. *Arch. Biochem. Biophys.* **129,** 42–48.

Pascher, I. (1976). Molecular Bonding of Ceramides and Their Implication on Membrane Stability and Permeability. *Biochim. Biophys. Acta* **455,** 433–451.

Pascher, I., & Sondell, S. (1977). Molecular Arrangement in Sphingolipids: The Crystal Structure of Cerebroside. *Chem. Phys. Lipids* **20,** 175–191.

Pearson, R. H., & Pascher, I. (1979). The Molecular Structure of Lecithin Dihydrate. *Nature (London)* **281,** 499–501.

Singer, S. J. (1974). The Molecular Organization of Membranes. *Annu. Rev. Biochem.* **43,** 805–833.

SECTION D. DNA-PROTEIN INTERACTION

Reprinted from Nature, Vol. 290, No. 5809, pp. 754-758, April 30 1981
© Macmillan Journals Ltd 1981

Structure of the cro repressor from bacteriophage λ and its interaction with DNA

W. F. Anderson*†, D. H. Ohlendorf*, Y. Takeda‡ & B. W. Matthews*

* Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oregon 97403, USA
† MRC Group on Protein Structure and Function, Department of Biochemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2H7
‡ Chemistry Department, University of Maryland, Baltimore County, Catonsville, Maryland 21228, USA

The three-dimensional structure of the 66-amino acid cro repressor protein of bacteriophage λ suggests how it binds to its operator DNA. We propose that a dimer of cro protein is bound to the B-form of DNA with the 2-fold axis of the dimer coincident with the 2-fold axis of DNA. A pair of 2-fold-related α-helices of the repressor, lying within successive major grooves of the DNA, seem to be a major determinant in recognition and binding. In addition, the C-terminal residues of the protein, some of which are disordered in the absence of DNA, appear to contribute to the binding.

INTERACTIONS between proteins and nucleic acids are of central importance in molecular biology. Some proteins bind single-stranded polynucleotides while others recognize double-stranded ones. None of these interactions is well understood in molecular terms. In particular, the way in which proteins recognize specific base sequences within double-stranded DNA has remained enigmatic.

The cro protein from bacteriophage λ (hereafter called 'cro') is a repressor molecule which recognizes and binds tightly certain sequence-specific operator regions within the DNA of the bacteriophage. Cro is the smallest repressor protein that has been characterized, and is of interest both because of its role in the delicately balanced regulatory system of bacteriophage λ, and as a general model for protein–DNA interaction.

We describe here the three-dimensional structure of the cro repressor and its apparent mode of interaction with operator DNA. The postulated interaction between repressor and operator is consistent with a variety of chemical and genetic evidence, and may provide a general model for the interaction of proteins with helical DNA.

The bacteriophage λ genome encodes two repressor proteins that play essential, contrasting parts in regulating the development of the phage. The product of the *cI* gene, the cI or λ repressor ('cI'), is required for the maintenance of the lysogenic state, whereas the product of the *cro* gene is necessary during the lytic cycle of phage development (for reviews see refs 1–4, 33, 34). Both the cro and the cI repressors compete for and bind two operator regions within the phage genome. These operator regions each consist of three binding sites for the repressor molecules, and the respective repressors have different relative affinities for each of the sites within the operator regions[5,6]. The use of these binding sites by cro and cI provides for a switch between the lysogenic and lytic pathway. The cro/cI interaction is the clearest example of such a developmental switch[33] and the structural results presented here begin an understanding of this switch at the molecular level.

The cro protein is well characterized[7,8]. It is a small basic protein 66-amino acid residues long, with a monomer molecular weight of 7,351. Both the amino acid sequence of the protein[9] and the nucleotide sequence of the *cro* gene have been determined[10].

Structure determination

The structure was determined by the method of isomorphous replacement[11–13]. Protein purification and crystallization from 1.1 to 1.3 M phosphate were as described previously[8,14]. The space group is R32 and the cell dimensions in the hexagonal system $a = b = 91.9$ Å, $c = 268.9$ Å.

Potential heavy-atom derivatives were initially screened by precession photography, and, later, by using a diffractometer to collect 5-Å-resolution data sets. The first heavy-atom derivative to be interpreted was obtained using $PtCl_4^{2-}$ and was initially solved by a 'brute force' approach in which all possible individual sites, and then combinations of sites, were explored (ref. 15 and unpublished results of B.W.M.). Using the phases obtained from this platinum derivative, other potential derivatives were then evaluated in the normal way[12,13]. In the end, five derivatives were selected for high-resolution data collection (Table 1). All the derivatives have multiple binding sites, as expected with multiple copies of the protein present in the crystallographic asymmetric unit (see below). Higher-resolution data were collected by oscillation photography[16] which permitted collection of data sets to a nominal resolution of 2.8 Å or better from a single crystal. The crystals often grow as long triangular prisms[14], in which case they can be translated between series of exposures. Some data collection statistics are included in Table 1.

Two preliminary electron density maps to 5 Å resolution were calculated, one using the data collected by diffractometry, and the other using a subset of the higher-resolution oscillation data. The two maps were quite similar, and, as discussed below, were important in working out the arrangement of the cro monomers within the crystallographic asymmetric unit.

Molecular symmetry

The initial analysis of the cro protein crystals showed them to contain multiple copies of the protein in the asymmetric unit. On the basis of the density measurement of two cross-linked crystals, there seemed to be six monomers, that is, three dimers, per asymmetric unit[14]. However, examination of the 5 Å-resolution electron density map revealed that there were four monomers per asymmetric unit, not six. The error in the previous estimate could be due either to the presence of the cross-linking agent or to errors in density due to the limited amount of material available.

We attempted to determine the relation between the monomers in two ways—by inspection of the 5-Å-resolution electron density maps, and by consideration of the relation between the sites of heavy-atom binding. Initial inspection of the 5-Å-resolution map revealed an obvious 2-fold symmetry

Table 1 Heavy-atom derivatives

Derivative	No. of sites	R_{Merge} (%)	R_{Der} (%)	Reflections above 3σ	R_{wc} (%)	$\langle f_H \rangle$	E
Native	—	9.5	—	9184			
$PtCl_4^{2-}$	6	9.6	14.5	5068	30.5	64.3	54.5
Hg^{2+}	16	8.9	12.9	3616	44.9	113.4	72.4
Au^{2+}	16	8.7	14.0	7273	47.1	59.5	46.8
EMTS	12	0.3	11.0	7856	48.7	39.0	32.2
Mersalyl	13	10.2	14.0	5985	46.1	62.5	51.5

Overall figure of merit[18] to 2.8 Å resolution, 0.45. $R_{Merge} = 100 \sum (I_i - \langle I \rangle) / \sum \langle I \rangle$, $R_{Der} = 200 \sum |I_{PH} - I_P| / \sum (I_{PH} + I_P)$, $R_{wc} = 100 \sum \omega ||F_{PH} - F_P| - f_H| / \sum \omega |F_{PH} - F_P|$. $\langle f_H \rangle$ is the mean heavy-atom scattering, E the average lack of closure for centric reflections[18]. EMTS, ethyl mercury thiosalicic acid. Hg^{2-} data set collected by precession photography.

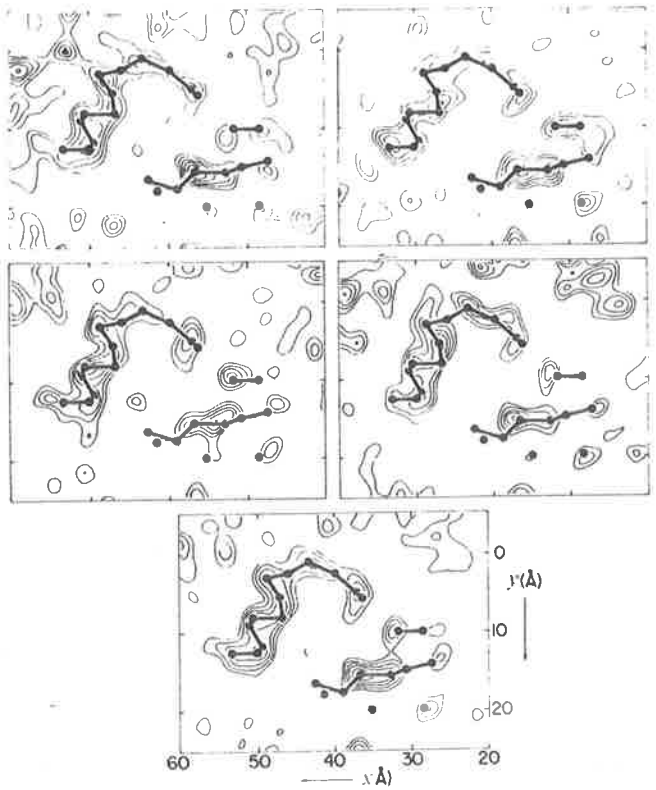


Fig. 1 Corresponding electron density sections through each of the four cro molecules in the crystallographic asymmetric unit, transformed into a common coordinate frame. Clockwise from top left the density is for molecules O, A, B and C, with the averaged density below. α-carbon atoms within ±2.5 Å of the section are indicated by the solid circles.

axis lying approximately in the plane $z = 1/12$. This 2-fold axis also related, in pairs, most of the major heavy-atom sites. Further analysis of the 5-Å map, facilitated by the use of an MMS-X computer graphics system[17], suggested that there were two additional 2-fold axes inclined to the first. The positions of these local symmetry elements were confirmed by calculating correlation coefficients between the electron density of the individual monomers, rotated and translated to superimpose one monomer on another (Table 2). The four monomers are related in pairs by mutually perpendicular intersecting (or nearly intersecting) 2-fold axes. In other words, the four monomers are arranged in the crystal with local 222-point symmetry. The symmetry of the molecules in the crystal does not itself imply anything about the state of aggregation in solution. In principle, the protein could exist as individual monomers, as dimers with 2-fold symmetry, or as tetramers. However, the specific interactions which we observed between the four monomers (see below) suggests that cro may exist in solution as a tetramer. The extent to which the molecular arrangement may depart from exact 222 symmetry, and the potential biological relevance of such deviations from exact symmetry, remain to be determined.

Most, but not all, of the heavy-atom sites are related to each other by the 222 local symmetry elements. For each 2-fold axis, related pairs of sites can be superimposed within ~1.9 Å (Table 2). Although the 2-fold symmetry was used as an additional check in the interpretation of the heavy-atom derivatives, it was not imposed on the heavy-atom coordinates.

The electron density map

An electron density map was calculated to a nominal resolution of 2.8 Å using the five heavy-atom derivatives described in Table 1 (ref. 18). Although the phase angles are well determined to about 3.5 Å resolution, their accuracy at higher angles is limited by the scattering intensity of the derivative crystals, which becomes very weak at higher resolution, and also, in part, by non-isomorphism.

To improve the accuracy of the electron density map and facilitate its initial interpretation, the density of the four monomers in the asymmetric unit was averaged (Table 2, Fig. 1). If the assumed transformations are correct and all four monomers are, in fact, identical, the averaging process should improve the accuracy of the electron density map. In practice, the averaging process clearly improved the overall 'cleanliness' of the map, but also resulted in a loss of detail and reduction in the effective resolution (Fig. 1). Presumably, application of more sophisticated averaging techniques[19] and refinement of the individual monomers will lead to a more precise description of the structure.

The averaged electron density map was displayed both as a 'mini-map' and on a scale of 2 cm = 1 Å in an optical comparator[20,21]. Several helices and a region of β-sheet were immediately apparent, and the course of the polypeptide backbone could be readily followed from one end to the other. The direction of the polypeptide chain was initially deduced from the observation that the major platinum binding site was at a position 12 residues from one terminus, obviously corresponding to Met 12. From this point the known amino acid sequence of the protein[9,10] could be compared with the electron density map. Although the interpretation of the electron density map is fully consistent with the sequence, note that the limited resolution of the current map, together with the averaging process described above, results in poorly defined density for the respective side chains. The backbone density is strong and unambiguous from the amino terminus to residue Lys 62, but at this point the density becomes very weak, indicating that the four carboxy-terminal residues are disordered in the crystals. This appears to be the case for each of the four cro monomers.

Conformation of the cro repressor

The conformation of a cro monomer is illustrated in Fig. 2. The structure is very simple, consisting of three strands of antiparallel β-sheet (residues 2–6, 39–45 and 48–55), and three α-helices (residues 7–14, 15–23 and 27–36). As described below, the C-terminal residues form an extended 'arm' (Fig. 2) which interacts with another monomer and seems to participate in DNA binding.

The detailed arrangement of the four repressor monomers in the crystals is shown in Fig. 7. The respective local 2-fold symmetry axes relating the monomers are designated P, Q and R, and the monomers themselves O ('original'), A, B and C (Table 2). It has been assumed for some time that the cro

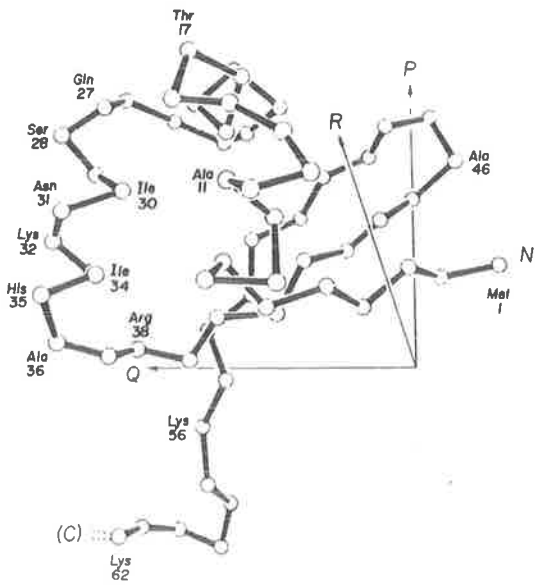


Fig. 2 Backbone of monomer O of the cro repressor protein. As discussed in the text, the four C-terminal residues 63–66 are disordered. The local orthogonal 2-fold symmetry axes P, Q and R relate the molecule shown to monomers A, B and C, respectively.

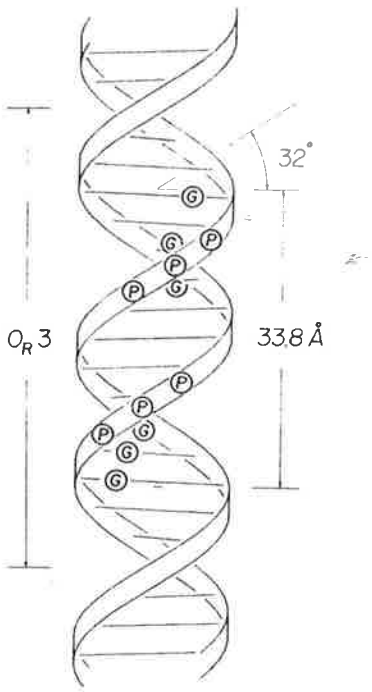


Fig. 3 DNA operator O_R3, drawn in the B form (after Ptashne *et al.*[4]). Ethylation of the phosphates, labelled P, hinders cro repressor binding; binding of cro protects the N^7 positions of guanines, labelled G, from methylation by dimethyl sulphate. The figure is based on experiments of Johnson, Ptashne and collaborators.[4,5,35].

repressor exists as a dimer in physiological conditions[8]. Therefore, we expected the four cro monomers in the crystal to consist of two dimers. However, inspection of the molecular packing does not obviously support this assumption. The monomers O and B (and, similarly A and C) are related by the Q axis, and have extensive interactions between their C-terminal arms, each of which lies against, and is, in part, hydrogen bonded to, its 2-fold-related partner. It seems likely that this interaction persists in solution. However, there are also specific interactions about the R symmetry axis, where the polypeptide backbone of one monomer hydrogen bonds to the 2-fold-related backbone of the other monomer so as to expand the three-strand β-sheet of the monomer into a six-strand antiparallel sheet in the dimer. In addition, the amino terminus of monomer O is linked by an ionic interaction to Glu 53 of monomer A, with similar interactions for the other amino termini. The four amino-terminal methionines are grouped together and participate in a distinct hydrophobic region at the centre of the cro tetramer. Thus, although the packing arrangement seen in the crystal does not prove that cro exists as a tetramer in solution, it does suggest this. Whether cro exists as a dimer or tetramer has no bearing on the proposed model for DNA binding.

Interaction of cro with DNA

Both cI and cro repressors compete for three sites within the left and the right operators of bacteriophage λ. In the right operator there are three binding sites (O_R1, O_R2 and O_R3) each consisting of 17-base pair regions separated by spacers of 6 or 7 base pairs[22]. The sequences of the three 17-base pair regions are similar to each other, but not identical, and, in addition, the sequence within each region is approximately symmetric. Mutations within the 17-base pair regions reduce the binding affinity of both cro and cI, whereas mutations in the spacer regions have no effect on binding (refs 4, 6, 23, 24 and refs therein). Chemical probe experiments[4,5,35] indicate that both cro and cI repressors bind primarily along one face of the DNA double helix and protect many of the same groups. The results of such experiments for cro binding to O_R3 are illustrated in Fig. 3. Methylation by dimethyl sulphate of the N^7 of six guanines in the major groove is prevented by the presence of cro, but methylation of N^1 of adenine, exposed in the minor groove of the DNA, is not prevented. Furthermore, ethylation of any of the six phosphates shown in Fig. 3 interferes with cro binding[4]. The pattern for cI is very similar, but is slightly more extended, and includes four additional phosphates. Such experiments suggest that both cro and cI contact the DNA in the major, but not the minor, groove[1,5]. Furthermore, it seems likely that both repressors bind to the DNA with two symmetry-related subunits of the protein in contact with the two symmetry-related halves of the operator, and the 2-fold axis of the protein coincident with the 2-fold axis of the DNA. In this way the complex of bound repressor and operator would have a common 2-fold symmetry axis. There would, of course, be slight deviations from exact 2-fold symmetry because of the inexact symmetry of the DNA base sequence, and also, possibly, due to the protein.

To determine possible modes of binding of cro to DNA, we looked for a feature of the cro monomer which, together with its 2-fold related mate, would form a structure complementary to that of the operator DNA, and which, when aligned on the DNA, would be consistent with the protection experiments summarized in Fig. 3. An obvious candidate is the helix from residues Gln 27 to Ala 36. This helix lies along the surface of the protein, with isoleucines 30 and 34 contributing to the hydrophobic interior of the monomer, but with the helix otherwise protruding from the surface of the protein (Fig. 2). The corresponding helix in monomer B, related by rotation about axis Q (Fig. 4) is parallel, and the respective helices have a centre-to-centre distance of 34 Å, the same as the separation between successive major grooves of B-form DNA (Fig. 3). Furthermore, the angle made between the axes of the two α-helices and the line connecting their midpoints is equal to the angle between the major groove of the DNA and its long axis (Figs 3, 4). The sense of the respective angles is such that the two helices can be accommodated, in a very natural way, within successive major grooves of DNA in its normal Watson–Crick B conformation[25,26].

This hypothetical arrangement was checked by fitting a model of the cro molecule to that of DNA. We required that the 2-fold axis relating the cro monomers had to coincide with the 2-fold axis through the centre of the DNA operator. Thus, the fitting of cro to the DNA is very restrictive, allowing only two degrees of freedom—one a rotation of cro relative to the DNA, about the common symmetry axis; the second a translation of cro along the symmetry axis, which varies the distance between cro and the long axis of the DNA. The results of this first, albeit crude, fitting process are shown in Figs 5, 6 and 7.

The complementarity between protein and DNA is striking. In addition to the overall correspondence between the respective backbones of the DNA and protein, residues Ser 28, Ala 29, Asn 31, Lys 32, Ala 33 and His 35 of the above-mentioned α-helix are in a position to interact either with the exposed hydrogen bonding groups of the major groove, conferring specificity, or with the negatively charged phosphates, promoting generalized binding. Nearby residues, including Arg 38 and Lys 39, are also suitably located to contribute to binding and recognition. Close to the 2-fold symmetry axis, the sequences Glu 54–Val 55–Lys 56 of the two monomers form a pair of antiparallel β-sheet strands parallel to the minor groove which possibly interact with the DNA in the manner proposed

Table 2 Relations between the four cro monomers

Symmetry axis	Pairs of heavy-atom sites	r.m.s. error in heavy-atom sites	Monomers related	Electron density correlation
P	26	1.80	O,A	0.58
Q	26	1.97	O,B	0.54
R	25	1.94	O,C	0.55

$C_{ij} = \sum \rho_i \rho_j / (\sum \rho_i^2 . \sum \rho_j^2)^{1/2}$ where the summations are over all values, ρ_i, of the 2.8 Å-resolution electron density map within 15 Å of the centre of the respective monomers.

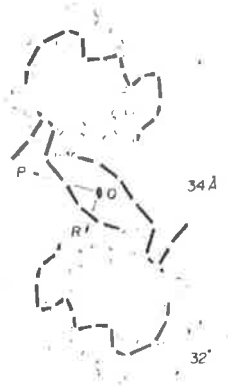


Fig. 4 Dimer of two cro molecules viewed along the Q symmetry axis. The P and R symmetry axes are also shown. Regions of the protein backbone closest to the viewer, and presumed to interact with DNA, are drawn solid. These regions include two α-helices, 34 Å apart, inclined at 32°, and also a pair of extended strands close to the Q-axis.

by Church *et al.*[27]. The four C-terminal residues of the 2-fold-related cro monomers, which are disordered in the crystals, and also, presumably, in solution, could also interact with the DNA (Figs 5, 6, 7). Such interactions could include Lys 62 and Lys 63. Presumably these C-terminal arms, which extend in solution like 'feelers', occupy well defined positions when complexed with the DNA. The finer details of the interactions between the cro repressor and its specific operator sequences are now being explored by more careful model building, and will be described elsewhere.

The proposed binding of cro to DNA explains many of the known properties of the complex. A tetramer or a dimer of cro bound to DNA extends over exactly 17 base pairs, the length of one operator (Figs 3, 5). The mode of interaction between protein and DNA is consistent with the approximate 2-fold symmetry of the operator sequence, and the symmetry of the protein. Also, the proposed binding explains, within the present level of detail, the chemical protection experiments (Figs 3, 6). Furthermore, there is a space of ~20 Å between cro molecules bound to adjacent operators, consistent with non-cooperativity of binding[28].

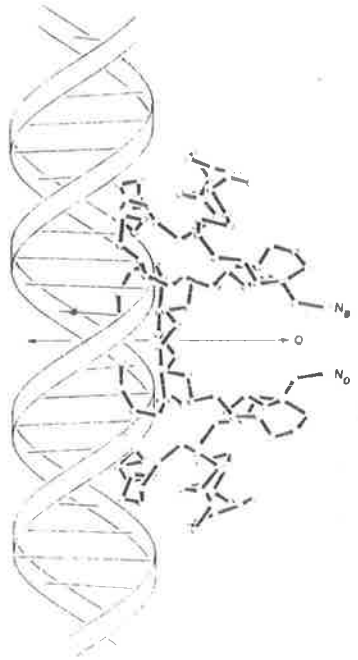


Fig. 5 Presumed interaction of cro repressor with DNA. Two monomers of cro (O and B), related by 2-fold symmetry axis Q, interact with the DNA so that axis Q coincides with the 2-fold symmetry axis of the DNA. The DNA has been rotated 90° relative to that in Fig. 3 so that its 2-fold symmetry axis is in the plane of the paper. The respective amino termini of the two cro molecules are labelled N_O and N_B. A pair of 2-fold-related α-helices occupy successive major grooves of the DNA and, closer to the symmetry axis, two extended polypeptide strands run parallel to the backbone of the DNA.

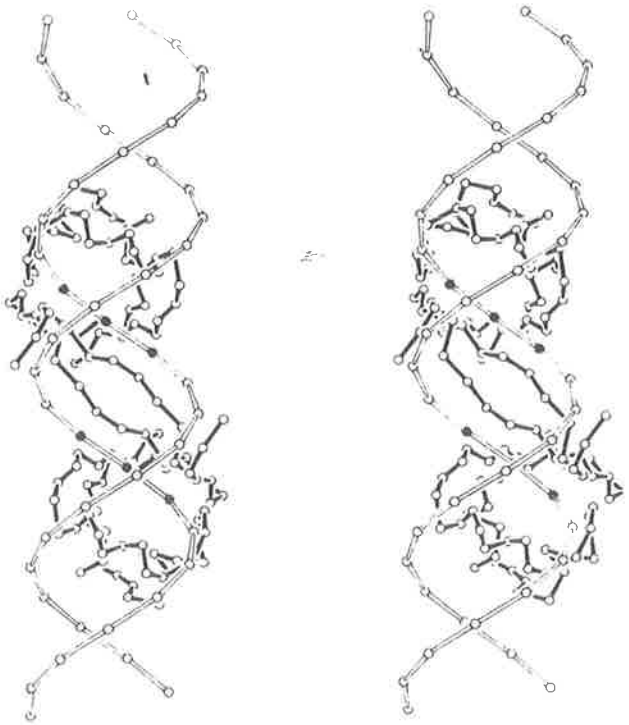


Fig. 6 Stereo view of the presumed interaction between two molecules of cro repressor and DNA, viewed along their common 2-fold symmetry axis. The open circles along the DNA backbone indicate the positions of the phosphates; those phosphates whose ethylation hinders cro binding (Fig. 3) are drawn solid.

Conclusions

This is the first structure determination of a specific DNA repressor, and the results obtained here may be of general relevance for interactions between proteins and DNA.

A principal feature of the proposed binding of cro to DNA is the use of symmetry. The 2-fold symmetry axis of the repressor coincides with the 2-fold symmetry axis of the DNA operator. This type of interaction is expected to be of general significance for proteins which recognize symmetric, sequence-specific regions of the DNA. Departures from exact symmetry could be important in 'fine tuning' of the strength of the DNA-protein complex. Our results indicate that the symmetric sequences often used for specific DNA–protein interactions are a consequence of the oligomeric structure of the regulatory proteins rather than the source of a special type of DNA structure.

Also, the α-helical conformation of proteins may presumably be used in both specific and nonspecific DNA binding. Sung and Dixon[29] were the first to propose that the histone H4 N-terminus adopts a helical conformation when interacting with chromatin nucleic acid. A related model has been proposed for *lac* repressor[30]. Also, Warrant and Kim[31] have presented direct evidence showing that protamine changes to an α-helical conformation on binding to a double-helical portion of crystalline tRNA. The cro repressor structure supports the notion that the binding of α-helices in the major groove of DNA may be a principal determinant in nonspecific protein binding to DNA, and is also of importance in specific recognition.

The structure of cro, and its complementarity to the B-form of DNA, suggest that the DNA retains the right-handed B-conformation when the repressor is bound. Of course, slight deformations of the DNA or protein could well occur on binding, and may be suggested by detailed model-building experiments, but the present results indicate that large adjustments in the DNA structure are not required to accommodate the protein. In contrast, McKay and Steitz[32] have recently proposed that the cyclic AMP receptor protein of *Escherichia coli*, which promotes the action of RNA polymerase, binds to left-handed DNA. In the case of the cro repressor, the protein is located

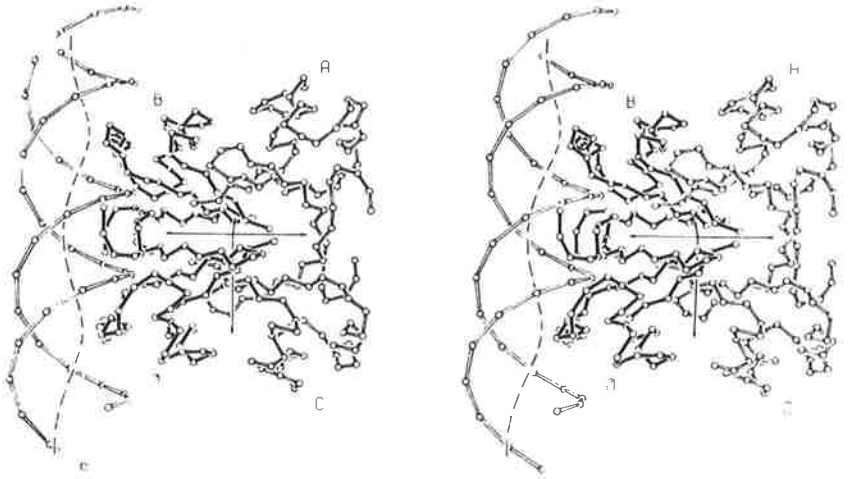


Fig. 7 Stereo view of the presumed binding of a tetramer of cro to DNA. Monomers O and B, seen interacting with the DNA, are drawn with solid bonds, whereas monomers A and C which form a second potential DNA binding region on the opposite side of the tetramer, are drawn with open bonds. The broken line follows the 'bottom' of the major groove of the DNA. The DNA and protein have been rotated 30° about the horizontal axis so that the direction of view is essentially along the major groove of the DNA.

against one side of the DNA, with most of the specific recognition occurring between side chains of the protein and the parts of the bases that are exposed in the major groove of the DNA. There is no apparent need to expose the bases further to facilitate recognition.

We thank Drs Howard Schachman and Harrison Echols for their help in instigating this project. We also thank our colleagues, in particular D. Bacon, R. G. Fisher, I. Herskowitz, M. F. Schmid, G. R. Smith, L. F. Ten Eyck and L. H. Weaver, for assistance and advice. This work was supported in part by grants from the NSF (PCM-8014311), the NIH (GM 20066; GM 28138), the MRC of Canada and the M. J. Murdock Charitable Trust. D.H.Ō. was supported by a Damon Runyan–Walter Winchell Cancer Fund Postdoctoral Fellowship (DRG-303-F).

Received 10 December 1980; accepted 18 February 1981.

1. Echols, H. *A. Rev. Genet.* **6**, 157–190 (1972).
2. Herskowitz, I. *A. Rev. Genet.* **7**, 289–323 (1973).
3. Ptashne, M. *et al. Science* **194**, 156–161 (1976).
4. Ptashne, M. *et al. Cell* **19**, 1–11 (1980).
5. Johnson, A., Meyer, B. J. & Ptashne, M. *Proc. natn. Acad. Sci. U.S.A.* **75**, 1783–1787 (1978).
6. Takeda, Y. *J. molec. Biol.* **127**, 177–191 (1979).
7. Folkamanis, A., Takeda, Y., Simuth, J., Gussin, G. & Echols, H. *Proc. natn. Acad. Sci. U.S.A.* **73**, 2249–2253 (1976).
8. Takeda, Y., Folkmanis, A. & Echols, H. *J. biol. Chem.* **252**, 6177–6183 (1977).
9. Hsiang, M. W., Cole, R. D., Takeda, Y. & Echols, H. *Nature* **270**, 275–277 (1977).
10. Roberts, T. M., Shimatake, H., Brady, C. & Rosenburg, M. *Nature* **270**, 274–275 (1977).
11. Green, D. W., Ingram, V. M. & Perutz, M. F. *Proc. R. Soc.* **A225**, 287–307 (1954).
12. Matthews, B. W. in *The Proteins* Vol. 3, 3rd edn (eds Neurath, H. & Hill, R. L.) 403–590 (Academic, New York, 1977).
13. Blundell, T. L. & Johnson, L. N. *Protein Crystallography* (Academic, London, 1976).
14. Anderson, W. F., Takeda, Y., Echols, H. & Matthews, B. W. *J. molec. Biol.* **130**, 507–510 (1979).
15. Anderson, W. F., Takeda, Y., Ohlendorf, D. H. & Matthews, B. W. *Proc. 7th Katzir-Katchalsky Conf. on Structural Aspects of Recognition and Assembly in Biological Macromolecules* March 1980 (in the press).
16. Schmid, M. F. *et al. Acta crystallogr.* (in the press).
17. Molnar, C. E., Barry, C. D. & Rosenberger, F. U. *Tech. Memo. No. 229* (Computer Systems Laboratory, Washington University, St. Louis, 1976).
18. Blow, D. M. & Crick, F. H. C. *Acta crystallogr.* **12**, 794–802 (1959).
19. Bricogne, G. *Acta crystallogr.* **A30**, 395–405 (1974).
20. Richards, F. M. *J. molec. Biol.* **37**, 225–230 (1968).
21. Colman, P. M., Jansonius, J. N. & Matthews, B. W. *J. molec. Biol.* **70**, 701–724 (1972).
22. Maniatis, T. *et al. Cell* **5**, 109–113 (1975).
23. Flashman, S. M. *Molec. gen. Genet.* **166**, 61–73 (1978).
24. Meyer, B. J., Maurer, R. & Ptashne, M. *J. molec. Biol.* **139**, 163–194 (1980).
25. Watson, J. D. & Crick, F. H. C. *Nature* **171**, 737 (1953).
26. Arnott, S. & Hukins, D. W. L. *Biochem. biophys. Res. Commun.* **47**, 1504–1509 (1972).
27. Church, G. M., Sussman, J. L. & Kim, S.-H. *Proc. natn. Acad. Sci. U.S.A.* **74**, 1458–1462 (1977).
28. Johnson, A., Meyer, B. J. & Ptashne, M. *Proc. natn. Acad. Sci. U.S.A.* **76**, 5061–5065 (1979).
29. Sung, M. T. & Dixon, G. H. *Proc. natn. Acad. Sci. U.S.A.* **67**, 1616–1623 (1970).
30. Adler, K. *et al. Nature* **237**, 322–327 (1972).
31. Warrant, R. W. & Kim, S. H. *Nature* **271**, 130–135 (1978).
32. McKay, D. & Steitz, T. A. *Nature* **(this issue)**.
33. Herskowitz, I. & Hagen, D. *A. Rev. Genet.* **14**, 399–445 (1980).
34. Echols, H. in *The Molecular Genetics of Development* (eds Loomis, W. & Leighton, T.) 1–16 (Academic, New York, 1980).
35. Johnson, A. thesis. Harvard Univ. (1980); and manuscript in preparation.

Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 1428–1432, March 1982
Biochemistry

Structure of the DNA-binding region of *lac* repressor inferred from its homology with *cro* repressor

(gene regulation/DNA–protein interaction/amino acid sequence/protein structure)

B. W. MATTHEWS*, D. H. OHLENDORF*, W. F. ANDERSON†, AND Y. TAKEDA‡

*Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oregon 97403; †Medical Research Council Group on Protein Structure and Function, Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 247, Canada; and ‡Chemistry Department, University of Maryland, Baltimore County, Maryland 21228

Communicated by Peter H. von Hippel, November 9, 1981

ABSTRACT **It is shown that the amino acid sequence and the DNA gene sequence of the 25 amino-terminal residues of the *lac* repressor protein of *Escherichia coli* are homologous with the sequences of five DNA-binding proteins: the *cro* repressor proteins from phage λ and phage 434, the *c*I and *c*II proteins from phage λ, and the repressor protein from *Salmonella* phage P22. The region of homology between *lac* repressor and the other proteins coincides with the principal DNA-binding region of *cro* repressor. In particular, residues Tyr-17 through Gln-26 of *lac* repressor correspond to the α-helix Gln-27 through Ala-36 of *cro* repressor, which we have postulated to bind within the major groove of the DNA and to be primarily responsible for the recognition of the DNA operator region by the protein [Anderson, W. F., Ohlendorf, D. H., Takeda, Y. & Matthews, B. W. (1981) *Nature (London)* 290, 754–758]. By analogy with *cro* repressor, we propose that residues 17–26 of *lac* repressor are α-helical and that this helix and a twofold-related α-helix in an adjacent subunit bind within successive major grooves of the *lac* operator, which is in a right-handed Watson–Crick B-DNA conformation. Also, by analogy with *cro* repressor, we suggest that residues Thr-5 through Ala-13 of *lac* repressor form a second α-helix and contribute, in part, to DNA binding. The proposed structure for the DNA-binding region of *lac* repressor is consistent with chemical protection data and with genetic experiments identifying the probable locations of a number of the residues of the repressor protein that either do or do not participate in DNA binding.**

The way in which certain proteins recognize specific base sequences within double-stranded DNA has been a central problem in molecular biology. From a biochemical and genetic point of view the best-characterized example of such an interaction is the *lac* repressor–*lac* operator system of *Escherichia coli* (1–3), but the lack of good crystals has prevented the determination of the three-dimensional structure of the *lac* repressor protein. Recently, we determined the structure of the *cro* repressor ("cro") from bacteriophage λ and proposed a model for its interaction with DNA (4). We have also shown on the basis of amino acid sequence and DNA gene sequence comparisons that the helical DNA-binding domain of cro probably occurs in a number of other DNA-binding proteins, including *cro* protein from bacteriophage 434, *c*I and *c*II proteins from bacteriophage λ, and the *Salmonella* phage P22 repressor protein (5). Here we suggest that the DNA-binding region of *lac* repressor also includes an α-helical region that interacts with its DNA operator in a manner similar to that proposed for cro. The suggested mode of interaction of *lac* repressor with *lac* operator is consistent with a variety of biochemical and genetic evidence and is also strikingly similar to that proposed previously by Adler *et al.* (6).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Comparison of DNA-binding proteins

Fig. 1 shows a comparison of the amino-terminal sequences of a series of proteins that bind to sequence-specific regions of double-stranded DNA (5). cro and 434-cro are small repressor proteins from bacteriophage λ and bacteriophage 434, respectively (10–13). "cI" (often referred to as λ repressor) and "P22" are larger repressor proteins from phage λ and from *Salmonella* phage P22, respectively, that, under different circumstances, can mediate positive or negative control of gene transcription (14–18). "cII" is also larger than cro and, in conjunction with another protein (cIII), acts as a positive regulator of transcription in bacteriophage λ (16, 17, 19, 20). With the exception of cro and cI, these five proteins all recognize different sequences on the DNA.

As can be seen in Fig. 1, there is extensive amino acid sequence homology between the five DNA-binding proteins. The correspondence between the five proteins can also be seen in the DNA gene sequences that code for the respective polypeptides, and, on the basis of this sequence homology, we have argued (5) that these proteins all have in common a region of tertiary structure corresponding to the segments labeled α_1, α_2, and α_3. In the cro structure, α_1 and α_2 are "structural" α-helices, whereas α_3 is the "DNA recognition" α-helix, which we have postulated to lie within the major groove of B-form DNA and to be primarily responsible for the specific recognition of the DNA by the protein (4).

A series of comparisons of both the amino acid sequence (21, 22) and the DNA coding sequence (23) indicates that the *lac* repressor protein from *E. coli* ("lac") may also have structural features in common with the other DNA-binding proteins. In Fig. 1 we have included the first 38 amino acids of lac, aligned to maximize the homology with the other proteins. The homology is most striking in the region 19–32 of cro (9–22 of lac), where Gly-24 is invariant and Ala-20 and Val-25 of cro, which are conserved in four of the five proteins, also occur in lac.

The homology between lac and the other proteins can also be seen at the level of the DNA sequences that code for the respective polypeptides. Table 1 summarizes comparisons of the DNA coding sequences for the 60 bases (i.e., 20 amino acids) corresponding to Met-12 to Asn-31 of cro (or Met-1 to Ser-21 of lac). Each entry in the table gives the fraction of the bases that are common for a given pair of proteins. If the respective DNA sequences were unrelated, then this ratio would be expected to equal approximately 0.25 (one base in four in common; each of the four bases occurs with approximately equal frequency in each of the gene sequences being compared). As can be seen, every entry in the table exceeds the random value, and in some cases more than 50% of the bases are the same. The

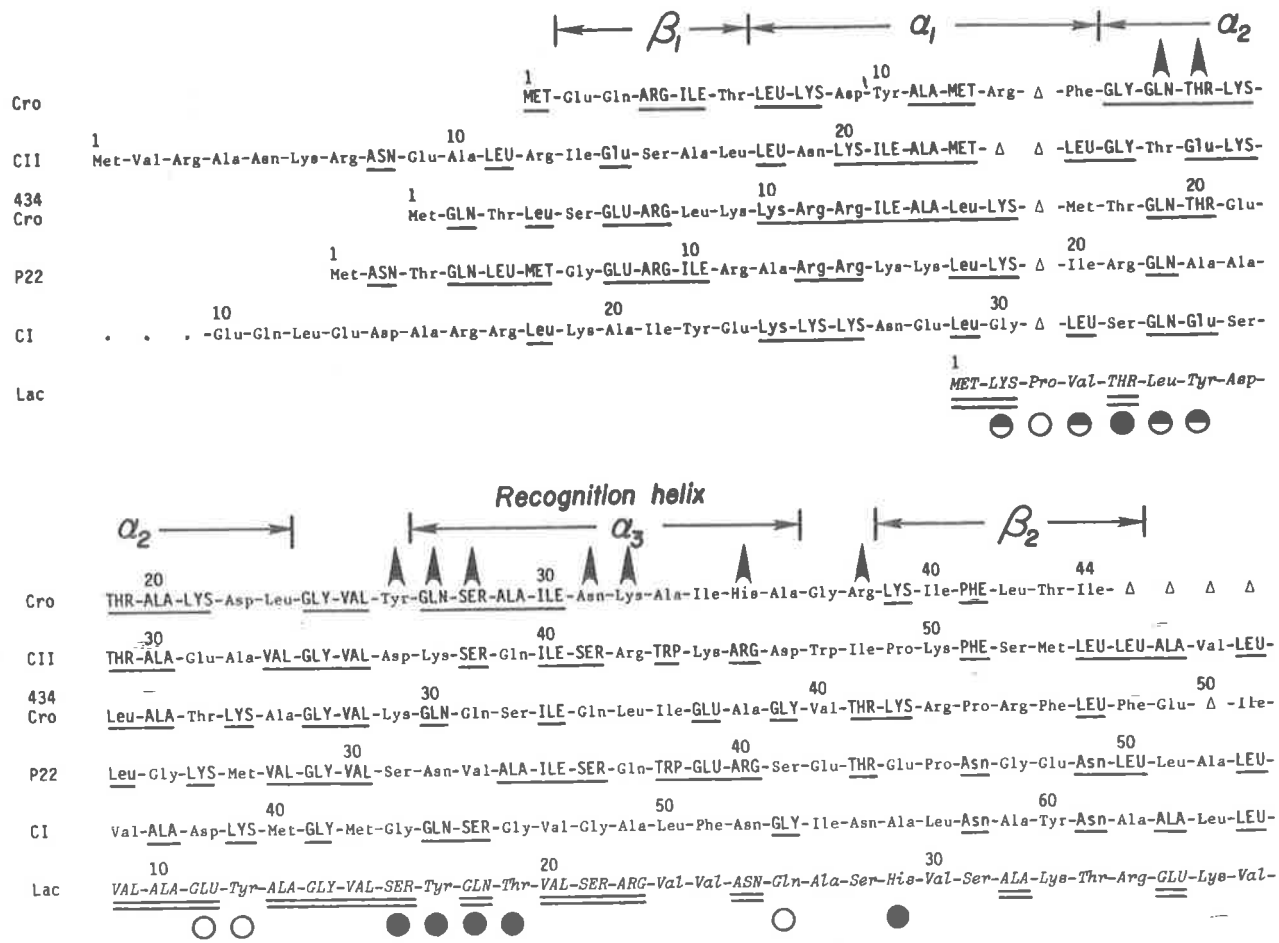


FIG. 1. Comparison of the amino-terminal amino acid sequence of *lac* repressor with five other DNA-binding proteins. Residues that are homologous within the five DNA-binding proteins are indicated by different typeface and a single underline. Residues of *lac* repressor that are common to one or more of the five proteins are capitalized and have a double underline. The letters α and β show the locations of the α-helices and β-sheet strands of *cro* protein. Residues of cro that are presumed, from model building, to interact with DNA are capped by an arrowhead. Δ indicates an assumed deletion. The solid circles underneath the lac sequence indicate those locations where known mutations dramatically reduce DNA-binding ability but do not interfere with inducer binding (i.e., they do not simply destabilize the whole protein). (Phenotypically similar sites of substitution also occur at positions 45, 47, and 54 but cannot be shown in Fig. 1.) The half-filled circles indicate locations where amino acid substitutions may reduce DNA-binding affinity or where the reduction in binding is weak. Open circles indicate sites where substitution does not interfere with DNA binding. (Data on mutants taken from refs. 2, 3, and 6–9).

table also includes the average value for each protein compared, in turn, with the other five. Somewhat surprisingly, the average agreement ratio for lac is the second highest in the series.

Significance of the sequence homology

It might be asked whether the amino acid homology between *lac* repressor and the other proteins listed in Fig. 1 is sufficiently good to provide convincing evidence that *lac* repressor is, in fact, related to the other proteins. In estimating the significance of the homology, the essential factor that has to be taken into account is that the first five sequences are already aligned among themselves so that the *lac* sequence is aligned *simultaneously* with each of the other five sequences. It is this simultaneous correspondence that increases the overall significance of the homology. For example, if one compares the first 22 residues of *lac* repressor with *cro* repressor alone, there are 4 equivalences (Fig. 1). A frequency of 4 in 22 is above the chance value of about 1 in 15 for a typical amino acid composition (24) but, on its own, is not unusual (significance of 2.2σ). However, in the case of simultaneous agreement between one protein and a number of others, it is necessary to sum all the equivalences between the one protein and each of the other proteins (24, 25). For the first 22 residues of *lac* repressor there are 110 (i.e., 22×5) possible matches with the other proteins, and, of these, 27 residues are identical (Fig. 1). An agreement of 27 amino acid residues out of 110 (significance of 7.5σ) is very unlikely to have occurred by chance and is usually taken as clear-cut evidence for significant sequence homology (24).

In addition to using the amino acid sequence homology, we also attempted to evaluate the significance of the homology between the gene sequences of *lac* repressor and the other proteins. To see whether the agreement ratios between lac and the other proteins listed in Table 1 are better than for other alignments, we took the first 21 amino acids of lac, translocated them to the left or the right of the alignment shown in Fig. 1, and calculated the number of corresponding bases. Table 1 summarizes the results for the ten cases in which the gene sequence beginning at Met-1 of lac was aligned, in turn, with gene sequences beginning at Leu-7 through Gln-17 of cro (and the corresponding alignments for the other proteins). In each case, the average base agreement ratio for the 10 possible alignments is close to the expected value of 0.25. The best agreement ratios for the 10 possible alignments are, with one exception (lac vs. cII), substantially poorer than the ratios obtained for the alignment of lac as in Fig. 1. The high agreement ratios for lac vs. 434-cro and lac vs. P22 repressor (52% and 45% of bases in com-

Table 1. DNA sequence comparisons for the genes of six DNA-binding proteins*

	cro	cII	434-cro	P22	cI	lac
	Sequences aligned as in Fig. 1					
cro	—	0.60	0.43	0.42	0.32	0.40
cII	0.60	—	0.28	0.35	0.35	0.37
434-cro	0.43	0.28	—	0.45	0.35	0.52
P22	0.42	0.35	0.45	—	0.37	0.45
cI	0.32	0.35	0.35	0.37	—	0.38
lac	0.40	0.37	0.52	0.45	0.38	—
Average ratio	0.43	0.39	0.41	0.41	0.35	0.42
	lac sequence aligned in 10 different ways (see text)					
Average of 10 ratios	0.24	0.29	0.24	0.25	0.25	—
Highest ratio	0.35	0.41	0.35	0.35	0.30	—

* The comparisons are made for the 60 bases corresponding to Met-1 to Ser-21 of *lac* repressor. Each entry gives the fraction of the 60 bases that are common for a given pair of proteins.

mon) relative to the corresponding highest control values (about 35%) are particularly noteworthy. Furthermore, it has to be emphasized that the "average ratios" in Table 1 were obtained for *lac* repressor aligned simultaneously with the five other DNA-binding proteins, whereas the "highest ratios" were selected from the 10 possible alignments of lac with each of the individual proteins in turn.

In another test we took the gene sequence for the whole *lac* repressor molecule and calculated the base agreement ratio as defined in Table 1 for every possible 60-base segment of lac aligned with the other five proteins. For these 1018 alignments, the average fraction of the 300 (i.e., 60 × 5) possible bases in common between lac and the other five proteins was 0.249 (compare the expected value of 0.25), the standard deviation, σ, was 0.036 and the highest value of 0.424 was for the alignment of the lac sequence shown in Fig. 1. The significance of the highest value is 4.9σ. Thus, this method of evaluating the significance of the sequence homology, which is essentially that of Fitch (26), also suggests that the proposed homology between *lac* repressor and the other proteins is very unlikely to be due to chance.

Discussion

Taken together, the amino acid homology and the gene sequence homology between *lac* repressor and the other five DNA-binding proteins strongly suggest that at least a part of *lac* repressor evolved from the same precursor as did the other proteins. Because of the apparent sequence homology, we postulate that the structure of the amino terminus of *lac* repressor is similar to that of residues 11–40 of cro. In particular, we suggest that residues Tyr-17 to Gln-26 of lac form an α-helix that binds in the major groove of right-handed B-form DNA and is primarily responsible for the recognition of the specific operator sequence by the repressor. The proposed conformation of the 30 amino-terminal residues of *lac* repressor and their interaction with operator DNA are shown in Figs. 2 and 3.

Convincing sequence homology between two proteins provides compelling evidence that they have similar three-dimensional structures (30), but it has to be noted that the apparent amino acid and DNA sequence correspondence for lac is restricted to a stretch of about 25 amino acid residues, so that the significance of the implied structural homology might be questioned. Clearly one has to ask if the postulated structural correspondence between lac and the other proteins (cro in particular) is consistent with the known properties of *lac* repressor.

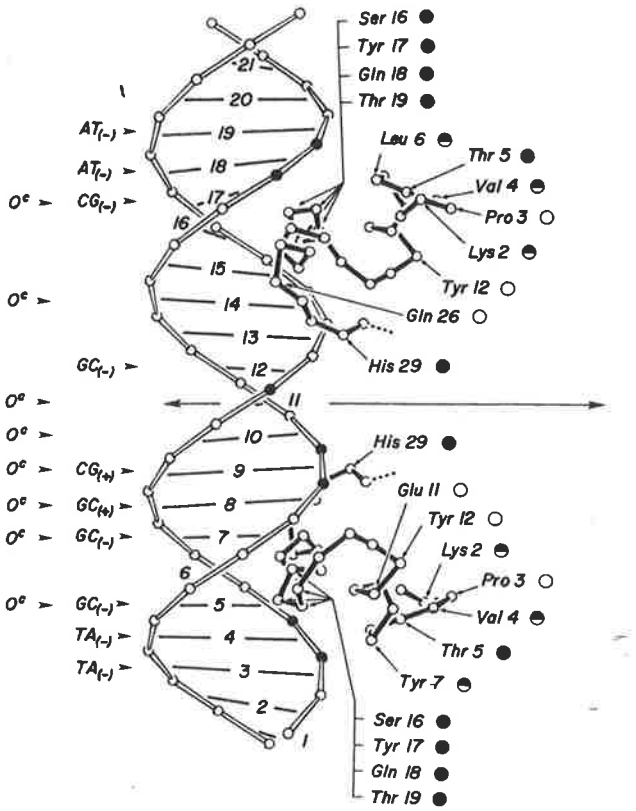


FIG. 2. Diagram showing the proposed interaction of a pair of two-fold-related DNA-binding regions of *lac* repressor with operator DNA. The twofold axis that relates the subunits of the repressor also passes through the central base pair (no. 11) of the operator. The backbone of the 30 amino-terminal residues of *lac* repressor (shown solid) is drawn as if it were identical with the backbone of residues 11–40 of *cro* repressor, although it is likely that the 5 or 6 amino-terminal residues of lac interact more closely with the DNA than is suggested by the figure (see text). Solid circles and half-solid circles indicate residues that are thought, on the basis of known mutations at these locations, to potentially interact with the DNA. Open circles show residues that are presumed, on the basis of genetic evidence, not to interact with the DNA (see caption to Fig. 1). Base pairs given at the left are those for which binding of the repressor modifies the availability of the purine to methylation by dimethyl sulfate (27). G·C or C·G methylation occurs in the major groove, whereas A·T or T·A methylation occurs in the minor groove. A minus indicates that *lac* repressor binding inhibits methylation; a plus indicates enhanced methylation in the presence of repressor. Also shown at the left are the locations of known O^c operator mutations, which reduce the affinity of the repressor for the operator (data from ref. 28). The phosphates of the DNA drawn solid are those for which alkylation with ethylnitrosourea interferes with repressor binding (data from W. Gilbert and A. Maxam, reported in ref. 29).

Both biochemical and genetic evidence shows that most, if not all, of the DNA-binding region of lac resides within the 60 amino-terminal residues of the protein (2, 3, 6, 31, 32). Under certain conditions, tryptic cleavage of the intact tetrameric repressor yields four amino-terminal "headpieces" of 51 or 59 residues and a tetrameric "core" with full inducer binding activity (31, 32). These headpieces bind nucleic acids, interact with the *lac* operator (33), and protect the same set of bases against methylation as does the intact repressor (34).

Predictions of the secondary structure of *lac* repressor give conflicting results (6–8, 35–38), some of which are in agreement with the proposed conformation.

A series of genetic analyses has pinpointed a number of residues of lac that may be directly involved with DNA binding

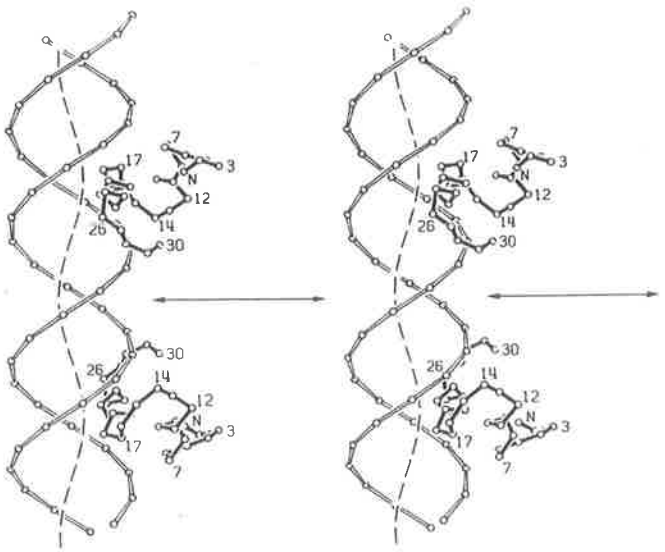


FIG. 3. Stereo drawing showing the proposed conformation of the 30 amino-terminal (N) residues of *lac* repressor (drawn solid) and their interaction with operator DNA (cf. Fig. 2). The horizontal line shows the location of the twofold axis that passes through the center of the operator and also relates the repressor subunits. The broken line extending along the DNA shows the location of the bottom of the major groove.

(2, 3, 6–9). In Fig. 1 the solid and half-filled circles below the lac sequence indicate residues that probably interact directly with the DNA or whose substitution modulates DNA binding. These sites can be compared with the locations of the residues of cro that we believe, from model building, to interact with DNA. [The model-building experiments are based on a model of cro partially refined to a crystallographic residual of 27% at 2.2-Å resolution (unpublished results).] The locations of the putative DNA-binding residues are also shown in Figs. 2 and 3. As can be seen, each of these residues is in a position where it could interact with the DNA, especially if one makes the reasonable assumption that the five or six amino-terminal residues undergo a slight conformational adjustment so as to lie against the DNA. Residues Ser-16 to Thr-19 of lac, which were presumed on the basis of the genetic studies to be critically involved in DNA binding, are located at the beginning of the postulated "recognition helix" (Figs. 1–3). Residues Thr-5 to Tyr-7 of lac are aligned with the DNA-binding region Gly-15 to Thr-17 of cro. Also, His-29 of lac closely corresponds to Arg-38 to Lys-39 of cro. Residues Glu-11 and Tyr-12 of lac, which were thought not to be directly involved with DNA binding, are located away from the DNA. Gln-26 of lac is at the end of the presumed recognition helix but is directed away from the DNA.

In 1972, Müller-Hill and coworkers (6) proposed that the DNA-binding region of lac consisted of a protruding α-helix that extended from amino acids 17 to 33. Our analysis suggests that, in the main, these workers (6) were very close to the truth, even though their ideas concerning the role of the four subunits of lac need modification. The structure we propose for the DNA-binding region of *lac* repressor is not consistent with models in which the operator-binding region adopts an extended β-strand conformation (35, 39) or a left-handed δ-helix (40, 41). However, as can be seen in Fig. 3, the proposed structure is consistent with the recent nuclear magnetic resonance study of Arndt *et al.* (42) indicating that Tyr-7 and Tyr-17 are very close to each other, possibly ring stacked, and that Tyr-12 and His-29 are at or near the surface of the protein.

On the basis of the knowledge of the cro structure we suggest that the *lac* repressor binds to its operator with a twofold symmetry axis of the repressor coincident with the twofold axis of the DNA operator (4, 43). Probably one pair of twofold-related subunits of the repressor tetramer makes contact with one operator (43–48) and, in addition, the second pair of twofold-related subunits constitutes a second operator-binding site, explaining the observed stoichiometry of two operators bound per repressor tetramer (49, 50). We postulate that a pair of twofold-related α-helices lies within successive major grooves of the DNA, as is the case for the cro protein (4). Recognition of a specific DNA sequence is mediated primarily by a complementary set of hydrogen bonds between the side chains of the protein and the parts of the base pairs that are accessible in the major grooves of the DNA. The action of inducer in reducing the affinity of *lac* repressor for its operator DNA is most easily envisaged as causing a slight conformational rearrangement of the subunits so that the relative alignment of the two DNA-binding helices is perturbed. This could be a general phenomenon for allosterically regulated DNA-binding proteins. The separation of the DNA-binding function into one domain and the regulatory function into a second domain also seems to be generally true for such proteins.

On the basis of the model-building experiments with cro, the 30 or so amino-terminal residues of lac would make interactions predominantly on one face of the DNA extending to seven or eight base pairs on either side of the central dyad. This is consistent with the locations of recognition sites of *lac* operator inferred from chemical protection (27, 51), genetic (28), and modification (52) experiments. The proposed structure of *lac* repressor as drawn in Figs. 2 and 3 would not explain the outermost methylation protection/modification sites, which are seven and eight base pairs from the center of the operator and located in the minor groove of the DNA. However, we have suggested above that the five or six amino-terminal residues of lac undergo a conformational adjustment and probably interact more closely with the DNA than is suggested by Figs. 2 and 3. These interactions might extend, for the respective lac subunits, to base pairs 1 and 21 in Fig. 2 and could well account for the outermost protection/modification sites. Another possibility is that the part of the lac headpiece including residues 45 and 47, which are possibly involved with DNA binding, interacts with the outermost part of the operator. It is also possible that the region of the headpiece including Lys-33, Arg-35, and Lys-37 interacts with the central part of the operator (Figs. 2 and 3). In our model for the lac headpiece structure, Lys-2, Arg-22, and His-29 are in a position to make contact with the DNA, and we have suggested that some or all of Lys-33, Arg-35, Lys-37, and the amino terminus of the protein could interact as well. Allowing for the combined interactions in the twofold-related subunits accounts for the experimental estimation that the nonspecific interaction of *lac* repressor with DNA involves about 11 ionic interactions with the sugar phosphate backbone and also that binding of the repressor is accompanied by uptake of two protons, probably on histidines or α-amino groups (53–56).

Although the proposed model for the DNA-binding region of *lac* repressor may not include all parts of the protein that interact with the DNA, it could well account for most, if not all, of the sequence-specific interactions. As such, it can be tested by matching models of the protein with native, mutant, and modified operator sequences. Detailed model building can be used to try to understand the structural basis for the sequence specificity of the recognition of operator DNA by the *cro* repressor. Similarly, the proposed structure for the DNA-binding region of *lac* repressor will permit the building of a detailed model of a putative *lac* repressor-*lac* operator complex. The consistency of this model with the wealth of experimental data

on the *lac* system should serve as a test for the validity of the structure that we have proposed.

We thank Dr. Peter H. von Hippel for helpful discussions and comments on the manuscript. This work was supported in part by grants from the National Science Foundation (PCM-8014311), the National Institutes of Health (GM 20066; GM 28138), the Medical Research Council Group on Protein Structure and Function of Canada, and the M. J. Murdock Charitable Trust. D.H.O. was supported in part by a Damon Runyon–Walter Winchell Cancer Fund Postdoctoral Fellowship (DRG-303-F).

1. Jacob, F. & Monod, J. (1961) *J. Mol. Biol.* **3,** 318–356.
2. Sund, H. & Blauer, G., eds. (1975) *Protein-Ligand Interactions* (de Gruyter, Berlin).
3. Miller, J. H. & Reznikoff, W. S., eds. (1978) *The Operon* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
4. Anderson, W. F., Ohlendorf, D. H., Takeda, Y. & Matthews, B. W. (1981) *Nature (London)* **290,** 754–758.
5. Anderson, W. F., Ohlendorf, D. H., Takeda, Y. & Matthews, B. W. *J. Mol. Biol.*, in press.
6. Adler, K., Beyreuther, K., Fanning, E., Geisler, N., Gronenborn, B., Klemm, A., Müller-Hill, B., Pfahl, M. & Schmitz, A. (1972) *Nature (London)* **237,** 322–327.
7. Müller-Hill, B. (1975) *Prog. Biophys. Mol. Biol.* **30,** 227–252.
8. Weber, K., Platt, T., Ganem, D. & Miller, J. H. (1972) *Proc. Natl. Acad. Sci. USA* **69,** 3624–3628.
9. Miller, J. H., Coulondre, C., Hofer, M., Schmeissner, U., Sommer, H. & Schmitz, A. (1979) *J. Mol. Biol.* **131,** 191–222.
10. Folkmanis, A., Takeda, Y., Simuth, J., Gussin, G. & Echols, H. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 2249–2253.
11. Takeda, Y., Folkmanis, A. & Echols, H. (1977) *J. Biol. Chem.* **252,** 6177–6183.
12. Hsiang, M. W., Cole, R. D., Takeda, Y. & Echols, H. (1977) *Nature (London)* **270,** 274–275.
13. Grosschedl, R. & Schwarz, E. (1979) *Nucleic Acids Res.* **6,** 867–881.
14. Sauer, R. T. & Anderegg, R. (1978) *Biochemistry* **17,** 1092–1100.
15. Ptashne, M., Jeffrey, A., Johnson, A. D., Maurer, R., Meyer, B. J., Pabo, C. O., Roberts, T. M. & Sauer, R. T. (1980) *Cell* **19,** 1–11.
16. Herskowitz, I. & Hagen, D. A. (1980) *Annu. Rev. Genet.* **14,** 399–445.
17. Echols, H. (1980) in *The Molecular Genetics of Development*, eds. Loomis, W. & Leighton, T. (Academic, New York), pp. 1–16.
18. Sauer, R. T., Pan, J., Hopper, P., Hehis, K., Brown, J. & Poteete, A. R. (1981) *Biochemistry* **20,** 3591–3598.
19. Schwarz, E., Scherer, G., Hobom, G. & Kossel, H. (1978) *Nature (London)* **272,** 410–414.
20. Shimatake, H. & Rosenberg, M. (1981) *Nature (London)* **292,** 128–132.
21. Beyreuther, K., Adler, K., Geisler, N. & Klemm, A. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 3576–3580.
22. Beyreuther, K. (1978) *Nature (London)* **274,** 767.
23. Farabaugh, P. J. (1978) *Nature (London)* **274,** 765–769.
24. Doolittle, R. F. (1981) *Science* **214,** 149–159.
25. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic, New York), Vol. 3, pp. 21–132.
26. Fitch, W. M. (1966) *J. Mol. Biol.* **16,** 9–16.
27. Ogata, R. & Gilbert, W. (1979) *J. Mol. Biol.* **132,** 709–728.
28. Gilbert, W., Gralla, J., Majors, J. & Maxam, A. (1978) in *The Operon*, eds. Miller, J. H. & Reznikoff, W. S. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 193–210.
29. Barkley, M. D. & Bourgeois, S. (1978) in *The Operon*, eds. Miller, J. H. & Reznikoff, W. S. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 177–220.
30. Matthews, B. W. (1977) in *The Proteins*, eds. Neurath, H. & Hill, R. L. (Academic, New York), Vol. 3, 3rd Ed., pp. 403–590.
31. Platt, T., Files, J. G. & Weber, K. (1973) *J. Biol. Chem.* **248,** 110–121.
32. Files, J. G. & Weber, K. (1976) *J. Biol. Chem.* **251,** 3386–3391.
33. Geisler, N. & Weber, K. (1977) *Biochemistry* **16,** 938–943.
34. Ogata, R. T. & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 5851–5854.
35. Chou, P. Y., Adler, A. J. & Fasman, G. D. (1975) *J. Mol. Biol.* **96,** 29–45.
36. Bourgeois, S., Jernigan, R. L., Shousun, C. S., Kabat, E. A. & Wu, T. T. (1979) *Biopolymers* **18,** 2625–2643.
37. Lim, V. I. (1974) *J. Mol. Biol.* **88,** 857–872.
38. Lim, V. I. (1974) *J. Mol. Biol.* **88,** 873–894.
39. Gursky, G. V., Tumanyan, V. G., Zasedatelev, A. S., Zhuse, A. L., Grokhovsky, S. L. & Gottikh, B. P. (1977) in *Nucleic Acid-Protein Recognition*, ed. Vogel, H. J. (Academic, New York), pp. 189–217.
40. Chandrasekaran, R., Jardetzky, T. S. & Jardetzky, O. (1979) *FEBS Lett.* **101,** 11–14.
41. Riberio, A. A., Wemmer, D., Bray, R. P., Wade-Jardetzky, N. G. & Jardetzky, O. (1981) *Biochemistry* **20,** 818–823.
42. Arndt, K. T., Boschelli, F., Lu, P. & Miller, J. H. (1981) *Biochemistry* **20,** 6109–6118.
43. Steitz, T. A., Richmond, T. J., Wise, D. & Engelman, D. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 593–597.
44. Müller-Hill, B., Gronenborn, B., Kania, J., Schlotmann, M. & Beyreuther, K. (1977) in *Nucleic Acid-Protein Recognition*, ed. Vogel, H. J. (Academic, New York), pp. 219–236.
45. Charlier, M., Maurizot, J. C. & Zaccai, G. (1980) *Nature (London)* **286,** 423–425.
46. Geisler, N. & Weber, K. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 3103–3106.
47. Kania, J. & Brown, T. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 3529–3533.
48. McKay, D. B., Pickover, C. A. & Steitz, T. A., *J. Mol. Biol.*, in press.
49. Culard, F. & Maurizot, J. C. (1979) *Nucleic Acids Res.* **9,** 5175–5184.
50. Dunaway, M. & Matthews, K. S. (1980) *J. Biol. Chem.* **255,** 10120–10127.
51. Ogata, R. & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 5851–5854.
52. Goeddel, D. V., Yansura, D. G. & Caruthers, M. H. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 3578–3582.
53. Revzin, A. & von Hippel, P. H. (1977) *Biochemistry* **16,** 4769–4776.
54. de Haseth, P. L., Gross, C. A., Burgess, R. R. & Record, M. T., Jr. (1977) *Biochemistry* **16,** 4777–4783.
55. de Haseth, P. L., Lohman, T. M. & Record, M. T., Jr. (1977) *Biochemistry* **16,** 4783–4790.
56. Record, M. T., Jr., de Haseth, P. L. & Lohman, T. M. (1977) *Biochemistry* **16,** 4791–4796.

nature
Vol 298 No 5876 19–25 August 1982 £1.60 $4.00



SEQUENCE SPECIFICITY OF DNA BINDING PROTEIN

Reprinted from Nature, Vol. 298, No. 5876, pp. 718-723, 19 August 1982
© Macmillan Journals Ltd., 1983

The molecular basis of DNA–protein recognition inferred from the structure of cro repressor

D. H. Ohlendorf*, W. F. Anderson†, R. G. Fisher*, Y. Takeda‡ & B. W. Matthews*

*Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oregon 97403, USA
†MRC Group on Protein Structure and Function, Department of Biochemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2H7
‡Chemistry Department, University of Maryland, Baltimore County, Catonsville, Maryland 21228, USA

Recognition by cro repressor protein of its specific DNA binding sites appears to occur via multidentate hydrogen bonds between amino acid side chains of the protein and base-pair atoms in the major groove of right-handed B-form DNA. Most of the sequence-specific interactions between cro and DNA, as well as a number of sequence-independent ones, are mediated by a two-α-helical unit which appears to be common to many proteins that regulate gene expression.

THE structures of three proteins that specifically bind to DNA have recently been determined, namely, the cro repressor protein (hereafter called 'cro') from bacteriophage λ[1], the catabolite gene activator protein (CAP) from *Escherichia coli*[2] and the amino-terminal fragment of the *c*I repressor protein (cI) from bacteriophage λ[3]. It has been proposed that these proteins bind to DNA with protruding α-helices penetrating[1,2] or partly penetrating[3] the major grooves of the DNA (in the case of CAP, the DNA has been postulated to be left-handed[2]).

In this article we propose a detailed model for the complex between cro and DNA. The model is consistent with the relative binding affinities of cro for the six operator sites within bacteriophage λ at which it binds specifically, and can also be used to rationalize known changes in binding affinity caused by mutations within these sites.

Cro is one of a number of proteins that participate in the commitment to either the lytic or lysogenic pathway during the development of bacteriophage λ (for review see refs 4–6). Cro recognizes six binding sites on the bacteriophage genome, three within the left operator region (O_L1, O_L2 and O_L3) and three within the right operator region (O_R1, O_R2 and O_R3). Each binding site consists of 17 base pairs with approximate twofold symmetry. The base sequences of all six sites are similar but not identical[7].

Starting parameters

Coordinates for the DNA were derived from the fibre diffraction analysis of B-DNA by Arnott and colleagues (refs 8, 9 and S. Arnott, personal communication). Initially, the rise per base-pair, z, was set at 3.4 Å and the twist per base-pair, θ, was set at 34.6° (10.4 base-pairs per turn). These parameters are representative of values obtained by others[10–12].

The coordinates for cro were based on the previously reported structure determination[1]. The quality of the electron density map has been improved by a process of electron density averaging and phase extension[13] in which the final agreement between observed and electron-density-averaged structure factors was 22.3% at 2.8 Å resolution (R.G.F., D.H.O. and B.W.M., unpublished results). Also the cro structure has been partially refined, the coordinates used here being taken from the present stage of refinement, for which the crystallographic residual is 27% at 2.2 Å resolution (23% at 2.8 Å resolution) (D.H.O. and B.W.M., unpublished). The estimated error in the protein coordinates is about 0.4 Å.

The protein dimer was placed so that its twofold axis of symmetry coincided with that of the DNA[1], leaving two adjustable parameters, one the separation, d, between the protein and the DNA, and the other the rotation, τ, of the protein dimer relative to the DNA about the common protein-DNA symmetry axis. In addition to the above four parameters, z, θ, d and τ, adjustments were allowed in the radius of curvature, r, of the DNA. Because of the twofold symmetry of the protein–DNA complex, bending was constrained to occur only about points on the dyad axis.

Model building

Our general approach was to carry out model building on an MMS-X computer graphics facility[14] and to supplement this with energy minimization using the program EREF of Jack and Levitt[15]. Use of computer graphics permitted the screening of a large number of potential interactions between the protein and the DNA while energy minimization ensured that the stereochemistry of a proposed DNA–protein complex was reasonable.

The energy minimization procedure gives the calculated energy for the system under consideration, and, as such, provides an indication of the relative strength of a postulated cro: DNA complex. However, the energy minimization algorithm does not take account of solvent, and has other limitations which prevent its use for the reliable estimation of absolute interaction energies. Therefore, in attempting to optimize a given cro:DNA complex, we used energy minimization as a guide, but supplemented this with additional criteria such as the adherence of postulated hydrogen bonds to acceptable stereochemistry.

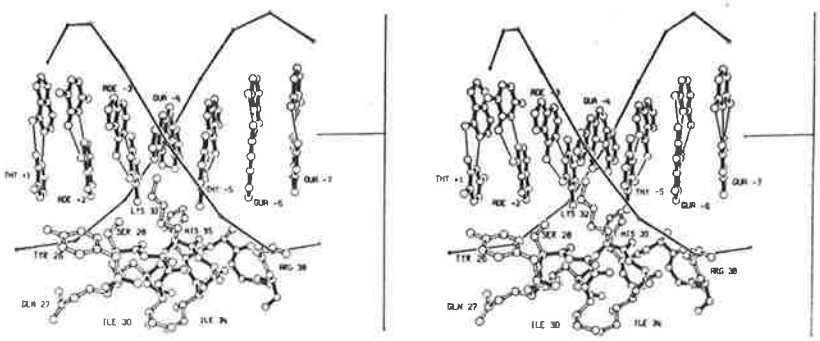


Fig. 1 Stereo drawing illustrating the initial alignment of residues 26–38 of cro repressor within the major groove of the DNA. Only the left half of the 17-base-pair binding site is shown. The vertical line at the right of the figure is the common protein and DNA twofold-symmetry axis which relates the other half of the binding site to that shown here. The horizontal line at the right coincides with the long axis of the DNA. The conformations of the individual protein side chains are those determined from the crystal structre of cro repressor protein. The base sequence is that for O_R3, with numbering as defined in the text. For clarity the details of the DNA backbone have been omitted and replaced by lines connecting the phosphates.

There has been considerable speculation that the free energy of stabilization of DNA:protein complexes comes primarily from relatively nonspecific interactions[16] while specificity derives largely from hydrogen bonds between protein side chains and the parts of the base pairs exposed in the grooves of the DNA[16–19]. Therefore, in model building we attempted to maximize the number of hydrogen bonds and other interactions between cro and the DNA, and assumed that each 'mispositioned' (that is, unsatisfied) hydrogen bond decreased the affinity of the DNA–protein complex, although it is understood that other factors such as shape complementary, hydrophobic and dipolar interactions and the role of solvent are also important. Since the binding constant, K, for the cro:DNA interaction is related to the free energy of the interaction, ΔG, by the formula $K = \exp(-\Delta G/RT)$, it follows that each improperly positioned hydrogen bond donor or acceptor, which reduces the free energy by about 1 kcal mol^{-1} (refs 16, 20), results in a reduction of the binding constant by a factor of approximately seven.

The presumed complexes between cro and DNA were initially subjected to energy minimization in which the protein backbone was kept rigid and only those side chains which appeared to interact with the DNA were allowed to move. A specialized version of EREF was used which allowed the explicit specification of hydrogen bonds. The five interaction parameters z, θ, d, τ and r (see above) were then systematically varied in small increments to determine which values gave the best value for the calculated energy of interaction. (This stepwise approach was necessary because EREF tends to make changes in local geometry, but not in the global interaction parameters.) In the final stages of the energy minimization, the local geometry of the DNA was allowed to vary, and, in addition, small adjustments were permitted in the protein backbone.

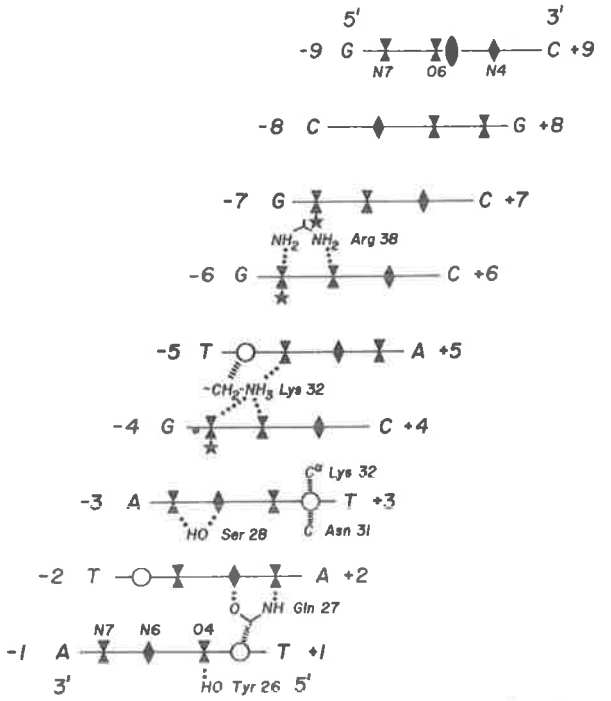


Fig. 2 Schematic representation, following Woodbury *et al.*[20], of the presumed sequence-specific interactions between cro and the parts of the base pairs exposed within the major groove of the DNA. The direction of view is imagined to be directly into the major groove of the DNA with the base-pairs seen edge-on. The dyad symbol within the topmost base pair indicates the center of the overall 17-base-pair binding region. The symbols are as follows: ⧗, hydrogen bond acceptor; ◆, hydrogen bond donor; ○, methyl group of thymine; ★, guanine N7 which is protected from methylation when cro is bound[22,23]. Presumed hydrogen bonds between cro side chains and the bases are indicated (. . . .). Apparent van der Waals contacts between cro and the thymine methyl groups are shown (||||||). Other van der Waals contacts are not shown.

Interaction of cro with DNA

The initial structure determination of cro suggested that it bound to the DNA with a pair of dyad-related helices penetrating the major groove of the DNA[1]. This overall complementarity between the surface of the protein and the DNA not only suggests the approximate alignment of the protein relative to the DNA, but (short of making major conformational changes in the protein which we shall consider below), makes it sterically difficult to devise alternative interaction schemes which seem plausible.

Since both cro and the DNA binding site have approximate twofold symmetry it is, in the first instance, sufficient to consider the interaction between one monomer of cro and half of the binding site. We have numbered the base pairs within each half-site from 1 to 9 starting at the distal base pair. Bases +1 and −1 constitute a base pair, and so on. The '+' base is the one for which the inward direction of numbering is from 5′ to 3′.

The initial alignment of cro (Fig. 1) suggested that the amino acid side chains most likely to participate in interactions within the major groove of the DNA were Tyr 26, Gln 27, Ser 28, Asn 31, Lys 32, His 35 and Arg 38. Based on the framework of the potential hydrogen bonds and the cro:DNA geometry suggested by the energy calculations, it is possible to construct a set of rules which can be used to predict the hydrogen bonds expected when cro interacts with a given sequence of DNA. The proposed rules for hydrogen bonding are as follows:

(1) Arg 38 forms two hydrogen bonds to O6 and N7 of guanine −6. If base −6 is an adenine then Arg 38 makes a hydrogen bond with either the N7 of this adenine or the N7 of a purine at −7. If base −6 is a thymine then Arg 38 can make no hydrogen bonds to this base because of the blocking methyl group on C6 but can form one hydrogen bond to N7 of a purine at −7 (that is, a net loss of one hydrogen bond).

(2) Lys 32 N^{ζ} forms one hydrogen bond with either a purine N7 or an O4 of thymine at position −5. In addition, it forms a second hydrogen bond with O6 of guanine −4 and possibly a third hydrogen bond with N7 of guanine −4.

(3) Ser 28 O^{γ} forms two hydrogen bonds with N6 and N7 of adenine −3 or, alternatively, one hydrogen bond to N7 of guanine, N4 of cytosine, or O4 of thymine at −3. In the case of a thymine at position −3 there is a second hydrogen bond from the N6 of the adenine at +3 to the carbonyl oxygen of the serine.

(4) Gln 27 forms two hydrogen bonds to N6 and N7 of an adenine at +2. In all other cases it forms a single hydrogen bond.

(5) Tyr 26 O^{η} forms a hydrogen bond with either O4 of thymine +1 or N7 of a purine at +1, but cannot be placed in a stereochemically acceptable position to accept a hydrogen bond from N6 of adenine or N4 of cytosine.

As shown in Table 1, this model for the interaction between cro and DNA is consistent with the known relative affinities of cro for its six binding sites and for mutant sites as well.

The presumed sequence-specific hydrogen-bonding between cro and O_R3 is shown in Fig. 2 using a schematic representation based on that proposed by Woodbury *et al.*[20] and is shown in stereo in Fig. 3. The apparent interactions between the protein and the DNA backbone, which are not sequence-specific, and, presumably, provide much of the overall energy of interaction, are summarized in Fig. 4. The overall conformation of cro repressor when bound to its specific binding site is illustrated in Fig. 5.

In the proposed cro:O_R3 complex, there are about 20 sequence-specific hydrogen bonds (Fig. 2) as well as many sequence-independent interactions (Fig. 4). These interactions could easily account for the observed dissociation constant of about 10^{-10} M (refs 21, 22).

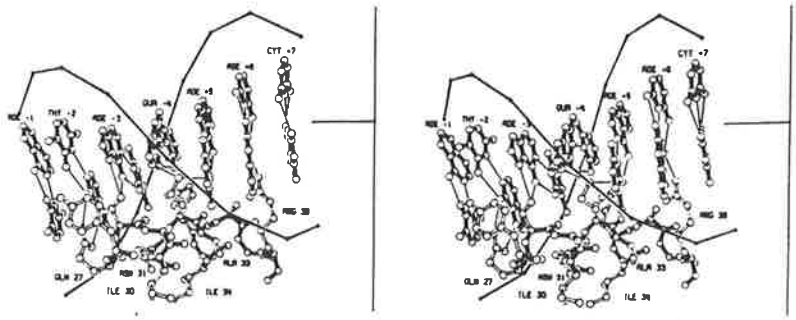


Fig. 3 Stereo drawing (Compare Fig. 1) showing the presumed hydrogen bonding of cro to base pairs in the major groove of O_R3, following model building and energy refinement.

Because cro is known to bind most tightly to O_R3 (refs 23–25) we used this for the initial model building and for most of the energy calculations. After refinement of the cro:O_R3 complex, the DNA had a mean z of 3.48 ± 0.09 Å with a range of 3.30 Å to 3.56 Å and a mean twist θ of 34.7 ± 3.2° with a range of 26.9° to 39.8°. These values compare well with previously reported values for these parameters[10,26,27].

The optimization procedure suggests that the DNA is bent to a radius of about 75 Å. This degree of curvature is not excessive, as might at first be supposed. In the crystallized B-DNA dodecamer CGCGAATTCGCG, the DNA bends to a radius of 110 Å (ref. 28). Also, when associated with proteins in the nucleosome, DNA adopts a radius of about 45 Å (ref. 29). It can be estimated that the energy required to bend a 17 base-pair segment of DNA to a radius of 75 Å is between 0.85 and 1.7 kcal mol^{-1} (refs 30, 31), an amount of energy approximately equivalent to the formation of one hydrogen bond per monomer. As an alternative to the bending of the DNA, an essentially equivalent result could be obtained by a 'hinge-bending' motion of the two cro monomers in the vicinity of the antiparallel β-sheet strands, Glu 54–Lys 56. These residues form part of the C-terminal 'arms' (residues 52–60) which appear to be the major determinant in stabilizing the cro dimer. Other than the contacts mediated by these two arms there are only very tenuous interactions between the respective monomers, suggesting that a 'hinge-bending' motion could easily occur (Fig. 5; see also Figs 3, 4 and 6 of ref. 1).

Specific and nonspecific cro:DNA interactions

The proposed model for the complex between cro and DNA suggests that interactions occur predominantly within two regions of the protein, namely residues 15–38, that is, essentially the α_2 and α_3 helices, and residues 54–66 which include part of the third β-sheet strand and the carboxyl terminus of the molecule.

The α_3 helix (residues 27–36) penetrates the major groove of the DNA and appears to contribute most of the sequence-specific interactions (Figs 2, 3). Interactions of Lys 32 and Arg 38 with bases in positions −4, −5 and −6 account for the observed protection from methylation of guanine N7 at these positions[4,23,24]. In the presumed complex between cro and O_R3 (Figs 2, 3) there is no hydrogen bonding at the N7 of guanine −7. This is, perhaps, surprising in view of the observed protection of this nitrogen from methylation. However, the close proximity of the guanidinium group of Arg 38 to guanine −7 (Fig. 3) might sterically inhibit the access of dimethyl sulphate to this base.

The α_2 helix (residues 15–23) appears to make a number of sequence-independent interactions with the outermost part of the DNA binding site (Fig. 4). Altogether, residues in the α_2 and α_3 helices may contribute as many as five hydrogen bonds to the phosphate backbone. In the model, there are no sequence-specific interactions between cro and the central three base pairs of the binding site, consistent with the observation that, for these three base-pairs, nitrogens N3 of the adenines exposed in the minor groove and N7 of the guanines exposed in the major groove are susceptible to methylation by dimethyl sulphate[4,23,24].

Residues Glu 54–Lys 56 of the third β-strand, together with the twofold-related strand in the adjacent cro monomer, are located close to the centre of the binding site and lie parallel to the minor groove, but, in part because of the bending of the DNA, do not lie within the minor groove (compare refs 1, 32). Rather, their contact with the DNA appears to be limited to interactions between the side chains and the phosphate backbone (Fig. 4). In the model for the presumed cro:DNA complex, the phosphate at position −6 (between base pairs 6 and 7 in Fig. 4) and possibly one or two solvent atoms or cations are surrounded by protein. The two neighbouring phosphates also appear to be partly 'covered' by protein (see Fig. 4), consistent with the observation that ethylation of these three phosphates interferes with the formation of the cro:DNA complex[4,23]. In the presumed cro:DNA complex, it appears that the seven C-terminal residues, including Lys 62 and Lys 63, could lie along the minor groove of the DNA with the side chains interacting with the phosphate backbone (Figs 4, 5). Since the five C-terminal residues are disordered in the crystals of the native protein[1] it has to be emphasized that the details of such interactions remain very speculative.

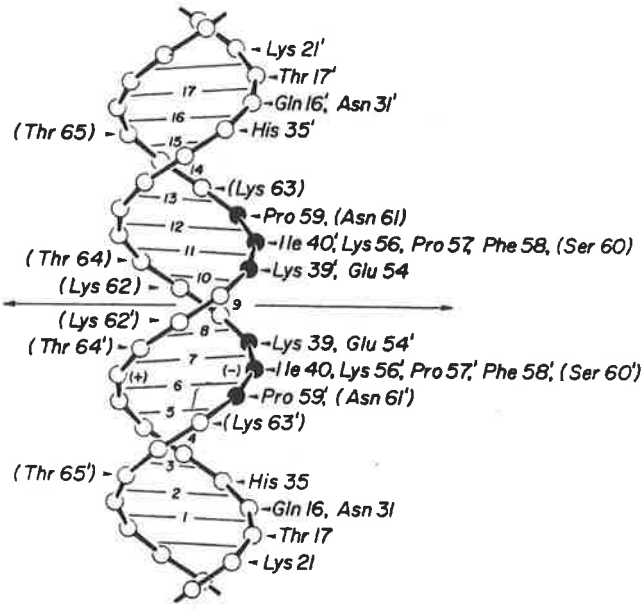


Fig. 4 Schematic representation of the contacts and interactions which are presumed to occur between cro and the phosphates in the DNA backbone. In the presumed cro: DNA complex, at least one close approach occurs between a given phosphate group and the named amino acids. Those phosphates for which ethylation interferes with cro binding[23] are drawn solid. Numbering of the base pairs and the signs (+) and (−) correspond to the identification used in the text. Amino acids belonging to 'lower' and 'upper' cro monomers related by the horizontal dyad axis are non-primed and primed, respectively. Amino acids in parentheses are those at the C-terminus of the molecule which are disordered in the crystal (residues 62–66) or for which the backbone conformation has been adjusted (residues 60, 61) to improve the contact with the DNA.

Table 1 Cro binding to different DNA sites

Site	Sequence	Mutant	Affinity relative to O_R3	Predicted H-bonds relative to O_R3
O_R3	1 5 9 13 17 T A T C A C C G C A A G G G A T A A T A G T G G C G T T C C C T A T		1	0
	G C	v3C	1/10	−1
	T A	r1	1/3	0
	T A A T	r2 r3	<1 <1	0 0
	G C C G	c12		−1
		c10	1/10	−1
O_R2	T A A C A C C G T G C G T G T T G A T T G T G G C A C G C A C A A C		1/8	−1
	G C	v3C	1/80	−2
	A T A A T T	vi or vH		−2
		virC23	1/400	−3
	T A	vN or vC34	<1/8	−3
O_R1	T A C C T C T G G C G G T G A T A A T G G A G A C C G C C A C T A T		1/8	−1
	G C	vR18		−2
	A T	vs326	1/40	−2
	A T	v3		−1
	C G T A	vs387	<1/8	−2
		vC1	1/40	−2
O_L1	T A C C A C T G G C G G T G A T A A T G G T G A C C G C C A C T A T		~1/2	−1
	G C	v101		−2
	T A A T	v2		−2
		v003		−2
O_L2	T A T C T C T G G C G G T G T T G A T A G A G A C C G C C A C A A C		~1/2	−1
	A T	v305		−1
O_L3	A A C C A T C T G C G G T G A T A T T G G T A G A C G C C A C T A T 1 5 9 13 17		~1/10	−2
Amino acid	26 Lys Arg Arg Lys 26 −28 32 38 38 32 −28			

DNA sequences and relative affinities are shown for the six 17-base pair sites to which cro binds specifically in phage λ. Mutations in these sites with known relative affinities for cro binding are also shown. (Data taken from refs 23, 24.) The relative binding affinities are to be compared with the change in the number of sequence-specific hydrogen bonds between cro and the DNA site as predicted from the model for cro binding proposed in the text. If the model is correct, and if hydrogen bonding is a major determinant in the specificity of cro: DNA recognition, then the loss of successive hydrogen bonds ought to correspond to successive reductions in the affinity of binding by about a factor of seven.

Inferences for protein–DNA recognition

The proposed complex between cro and DNA has a numnber of features which may be of general significance for recognition by proteins of specific sites on DNA.

The initial determination of the cro structure[1] revealed a pair of twofold-related α-helices arranged in such a way that they could be placed within successive major grooves of right-handed Watson–Crick B-form DNA. This result gave credence to previous suggestions that α-helices of proteins might bind within the major groove of B-DNA[33–36].

The present study strongly suggests that sequence-specific recognition of DNA is mediated by hydrogen bonds between protein side chains and the parts of the base pairs exposed within the grooves of the DNA as has been advocated by a number of authors[16,19,35–38]. In particular, it is of interest to

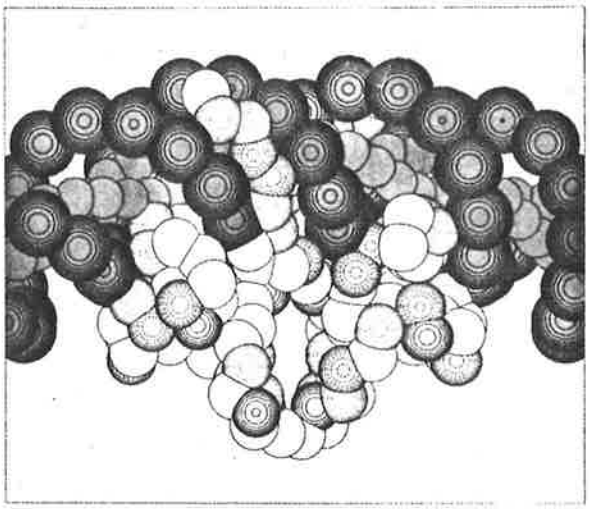


Fig. 5 Schematic drawing of the proposed sequence-specific complex of cro with DNA. For the stylized DNA (shaded) the large concentric circles indicate the positions of the phosphate groups and the smaller circles follow the bottom of the major and minor grooves. For the cro dimer, one circle is drawn for each amino acid; continuous concentric circles indicate acidic residues, broken circles indicate basic residues, dotted circles indicate non-charged hydrophilic residues and open circles show hydrophobic residues.

note that the specificity of recognition appears to be enhanced by multiple interactions between a given side chain and appropriate hydrogen bond donors and acceptors on the base pairs. The apparent bidentate interaction of Gln 27 with nitrogens N6 and N7 of adenine +2 as well as the interaction of Arg 38 with O6 and N7 of guanine −6 (Figs 2, 3) are as anticipated[18,39]. Ser 28 and Lys 32 are also presumed to participate in multiple interactions. Such interactions had not been postulated previously, although von Hippel and colleagues[16,19] have suggested that bridging interactions could occur between atoms on successive base pairs, as seems to be the case with cro, in some instances. Also Drew and Dickerson[40] have observed a solvent molecule (molecule 66 in their numbering scheme) which forms hydrogen bonds with both the N6 and N7 nitrogens of an adenine, precisely as is proposed to occur between O^{γ} of Ser 28 and adenine −3.

In addition to the above hydrogen bonding within the major groove of the DNA, there are a number of apparent van der Waals contacts which are presumably also important in sequence-specific recognition. In particular, the C7 methyl groups of thymines +1, +3 and −5 appear to interact, respectively, with C^{δ} of Gln 27, with the carbonyl carbon of Asn 31 and the α-carbon of Lys 32 and with C^{γ} and C^{ε} of Lys 32 (Fig. 2). The importance of such hydrophobic interactions in the recognition of *lac* operator by lac repressor has been shown by Caruthers[41].

Another feature of the model which may be of general relevance is the combination of sequence-specific interactions, mediated by the more rigid parts of the protein structure, together with non-specific interactions that are assumed to derive, in part, from flexibility in the cro structure. Two distinct types of flexible motion seem likely. First, the C-terminal residues (62–66) are not ordered in the crystals of the native protein, and are presumably free to move in solution. Because of this freedom of motion, and also because the C-terminal residues do not seem to be involved in sequence-specific interactions, we suggest that these residues are involved in the initial binding of the protein to the DNA. The second type of flexible motion suggested by the cro structure is the 'hinge bending' motion described above. We suggest that such flexibility may be of general importance in facilitating 'sliding' along the DNA as has been proposed to occur for lac repressor[42–47]. Inspection of the structures of CAP[2] and *c*I repressor[3] indicates that 'hinge' motion seems possible for these structures as well, suggesting that this type of flexibility may be of general significance. One could imagine that a hinge motion of the protein, as well as flexibility in the DNA, would give each of the cro monomers some freedom to 'sample' adjacent base pair atoms in the major groove of the DNA. Only at the correct twofold-symmetric binding site would the full DNA-binding specificity be satisfied for both cro subunits simultaneously. Because a dimer rather than a monomer of cro binds to DNA, the overall recognition site extends over 17 base pairs (~60 Å). However, the cro:DNA complex is presumably optimized for control rather than tightness of binding so that not all of the 17 base pairs need contact the protein and, in addition, departures from exact twofold symmetry of the base sequence could be used to 'fine tune' the strength of the cro:DNA complex.

Finally, we note that in the proposed cro:DNA complex, most of the sequence-specific interactions, as well as a number of the sequence-independent ones, come from the α_2–α_3 helical unit. It has been striking to find, on the basis of recent sequence and structure comparisons, that a similar two-helical fold appears to occur in a number of other DNA-binding proteins[48–54] indicating that the principles of specific protein–DNA recognition suggested by the proposed model may be of general significance.

We thank Alan Bender for help with model-building and Dr. P. H. von Hippel for discussions. This work was supported in part by an NIH postdoctoral fellowship to D.H.O. (GM08403), a Damon Runyon-Walter Winchell Cancer Fund postdoctoral fellowship to R.G.F. (DRG-388-FT), by the MRC of Canada through the Group on Protein Structure and Function (to W.F.A.), by NIH grants to B.W.M. (GM20066) and Y.T. (GM28138) and by grants to B.W.M. from the NSF (PCM-8014311) and the M. J. Murdock Charitable Trust.

Received 27 April; accepted 2 July 1982.

1. Anderson, W. F., Ohlendorf, D. H., Takeda, Y. & Matthews, B. W. *Nature* **290,** 754–758 (1981).
2. McKay, D. B. & Steitz, T. A. *Nature* **290,** 744–749 (1981).
3. Pabo, C. O. & Lewis, M. *Nature* **298,** 443–447 (1982).
4. Ptashne, M. *et al. Cell* **19,** 1–11 (1980).
5. Echols, H. in *The Molecular Genetics of Development* (eds Loomis, W. & Leighton, T.) 1–16 (Academic, New York, 1980).
6. Herskowitz, I. & Hagen, D. A. *Rev. Genet.* **14,** 399–445 (1980).
7. Maniatis, T. *et al. Cell* **5,** 109–113 (1975).
8. Watson, J. D. & Crick, F. H. C. *Nature* **171,** 737–738 (1953).
9. Arnott, S. & Hukins, D. W. L. *Biochem. biophys. Res. Commun.* **47,** 1504–1509 (1972).
10. Drew, H. R. *et al. Proc. natn. Acad. Sci. U.S.A.* **78,** 2179–2183 (1981).
11. Peck, L. J. & Wang, J. C. *Nature* **292,** 375–378 (1981).
12. Rhodes, D. & Klug, A. *Nature* **292,** 378–390 (1981).
13. Bricogne, G. *Acta crystallogr.* **A30,** 395–405 (1974).
14. Molnar, C. E., Barry, C. D. & Rosenberger, F. U. *Tech. Memo.* No. **229** (Computer Systems Laboratory, Washington University, St Louis, 1976).
15. Jack, A. & Levitt, M. *Acta crystallogr.* **A34,** 931–935 (1971).
16. von Hippel, P. H. in *Biological Regulation and Development* Vol. 1 (ed. Goldberger, R. F.) 279–347 (Plenum, New York, 1979).
17. Yarus, M. *A. Rev. Biochem.* **38,** 841–880 (1969).
18. Seeman, N. C., Rosenberg, J. M. & Rich, A. *Proc. natn. Acad. Sci. U.S.A.* **73,** 804–808 (1976).
19. Woodbury, C. P., Hagenbuchle & von Hippel, P. H. *J. biol. Chem.* **255,** 11534–11546 (1980).
20. Levitt, M. *J. molec. Biol.* **82,** 393–420 (1974).
21. Takeda, Y., Folkmanis, A. & Echols, H. *J. biol. Chem.* **252,** 6177–6183 (1977).
22. Johnson, A. D., Pabo, C. O. & Sauer, R. T. *Meth. Enzym.* **65,** 839–856 (1980).
23. Johnson, A. thesis, Harvard Univ. (1980).
24. Johnson, A., Meyer, B. J. & Ptashne, M. *Proc. natn. Acad. Sci. U.S.A.* **76,** 5061–5065 (1979).
25. Takeda, Y. *J. molec. Biol.* **127,** 177–191 (1979).
26. Wang, J. C. *Proc. natn. Acad. Sci. U.S.A.* **76,** 200–203 (1979).
27. Mandelkern, M., Elias, J. G., Eden, F. & Crothers, D. M. *J. molec. Biol.* **152,** 153–161 (1981).
28. Dickerson, R. E. & Drew, H. R. *J. molec. Biol.* **149,** 761–786 (1981).
29. Finch, J. T. *et al. Nature* **269,** 29–36 (1977).
30. Schellman, J. A. *Biopolymers* **13,** 217–226 (1974).
31. Hats, J. B., Magar, M. E. & Zimm, B. H. *Biopolymers,* **8,** 531–536 (1969).
32. Church, G. M., Sussman, J. L. & Kim, S.-H. *Proc. natn. Acad. Sci. U.S.A.* **74,** 1458–1462.
33. Zubay, G. & Doty, P. *J. molec. Biol.* **1,** 1–2 (1959).
34. Sung, M. T. & Dixon, G. H. *Proc. natn. Acad. Sci. U.S.A.* **67,** 1616–1623 (1970).
35. Adler, K. *et al. Nature* **237,** 322–327 (1972).
36. Warrant, R. W. & Kim, S. H. *Nature* **271,** 130–135 (1978).
37. von Hippel, P. H., Bear, D. G., Winter, R. B. & Berg, O. G. *Promoters: Structure and Function* (Eds Rodriguez, R. & Chamberlin, M.) (Praeger Scientific, New York, in the press).
38. Woodbury, C. P. & von Hippel, P. H. in *Gene Amplification and Analysis* Vol. 1 (ed. Chirikjian, J. C.) (Elsevier, New York, 1981).
39. Helene, C. *FEBS Lett.* **74,** 10–13 (1977).

40. Drew, H. R. & Dickerson, R. E. *J. molec. Biol.* **151**, 535–556 (1981).
41. Caruthers, M. H. *Accts Chem. Res.* **13**, 155–160 (1980).
42. Richter, R. H. & Eigen, M. *Biophys. Chem.* **2**, 255–263 (1974).
43. Berg, O. G. & Blomberg, C. *Biophys. Chem.* **4**, 367–381 (1976).
44. Berg, O. G., Winter, R. B. & von Hippel, P. H. *Biochemistry* **20**, 6929–6948 (1981).
45. Winter, R. B. & von Hippel, P. H. *Biochemistry* **20**, 6948–6960 (1981).
46. Winter, R. B., Berg, O. G. & von Hippel, P. H. *Biochemistry* **20**, 6961–6977 (1981).
47. Berg, O. G., Winter, R. B. & von Hippel, P. H. *Trends biochem. Sci.* **7**, 52–55 (1982).
48. Anderson, W. F., Takeda, Y., Ohlendorf, D. H. & Matthews, B. W. *J. molec. Biol.* (in the press).
49. Matthews, B. W., Ohlendorf, D. H., Anderson, W. F. & Takeda, Y. *Proc. natn. Acad. Sci. U.S.A.* **79**, 1428–1432 (1982).
50. Steitz, T. A., Ohlendorf, D. H., McKay, D. B., Anderson, W. F. & Matthews, B. W. *Proc. natn. Acad. Sci. U.S.A.* **79**, 3097–3100 (1982).
51. Ohlendorf, D. H., Anderson, W. F. & Matthews, B. W. *J. molec. Evol.* (submitted).
52. Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M. & Pabo, C. O. *Nature* **298**, 447–451 (1982).
53. Weber, I. T., McKay, D. B. & Steitz, T. A. *Nucleic Acids Res.* (submitted).
54. Matthews, B. W., Ohlendorf, D. H., Anderson, W. F., Fisher, R. G. & Takeda, Y. *Cold Spring Harbor Symp. quant. Biol.* **47** (in the press).

SECTION E. PROTEIN EVOLUTION

J. Mol. Biol. (1980) **140**, 77–99

A Systematic Approach to the Comparison of Protein Structures

S. J. Remington† and B. W. Matthews‡

Institute of Molecular Biology and
Department of Physics
University of Oregon
Eugene, Ore. 97403, U.S.A.

(Received 13 November 1979, and in revised form 18 February 1980)

A systematic method has been developed for comparing the backbone conformations of proteins (Remington & Matthews, 1978). Two proteins are compared by successively optimizing the agreement between all possible segments of a chosen length from one protein, and all possible segments of the same length from the other protein. The method reveals any similarities between the two proteins, and provides an estimate of the statistical significance of any given structure agreement that is obtained.

The method has been tested in a number of cases, including comparisons of the dehydrogenases and of the pancreatic and bacterial serine proteases. These examples were chosen to test the ability of the comparison method to detect structural similarities in the presence of large insertions and deletions. The results suggest that the detection of the "nucleotide binding fold" in the dehydrogenases is at the limit of the capability of the comparison technique in its original form, although it may be possible to generalize the method to allow for insertions and deletions in proteins.

The results of many protein comparisons, made with different probe lengths, are summarized. For medium and long probe lengths, the average value of the structural agreement does not depend very much on the type of protein being compared. The average value of the structure agreement increases with the square root of the probe length, but for probe lengths above about 40 residues, the standard deviation is independent of probe length. From these observations it is possible to construct a generalized probability diagram to evaluate the significance of any structure agreement that might be obtained in comparing two proteins.

1. Introduction

As the number of known three-dimensional structures of proteins has increased, it has become increasingly apparent that similar patterns of folding often occur in different proteins. These similarities occur within families of homologous proteins such as the pancreatic serine proteases, or the globins, which are obviously derived from a

† Present address: Max-Planck-Institut für Biochemie und Physikalisch-Chemisches, Institut der Technischen Universität, D-8033 Martinsried bei München, Germany.

‡ Author to whom correspondence should be addressed.

0022–2836/80/170077–23 $02.00/0

© 1980 Academic Press Inc. (London) Ltd.

common precursor, and also between proteins that have no obvious evolutionary relationship. For example, the "immunoglobulin fold" has also been found to occur in superoxide dismutase (Richardson *et al.*, 1976). In some cases it has been easy to recognize common patterns of folding in different proteins, but in other cases the intrinsic complexity of many proteins has masked underlying similarities.

It is clearly of considerable interest to identify similarities between different protein structures, even though the origin of such similarities may be open for debate. On the one hand, a common pattern of folding seen in two proteins may suggest that they evolved from the same precursor. On the other hand, it may be that the observed folding is favored energetically, and tends to occur spontaneously or is dictated, perhaps, by certain functional requirements. There is still a need for effective methods of comparing protein structures. As more protein structures are known and are compared with each other, it may become possible to determine the basis of structural similarity.

Two general approaches have been used to compare the folding of different proteins. In the *topological* approach, protein structures are reduced to simplified schematic diagrams that show the sequential location of α-helices and β-sheets. By visual inspection of such diagrams, it is possible to recognize potentially similar structural domains (Schulz & Schirmer, 1974; Sternberg & Thornton, 1977; Richardson, 1977; Levitt & Chothia, 1976). These methods are simple to apply but, except in a limited sense, are non-quantitative. They give only limited information about spatial equivalence, and may not detect potential regions of similarity, especially those that do not include α-helices and β-sheets. In the *spatial* method of structure comparison, a set of equivalent points in the two structures is identified, and their co-ordinates determined. Then, the set of reference points from one structure is rotated and translated to determine the optimum agreement with the other structure (see e.g. Matthews *et al.*, 1968; Freer *et al.*, 1970; Huber *et al.*, 1971). Rossmann and co-workers have extended this method to permit the comparison of the backbones of pairs of proteins in which extensive "insertions" or "deletions" may have occurred (Rao & Rossmann, 1973; Rossmann & Argos, 1976,1977).

Recently, we proposed a new method of spatial comparison (Remington & Matthews, 1978) based on Fitch's (1966,1970) method of comparing amino acid sequences. Both techniques permit exhaustive searches to be made for structural similarity, although the Rossmann–Argos procedure requires arbitrary assumptions on the values of various constants. On the other hand, the proposed method, in its present form, may not detect structural similarities that are interspersed with extensive insertions and deletions. The method provides an estimate of the statistical significance of any apparent structural similarity by evaluating it relative to the frequency distribution of all other comparisons. In the Rossmann–Argos procedure, the significance of the agreement between two proteins is estimated by comparing the number of "equivalent" residues obtained for the best alignment with the number of equivalent residues obtained for all other alignments of the two proteins (Rossmann & Argos, 1977). The method we have proposed is useful for comparing different proteins, and for correlating repeated structural elements within a single polypeptide chain, and has been used in this context by McLachlan (1979).

In our preliminary communication (Remington & Matthews, 1978) we applied the

structure comparison method to comparisons of hemoglobin and myoglobin, phage lysozyme and carp calcium-binding protein, and phage and hen egg-white lysozyme. In this paper we describe further tests of the method with the dehydrogenases and the pancreatic and bacterial serine proteases. These were chosen because they represent families of proteins in which there are regions of structural similarity interspersed with large insertions and deletions. We showed previously that the comparison method will clearly reveal close structural correspondence, as is found in the different globins. Here we are more concerned with testing the method in cases of limited similarity. The tests reported here also provide a representative set of structure comparison statistics for proteins with different types of secondary structure.

2. Methods

To compare two proteins, a suitable probe length L is chosen; e.g. 40 residues. Then, each possible backbone segment of length L residues from the first protein, 1 to 40, 2 to 41, etc. is compared in turn with each possible backbone segment of length L from the second protein. In each case the 40-alpha-carbon segment of one protein is translated and rotated so as to minimize $R_{C\alpha}$, the root-mean-square distance between the respective alpha-carbon atoms of the 2 segments:

$$R^{ij}_{C\alpha} = \left(\frac{1}{L}\sum_{k=0}^{L-1} |X_{i+k} - A(\alpha, \beta, \gamma)\, X'_{j+k}|^2\right)^{\frac{1}{2}}, \tag{1}$$

where X_i and X'_j are, respectively, the co-ordinate vectors of the ith alpha-carbon atom and the jth alpha-carbon atom of the 2 proteins relative to the center of mass of the L atoms being compared. The minimization of $R_{C\alpha}$ requires that the centers of mass of the 2 sets of atoms coincide, so that minimization reduces to the determination of the rotation matrix $A(\alpha, \beta, \gamma)$. This matrix is the resultant of successive rotations of the angle γ about the Z axis of protein 2, the angle β about the new Y axis, and the angle α about the resultant X axis.

Several methods are available for the minimization of $R_{C\alpha}$, one being to iteratively adjust the rotation angles (α, β, γ) until a minimum value of $R_{C\alpha}$ is obtained (cf. Matthews *et al.*, 1968). This was the procedure used in our original paper (Remington & Matthews, 1978), and has been used for many of the comparisons reported here. The method has the advantage that it provides both the minimum of $R_{C\alpha}$, and the corresponding rotation matrix $A(\alpha, \beta, \gamma)$. Also, as judged by a variety of tests, it satisfactorily minimizes $R_{C\alpha}$ for both good and bad structural agreements, and can be relied on to locate the true minimum. On the other hand, the iterative method is rather slow, which is a disadvantage when comparing large proteins (Remington & Matthews, 1978). The number of minimizations required to compare 2 proteins of length M and N residues, with a probe length L, is $(M - L + 1) \times (N - L + 1)$, which can approach 10^5 minimizations (see Table 1).

Alternatively, $R_{C\alpha}$ can be minimized by a number of matrix methods. McLachlan (1979) has recently reviewed these methods and proposed a new minimization technique that is both fast and reliable, and which will, with some additional computing, provide the optimum rotation matrix $A(\alpha, \beta, \gamma)$. In practice, we found McLachlan's method to be about 100 times faster than the angle adjustment method, when both algorithms were tested on a Varian V76 minicomputer.

After the values of $R_{C\alpha}$ have been obtained for all segments of the proteins being compared, the mean and standard deviation of $R_{C\alpha}$ are obtained, and the values of $R_{C\alpha}$ plotted as a contoured "structure comparison map". When comparing 2 very similar proteins, such as myoglobin and hemoglobin, the comparison map will have an obvious

band of good structural agreement down the diagonal (see e.g. see Fig. 1 of Remington & Matthews, 1978). The peaks tend to be elongated parallel to the diagonal because, as one moves in this direction, there is a substantial overlap between the successive structure comparisons.

One of the features of the structure comparison method is that it provides a large sample of values of $R_{C\alpha}$, against which an unusually good agreement can be compared. In comparing a number of different proteins we have found that the distribution of the observed values of $R_{C\alpha}$ is approximately Gaussian, excepting those values of $R_{C\alpha}$ that correspond to unusually good structural agreement. It is the departure of these low $R_{C\alpha}$ values from a Gaussian distribution that shows them to be statistically "unusual".

In the structural comparisons quoted below, we compare the observed distribution of $R_{C\alpha}$ with the best-fit Gaussian curve, and also plot the results as a cumulative probability distribution (see e.g. Fig. 2(b)). In this case a Gaussian distribution appears as a straight line, and any unusually good structural correspondence will cause an increase in the frequency of observed values of $R_{C\alpha}$ at the left-hand end of the distribution.

All co-ordinates used in this paper were taken from the Protein Data Bank (Bernstein *et al.*, 1977).

3. Results

(a) *Glycolytic enzymes*

One of the best known examples of a common structural fold is the "nucleotide binding domain", first observed in lactate dehydrogenase and malate dehydrogenase, and subsequently in glyceraldehyde-3-phosphate dehydrogenase and horse liver alcohol dehydrogenase (Rao & Rossmann, 1973; Rossmann *et al.*, 1974; Webb *et al.*, 1973; Ohlsson *et al.*, 1974). With the exception of lactate and malate dehydrogenase, none of the dehydrogenases has obviously homologous amino acid sequences.

The "idealized" nucleotide binding fold consists of six strands of parallel β-sheet, connected *via* two helices on one side of the sheet and two helices on the other, including about 150 residues in all. In the actual structures, the respective coenzyme binding domains differ from each other somewhat and it is necessary to "ignore" these variations in superimposing one domain on another. In comparing the nucleotide binding domains of lactate dehydrogenase and glyceraldehyde-3-phosphate de-

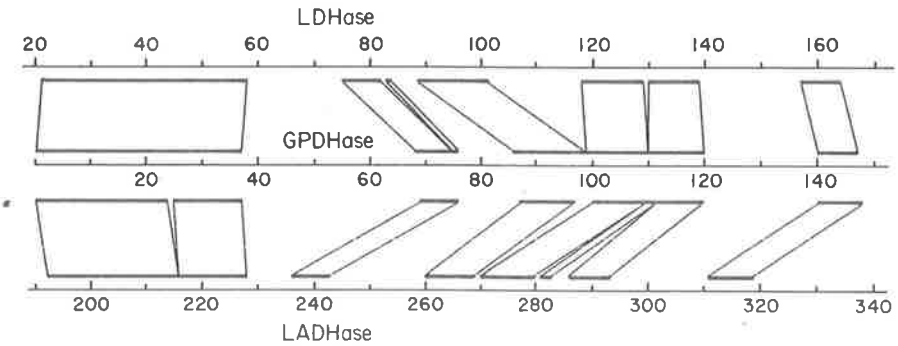


FIG. 1. Alignment of the nucleotide binding folds of lactate dehydrogenase (LDHase), glyceraldehyde-3-phosphate dehydrogenase (GPDHase) and liver alcohol dehydrogenase (LADHase), proposed by Rossmann *et al.* (1974). The connected bars indicate α-carbon atoms in the respective dehydrogenases that are structurally equivalent.

hydrogenase, for example, Rossmann *et al.* (1974) found that 75 "equivalent" alpha-carbon atoms could be superimposed within a distance of 3·8 Å. This structural alignment is illustrated diagrammatically in Figure 1. As can be seen, there are blocks of residues that coincide, interspersed with regions where the structural superposition is poor (greater than 3·8 Å). It seemed that this would provide a good test of the structure comparison method, since one would be attempting to locate a region of structural agreement in the presence of "insertions" and "deletions".

The comparison of LDHase† and GPDHase with a probe length of 80 residues is illustrated in Figure 2(a), and the corresponding probability plots are shown in Figure 2(b). In Figure 2(b) we show the observed frequency distribution and the best-fit Gaussian, as well as the same data plotted as a cumulative distribution. The average structural agreement for all possible alignments is 14·72 Å, with a standard deviation $\sigma = 1{\cdot}72$ Å. The best agreement, $R_{C\alpha} = 6{\cdot}1$ Å, aligns residues 22 to 101 of LDHase with 1 to 80 of GDPHase. This is essentially the same alignment as proposed by Rossmann *et al.* (1974) to superimpose the nucleotide binding domains of these two dehydrogenases (Fig. 1). The agreement $R_{C\alpha} = 6{\cdot}1$ Å, corresponding to $5{\cdot}0\sigma$, is statistically significant. Also, the cumulative probability plot for $L = 80$ residues (Fig. 2(b)) is distinctly non-linear, showing an increase in the frequency of low agreements relative to that expected for a Gaussian distribution. We conclude, therefore, that the structure comparison method can successfully reveal the agreement between the "nucleotide binding domains" of LDHase and GPDHase, notwithstanding the insertions and deletions.

The next-highest peak in Figure 2(a) indicates an agreement of $R_{C\alpha} = 7{\cdot}7$ Å between residues 139 to 218 of LDHase and residues 178 to 257 of GPDHase. These two segments include residues of the respective active sites of LDHase and GPDHase, and it is interesting to note that the above structural superposition of the two backbone segments also results in a superposition of the substrate binding sites of the two enzymes. On the other hand, the above superposition does not align Arg171 and His196 of LDHase with Cys149 and His176 of GPDHase. Garavito *et al.* (1977) have suggested that these pairs of residues play analogous roles, albeit with opposite hand, in the two dehydrogenases. Although the agreement $R_{C\alpha} = 7{\cdot}7$ Å corresponds to $4{\cdot}1\sigma$, it is to be expected that in a comparison of proteins of this size, such an agreement will occur at least once by chance alone. Therefore, the significance of the apparent structural agreement within the catalytic domains of LDHase and GPDHase is uncertain. Nevertheless, the suggestion remains that there is an heretofore undetected structural similarity between parts of the catalytic domains of LDHase and GPDHase.

We show in Figure 3(a) and (b) the results of comparing LDHase and GPDHase with a probe length of 120 residues. Here, the agreement between the nucleotide binding regions is quite poor ($2{\cdot}8\sigma$). In this case the insertions and deletions seen in Figure 1 are located in such a way that they prevent significant structural similarity occurring between any corresponding polypeptide segments of length 120 residues.

† Abbreviations used: LDHase, lactate dehydrogenase; GPDHase, glyceraldehyde-3-phosphate dehydrogenase; LADHase, liver alcohol dehydrogenase; SGPBase, protease type B from *Streptomyces griseus*.

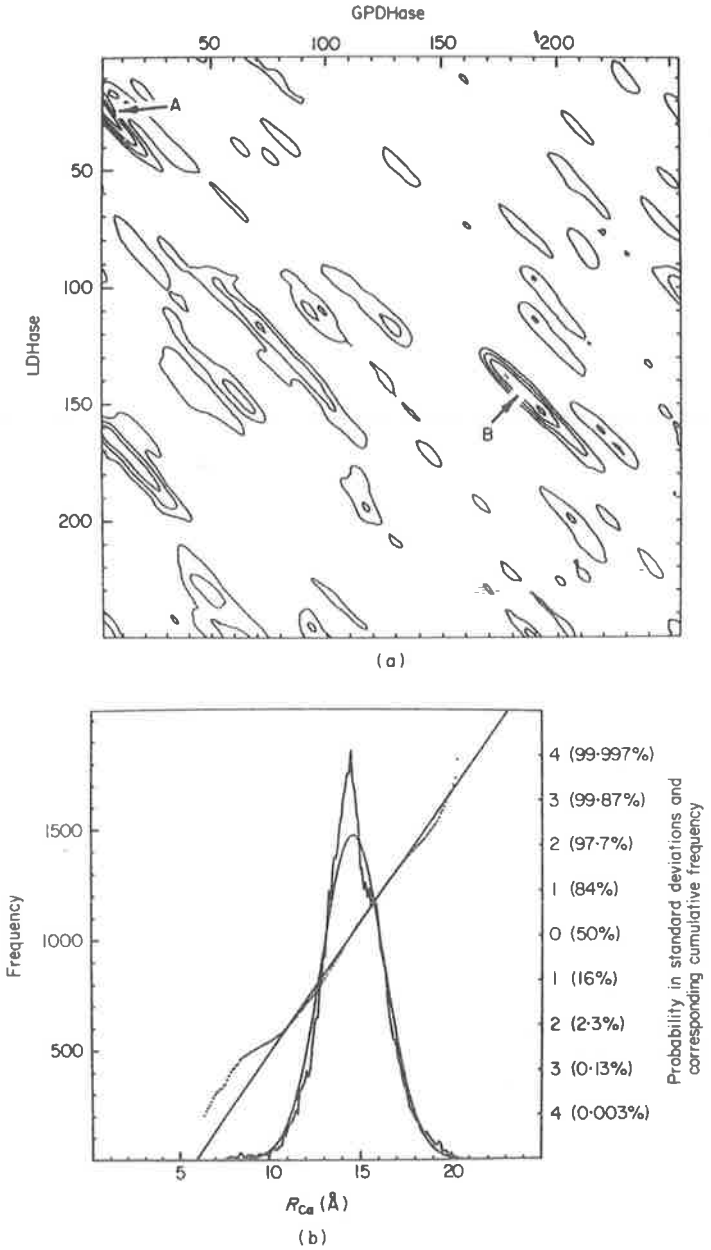


FIG. 2.(a) Structure comparison map for LDHase and GPDHase with a probe length of 80 residues. Successive contour levels indicate values of R_{Ca} equal to 13·0 Å, 11·3, . . . and are at intervals of 1 standard deviation (1·72 Å) below the mean value of 14·73 Å. Peak A indicates the alignment of the respective nucleotide binding folds, and peak B indicates similarity in the structures of the catalytic domains.

(b) Frequency distribution of R_{Ca} for the comparison of LDHase and GPDHase with a probe length of 80 residues. The best-fit Gaussian distribution is superimposed. The Figure also includes the same data plotted as a cumulative distribution. Values of R_{Ca} are grouped in increments of 0·1 ångström unit. The ordinate on the right gives the probability in units of 1σ and the corresponding cumulative frequency of R_{Ca}, expressed as a percentage.

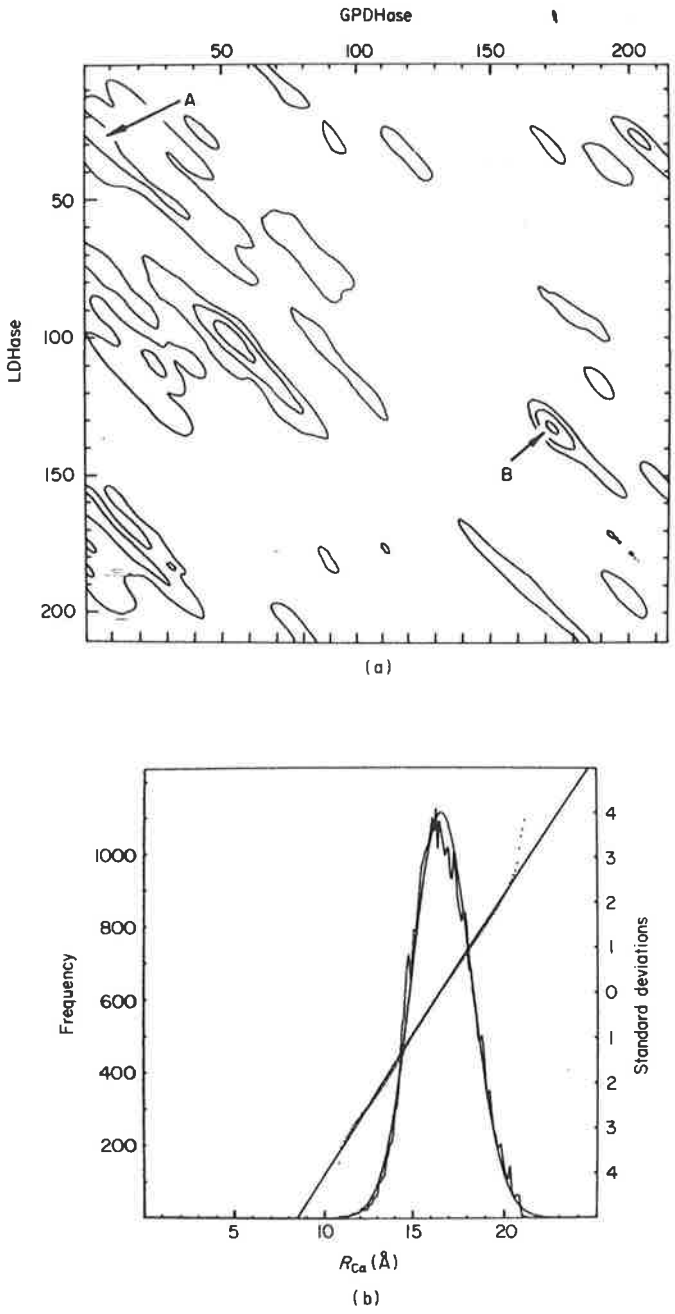


FIG. 3. (a) Structure comparison map for LDHase and GPDHase with a probe of length 120 residues. Contours are drawn at 1σ (1·61 Å) below the mean of 16·6 Å. Peak A corresponds to alignment of the nucleotide binding domains, and peak B to superposition of part of the catalytic domains.

(b) Probability distribution corresponding to (a).

Figure 3(a) indicates that there are two alignments with agreement better than 3σ, one of which superimposes the active site regions, but in Figure 3(b) the cumulative distribution is a straight line, indicating that these superpositions are no better than might be expected by chance alone.

On comparing LDHase and GPDHase with a probe length $L = 40$ residues, the best agreement, $R_{C\alpha} = 2{\cdot}8$ Å ($3{\cdot}9\sigma$), is for the superposition of residues 21 to 60 of LDHase on 1 to 40 of GPDHase. Although this alignment superimposes the nucleotide binding domains of the two dehydrogenases, it is not better than the statistically expected best value, and the cumulative probability plot is a straight line. As discussed in more detail below, the level of structural agreement obtained with a short probe must be substantially better than with a long probe, in order to have the same statistical significance.

We have also compared the structure of liver alcohol dehydrogenase with both LDHase and GPDHase. In the comparison of LADHase and LDHase with a probe length $L = 80$ residues, the best agreement, $R_{C\alpha} = 9{\cdot}0$ Å, superimposes residues 192 to 271 of LADHase on 22 to 101 of LDHase. This is the superposition of the nucleotide binding domains (Rossmann *et al.*, 1974), but the level of significance ($3{\cdot}3\sigma$) is not enough to distinguish this superposition as being unusually good. In the comparison of LADHase and GPDHase with a probe length of 80 residues, the superposition of the nucleotide binding domains shows up as one of the higher peaks on the comparison map, but is by no means obvious (Fig. 4(a) and (b)). As can be seen in Figure 4(a), the region corresponding to the respective nucleotide binding domains (1 to 140 in GPDHase and 190 to 330 in LADHase) contains all of the best alignments in the two structures, but none of these shows out as a single dominant feature. Presumably these "satellite" agreements arise in part from the fact that the nucleotide-binding domain is rather repetitive, with alternating β-strands and α-helices, and also has an overall twofold repeat (Rossmann *et al.*, 1974) so that different parts of the nucleotide binding domains can superimpose partially to give some structural correspondence. Also, the large insertion between residues 40 and 60 of GPDHase, relative to LADHase, which prevents overall good agreement between the two domains, contributes to the satellite peaks by separating regions of partial agreement (see Fig. 1). It may be noted that Eventoff & Rossmann (1975) also found the nucleotide binding folds of LADHase and GPDHase to be the most different among the dehydrogenases tested.

In summary, the above comparisons of the dehydrogenases suggests that the detection of the nucleotide binding fold is at the limit of the capability of the comparison technique in its present form. In the case of LDHase and GPDHase, the comparison method clearly confirms that the structural correspondence of the respective coenzyme binding domains of the two enzymes is statistically significant. On the other hand, the agreement between the nucleotide binding domains of, for example, LADHase and GPDHase, is not unusually significant.

(b) *Microbial and pancreatic serine proteases*

We wished to test the comparison method in a case where two structures were known to be related evolutionarily, yet had substantial differences in structure due to large insertions and deletions. Such an example is provided by the microbial and

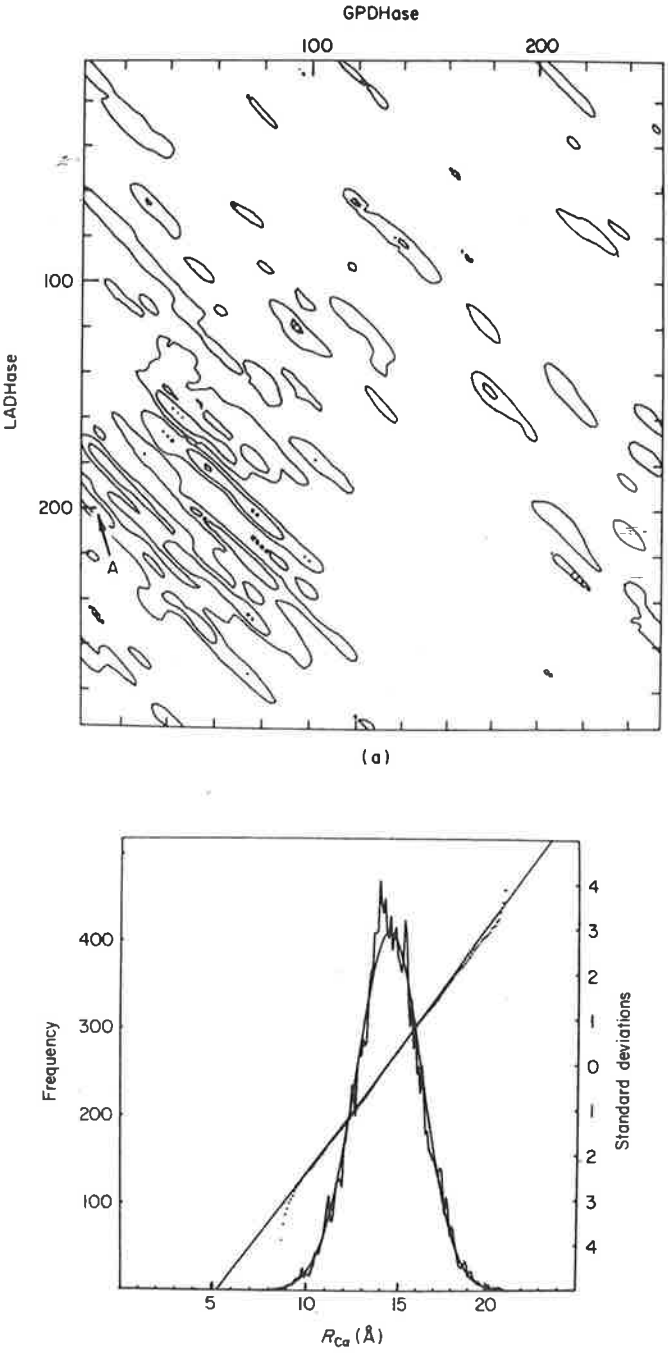


FIG. 4. (a) Structural comparison of LADHase and GPDHase with a probe of 80 residues. Peak A indicates the (imperfect) superposition of the nucleotide binding regions. Contours drawn at intervals of 1σ (1·83 Å) below the mean agreement of 14·47 Å.
(b) Probability distribution corresponding to (a).

pancreatic serine proteases. The microbial serine proteases are smaller than the pancreatic enzymes, with molecular weights of about 20,000, compared to about 25,500 for α-chymotrypsin and elastase. The two classes of enzymes have only about 18% amino acid sequence homology, yet they have obviously derived from a common precursor. The three-dimensional structures of the microbial enzymes clearly resemble those of the pancreatic enzymes, although there are many differences, and only about two-thirds of the residues adopt topologically equivalent positions (Delbaere *et al.*, 1975,1979; James *et al.*, 1978). In the following comparisons we have used the co-ordinates for the protease type from *Streptomyces griseus* (Delbaere *et al.*, 1979), and for elastase (Shotton & Watson, 1970).

In Figure 5 we show the residues in SGPBase and elastase that are topologically equivalent, according to the recent alignment reported by James *et al.* (1978). The results of the comparison of SGPBase and elastase with probe lengths of $L = 40$ and 80 residues are shown in Figures 6 and 7. In both cases the comparison maps have peaks along the diagonal, suggesting some structural correspondence, but in neither case is the structural agreement of high significance. Also, the probability plots (Figs 6(b) and 7(b)) show the distribution of the agreements to be essentially Gaussian. The reason that the comparison method does not detect the structural correspondence of SGPBase and elastase can be seen in Figure 5. The regions of structural homology are short, typically about ten residues, and are interspersed with large insertions or deletions. Because of these insertions and deletions, there is no case where 40 (or 80) consecutive residues of one structure remain in alignment with 40 (or 80) consecutive residues of the other structure. Ordinarily, one might expect to have insertions and deletions in the two structures that would compensate, but this is not the case for SGPBase, which is much shorter than elastase. Here there are 60 insertions in the elastase sequence, but only seven compensating insertions in SGPBase (see Fig. 5; and Table 1 of James *et al.*, 1978).

As a crude attempt to compensate for the difference in length of elastase and SGPBase, we ran a comparison of the two structures in which every fourth residue of elastase was deleted. This reduces the number of elastase residues to about the same as SGPBase. The results are shown in Figures 8 and 9. Here the similarity between elastase and SGPBase is obvious. In both cases (Figs 8(a) and 9(a)) there is the characteristic band down the diagonal that one sees in comparing very similar proteins such as trypsin and chymotrypsin, or myoglobin and hemoglobin. Also,

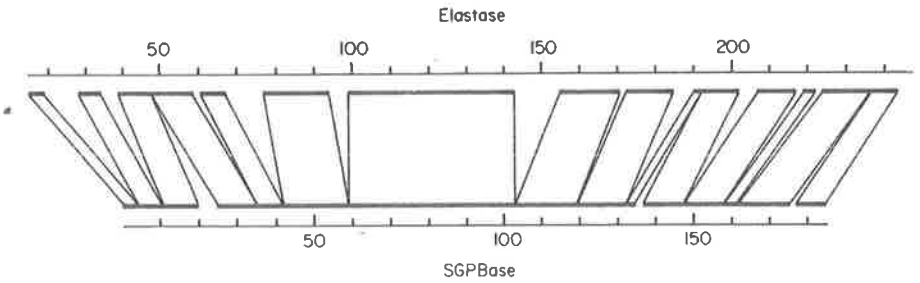


FIG. 5. Structural correspondence of elastase and *S. griseus* protease type B (SGPBase) proposed by James *et al.* (1978).

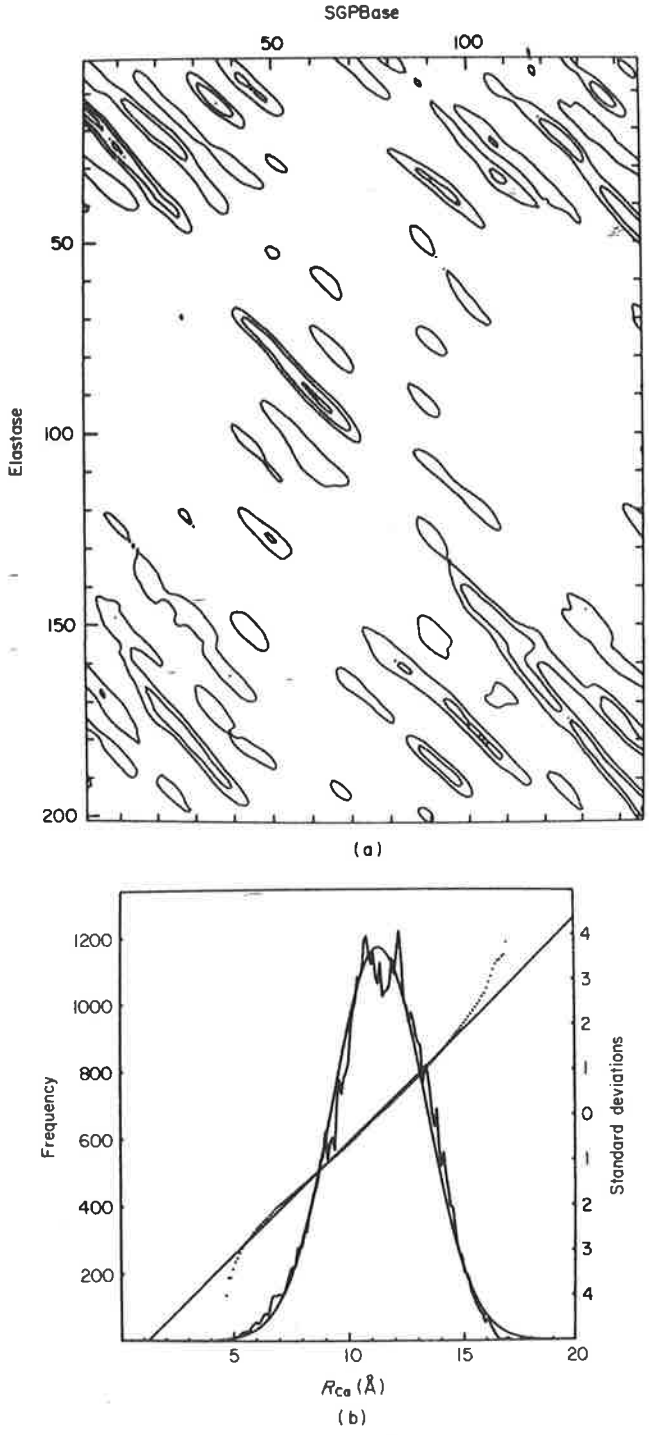


FIG. 6. (a) Structure comparison of elastase and SGPBase with a probe of length 40 residues. Contours drawn at intervals of 1σ (2·00 Å) below the mean (11·32 Å).
(b) Probability distribution corresponding to (a).

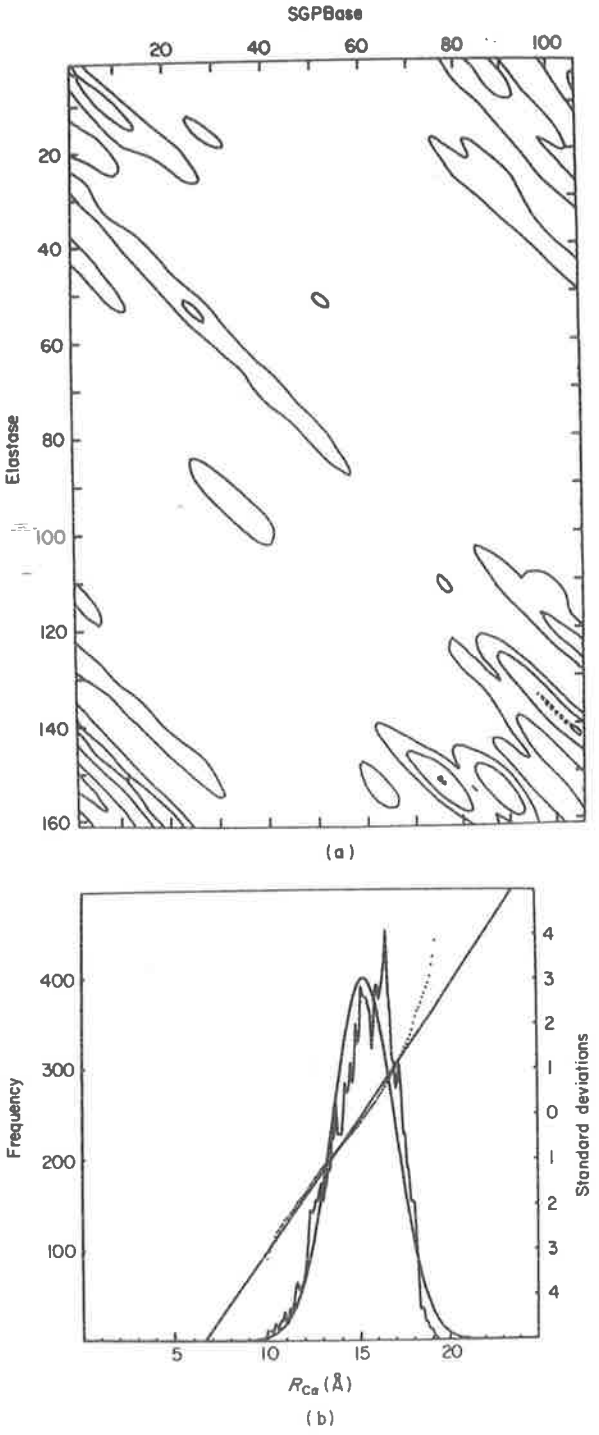


FIG. 7. (a) Structure comparison of elastase and SGPBase with a probe of length 80. Contours drawn at intervals of 1σ (1·71 Å) below the mean (15·23 Å).

(b) Probability distribution corresponding to (a).

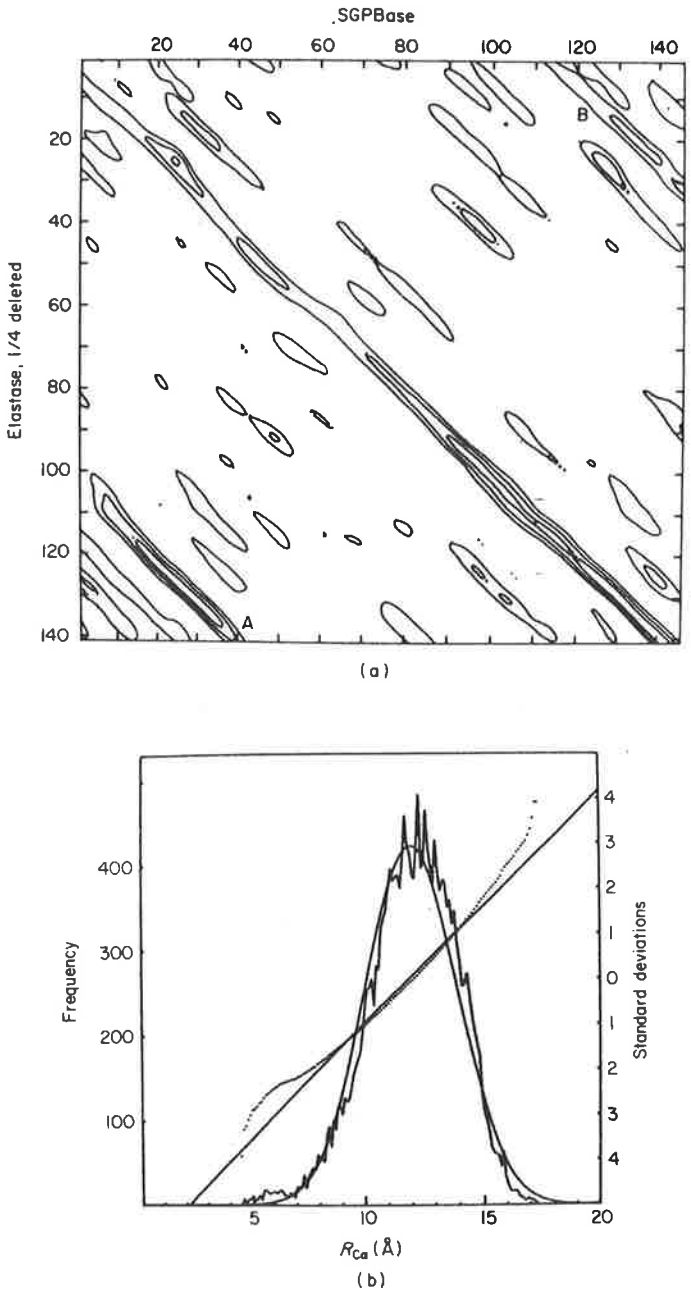


FIG. 8. (a) Structure comparison of SGPBase with a concatenated elastase in which every 4th α-carbon was deleted (see the text). Probe length 40 residues and contours at levels of 1·95 Å below the mean (11·95 Å). The continuous peak down the diagonal indicates the overall agreement of the two structures, and peaks A and B indicate agreement between one domain of elastase and the other domain of SGPBase.

(b) Probability distribution corresponding to (a).

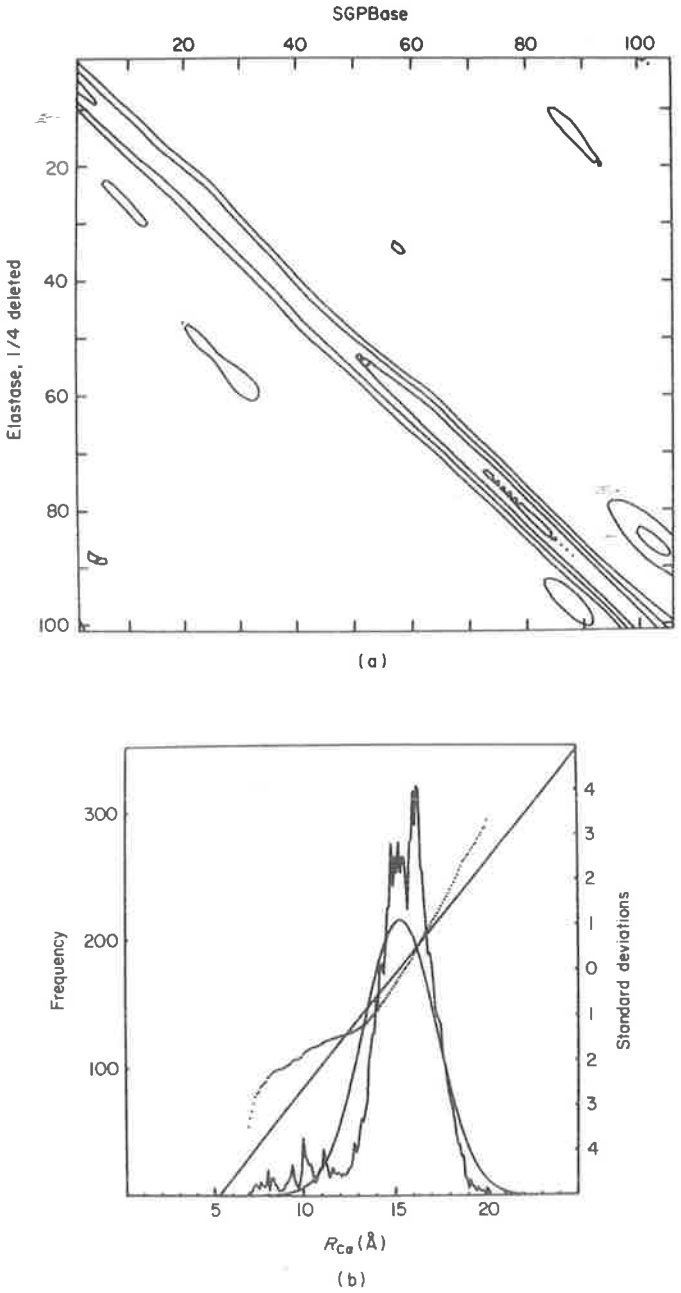


FIG. 9. (a) Structure comparison of SGPBase and concatenated elastase with every 4th α-carbon deleted. Probe length 80 residues and contours drawn at increments of 1·99 Å below the mean (15·28 Å).

(b) Probability distribution for (a).

the probability plots (Figs 8(b) and 9(b)) show obvious departures from Gaussian distribution, confirming the good structural correspondence. Deleting the elastase residues as described above clearly allows the two structures to remain sufficiently "in register" that their overall structural similarity becomes obvious, despite the local perturbations resulting from the deletion of every fourth residue. This preliminary test indicates that it may be possible to increase the ability of the comparison method to deal with insertions by making comparisons in which one or the other protein is "condensed", by systematically deleting residues.

We do not suggest that "condensation" in the form used here will always allow one to detect structural similarities in the presence of insertions and deletions. In particular, "condensation" would not be expected to work if the proteins being compared were similar except for one or two long insertions or deletions. (Although in such an idealized case, the normal comparison method would be expected to detect the structural similarity away from the insertions or deletions.) The successful result in the case of elastase and SGPBase shows that "condensation" is at least one possible way in which the comparison method may be generalized. Other approaches are possible, and need to be tested in a number of real cases.

The fact that the region of good structural agreement between SGPBase and condensed elastase extends continuously along the diagonal of Figure 8(a) and 9(a) reveals something about the relation between these two molecules. Clearly, elastase is "derived" from SGPBase by making a number of insertions that are distributed fairly uniformly along the length of the molecule. Continuous structural agreement along the diagonal would not be expected if there was a large insertion in SGPBase, as was thought to occur in the region 164 to 182 (Olson *et al.*, 1960; McLachlan & Shotton, 1971; Delbaere *et al.*, 1975). In the revised alignment of the two sequences (Fig. 5; James *et al.*, 1978) there are only three short insertions in the SGPBase molecule, relative to elastase, these being of length one, two and four residues.

In Figure 8(a) the strong elongated peak in the bottom left corner corresponds to superposition of part of the first β-structure domain of SGPBase on the second β-structure domain of elastase (Blow, 1969; McLachlan, 1979). The weaker peak in the top right corner arises from the superposition of the second SGPBase domain on the first elastase domain.

(c) *Comparisons of different structural types*

In the comparison of two structures, statistical analysis of the data is based on the method proposed by Fitch (1966,1970) for amino acid sequence comparison. In essence, the many individual comparisons of different structural segments within the two proteins are used as a data base against which to assess the significance of unusually good agreements. In the statistical analysis, it is assumed that the frequency distribution of structure agreements for any two proteins will be Gaussian. Also, it is anticipated (although not strictly required) that the frequency distribution will be approximately the same when comparing different types of proteins. In this section we discuss the validity of these assumptions.

The assumption of a Gaussian distribution $R_{C\alpha}$ is supported by the experimental data. In many comparisons of proteins of different types (helical, sheet and mixed) we have found that the distribution of $R_{C\alpha}$ can be fitted by a Gaussian curve (e.g.

Figs 3(b), 4(b) and 6(b)), except for those cases where there is unusually good structural agreement between the structures being compared (e.g. Fig. 9(b)).

The distributions obtained in the comparisons of a number of different structures, and for different probe lengths, are summarized in Table 1. Figure 10 shows the

TABLE 1

Distribution of R_{Ca} *for different structure comparisons*

Proteins compared	Probe length	Number of R_{Ca}	Best R_{Ca} (Å)	Average R_{Ca} (Å)	Standard deviation (Å)
T4 lysozyme; hen egg-white lysozyme	20	15,950	1·8	5·70	1·21
Elastase; microbial protease SGPBase	20	36,686	2·6	7·42	1·87
T4 lysozyme; carp Ca -binding protein	30	10,665	3·0	6·89	1·30
T4 lysozyme; hen egg-white lysozyme	40	11,250	3·8	8·71	1·39
Myoglobin; hemoglobin β chain	40	12,198	0·9	8·19†	1·88†
Lactate dehydrogenase; glyceraldehyde-3-P-dehydrogenase	40	85,260	2·8	10·81	2·04
Lactate dehydrogenase; alcohol dehydrogenase	40	24,360‡	3·3	10·75	2·02
T4 lysozyme; concanavalin A	40	24,750	6·3	11·54	1·91
Elastase; microbial protease SGPBase	40	29,346	4·7	11·32	2·00
Elastase (3/4); microbial protease SGPBase	40	20,586	4·5	11·95	1·95
Alcohol dehydrogenase; glyceraldehyde-3-P-dehydrogenase	40	24,696	2·6	10·94	2·00
T4 lysozyme; hen egg-white lysozyme	60	7350	5·7	10·75	1·47
T4 lysozyme; carp Ca-binding protein	60	5145	5·6	10·70	1·49
Cytochrome b_5; myoglobin	60	2444	9·0	11·85	1·07
T4 lysozyme; hen egg-white lysozyme	80	4250	6·1	12·32	1·63
Myoglobin; hemoglobin β chain	80	4958	1·5	11·42†	3·16†
Lactate dehydrogenase; glyceraldehyde-3-P-dehydrogenase	80	63,500	6·4	14·73	1·72
Lactate dehydrogenase; alcohol dehydrogenase	80	18,500‡	9·0	14·47	1·64
Alcohol dehydrogenase; glyceraldehyde-3-P-dehydrogenase	80	18,796‡	8·7	14·47	1·83
Hexokinase; glyceraldehyde-3-P-dehydrogenase	80	23,622‡	7·9	15·32	2·05
Hexokinase; lactate dehydrogenase	80	92,750	7·9	15·10	2·09

Proteins compared	Probe length	Number of R_{Ca}	Best R_{Ca} (Å)	Average R_{Ca} (Å)	Standard deviation (Å)
Elastase; microbial protease SGPBase	80	17,066	10·0	15·23	1·71
Elastase (3/4)§; microbial protease SGPBase	80	10,706	6·9	15·28	1·99
Lactate dehydrogenase; glyceraldehyde-3-P-dehydrogenase	120	44,940	10·8	16·58	1·61
Lactate dehydrogenase; alcohol dehydrogenase	120	13,440‡	11·9	16·40	1·65
Alcohol dehydrogenase; glyceraldehyde-3-P-dehydrogenase	120	13,696‡	10·7	16·54	1·78
Elastase; microbial protease SGPBase	120	7986	13·6	17·05	1·16

† Whenever there is very good structural agreement, as for hemoglobin and myoglobin, the low values of R_{Ca} will tend to decrease the average value of R_{Ca} and increase the standard deviation. This is particularly obvious for the hemoglobin/myoglobin comparison with a probe length of 80, and these values are not included in Figs 10 to 13. Strictly, the mean and standard deviation should be recalculated omitting the unusually low values of R_{Ca}.

‡ In these comparisons R_{Ca} was calculated for every 2nd residue of each protein. This reduces the number of R_{Ca}, and the computing time, by a factor of approx. 4.

§ See the text.

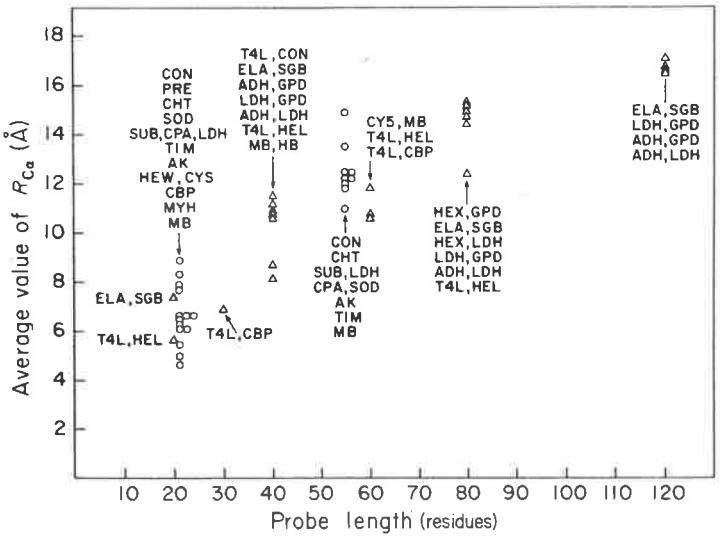


FIG. 10. Average structure agreement for different protein comparisons plotted as a function of the probe lengths. The Figure includes our comparisons of different proteins (△) together with McLachlan's (1979) comparisons of a protein with itself (○). Abbreviations are as follows: CON, concanavalin A; PRE, prealbumin; CHT, α-chymotrypsin; SOD, superoxide dismutase; SUB, subtilisin; CPA, carboxypeptidase A; LDH, lactate dehydrogenase; TIM, triose phosphate isomerase; AK, adenylate kinase; HEL, hen egg-white lysozyme; CY5, cytochrome b_5; CBP, carp calcium binding parvalbumin; MYH, myohemerythrin; MB, myoglobin; ELA, elastase, T4L, bacteriophage T4 lysozyme; SGB, microbial protease B; ADH, alcohol dehydrogenase; GPD, glyceraldehyde-3-phosphate dehydrogenase.

average structure agreement; i.e. the center of the observed distribution for many different protein comparisons, plotted as a function of probe length. The Figure also includes data from McLachlan (1979), who has made a series of comparisons of a protein with itself, using the comparison method to look for repeated structure elements within a single polypeptide chain.

As expected, the average structure agreement increases with the length of the probe. With very short probe lengths ($L = 20$), the average value of $R_{C\alpha}$ for different proteins varies by almost a factor of two, but with longer probes the relative spread is much less. It might be expected that the average agreement when comparing two small, compact α-helical proteins would be less than for a comparison of two large, extended β-type proteins (McLachlan, 1979). This dependence on structural type is significant with short probes, but becomes less noticeable with longer probes. For example, McLachlan (1979) found that the average structure agreement for a comparison of myoglobin with itself with a probe of 21 residues was 4·7 Å, whereas a comparison of concanavalin A with itself yielded a value of 8·9 Å. With a probe of 55 residues, the respective values become 10·9 Å and 14·9 Å, and all other proteins tested are well within these limits (see Fig. 10).

The average structure agreement increases in proportion to the square-root of the probe length, and the data in Figure 10 can be fitted reasonably well by the equation:

$$\overline{R_{C\alpha}} = 1{\cdot}55\sqrt{L}\,, \tag{2}$$

where $\overline{R_{C\alpha}}$ is the average structure agreement for comparing any pair of proteins, or a protein with itself, and L is the probe length. For the reasons discussed above, this equation is less reliable for short probes.

In contrast to $\overline{R_{C\alpha}}$, which depends on L, the standard deviation of the distribution of $R_{C\alpha}$ is essentially independent of the probe length (Fig. 11), at least for probe lengths greater than 20 residues. This is an interesting and, perhaps, unexpected result. It

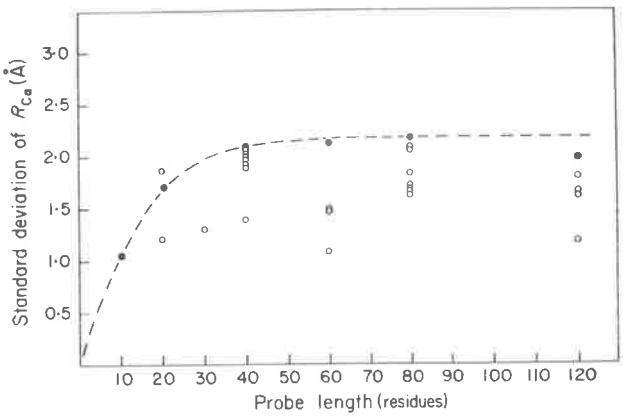


FIG. 11. Standard deviation of $R_{C\alpha}$ for different protein comparisons plotted as a function of probe length (open circles). The filled circles are the standard deviations obtained from 10^6 random comparisons of 32 proteins (see the text) and the broken line is given by eqn (3). (Because McLachlan (1979) does not quote standard deviations, his data cannot be included here.)

might be anticipated that structure agreements obtained with a long probe would scatter over a much wider range than for a small probe, but this is not the case. The spread is essentially the same for a probe length of 20 residues as it is for 120.

As will be discussed below, we wished to combine the average structure agreement (eqn (2)) with the average standard deviation of $R_{C\alpha}$ (Fig. 11) to obtain a generalized probability distribution of $R_{C\alpha}$. However, the generality of equation (2) and of Figure 11 might be suspect, since they were obtained from comparisons of a limited sample of proteins. Therefore, we carried out the following, more general, survey of the known protein structures. The alpha-carbon co-ordinates of 32 proteins were taken from the Protein Data Bank and used to construct a single list of about 7000 numbered co-ordinates. The 32 proteins included most of the known structures for which co-ordinates were available, excluding obviously homologous structures. Then, for a chosen probe length, a random number generator was used to choose from this list one million different pairs of structural segments for which $R_{C\alpha}$ was calculated. Any chosen segment that happened to overlap the beginning or end of a protein was ignored. This survey was made for probe lengths of $L = 10$, 20, 40, 60, 80 and 120 residues. Thus, for each of these probe lengths, we obtained a distribution of $R_{C\alpha}$ that was based on a large number of comparisons (10^6), which included data for many different proteins, and which avoided the redundancy and non-randomness that unavoidably occur when one compares a single protein with another.

The results of these comparisons are summarized in Table 2. In Figure 12 we show the distributions for $L = 10$ and $L = 60$ residues. The distribution for $L = 10$ residues is atypical in having a pronounced maximum at an $R_{C\alpha}$ value of about 0·6 Å. This is due primarily to the alignment of α-helical segments from different proteins, and presumably also includes some other localized secondary structure elements such as "hairpin bends". The plot illustrates quite clearly the limited

TABLE 2

Distribution of $R_{C\alpha}$ *for* 10^6 *random comparisons from 32 proteins*

Probe length	Average $R_{C\alpha}$ (Å)	Standard deviation of $R_{C\alpha}$ (Å)
10	3·88	1·05
20	6·75	1·72
40	10·47	2·06
60	12·76	2·12
80	14·30	2·19
120	16·70	1·96

The proteins included in this calculation, with the number of residues in parenthesis, are as follows: rubredoxin (53); ferrodoxin (54); pancreatic trypsin inhibitor (59); cytochrome b_5 (85); high potential iron protein (85); cytochrome *c* (103); Bence–Jones dimer (107); carp parvalbumin (109); cytochrome c_2 (112); hen lysozyme (129); cytochrome c550 (134); flavodoxin (138); staphylococcal nuclease (142); superoxide dismutase (151); myoglobin (153); T4 phage lysozyme (164); *Streptomyces griseus* protease type B (185); adenylate kinase (194); immunoglobulin Fab (208); papain (212); concanavalin A (237); triose phosphate isomerase (247); carbonic anhydrase B (258); carboxypeptidase A (307); thermolysin (316); malate dehydrogenase (325); lactate dehydrogenase (329); glyceraldehyde-3-phosphate dehydrogenase (333); alcohol dehydrogenase (374); phosphoglycerate kinase (408); hexokinase (455); D-glucose-6-phosphate isomerase (514).

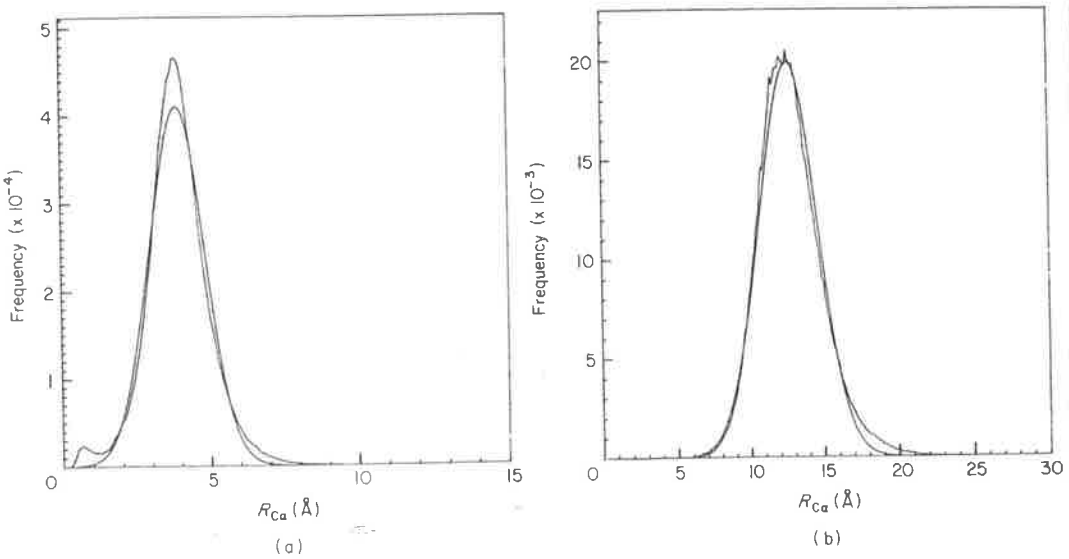


FIG. 12. Distribution of $R_{C\alpha}$ for 10^6 comparisons of structural segments chosen at random from 32 different proteins. The best-fit Gaussian is superimposed. (a) Probe length $L = 10$ residues. (b) Probe length $L = 60$ residues.

significance that can be attached to "good structural agreement" for a short probe length. Such agreement between two proteins may indicate no more than that they both contain α-helices.

As can be seen from Figure 13, the average values of $R_{C\alpha}$ obtained from the extended data base agree well with equation (2) and confirm the validity of this empirical relation. The standard deviations of $R_{C\alpha}$ obtained from the extended data base agree moderately well with those obtained for the individual comparisons, although they are at the high end of the ranges of individual values (Fig. 11). It will be noted in Figure 12(b), which shows the distribution for $L = 60$ residues for the extended data base, that there are more large values of $R_{C\alpha}$ than expected for a Gaussian distribution. This is also observed for the extended data base with the other probe lengths, but is not readily apparent in the individual comparisons. A survey of the structure segments that give rise to these "bad" agreements shows that usually they are due to a superposition of a very compact region from a small protein (e.g. porcine trypsin inhibitor) on a very extended region from a larger protein (e.g. an immunoglobulin). Such comparisons will occur fairly often with the extended data base, which contains a number of very small proteins in addition to larger structures. In contrast, most of the individual comparisons are between proteins of comparable size. As is illustrated in Figure 11, the dependence of the standard deviation of $R_{C\alpha}$ on probe length, based on the extended data base, can be approximated by the empirical relation:

$$\sigma(R_{C\alpha}) = 2{\cdot}2 \tanh (L/19). \tag{3}$$

For probe lengths of 40 residues or longer, $\sigma(R_{C\alpha})$ has essentially reached its asymptotic value of 2·2 Å.

By combining equations (2) and (3) it is possible to construct a generalized probability distribution of $R_{C\alpha}$ that can be used to estimate the significance of a given structural correspondence for any probe length. Such a formulation would be useful to quantitate the significance of a particular structural resemblance that might have been detected between two proteins. Also, the same measure of agreement could be used to calibrate the agreement between a predicted structure of a protein and the observed conformation.

The generalized probability distribution is illustrated in Figure 13. By using a scale proportional to the square-root of the probe length along the abscissa, equation (2) becomes a straight line giving the average value to be expected when comparing two protein backbone segments of length L residues. The observed average values are

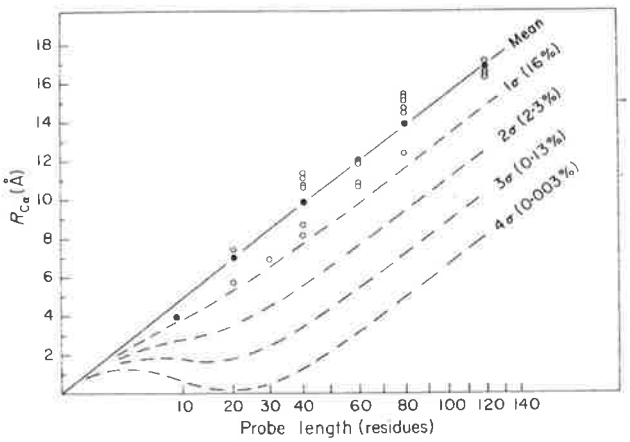


FIG. 13 Generalized structure agreement probability diagram. The solid line gives the average structure agreement as a function of probe length (eqn (2)). For comparison, the observed values for individual comparisons are also shown as open circles (cf. Fig. 10). The filled circles were obtained from 10^6 random comparisons of 32 proteins. Successive broken lines give structure agreements that are better than average by 1σ, 2σ, 3σ . . . The frequencies with which these levels of agreement are expected to occur by chance in a random population are also shown. The scale of the abscissa is proportional to the square-root of the probe length.

included in the Figure for comparison. The broken line drawn below the average value line shows values of $R_{C\alpha}$ that are 1σ better than the average, for a given probe length. Successive lines for 2σ, 3σ, 4σ, are also shown. The corresponding frequencies for these four lines are also included in the Figure. For example, if a value of $R_{C\alpha}$ is at the 3σ level, then the probability of this occurring by chance is 0·13% or 1 in 800. It has to be emphasized that Figure 13 is empirical, and will not give precise values for probabilities. Rather, it is intended as a guide to obtain the approximate expectation for a given value of $R_{C\alpha}$ that has been obtained as a result of some comparison. As discussed above, the average value of $R_{C\alpha}$ is the same for individual comparisons as for comparisons made from the extended data base, but the standard deviation of $R_{C\alpha}$ is somewhat larger in the latter case. The value of $\sigma(R_{C\alpha})$ for L greater than 40 is about 2·2 Å for the extended data base, but has an average value of 1·8 Å for the individual comparisons. Since the lines in Figure 13 for 1, 2, 3, . .

standard deviations are based on the extended data base, they are, if anything, drawn conservatively and in any case are intended only as a guide to the approximate signification of a particular observation.

One of the interesting implications of Figure 13 is that a "good" structure agreement between two short segments may not be very significant. For example, an agreement of 3·5 Å between two 20-residue segments is only at the level of about 2σ (1 in 44). Two 40-residue segments have to agree within about 3·4 Å to be at the 3σ level. On the other hand, very high significance can be attached to moderate agreement over an extended length. For example, an agreement of 6·5 Å over 100 residues is at the 4σ level.

Several attempts have been made to predict the three-dimensional structures of small proteins from their amino acid sequences. It is possible to use Figure 13 to evaluate the success of such methods. For example, Levitt & Warshel (1975) used a simplified representation of protein conformation to simulate the folding of pancreatic trypsin inhibitor and, starting from a fully extended structure, obtained a conformation that differed from the native structure by an average value of 7·7 Å for the 58 residues in the protein. From Figure 13, this corresponds to about 2σ; i.e. better than the value of 11·5 Å expected for a structure drawn at random, but better with only medium significance. (By requiring that residues 48 to 58 of trypsin inhibitor be in a helix, Levitt & Warshel obtained a structure agreement of 6·5 Å, but, since this number includes some prior knowledge of the structure, its significance is unclear.) As emphasized by Hagler & Honig (1978) a "good" prediction of a protein structure should not only have a low root-mean-square discrepancy, but also have the same topology as the native structure. Neither the best Levitt–Warshel structure nor a structure predicted by Hagler & Honig with a discrepancy of 6·2 Å (corresponding to $2{\cdot}5\sigma$ in Fig. 13) meets both criteria.

4. Conclusion

The comparison method readily detects similar structural segments that are not interrupted by large insertions and deletions. If the insertions and deletions in two proteins "compensate" for each other, it may still be possible to detect structural similarities by using a long probe. However, if all the deletions are in one protein, and none in the other, then regions of structural agreement may not extend over long enough stretches to be detected. In principle, it should be possible to extend the method to allow for insertions and deletions, and methods of doing this are being tested.

For medium and long probes, the average value of the structure agreement does not depend very much on the type of structure being compared. The average value of the structure agreement increases with the square-root of the probe length but, for probe lengths above about 40 residues, the standard deviation of the observed structure agreements is independent of probe length. From these observations it is possible to construct a generalized probability diagram to evaluate the significance of structure agreements that are obtained in comparing any two protein structure segments. The probability diagram shows that it is relatively easy to find "good" structural equivalence between short backbone segments of 10 to 30 residues. In

contrast, good agreement over extended pieces of backbone of 60 to 100 residues is much harder to find and can, therefore, be ascribed higher significance.

We thank Dr Andrew McLachlan for helpful discussions on comparison methods, and for providing a copy of this comparison algorithm. Also, we thank Dr William Bennett for a number of helpful comments on the first draft of this manuscript. This work was supported in part by grants from the National Institutes of Health (GM21967, GM20066) and the National Science Foundation (PCM77–19310).

REFERENCES

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Blow, D. M. (1969). *Biochem. J.* **112**, 261–268.

Delbaere, L. T. J., Hutcheon, W. L. B., James, M. N. G. & Theissen, W. E. (1975). *Nature (London)*, **257**, 758–763.

Delbaere, L. T. J., Brayer, G. D. & James, M. N. G. (1979). *Canad. J. Biochem.* **57**, 135–144.

Eventoff, W. & Rossmann, M. G. (1975). *CRC Crit. Rev. Biochem.* **3**, 111–140.

Fitch, W. M. (1966). *J. Mol. Biol.* **16**, 9–16.

Fitch, W. M. (1970). *J. Mol. Biol.* **49**, 1–14.

Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T. & Xuong, Ng.H. (1970). *Biochemistry*, **9**, 1997–2009.

Garavito, R. M., Rossmann, M. G., Argos, P. & Eventoff, W. (1977). *Biochemistry*, **16**, 5065–5069.

Hagler, A. T. & Honig, B. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 554–558.

Huber, R., Epp. O., Steigemann, W. & Formanek, H. (1971). *Eur. J. Biochem.* **19**, 42–50.

James, M. N. G., Delbaere, L. T. J. & Brayer, G. D. (1978). *Canad. J. Biochem.* **56**, 396–402.

Levitt, M. & Chothia, C. (1976). *Nature (London)*, **261**, 552–558.

Levitt, M. & Warshel, A. (1975). *Nature (London)*, **253**, 694–698.

Matthews, B. W., Cohen, G. H., Silverton, E. W., Braxton, H. & Davies, D. R. (1968). *J. Mol. Biol.* **30**, 179–183.

McLachlan, A. D. (1979). *J. Mol. Biol.* **128**, 49–79.

McLachlan, A. D. & Shotton, D. M. (1971). *Nature New Biol.* **229**, 202–205.

Ohlsson, I., Nordström, B. & Brändén, C.-I. (1974). *J. Mol. Biol.* **89**, 339–354.

Olson, M. O. J., Nagabhushan, N., Dzwiniel, M., Smillie, L. B. & Whittaker, D. R. (1960). *Nature (London)*, **228**, 438–442.

Rao, S. T. & Rossmann, M. G. (1973). *J. Mol. Biol.* **76**, 241–256.

Remington, S. J. & Matthews, B. W. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 2180–2184.

Richardson, J. S. (1977). *Nature (London)*, **268**, 495–500.

Richardson, J. S., Richardson, D. C., Thomas, K. A., Silverton, E. W. & Davies, D. R. (1976). *J. Mol. Biol.* **102**, 221–235.

Rossmann, M. G. & Argos, P. (1976). *J. Mol. Biol.* **105**, 75–96.

Rossmann, M. G. & Argos, P. (1977). *J. Mol. Biol.* **109**, 99–129.

Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). *Nature (London)*, **250**, 194–199.

Schulz, G. E. & Schirmer, R. H. (1974). *Nature (London)*, **250**, 142–164.

Shotton, D. M. & Watson, H. C. (1970). *Phil. Trans. Roy. Soc. ser. B*, **257**, 111–118.

Sternberg, M. J. E. & Thornton, J. M. (1977). *J. Mol. Biol.* **110**, 269–283.

Webb, L. E., Hill, E. J. & Banaszak, L. J. (1973). *Biochemistry*, **12**, 5101–5109.

Reprinted from Nature, Vol 290, No. 5804, pp. 334-335, March 26 1981
© *Macmillan Journals Ltd., 1981*

Common precursor of lysozymes of hen egg-white and bacteriophage T4

B. W. Matthews, M. G. Grütter, W. F. Anderson* & S. J. Remington*

Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oregon 97403, USA

The lysozymes of hen egg-white and bacteriophage T4 have similar catalytic properties but their amino acid sequences are not homologous. The question therefore arises whether they are derived from a common ancestral protein or have arisen independently. On the basis of the data we have gathered, it is shown here that the two enzymes are similar in the conformation of their backbones, in their modes of binding substrates, in specific protein–substrate interactions and in their presumed modes of action. We conclude that the two enzymes have diverged from a common precursor. This seems to be the most convincing example to date of the divergence of proteins with nonhomologous amino acid sequences.

If anything, the large number of protein structures and amino acid sequences now known has confused rather than simplified the main issue of whether extant proteins have evolved from a small number of ancestral precursors or have arisen independently. When the amino acid sequences are closely homologous, there is no difficulty: three-dimensional structures are always similar, and it is reasonable to infer that there is a common ancestral protein. In the absence of sequence homology, on the other hand, the origin of two proteins is logically ambiguous. Functional and/or structural similarity may indicate either divergent evolution from a common precursor or convergent evolution from amino acid sequences arising independently.

In many circumstances, tell-tale sequence homology is not readily apparent. The most notable example is the 'nucleotide-binding domain' of four NAD-dependent dehydrogenases and certain other enzymes[1–5]. Although the amino acid sequences of lactate, alcohol and glyceraldehyde-3-phosphate dehydrogenases are not obviously homologous, when the sequences are aligned in such a way that common structural features coincide, a degree of sequence homology becomes evident, suggesting divergent evolution. In other cases of enzymes with similar structures but no apparent sequence homology, the question of divergent or convergent evolution has remained open[6–21]. The greater the structural correspondence between two proteins, as estimated by a variety of methods[4,16,22–28], the more likely are the two proteins to have evolved from the same precursor, but structural similarity *per se* is no proof of divergence. In some instances, two domains of a single protein have been observed to have structural similarity but not sequence homology. The fact that the polypeptide chains of the two domains are contiguous, even though their sequences are not related, has been taken as evidence for gene duplication, that is, a special case of divergent evolution[9,28–30].

In the case of bacteriophage T4 lysozyme (T4L) and hen egg-white lysozyme (HEWL), there are additional critical features common to the respective structures which strongly suggest that the two enzymes have diverged from a common precursor. Correspondence between HEWL and T4L can be seen at three levels: (1) the respective backbones, (2) the locations of saccharides bound in the respective active site clefts and (3) specific protein–substrate interactions and the probable mechanisms of catalysis.

* Present addresses: Department of Biochemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2H7 (W.F.A.); Max-Planck-Institut für Biochemie, D-8033 Martinsried bei München, FRG (S.J.R.).

Rossmann and Argos[4] showed that, allowing for certain insertions and deletions, the amino-terminal half of T4L is structurally similar to HEWL. Of the 129 residues in HEWL and 164 in T4L, 78 α-carbon atoms were found to be 'equivalent' and could be superimposed within a r.m.s. distance of 4.1 Å. Notwithstanding the structural correspondence, the minimum base change per codon for the 78 'equivalent' residues is 1.53, a value expected for two amino acid sequences chosen at random. Remington and Matthews[25] used a different method of comparing the two lysozyme structures and concluded that the agreement between them was significantly better than one would expect by chance. The part of T4L 'equivalent' to HEWL is shown in Fig. 1.

Recently, we have completed a study of the binding of a series of mono- and oligosaccharides to phage lysozyme which has permitted a detailed comparison of the respective active sites of T4L and HEWL[31,32]. We took the coordinates of the tetrasaccharide δ-lactone bound to HEWL[31], and, using the respective transformations between HEWL and T4L proposed by Rossmann and Argos[4] and Remington and Matthews[25], transformed these coordinates into the T4L coordinate frame. As summarized in Table 1, the centres of the four sugars agree within 1–2 Å with the saccharide binding sites determined experimentally for T4L. The experimental data provide striking confirmation of the sugar binding sites in T4L predicted by Rossmann and Argos, and support their conclusions concerning the relationship between the two lysozymes. Furthermore, there is remarkable agreement (Fig. 2) between parts of T4L and HEWL which are critically important for substrate binding and catalysis.

In the case of HEWL, Phillips and co-workers proposed a mechanism of action in which substrate strain in subsite D favours the formation of a carbonium ion intermediate which is stabilized by Asp 52, with Glu 35 acting as the proton donor[33,34]. Subsequent work suggests that electronic stabilization rather than strain may be more important in promoting catalysis[35,36]. In the case of T4L, we see a close parallel[31,32]. Here, model building suggests that, as with HEWL, a saccharide with the normal chair conformation cannot be accommodated in site D because of steric interference, but a saccharide in the sofa conformation, as in the proposed transition state, will fit into the D site. For HEWL, the critical close contact in site D is between $C_{(6)}$ and the backbone C=O of Gln 57. As can be seen in Fig. 2, an exactly analogous close approach occurs for T4L between $C_{(6)}$ and the backbone C=O of Gly 30. In addition, for both enzymes the NH of the adjacent peptide forms a characteristic hydrogen bond to the acetamido group of saccharide C which is thought to help promote the preferential binding of monosaccharides in site C.

Finally, Asp 20 of T4L is in virtually the identical position, relative to substrate, as Asp 52, and is ideally placed to stabilize the developing positive charge on the postulated carbonium ion intermediate. Glu 11 of T4L is close to, but not in exactly the same relative position, as Glu 35 of HEWL, and is salt linked to Arg 45, in contrast to the nonpolar environment of Glu 35. Assuming a slight conformational adjustment when substrate is bound, Glu 11 could move into the required position to act as

Table 1 Comparison of predicted and observed saccharide binding sites for phage lysozyme

	Distance between predicted and observed site (Å)	
Saccharide subsite	Prediction based on 78 'equivalent' α-carbon atoms[4,32]	Prediction based on the best agreement between 80 contiguous α-carbon atoms[25,32]
A	1.8	2.6
B	1.5	1.6
C	1.5	1.1
D	1.5	1.2

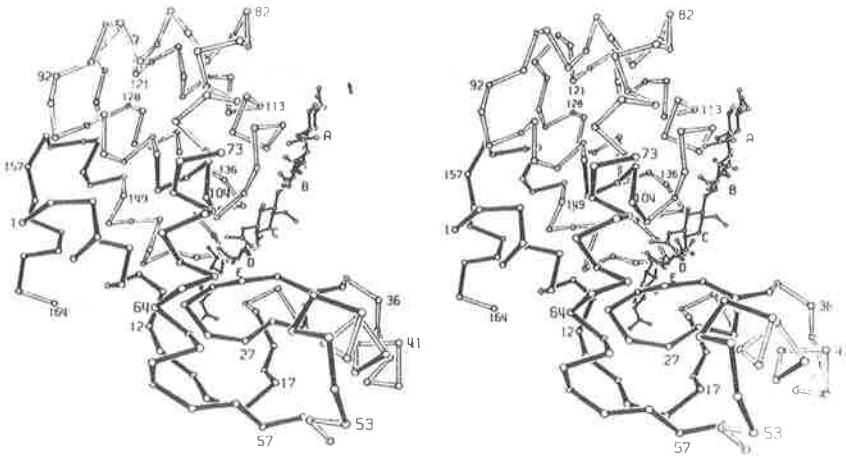


Fig. 1 Stereo view of the α-carbon backbone of bacteriophage T4 lysozyme showing the location of a bound oligosaccharide. The solid lines connect the 78 α-carbon atoms of phage lysozyme which have equivalent α-carbons in hen egg-white lysozyme[4].

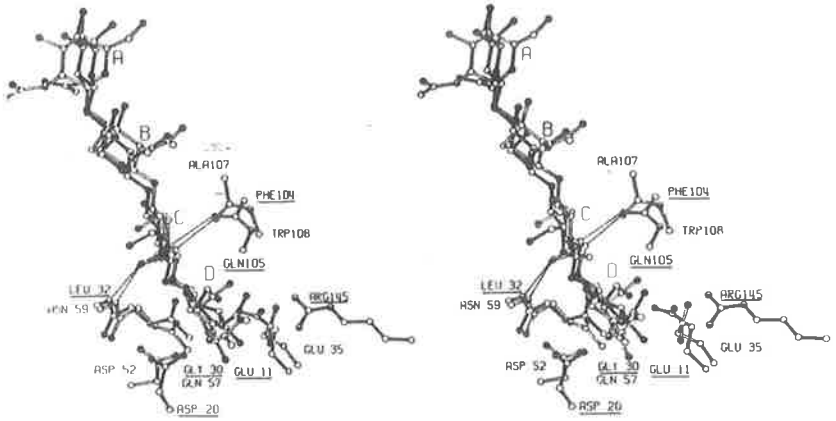


Fig. 2 Superposition of elements of the active site of hen egg-white lysozyme (open bonds), including a bound tetrasaccharide, on the corresponding elements of T4 phage lysozyme (solid bonds and names underlined).

the proton donor, as Glu 35 is postulated to do in HEWL. (Additional details are given in ref. 32.)

Thus, phage lysozyme and HEWL have similarities in their backbone conformations, in the locations of bound substrates and in the specific details of substrate binding and catalysis. These similarities strongly suggest that the two lysozymes have diverged from a common precursor. To argue for convergent, rather than divergent evolution, one would have to postulate that the similarities between the two lysozymes which we have enumerated all arose independently. We consider this unlikely.

The two lysozymes provide an example of the changes in proteins which can occur during evolution. In this case the only amino acids which seem to have retained their identities are Glu 11 (Glu 35) and Asp 20 (Asp 52). Other similarities between the respective lysozymes (Figs 1, 2) are at the level of the protein backbone rather than side-chain identity. As will be described in more detail elsewhere, the part of the upper lobe of T4L (Fig. 1), which has no counterpart in HEWL, seems to be involved in binding the cross-linking peptide of *Escherichia coli* cell walls. This peptide is required for substrates of T4L, but is not essential for HEWL.

After the submission of this manuscript, Jung *et al.*[37] reported the locations of the four exons of HEWL. Exon 2 codes for amino acids 28–32 and exon 3 for amino acids 82–108. Taken together, these two regions include the part of HEWL which is most similar to T4L and support the hypothesis[38] that exons may have been involved in the evolution of proteins. The same suggestion has been made independently by P. J. Artymiuk, C. C. F. Blake and A. E. Sippel (personal communication).

This work was supported in part by a Swiss NSF Fellowship to M.G.G., a NIH Postdoctoral Fellowship (GM 05972) to W.F.A., NIH research grants GM 20066 and GM 21967, NSF grant PCM-8014311 and a grant from the M. J. Murdock Charitable Trust. We thank Dr Colin Blake for communicating his manuscript before publication.

Received 15 September 1980; accepted 5 January 1981.

1. Blake, C. C. F. *Nature* **250,** 284–285 (1974).
2. Blake, C. C. F. *Nature* **267,** 482–483 (1977).
3. Rossmann, M. G., Moras, D. & Olsen, K. W. *Nature* **250,** 194–199 (1974).
4. Rossmann, M. G. & Argos, P. *J. molec. Biol.* **105,** 75–96 (1976).
5. Rossmann, M. G., Liljas, A., Branden, C. I. & Banaszak, L. J. in *The Enzymes* Vol. 11, 3rd edn (ed. Boyer, P. E.) 61–102 (Academic, New York, 1975).
6. Blake, C. C. F. & Evans, P. R. *J. molec. Biol.* **84,** 585–601 (1974).
7. Banks, R. D. *et al. Nature* **279,** 773–777 (1979).
8. Campbell, J. W., Watson, H. C. & Hodgson, G. I. *Nature* **250,** 301–303 (1974).
9. Ploegman, J. H., Drent, G., Kalk, K. H. & Hol, W. G. J. *J. molec. Biol.* **123,** 557–594 (1978).
10. Fletterick, R. J., Syguch, J., Semple, H. & Madsen, N. B. *J. biol. Chem.* **251,** 6142–6146 (1976).
11. Weber, I. T. *et al. Nature* **274,** 433–437 (1978).
12. Levine, M., Muirhead, H., Stammers, D. K. & Stuart, D. I. *Nature* **271,** 626–630 (1978).
13. Eichele, G. *et al. J. molec. Biol.* **133,** 161–180 (1979).
14. Borisov, V. V., Borisova, S. N., Sosfenov, N. I. & Vainshtein, B. K. *Nature* **284,** 189–190 (1980).
15. Matthews, D. A. *et al. J. biol. Chem.* **253,** 6946–6954 (1978).
16. Rossmann, M. G. & Argos, P. *J. molec. Biol.* **109,** 99–129 (1977).
17. Rossmann, M. G. & Argos, P. *J. biol. Chem.* **250,** 7525–7532 (1975).
18. Richardson, J. S., Richardson, D. C., Thomas, K. A., Silverton, E. W. & Davies, D. R. *J. molec. Biol.* **102,** 221–235 (1976).
19. Phillips, D. C., Sternberg, M. J. E., Thornton, J. M. & Wilson, I. A. *J. molec. Biol.* **119,** 329–351 (1978).
20. Argos, P., Rossmann, M. G. & Johnson, J. E. *Biochem. biophys. Res. Commun.* **75,** 83–86 (1977).
21. Blow, D. M., Irwin, M. J. & Nyborg, J. *Biochem. biophys. Res. Commun.* **76,** 728–734 (1977).
22. Schulz, G. E. & Schirmer, R. H. *Nature* **250,** 142–144 (1974).
23. Richardson, J. S. *Proc. natn. Acad. Sci. U.S.A.* **73,** 2619–2623 (1976).
24. Sternberg, M. J. E. & Thornton, J. M. *J. molec. Biol.* **105,** 357–382 (1976).
25. Remington, S. J. & Matthews, B. W. *Proc. natn. Acad. Sci. U.S.A.* **75,** 2180–2184 (1978).
26. Remington, S. J. & Matthews, B. W. *J. molec. Biol.* **140,** 77–99 (1980).
27. Schulz, G. E. *J. molec. Evol.* **9,** 339–342 (1977).
28. Schulz, G. E. *J. molec. Biol.* **138,** 335–347 (1980).
29. McLachlan, A. D. *J. molec. Biol.* **128,** 49–79 (1979).
30. McLachlan, A. D. in *Protein Folding* (ed. Jaenicke, R.) 79–99 (Elsevier, Amsterdam, 1980).
31. Anderson, W. F., Grütter, M. G., Remington, S. J. & Matthews, B. W. *J. molec. Biol.* (in the press).
32. Matthews, B. W., Remington, S. J., Grütter, M. G. & Anderson, W. F. *J. molec. Biol.* (in the press).
33. Ford, L. O., Johnson, L. N., Machin, P. A., Phillips, D. C. & Tjian, R. *J. molec. Biol.* **88,** 349–371 (1974).
34. Blake, C. C. F. *et al. Proc. R. Soc.* **B167,** 378–388 (1967).
35. Warshel, A. Levitt, M. *J. molec. Biol.* **103,** 227–249 (1976).
36. Warshel, A. *Proc. natn. Acad. Sci. U.S.A.* **75,** 5250–5254 (1978).
37. Jung, A., Sippel, A. E., Grez, M. & Schutz, G. *Proc. natn. Acad. Sci. U.S.A.* **77,** 5759–5763 (1980).
38. Gilbert, W. *Nature* **271,** 501 (1978).

Printed in Great Britain by Macmillan Production Ltd., Basingstoke, Hampshire.

Reprinted from Nature, Vol. 303, No. 5920, pp. 828-831, 30 June 1983
© Macmillan Journals Ltd., 1983

Goose lysozyme structure: an evolutionary link between hen and bacteriophage lysozymes?

M. G. Grütter*, L. H. Weaver & B. W. Matthews

Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oregon 97403, USA

During evolution, the amino acid sequence of a protein is much more variable and changes more rapidly than its tertiary structure. Given sufficient time, the amino acid sequences of proteins derived from a common precursor may alter to the point that they are no longer demonstrably homologous. The ability to make meaningful comparisons between such distantly related proteins must therefore come primarily from structural homology, and only secondarily (if at all) from sequence homology[1]. On the other hand, structural homology in the absence of sequence homology might be attributed to convergent rather than divergent evolution. (A common fold might be dictated by functional or folding requirements.) We have previously argued, on the basis of structural and functional similarities, that the lysozymes of hen egg-white and bacteriophage T4 have a common evolutionary precursor, even though their amino acid sequences have no detectable similarity[2,3]. Here we report the structure of the lysozyme from Embden goose, a representative of a third class of lysozymes[4-10] that has no sequence homology[11,12] (or perhaps very weak homology[13]) with either the hen egg-white or the phage enzyme. The structure of goose egg-white lysozyme has striking similarities to the lysozymes from hen egg-white and bacteriophage T4. However, some parts of goose lysozyme resemble hen lysozyme while other parts correspond only to the phage enzyme. The nature of the structural correspondence strongly suggests that all three lysozymes evolved from a common precursor.

The structure determination of goose lysozyme[14] by isomorphous replacement[1,15] was based on four heavy-atom replacements supplemented by an additional data set collected[16,17] for one of the derivatives at lower substitution. Electron density maps calculated at 2.8 Å and 3.2 Å resolution showed several helices and allowed extended segments of the polypeptide backbone to be followed, but also contained many ambiguous regions. In order to trace the complete backbone of the molecule we alternated model building[18,19] with cycles of crystallographic refinement. Details will be given elsewhere. The model of the structure used for the structure comparisons described here was refined to a crystallographic residual of 0.28 at 2.8 Å resolution. Subsequent refinement has reduced the residual to 0.25 at 2.1 Å resolution.

The polypeptide backbone of goose egg-white lysozyme (GEWL) was compared with those of hen egg-white lysozyme[20,21] (HEWL) and T4 phage lysozyme (ref. 22, and L.H.W., T. Gray & B.W.M., unpublished results) (T4L) by the methods of Rossmann and Argos[23,24] and Remington and Matthews[25,26]. Application of these methods to HEWL and T4L have been described previously[2,3,23,25].

The three-dimensional structure of GEWL has striking similarities to both HEWL and T4L (Table 1, Figs 1, 2). Apart from the amino-terminal region, residues 1–46, essentially every residue of GEWL has a counterpart either in HEWL or T4L or both (Fig. 1). There are 90 spatially equivalent α-carbons in GEWL and HEWL, and 91 in GEWL and T4L. For the residues that structurally correspond, the minimum base change per codon is, in each case, close to the random-sequence value of 1.5, confirming the overall lack of amino acid sequence homology between any pair of the proteins.

* Present address: Abt. Biophysikalische Chemie, Biozentrum der Universität Basel, CH-4056 Basel, Switzerland.

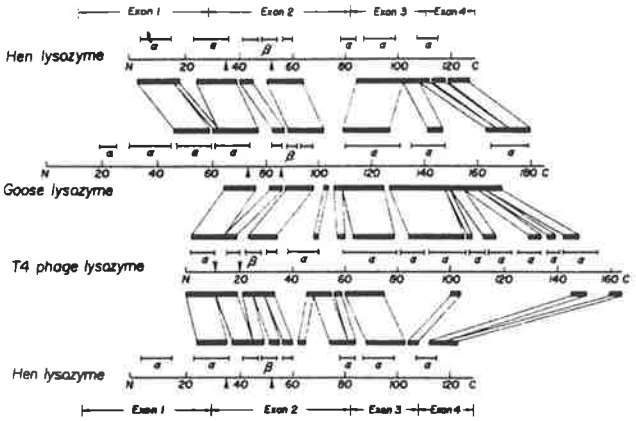


Fig. 1 Structural correspondence between goose, hen and phage lysozymes. The connected solid bars indicate parts of the polypeptide backbones that structurally correspond when the lysozymes are compared in pairs by the method of Rossmann and Argos[23,24]. The locations of the α-helices and β-strands are shown as are the exons of hen egg-white lysozyme. The arrowheads show the locations of the residues that are presumed to be involved in catalysis, *viz* Glu 35 and Asp 52 of HEWL, Glu 11 and Asp 20 of T4L, and Glu 73 and Asp 86 of GEWL.

There are two amino acids in HEWL that are thought to be of prime importance in catalysis, namely Glu 35 and Asp 52[27,28]. The respective counterparts in T4L are Glu 11 and Asp 20[29,30]. In the superposition of HEWL on GEWL (Figs 1, 2*a*) the α-carbon of Glu 35 (HEWL) is 0.8 Å from the α-carbon of Glu 73 (GEWL). Also, for T4L aligned with GEWL (Figs 1, 2*b*) the α-carbons of Glu 11 (T4L) and Glu 73 (GEWL) are 1.7 Å apart. Obviously, Glu 73 of GEWL is a counterpart to Glu 35 in HEWL and Glu 11 in T4L. Similarly, Asp 86 of GEWL corresponds to Asp 52 of HEWL and Asp 20 of T4L, although here the agreement is not so precise (see also ref. 31). According to the Rossmann–Argos procedure, Asp 86 of GEWL is spatially equivalent to Lys 19 rather than Asp 20 of T4L. In the case of HEWL, Asp 86 (GEWL) is topologically equivalent to Asn 44 (Fig. 1) rather than Asp 52, but its α-carbon is only 5.7 Å from that of the latter residue. All three lysozymes have their catalytic aspartates within their common three β-strand unit. For GEWL and T4L, the aspartate is in the first β-strand, whereas Asp 52 of HEWL is in the middle strand. Glu 86 and Asp 73 of GEWL are located on opposite sides of the presumptive active-site cleft. Their carboxyls are about 8 Å apart in the present mode (7–8 Å in HEWL and T4L). Supporting evidence for the location of the active site comes from ($hk0$) and difference Fourier projections (data not

Table 1 Lysozyme structural homologies

	GEWL and HEWL	GEWL and T4L	T4L and HEWL
No. of equivalent residues (N_e)	90	91	74
$100\,N_e/N_1$	70%	55%	57%
$100\,N_e/N_2$	49%	49%	45%
R.m.s. discrepancy	3.2 Å	3.2 Å	3.8 Å
Minimum base change per codon for equivalent residues	1.41	1.44	1.46

The table summarizes the comparisons of the three lysozymes by the method of Rossmann and Argos[23,24]. The procedure uses a complicated algorithm that can result in slightly different sets of equivalent residues, depending on the starting parameters. For this reason, little significance should be attached to small differences in the number of equivalent residues. N_1 and N_2 are, respectively, the number of amino acids in the smaller and larger protein being compared. HEWL has 129 residues, T4L 164 residues and GEWL 185 residues.

SECTION F. METHODS FOR PROTEIN STRUCTURE DETERMINATION

Reprinted from *Acta Crystallographica*, Vol. 20, Part 1, January 1966

PRINTED IN DENMARK

Acta Cryst. (1966). **20**, 82

The Extension of the Isomorphous Replacement Method to Include Anomalous Scattering Measurements

BY B. W. MATTHEWS

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge, England

(*Received* 3 *May* 1965)

The combination of isomorphous replacement and anomalous scattering data in phase determination of non-centrosymmetric reflexions is discussed. Expressions are derived which take into account the relative reliabilities of the experimental observations and enable any combination of replacement atoms to be used in the phase determination.

North (1965) has proposed an improved method for the combination of isomorphous replacement and anomalous scattering data in phase determination of non-centrosymmetric reflexions which makes use of the greater intrinsic accuracy of the anomalous scattering measurements. North's formulation leads to alternative expressions for the phase probability distribution, and North leaves it open as to which of these expressions should be used in practice. The purpose of this communication is to show that the problem may be exam-

ined in a general way, and to extend the method so that anomalous scattering measurements from isomorphous crystals differentiated by any combination of 'heavy atoms' may be included in the phase determination.

Blow & Crick's (1959) treatment of errors in the isomorphous replacement method can be illustrated by a phase diagram in which the three vectors **F**, $\mathbf{F}_H$ and $\mathbf{f}_c$ do not, in general, form a closed triangle. (F and F_H are the observed structure amplitudes of the isomorphous 'parent' and 'derivative' compounds, and $\mathbf{f}_c$ is the calculated structure factor of the atoms which differentiate the two compounds). By considering the lack of closure of this triangle and estimating the error in its determination, the probability of F having a given phase may be obtained. Similarly, to illustrate the treatment of errors in phase determination using both isomorphous replacement and anomalous scattering data the phase diagram shown in Fig. 1 can be constructed. F_H+ and F_H- are the observed structure amplitudes of a Friedel pair of derivative reflexions; F is the corresponding structure amplitude of the parent compound which is assumed to have negligible anomalous scattering; **f** and **δ** are the calculated components of the combined 'heavy atom' scattering and are discussed in more detail later, and $\mathbf{F}_c$ is defined as the resultant of **F** and **f**. ε_+ represents the lack of closure of the $F_H{}^+$ vector and ε_- the lack of closure of F_H-. The phase of the heavy atom structure factor has been denoted by ψ in place of North's α. This is in keeping with the nomenclature used in a related paper (Matthews, 1966) dealing with the location of anomalously scattering heavy atoms in protein structures. Blow & Rossmann (1961) have shown that the overall phase probability distribution using both isomorphous replacement and anomalous scattering data can be written as

$$P(\varphi) \propto \exp\{-(\varepsilon_+^2+\varepsilon_-^2)/2E^2\} \qquad (1)$$

where E is the total r.m.s. error in determining ε_+ and ε_-. As North points out, this procedure does not take into account the inherently greater accuracy of the anomalous differences. This accuracy can be taken into account by rewriting the lack of closure terms in a different form, *viz.*

$$\varepsilon_+^2+\varepsilon_-^2 = \tfrac{1}{2}[(\varepsilon_++\varepsilon_-)^2+(\varepsilon_+-\varepsilon_-)^2] = \tfrac{1}{2}[x_1(\varphi)^2+x_1'(\varphi)^2] . \qquad (2)$$

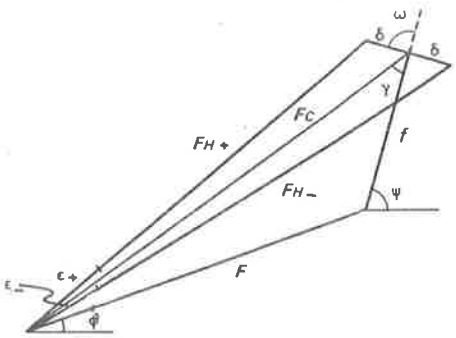


Fig. 1. Vector diagram showing lack of closure with combined isomorphous replacement and anomalous scattering data.

Having thus defined $x_1(\varphi)$ and $x_1'(\varphi)$ it will now be shown that these quantities can be given simple physical interpretations. Let us define $F_H'=\frac{1}{2}(F_H+ + F_H-)$. If there were no anomalous scattering F_h' would equal F_H the derivative structure amplitude. Since δ is normally small compared with F_c, it follows from Fig. 1 that

$$x_1(\varphi) = (\varepsilon_+ + \varepsilon_-) = 2(F_c - F_H') . \qquad (3)$$

$x_1(\varphi)$ can therefore be considered as the 'total isomorphous replacement lack of closure'. Analogously, $x_1'(\varphi)$ which depends on the anomalous scattering difference $\Delta_H=(F_H+ - F_H-)$ can be considered as the 'anomalous scattering lack of closure'. From the definitions of $x(\varphi)$ and $x'(\varphi)$ given by North, it follows that $x_1'(\varphi) = -x'(\varphi)$ and $x_1(\varphi)=2x(\varphi)$. As discussed by North, the error in determining $x_1'(\varphi)$ will tend to be less than that for $x_1(\varphi)$. If E_1 and E_1' are defined as estimates of the total r.m.s. error in determining $x_1(\varphi)$ and $x_1'(\varphi)$ respectively, then the overall relative probability $P(\varphi)$ of any particular phase φ being correct is

$$P(\varphi) = \exp\left\{-\left(\frac{x_1(\varphi)^2}{2E_1^2}+\frac{x_1'(\varphi)^2}{2E_1'^2}\right)\right\} . \qquad (4)$$

Clearly $P(\varphi)$ may be written

$$P(\varphi) = P_{\text{is}}(\varphi) \, . \, P_{\text{an}}(\varphi) \qquad (5)$$

where $P_{\text{is}}(\varphi)$ is the phase probability distribution obtained using isomorphous replacement data, and $P_{\text{an}}(\varphi)$ is the probability distribution obtained from anomalous scattering data.

$$P_{\text{is}}(\varphi) = \exp\{-2(F_c - F_H')^2/E_1^2\} . \qquad (6)$$

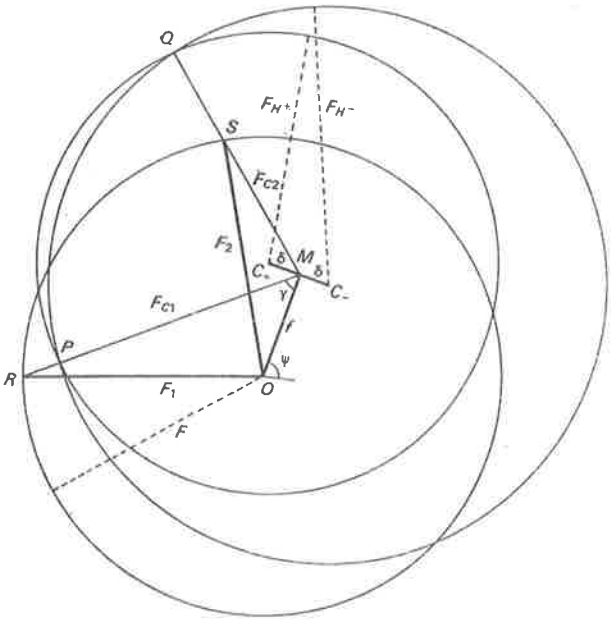


Fig. 2. Phase diagram illustrating phase determination using anomalous scattering data. F_{c_1} and F_{c_2} give the most probable phases of F_c, and F_1 and F_2 are the corresponding most probable values of F.

Since $E_1=2E$, this is equivalent to North's equations (1) and (2).

In evaluating $P_{an}(\varphi)$ we first consider the case in which f and δ in Fig. 1 are orthogonal. It will be shown later that $P_{an}(\varphi)$ may be evaluated without this assumption; however, this special case illustrates the general method and can be used if all the heavy atoms have the same relative anomalous scattering. If then $\omega=\pi/2$ and we again use the fact that δ is normally small compared with F_c, we have

$$\varepsilon_+-\varepsilon_-=-F_{H-}+F_{H+}-2\delta\sin\gamma$$
$$=-\Delta_H-2\delta\sin\gamma\,. \quad (7)$$

Now

$$\sin\gamma=(F/F_c)\sin(\psi-\varphi)$$
$$=(F/F_cf)(b\cos\varphi-a\sin\varphi) \quad (8)$$

where $a=f\cos\psi$ and $b=f\sin\psi$ [see also (14) and (14a)]. Substitution in (7) leads to

$$\varepsilon_+-\varepsilon_-=-\Delta_H-(2F\delta/F_cf)(b\cos\varphi-a\sin\varphi)\,. \quad (9)$$

This result is similar to North's equation (6), but the assumption that the phase triangle should be closed at phase φ has not been needed. From North's derivation it is not clear whether F_c should be replaced by F'_H, and although North suggests that it appears to be valid to do so, a more correct procedure in evaluating the probability distribution for the phase of F is to use

$$P_{an}(\varphi)=\exp\left\{-\frac{1}{2E_1'^2}[-\Delta_H-(2F\delta/F_cf)(b\cos\varphi-a\sin\varphi)]^2\right\} \quad (10)$$

where E'_1 equals North's E' and $F_c=[F^2+f^2+2F(a\cos\varphi+b\sin\varphi)]^{\frac{1}{2}}$. The physical interpretation of this result is illustrated by the construction of Fig. 2. The calculated vectors $\mathbf{f}$ and $\boldsymbol{\delta}$ are plotted first, and then circles of radii F_{H+} and F_{H-} are drawn with centres C_+ and C_-. The points of intersection of these circles give the most probable directions of F_c (*i.e.* along MP and MQ, the alternative solutions symmetrical about $\boldsymbol{\delta}$). It may be noted that if the magnitude of $F'_H=\frac{1}{2}(F_{H+}+F_{H-})$ were to change, $\Delta_H=(F_{H+}-F_{H-})$ remaining constant, P and Q would lie on a hyperbola the direction of whose asymptotes is given by $\sin\gamma=-\Delta_H/2\delta$. Using the approximation that δ is small compared with MP and MQ is equivalent to assuming that the hyperbola can be replaced by its asymptotes. The directions of the asymptotes are determined only by the *difference* between F_{H+} and F_{H-}, and are independent of their sum, *i.e.* the solutions are independent of any experimental error in the magnitude of F'_H. The most probable directions of F_c having been found, the most probable phases of F can now be found by drawing a circle of radius F about the origin. This circle represents the locus of all possible phases of F, and the most probable phases are given by OR and OS where R and S are the intersections of the F_c direction vectors with the F circle.

To show the effect of using F'_H instead of F_c in calculating $P_{an}(\varphi)$ and hence the joint phase probability distribution, two examples are given. In the first (Fig. 3) the isomorphous replacement and anomalous scattering data give a 'good' phase determination in the sense that $P_{is}(\varphi)$ and $P_{an}(\varphi)$ have maxima which approximately coincide. In the second (Fig. 4) the isomorphous replacement and anomalous scattering data give somewhat conflicting phase determinations and a compromise has to be made. In the calculations E' was put equal to a third of E. As might be expected the alternative calculations using F'_H and F_c do not lead to widely diverging results; however, in both cases the form of the $P_{an}(\varphi)$ distribution changes appreciably. The difference is rather striking in the case of Fig. 3(b) where the use of F'_H results in a bimodal distribution for $P_{an}(\varphi)$ whereas using F_c gives a unimodal distribution as would be expected from Fig. 3(a). For comparison, the most probable phases of F *obtained with the use of* $P_{an}(\varphi)$ *only* are shown on the phase diagrams. Clearly the use of F_c gives a physically more acceptable result; *i.e.* the direction of the most probable F is such that F_c is directed toward the points of closure of the F_{H+} and F_{H-} phase circles. In the 'good' phase determination neither the overall most probable phase nor the 'best' phase is changed very much, but in the poorer determination the 'best' phase is changed by 11 degrees and the most probable phase by 20 degrees. It may be noted that the treatment here, following that of Blow & Crick, assumes that any errors in F may be considered as lying in F'_H. Using F'_H in place of F_c in evaluating $P_{an}(\varphi)$ is *not* equivalent to assuming that all the errors reside in F rather than in F'_H. If this assumption were made, then in Fig. 2 the most probable phases of F would be given by OP and OQ rather than OR and OS. From Fig. 2 [and from Figs. 3(a) and 4(a)] it is clear that the use of either assumption would generally lead to much the same values for the most probable phases.

Phase determination with the use of heavy atoms of different types

Suppose that two isomorphous crystal structures are differentiated by N 'heavy atoms' which may or may not exhibit significant anomalous scattering. Let the position vector of the nth atom be $\mathbf{r}_n$ and its scattering factor

$$f_n=f'_n+if''_n\,.$$

For a reflexion with indices (h, k, l) the calculated structure factor of the N atoms is

$$\mathbf{f}_c(\mathbf{h})=\sum_{n=1}^{N}f'_n(\mathbf{h})\exp(2\pi i\mathbf{h}\,.\,\mathbf{r}_n)+i\sum_{n=1}^{N}f''_n(\mathbf{h})\exp(2\pi i\mathbf{h}\,.\,\mathbf{r}_n) \quad (11)$$
$$=\mathbf{f}(\mathbf{h})+\boldsymbol{\delta}(\mathbf{h})\,. \quad (12)$$

For simplicity we write

$$\mathbf{f}(\mathbf{h})+\boldsymbol{\delta}(\mathbf{h})=\mathbf{f}+\boldsymbol{\delta}$$

and this defines f and δ in Fig. 1.

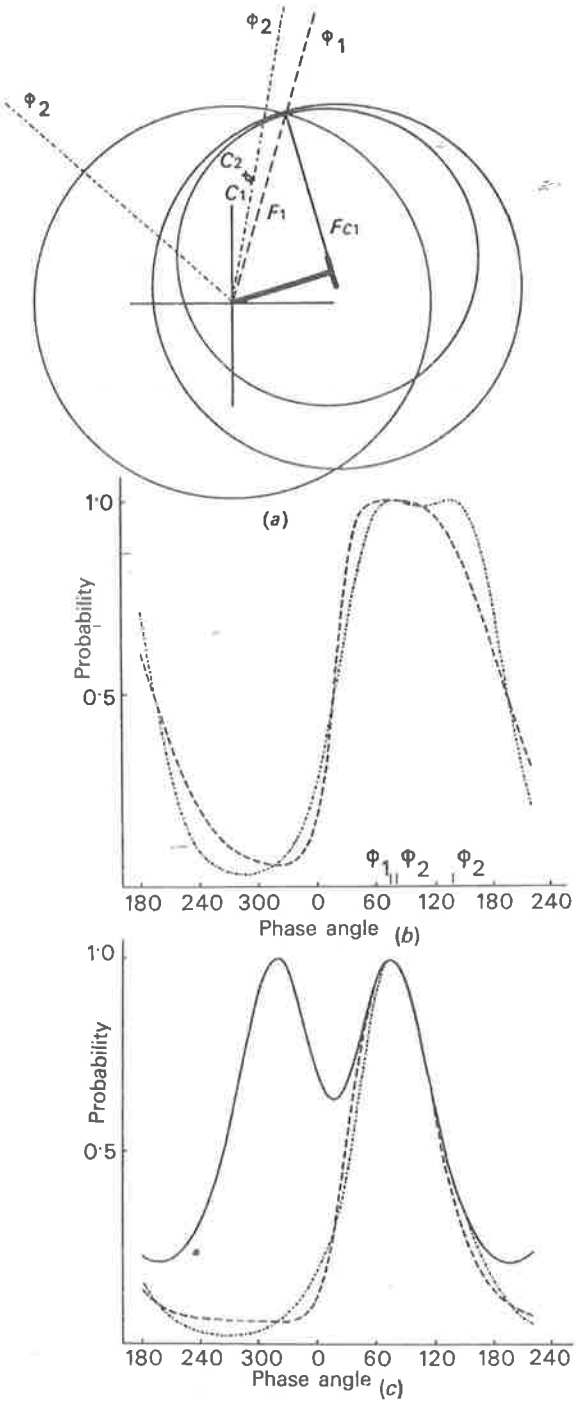


Fig. 3. (*a*) Phase diagram for a 'good' phase determination. φ_1 and φ_2 are the most probable phases of F obtained with the use of $P_{an}(\varphi)$ only. C_1 and C_2 are the centroids of the joint probability distributions $P(\varphi)$. The subscripts 1 and 2 refer to calculations using respectively F_c and F_H' in evaluating $P_{an}(\varphi)$. (*b*) Alternative anomalous scattering phase probability distributions $P_{an}(\varphi)$. The broken line is obtained with F_c, the chain line with F_H'. (*c*) Combination of isomorphous replacement probability distribution $P_{is}(\varphi)$ and anomalous scattering probability distribution $P_{an}(\varphi)$ to give joint phase probability $P(\varphi)$. $P_{is}(\varphi)$ is drawn solid and the alternative joint distribution $P(\varphi)$ is drawn similarly to the $P_{an}(\varphi)$ curve in (*b*) from which it was derived.

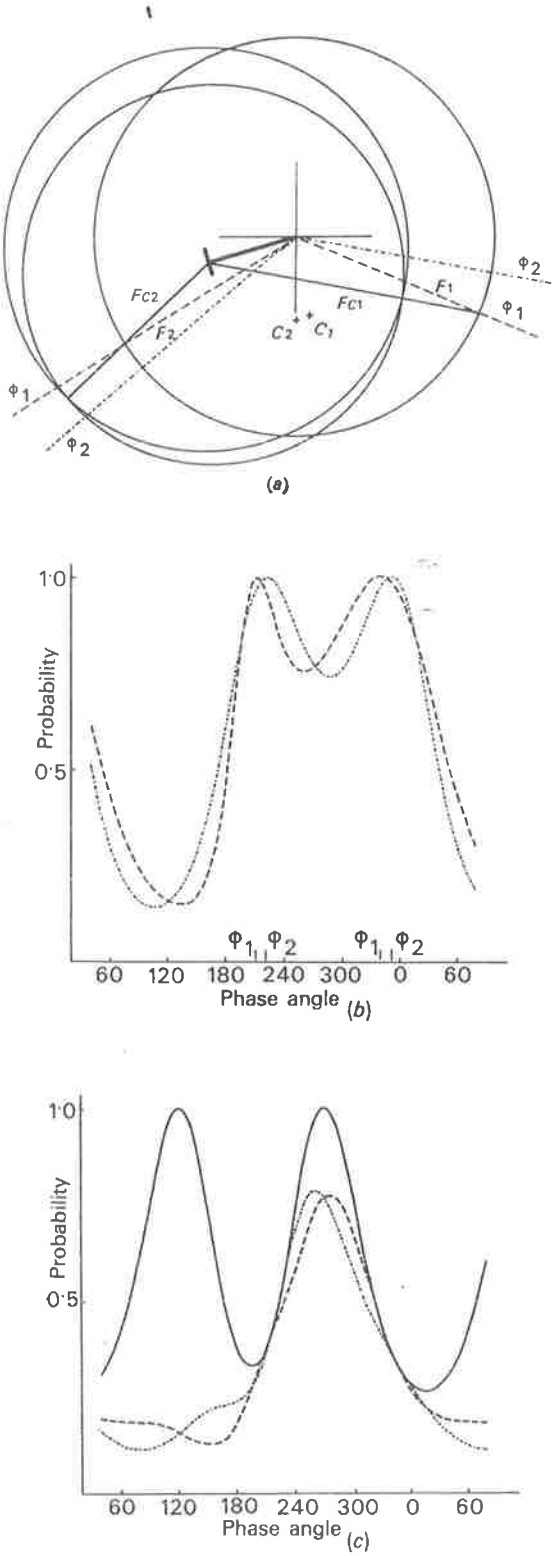


Fig. 4. (*a*) Phase diagram for a 'poor' phase determination. Otherwise as for Fig. 3(*a*). (*b*) Alternative anomalous scattering phase probability distribution $P_{an}(\varphi)$ for the 'poor' phase determination. (*c*) Combination of $P_{is}(\varphi)$ with alternative $P_{an}(\varphi)$ distributions shown in (*b*) to give alternative combined phase probability distributions $P(\varphi)$.

If, for this reflexion, the N atoms all have the same ratio $k=f'_n/f''_n$, *i.e.* they are all 'of the same type' (Rossmann, 1961) then

$$\mathbf{f}_c(\mathbf{h})=\mathbf{f}+\boldsymbol{\delta}=\mathbf{f}+ik^{-1}\mathbf{f} \tag{13}$$

and $\mathbf{f}$ and $\boldsymbol{\delta}$ are orthogonal. In this case the magnitude of $\boldsymbol{\delta}$ can be found directly from that of $\mathbf{f}$ and the phase determination carried out as outlined in the previous section. In a more general case the isomorphous derivative may differ from the parent in such a way that the ratio f'_n/f''_n will not be a constant. For example, the derivative may contain heavy atoms of different atomic species, or it may contain heavy atoms covalently bonded to the parent structure through atoms of known position but lower atomic number. Also, in the case of complex ions, which are often used as 'heavy atoms', the magnitude of the ratio f'/f'' may change as a function of the angle of scattering (Matthews, 1966). A method of treating such cases will now be described.

It was shown above [equation (12)] that the calculated heavy atom structure factor $\mathbf{f}_c$ could be written as the sum of the vectors $\mathbf{f}$ and $\boldsymbol{\delta}$. Let ω be the angle between these components (see Fig. 1). It is customary to express $\mathbf{f}$ in the form

$$\mathbf{f}=a+ib \tag{14}$$

where

$$a=\sum_{n=1}^{N} f'_n(\mathbf{h}) \cos (2\pi i\, \mathbf{h} \,.\, \mathbf{r}_n)$$

and

$$b=\sum_{n=1}^{N} f'_n(\mathbf{h}) \sin (2\pi i\, \mathbf{h} \,.\, \mathbf{r}_n)\,. \tag{14a}$$

Similarly, $\boldsymbol{\delta}$ may be written

$$\boldsymbol{\delta}=a'+ib' \tag{15}$$

where, from the definition of $\boldsymbol{\delta}$,

$$a'=-\sum_{n=1}^{N} f''_n(\mathbf{h}) \sin (2\pi i\mathbf{h} \,.\, \mathbf{r}_n)$$

and

$$b'=\sum_{n=1}^{N} f''_n(\mathbf{h}) \cos (2\pi i\mathbf{h} \,.\, \mathbf{r}_n)\,. \tag{15a}$$

In deriving the expression for $P_{\mathrm{is}}(\varphi)$, the phase probability distribution using isomorphous replacement, no assumption was made about the value of ω; therefore equation (6) can still be used, and we only need reconsider the derivation of $P_{\mathrm{an}}(\varphi)$.

Previously the anomalous scattering lack of closure was given by equation (7). This result can be rewritten in the more general form

$$\varepsilon_+-\varepsilon_-=-\Delta_H+2\delta\cos(\gamma+\omega)\,. \tag{16}$$

Thus, in the most general case, the expression for $P_{\mathrm{an}}(\varphi)$ the phase probability distribution using anomalous dispersion data is

$$P_{\mathrm{an}}(\varphi)=\exp\left\{-\frac{1}{2E'^2_1}[-\Delta_H+2\delta\cos(\gamma+\omega)]^2\right\}. \tag{17}$$

This may be evaluated by using the following relationships which are readily verified.

$$\begin{aligned}
\sin\gamma &= [F(b\cos\varphi-a\sin\varphi)]/(F_c f)\\
\cos\gamma &= [f^2+F(a\cos\varphi+b\sin\varphi)]/(F_c f)\\
\sin\omega &= (ab'-a'b)/(\delta f)\\
\cos\omega &= (aa'+bb')/(\delta f)\,.
\end{aligned} \tag{18}$$

It may be noted that in practice if a' and b' are determined at the same time as a and b, little extra computing will be needed, and in fact the evaluation of the phase probability distribution in the most general case involves a comparatively small increase in computation above that involving in using the isomorphous replacement method alone.

It is a pleasure to acknowledge helpful discussions with Dr D. M. Blow and Dr R. Diamond. I should also like to thank Miss J. Collard for her assistance and for preparing the diagrams.

References

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794.
Blow, D. M. & Rossmann, M. G. (1961). *Acta Cryst.* **14**, 1195.
Matthews, B. W. (1966). In the press.
North, A. C. T. (1965). *Acta Cryst.* **18**, 212.
Rossmann, M. G. (1961). *Acta Cryst.* **14**, 383.

Reprinted from *Acta Crystallographica*, Vol. 20, Part 2, February 1966

PRINTED IN DENMARK

Acta Cryst. (1966). **20**, 230

The Determination of the Position of Anomalously Scattering Heavy Atom Groups in Protein Crystals

BY B. W. MATTHEWS

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge, England

(*Received* 6 *May* 1965)

Methods are proposed by which isomorphous replacement and anomalous scattering measurements may be combined to locate anomalously scattering heavy atoms in protein structures. Both Patterson and Fourier methods are discussed and examples are given to illustrate the use of new methods.

The examples show how the relative coordinates of heavy atom groups in different derivatives may be determined, and how the absolute configuration of these coordinates may be established.

Introduction

The only proven method for determining the detailed structure of protein crystals is the technique of isomorphous replacement with heavy atom groups, first introduced by Green, Ingram & Perutz (1954).

In practice it is difficult to find satisfactory isomorphous derivatives and it therefore becomes important to make full use of any derivative which may be available. In particular, it may be worth while making separate measurements of the intensities of Friedel related reflexions so that the differences due to the anomalous scattering of the heavy atoms may be used to supplement the phase information obtained from the normal isomorphous replacement method (Blow, 1958; Blow & Rossmann, 1961; North, 1965; Matthews, 1966). The purpose of this communication is to show how such anomalous scattering measurements may be combined with the isomorphous replacement data to assist in the determination of the position of the heavy atom groups in the crystal (see also Kartha & Parthasarathy, 1965)*. Both Patterson and Fourier methods are discussed. Examples are given to illustrate the use of the new methods and to compare the results with those obtained by previous methods.

Theory

(*a*) *The difference Patterson function*

Let $F(\mathbf{h})$ be the observed *hkl* structure factor of the native protein crystal. In general, proteins consist of light atoms which give rise to very small anomalous scattering effects, and it will be assumed that such effects may be ignored. For the isomorphous derivative crystal which contains heavy atoms, the anomalous scattering effects will in general be significant and we define $F_H+=F_H(\mathbf{h})$ and $F_H-=F_H(\bar{\mathbf{h}})$ to be the observed structure amplitudes of a Friedel pair of reflexions of the substituted crystal.

Suppose that there are a total of N heavy atoms which differentiate the native and derivative structures. If the nth of these atoms has position vector $\mathbf{r}_n$ and scattering factor

$$f_n=f'_n+if''_n$$

then for the (h, k, l) plane the structure factor of the N atoms is

$$\mathbf{f}_c(\mathbf{h})=\sum_{n=1}^{N} f'_n(\mathbf{h}) \exp(2\pi i\mathbf{h}\,.\,\mathbf{r}_n) + i\sum_{n=1}^{N} f''_n(\mathbf{h}) \exp(2\pi i\mathbf{h}\,.\,\mathbf{r}_n) \quad (1)$$

$$=\mathbf{f}(\mathbf{h})+\boldsymbol{\delta}(\mathbf{h}) \quad (2)$$

If the N heavy atoms are 'of the same type' (Rossmann, 1961), *i.e.* the ratio $k=f'_n/f''_n$ is a constant, then

$$\mathbf{f}_c(\mathbf{h})=\mathbf{f}(\mathbf{h})+ik^{-1}\mathbf{f}(\mathbf{h})\,. \quad (3)$$

Thus, in this case, $\mathbf{f}$ and $\boldsymbol{\delta}$ are orthogonal, and $\delta=k^{-1}f$. Let us define $\mathbf{F}_H$ to be the structure factor of the derivative crystal in the absence of anomalous scattering. The relationship between $\mathbf{F}$, $\mathbf{F}_H$ and $\mathbf{f}$ is shown by the simple vector triangle in Fig. 1. When there is appreciable anomalous scattering the phase diagram can be conveniently represented by Fig. 2 (Ramachandran & Raman, 1956; Rossmann, 1961). In general, δ will be small compared with F_H, and therefore to a good approximation

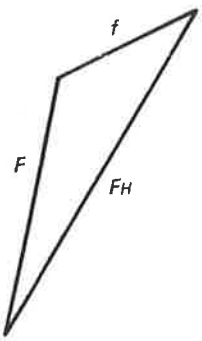


Fig. 1. Vector diagram showing the relationship between structure factors in the absence of anomalous scattering.

* The methods proposed here were formulated independently of this work.

$$F_H = \tfrac{1}{2}(F_{H+} + F_{H-}) .$$

Several methods of determining the positions of heavy atoms in isomorphous protein structures have been suggested (see Rossmann, 1960, for a summary). Rossmann favours the use of a Patterson-like function with coefficients $(F_H - F)^2$ (Blow, 1958).

In a non-centrosymmetric structure the phases of **F** and **f** will, in general, not be correlated and therefore in many cases the approximation that

$$(F_H - F)^2 = f^2 \qquad (4)$$

will be a poor one. However, if $(F_H - F)$ is large it follows that **F**, $\mathbf{F}_H$ and **f** must be nearly collinear and in such cases the approximation will be most accurate. Therefore the largest terms in the summation are relatively error-free. In practice, if a large number of terms are included, this method can give very satisfactory results. On the other hand there will be many cases when the $(F_H - F)^2$ function is difficult to interpret, and in such cases it would clearly be desirable to have a better approximation to f^2 than that given by (4).

The location of anomalous scatterers has also been discussed by Blow (1957) and Rossmann (1961). In this case similar arguments to those given above are used to show that the positions of the anomalous scatterers can be determined by using another Patterson-like function with coefficients $(F_{H+} - F_{H-})^2$. This function will suffer from the same limitations as the $(F_H - F)^2$ function but will again have the property that the largest coefficients will tend to be most accurate.

A method will now be described by which the observed isomorphous replacement and anomalous scattering data may be combined to give an accurate estimate of f, the structure factor of the heavy atom replacement group.

Fig. 2 illustrates the relationship between **F**, $\mathbf{F}_{H+}$ and $\mathbf{F}_{H-}$ and the scattering of the heavy atoms. Let α, β and γ be the angles shown. It was shown that if the heavy atoms are 'of the same type' then $\delta = k^{-1}f$ and $\boldsymbol{\delta}$ and **f** are at right angles. Since δ is generally small compared with F_H, it follows from Fig. 2 that

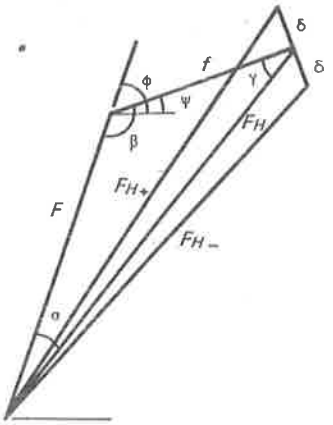


Fig. 2. Vector diagram showing the relationship between structure factors in the presence of anomalous scattering.

$$F_{H+} - F_{H-} = 2\delta \sin\gamma = 2k^{-1}f \sin\gamma . \qquad (5)$$

Now

$$\sin\alpha = (f \sin\gamma)/F$$

so that

$$\sin\alpha = f(F_{H+} - F_{H-})/2k^{-1}fF$$
$$= k(F_{H+} - F_{H-})/2F . \qquad (6)$$

Hence

$$f^2 = F^2 + F_H^2 - 2FF_H \cos\alpha$$
$$= F^2 + F_H^2 - 2FF_H\{1 - [k(F_{H+} - F_{H-})/2F]^2\}^{\frac{1}{2}} . \qquad (7)$$

Thus, if the ratio $k = f'_n/f''_n$ is known it is in principle possible to use the observed data to calculate a difference Patterson synthesis with coefficients equal to f^2. In practice the accuracy with which f can be determined will depend on the reliability of the various experimental measurements.

To show the relation between (7) and the result of Kartha & Parthasarathy (1965) the assumption is made that α in Fig. 2 is small. From (6) $k(F_{H+} - F_{H-})/2F$ will be small and so from (7)

$$f^2 \simeq F^2 + F_H^2 - 2FF_H\{1 - \tfrac{1}{2}[k(F_{H+} - F_{H-})/2F]^2\}$$
$$= (F_H - F)^2 + FF_H[k(F_{H+} - F_{H-})/2F]^2 . \qquad (8)$$

Making the further approximation that $(F_H/F) = 1$, (8) leads to

$$f^2 \simeq (F_H - F)^2 + (k/2)^2 (F_{H+} - F_{H-})^2 . \qquad (9)$$

This is the expression used by Kartha & Parthasarathy to combine isomorphous replacement and anomalous scattering measurements. The present treatment does not rely on the assumption that α in Fig. 2 is small, but uses the observed data to estimate the value of this angle. Although the use of (7) instead of (9) would be expected to give only marginal improvement in the difference Patterson function, an estimate of α is essential for the difference Fourier method which is discussed later.

In practice, in combining anomalous scattering and isomorphous replacement measurements we have found it desirable to take into account their relative accuracies.

Generally, the $(F_H - F)$ differences are larger than the $(F_{H+} - F_{H-})$ ones and can therefore be measured with greater relative accuracy. However, as discussed by North (1965), the $(F_H - F)$ measurements are influenced by lack of isomorphism, by scaling errors, and by errors due to absorption. In favourable cases the $(F_{H+} - F_{H-})$ measurements will be free from all these sources of error and will therefore have a greater intrinsic accuracy. We have attempted to allow for the relative accuracies of the experimental measurements by using, instead of (7), the analogous expression

$$f_e^2 = F^2 + F_H^2 - 2FF_H\{1 - [wk(F_{H+} - F_{H-})/2F]^2\}^{\frac{1}{2}} , \qquad (10)$$

where w is an estimate of the relative reliability of the $(F_{H+} - F_{H-})$ measurements compared with $(F_H - F)$, and f_e^2 is an estimate of the best coefficient to use (*i.e.* the one which will result in the highest peak to back-

ground ratio). Although this system of weighting is rather arbitrary it has been found to be satisfactory in practice. The procedure adopted has been to use the method described in the next section to estimate the value of k as a function of $\sin^2\theta/\lambda^2$ and to assume that w is a constant. The value $w=0{\cdot}75$ was found satisfactory and was used for all the calculations described here. In other cases the use of a different value of w might be more appropriate. For example, in the case of heavy atoms with relatively small anomalous scattering (k large) the observed differences between F_H+ and F_H- become small and relatively inaccurate. In such cases a smaller value of w would be used.

(*b*) *Determination of k*

$k=f''/f'$ is the ratio of the real to anomalous scattering of the heavy atom groups in the derivative crystal. If the replacement groups were single heavy atoms and were all 'of the same type' then the value of k as a function of the angle of scattering could be looked up from tables. However, for many protein crystals the useful heavy atom replacement groups are complex ions which can be thought of as 'lumps' of high electron density. The $[PtCl_4]^{2-}$ derivative of α-chymotrypsin, used later as an illustration, is a typical example. In such cases it would be more difficult to predict the value of k, for the anomalous component of the scattering of the complex ion results almost entirely from the inner electrons of the platinum atom, whereas the real scattering is derived from all the electrons in the 'lump'. A simple method for determining the effective value of k in such cases will now be described. It will be assumed that the 'lumps' are of the same type and that they may be regarded as spherically symmetrical. The latter assumption will be valid unless one is working to a resolution comparable to the dimensions of the heavy atom group.

For many reflexions F and F_H are large compared with f, and for these reflexions α in Fig. 2 will be small. The only possible exceptions to this general rule are the weak reflexions, and if these are ignored, it will always be a good approximation that

$$|f\cos\gamma|=|F_H-F|\,. \tag{11}$$

Similarly, δ will be small compared with F_H so that

$$|\delta\sin\gamma|=\tfrac{1}{2}|F_H+-F_H-|\,,$$

or

$$|k^{-1}f\sin\gamma|=\tfrac{1}{2}|F_H+-F_H-|\,. \tag{12}$$

By averaging (11) and (12) over a large number of reflexions, we have

$$f\,.\,\overline{|\cos\gamma|}=f\,.\,(2/\pi)=\overline{|F_H-F|} \tag{13}$$

and

$$\overline{k^{-1}f}\,.\,\overline{|\sin\gamma|}=\overline{k^{-1}f}\,.\,(2/\pi)=\tfrac{1}{2}\overline{|F_H+-F_H-|}\,. \tag{14}$$

Since the heavy atom groups are assumed to have spherical symmetry, k will be a constant for all reflexions with the same Bragg angle. Therefore, if the average is made within a small range of scattering angles and the weak reflexions are omitted, the value of k appropriate to that range is given by

$$k=2\overline{|F_H-F|}/\overline{|F_H+-F_H-|}\,. \tag{15}$$

It would be expected that, in practice, the various types of error would increase these averages over their 'theoretical' values. In particular, non-isomorphism would increase the numerator but not the denominator. However, it has been found satisfactory to use (15) as it stands.

(*c*) *The difference Fourier synthesis*

The first section dealt with the location of heavy atom replacement groups in proteins by using difference Patterson functions. Such a method requires no prior knowledge of any of the phases of the native protein. If such information is already available, for example from other heavy atom replacements, it may be preferable to locate new heavy atom groups by difference Fourier techniques. This procedure automatically gives the coordinates of the new heavy atom sites relative to the previously chosen origin.

Let us suppose that φ is the given phase of a reflexion of the native protein and that m is an estimate of the reliability of the determination of φ. m can be defined rigorously as the 'figure of merit' of the phase determination (Blow & Crick, 1959; Dickerson, Kendrew & Strandberg, 1961).

If F and F_H are the structure amplitudes of the parent crystal and an isomorphous derivative, the Fourier synthesis used to locate the atoms differentiating the two structures has coefficients

$$m(F_H-F)\exp(-i\varphi) \tag{16}$$

(*e.g.* see Stryer, Kendrew & Watson, 1964). These coefficients give the vector contribution of f in the direction specified by φ, suitably weighted by m, but include no contribution perpendicular to this direction. The necessary information is not supplied by the isomorphous replacement method, but can be obtained from anomalous scattering measurements (Ramachandran & Raman, 1956). If the phase ψ of f can be determined (with figure of merit m') the corresponding 'best' difference Fourier function is defined to be one with coefficients

$$m'f\exp(-i\psi)\,. \tag{17}$$

From Fig. 2

$$\psi=\beta+\varphi-\pi\,. \tag{18}$$

From this relation it follows that the reliability with which ψ may be determined is limited by the reliability of the determination of φ and it is therefore reasonable to use m in place of m' as an estimate of the reliability of ψ. The coefficients of the difference Fourier function are then given by

$$mf\exp(-i\psi)=-mf\exp(-i\beta)\exp(-i\varphi). \tag{19}$$

Also from Fig. 2

$$\sin\beta=(F_H\sin\alpha)/f$$

and

$$\cos\beta=(F^2+f^2-F_H^2)/2Ff\,.$$

Therefore by using (6) and (7) to evaluate $\sin\alpha$ and f, and from the given values of $m\sin\varphi$ and $m\cos\varphi$, the coefficients of the 'best' difference Fourier function can be determined. In practice, the weighting system described before was again used to reduce the influence of the less accurate anomalous scattering differences.

Application to α-chymotrypsin

α-Chymotrypsin is a proteolytic enzyme of molecular weight 23000 crystallizing in the space group $P2_1$ with two molecules per asymmetric unit. The unit-cell parameters are $a=49{\cdot}6$, $b=67{\cdot}8$, $c=66{\cdot}5$ Å, $\beta=102°\,10'$. This crystal form was first investigated by Bernal, Fankuchen & Perutz (1938) and has subsequently been studied by Blow, Rossmann & Jeffery (1964). Blow *et al.* prepared crystals isomorphous with α-chymotrypsin but containing the planar complex chloroplatinite $[PtCl_4]^{2-}$. Following the work of Sigler, Skinner, Coulter, Kallos, Braxton & Davies (1964) with γ-chymotrypsin, another heavy atom derivative has been made by preparing α-chymotrypsin crystals containing *p*-iodobenzenesulphonyl fluoride (pipsyl fluoride) (Sigler, Jeffery, Matthews & Blow, 1966). Pipsyl-α-chymotrypsin is closely isomorphous with the analogous derivative benzenesulphonyl-α-chymotrypsin and the two compounds are differentiated by one iodine atom per chymotrypsin molecule.

By using these derivatives it was possible to test the new method in two rather differing situations. The pipsyl and benzenesulphonyl derivatives provide an isomorphous pair of compounds differentiated by single heavy atoms at well defined sites. In contrast, the $[PtCl_4]^{2-}$ derivative is isomorphous with the native enzyme but contains heavy atom groups which are complex and which occupy several different sites with varying occupancy.

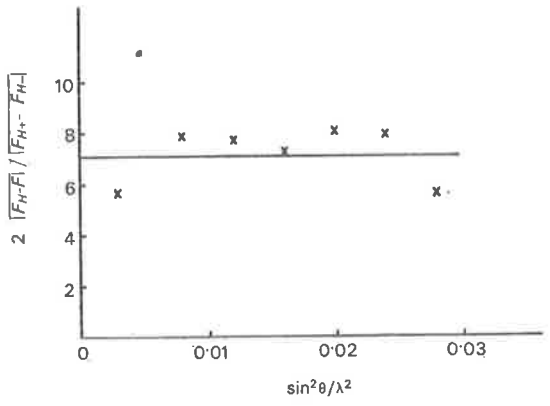


Fig. 3. Determination of k, the ratio of the real to anomalous scattering of the iodine atoms differentiating the isomorphous compounds pipsyl-α-chymotrypsin and benzenesulphonyl-α-chymotrypsin.

The non-centrosymmetric (100) projection is used to illustrate the use of the Patterson and Fourier methods. The corresponding $0kl$ spectra were recorded photographically with a Buerger precession camera using Cu $K\alpha$ radiation and a Joyce–Loebl microdensitometer was used to measure the intensities of the reflexions. The data were recorded to a resolution of 2·8 Å, this corresponding to about 900 independent $0kl$ reflexions.

(a) Difference Patterson syntheses

The difference Patterson syntheses obtained from the combined pipsyl and benzenesulphonyl data will be discussed first. As mentioned before, these two isomorphous compounds are differentiated only by single iodine atoms and it would be expected that the ratio of real to anomalous scattering of the iodine atoms would change only slightly with the small range of scattering angles used here. Using (15), k was estimated by plotting the ratio $2\overline{|F_H-F|}/\overline{|F_H+-F_H-|}$ as a function of $\sin^2\theta/\lambda^2$ as shown in Fig. 3. It is noteworthy that not only does k appear to be a constant over this range of scattering angles but also that the mean value of $k=7{\cdot}1$ is equal to the value predicted from tables. Using this value of k and assuming that the weighting factor w has the constant value of 0·75, equation (10) was used to calculate the 'combined' difference Patterson coefficients. The resultant synthesis is shown in Fig. 4(*d*). For comparison the corresponding $(F_H+-F_H-)^2$ anomalous scattering function, the $(F_H-F)^2$ isomorphous replacement function, and the sum of these two, weighted according to equation (9) (Kartha & Parthasarathy, 1965) with $k=7{\cdot}1$ are shown in Fig. 4(*a*), (*b*) and (*c*).

To assist in the comparison of the different results, Table 1 gives a summary of the peak and background heights in the four syntheses. In both the 'combined' and 'sum' functions the four iodine–iodine vector peaks show up clearly above the background, the contrast being slightly better for the 'combined' difference Patterson synthesis.

Turning now to the native enzyme and the isomorphous $[PtCl_4]^{2-}$ derivative, the behavior of k as a function of $\sin^2\theta/\lambda^2$ is shown in Fig. 5. As expected, the anomalous scattering becomes relatively stronger at higher angles. Over the rather small range of scattering angles used, the change of k can be approximated equally well by either the straight line shown in Fig. 5, or the exponential approximation

$$k=11{\cdot}5\exp(-16\sin^2\theta/\lambda^2)\,.$$

The linear approximation was the one used, and by again assuming the constant value of 0·75 for w, the 'combined' (100) difference Patterson projection was calculated [Fig. 6(*d*)]. For comparison the $(F_H+-F_H-)^2$, $(F_H-F)^2$ and 'sum' functions are shown in Fig. 6(*a*), (*b*) and (*c*). (In using (9) to evaluate the coefficients of the 'sum' function, k was put equal to 9·1, the expected value for platinum.) In this case the Pat-

Table 1. *Comparison of difference Patterson syntheses shown in Fig.* 4

(*a*) 'Anomalous differences' synthesis		(*b*) 'Isomorphous replacement' synthesis		(*c*) 'Sum' function		(*d*) 'Combined' difference Patterson function	
Peak height	Highest background	Peak	Background	Peak	Background	Peak	Background
4·0	3·6	60	33	103	49	87	40
2·8	2·2	42	33	70	45	57	38
2·6	2·1	38	31	70	40	52	32
2·5	2·1	34	27	59	40	48	31

	(*a*)	(*b*)	(*c*)	(*d*)
Average peak height	3·0	43·0	75·5	61·0
Average background	0·87	7·6	10·7	8·3
Peak to background ratio	4·3	5·7	7·1	7·3
Relative improvement	0·77	1·0	1·25	1·30

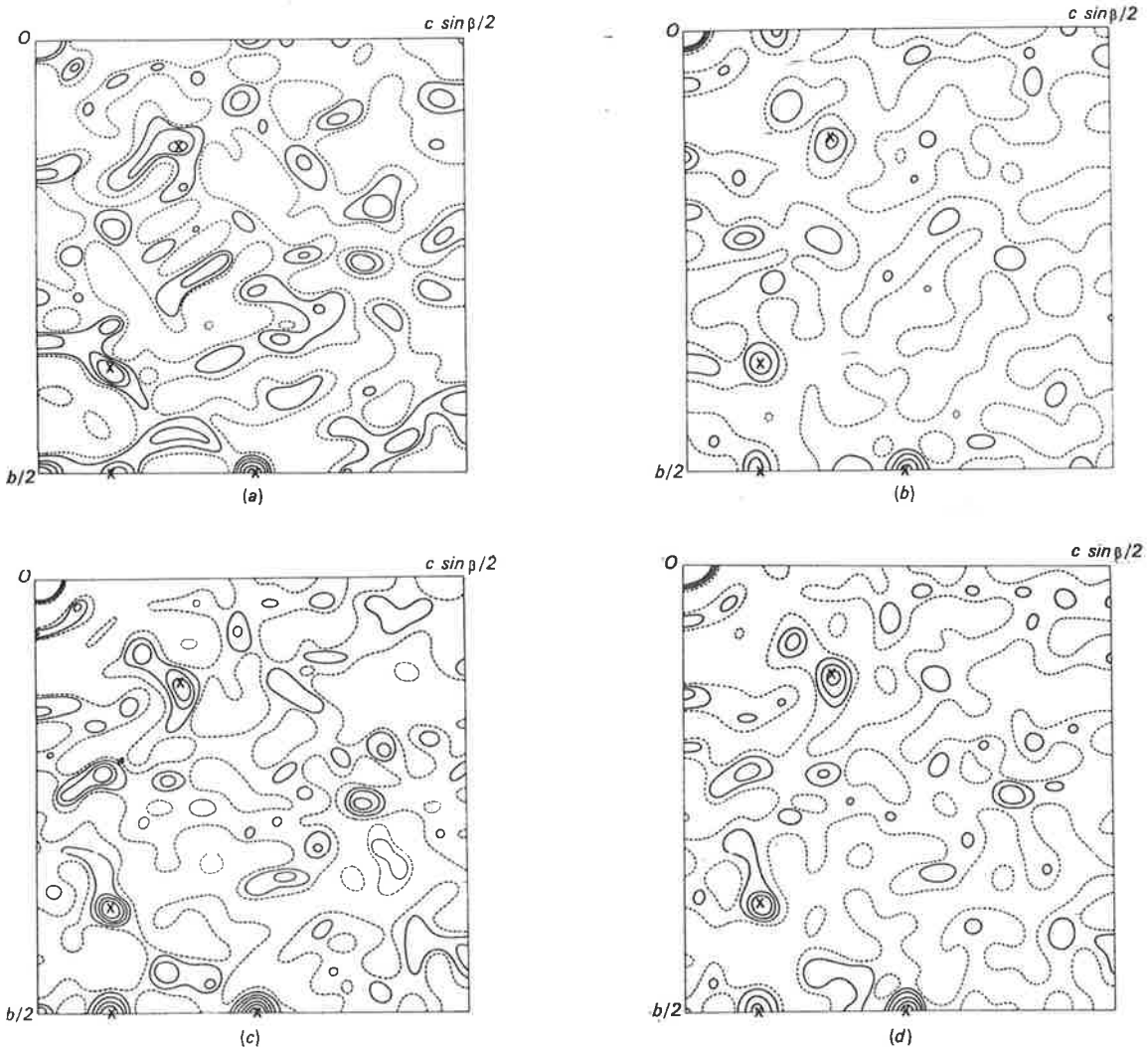


Fig. 4. Comparison of different (0*kl*) difference Patterson syntheses used to locate the iodine atoms differentiating pipsyl-α-chymotrypsin and benzenesulphonyl-α-chymotrypsin. The crosses indicate the expected positions of the four equal-weight iodine–iodine vector interactions. In these and all succeeding diagrams the contours are drawn at arbitrary equal intervals, the zero contour being dashed. (*a*) 'Anomalous differences' synthesis using coefficients $(F_H+ - F_H-)^2$. Contours drawn at one twentieth the interval used for (*b*), (*c*) and (*d*). (*b*) 'Isomorphous replacement' synthesis using coefficients $(F_H - F)^2$. (*c*) 'Sum' synthesis (see text). (*d*)'Combined' difference Patterson function (see text).

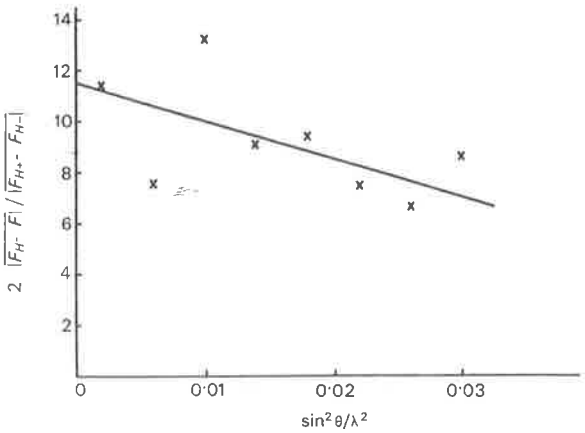


Fig. 5. Determination of the effective value of k for the chloroplatinite groups differentiating $[PtCl_4]^{2-}$-substituted α-chymotrypsin and native α-chymotrypsin.

terson functions are more difficult to interpret as there are five sites occupied with varying occupancy by the $[PtCl_4]^{2-}$ ions. Nevertheless, from the data of Blow *et al.* (1964) and from subsequent work at higher resolution (Sigler *et al.*, 1966), the coordinates and occupancies of the various sites have been determined with some precision, and it was possible to construct the 'expected' set of weighted vector points as shown on the various maps. Because many of the vector peaks overlap it is not possible to list the individual peak heights. Instead, Table 2 gives the heights of the major composite vector peaks and the highest background peaks. It may be noted that by use of the 'combined' difference Patterson function rather than the $(F_H - F)^2$ function, the peaks are enhanced by an average of 27% and the highest background peaks are in fact reduced. In this case the 'combined' Patterson function gives appreciably better results than the 'sum' function.

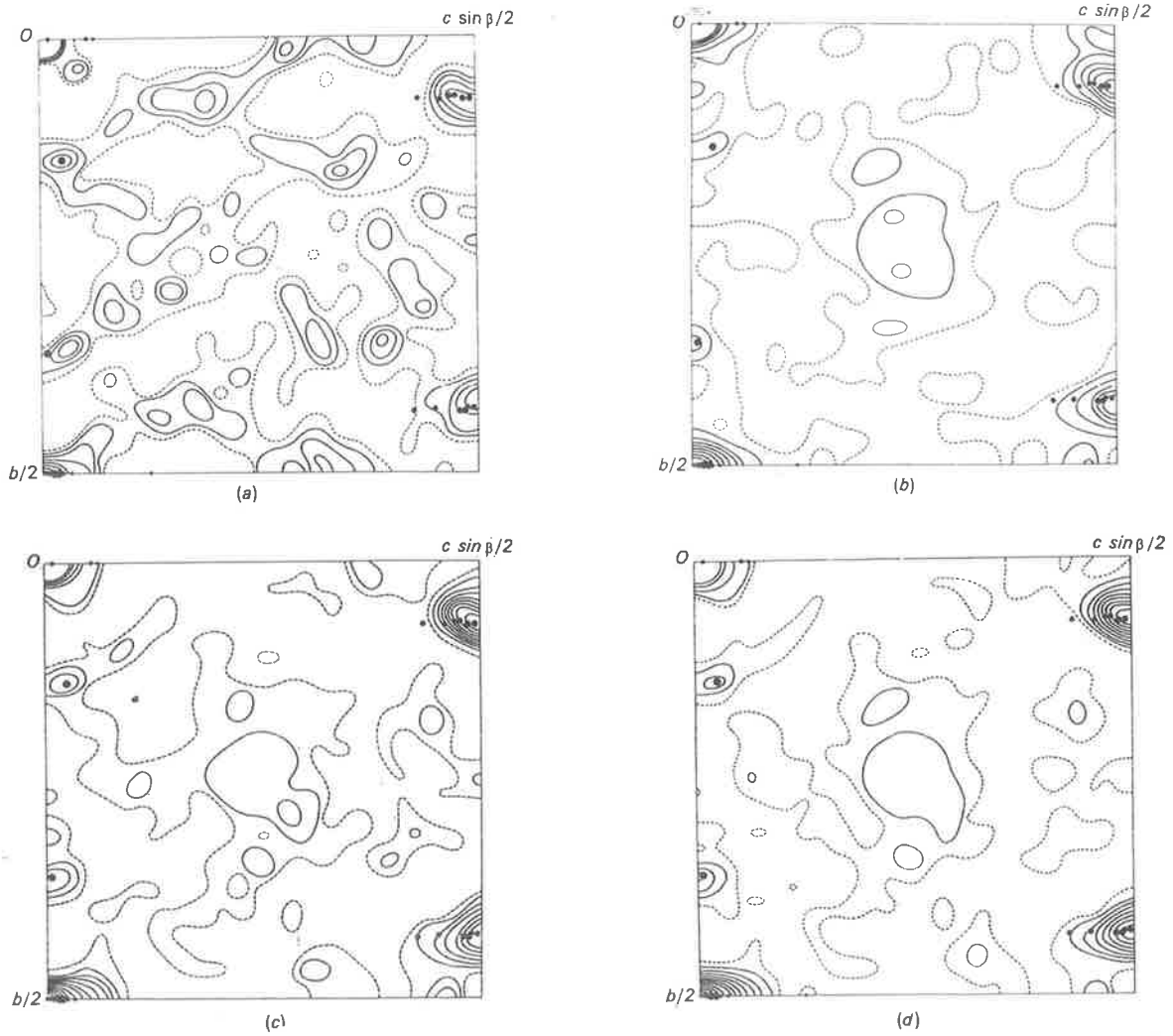


Fig. 6. Comparison of different ($0kl$) difference Patterson syntheses used to locate the heavy atom groups in $[PtCl_4]^{2-}$-substituted α-chymotrypsin. The position and size of the small solid circles indicate the position and approximate height of the expected vector peaks. (*a*) 'Anomalous differences' synthesis. Contours drawn at one twenty-fifth the interval used for (*b*), (*c*) and (*d*). (*b*) 'Isomorphous replacement' synthesis. (*c*) 'Sum' synthesis. (*d*) 'Combined' synthesis.

To compare further the relative merits of the 'sum' and 'combined' functions, several other syntheses similar to those shown in Fig. 4(*c*) and (*d*) and Fig. 6(*c*) and (*d*) were calculated with different values for k and w. The first conclusion drawn from these comparisons is that if k in equation (9) is put equal to the value of (wk) used in equation (10), the resultant syntheses are very similar, and, within our experience, give equally good results. The second conclusion is that in all cases the peak to background contrast can be improved, in some cases markedly, by allowing for the change in the ratio of the real to anomalous scattering of the heavy atom groups, and the relative reliability of the experimental measurements. In other words, the 'combined' synthesis is superior to the 'sum' synthesis mainly because of the use of the weighting factor w rather than the abandonment of the assumption that α is small.

(*b*) *Difference Fourier syntheses*

To test the new difference Fourier method it was necessary to have an estimate of the $0kl$ phases. In phase determination of non-centrosymmetric reflexions the conventional isomorphous replacement method leads to an ambiguous result for the phase angle (Bokhoven, Schoone & Bijvoet, 1951). However, by including anomalous scattering information in the phase determination, this ambiguity may be resolved and a unique determination made (Ramachandran & Raman, 1956). Bearing this in mind it was decided to calculate two independent sets of $0kl$ phases, the first from the isomorphous native and $[PtCl_4]^{2-}$ compounds, and the second from the pipsyl and benzenesulphonyl compounds. Since the four different crystals are almost exactly isomorphous, the '$PtCl_4$' phases could then be used to calculate an 'iodine' difference Fourier synthesis, and the 'iodine' phases could be used to calculate a '$PtCl_4$' difference Fourier synthesis. This procedure not only enabled the new method to be tested, but also served as a test of the effectiveness of using iodine atoms to provide phase information. Furthermore, it allowed the relative y coordinates of the $[PtCl_4]^{2-}$ groups and the iodine atoms to be determined, and was used to establish the absolute configuration of these coordinates.

In the phase angle calculations, proper account was taken of the relative reliabilities of the isomorphous replacement and anomalous scattering measurements (North, 1965; Matthews, 1966).

The first difference Fourier synthesis used to locate the iodine atoms is shown in Fig. 7(*a*). The coefficients are those defined in equation (16), the observed structure amplitudes for the pipsyl and benzenesulphonyl derivatives being substituted for F_H and F respectively, and the protein phases φ being those determined by considering the native crystals and the $[PtCl_4]^{2-}$ derivative as a single isomorphous pair. In calculating the $[PtCl_4]^{2-}$ structure factors for the phase determination, the plane $y=0$ was chosen to lie midway between the $[PtCl_4]^{2-}$ sites shown in Fig. 7(*a*) as *A* and *B*. From the (100) Patterson projections [Fig. 6(*a*) to (*d*)] and from other evidence (Sigler *et al.*, 1966), the three other sites *C*, *D* and *E* also lie in or near this plane. Blow *et al.* (1964) have shown that there is a non-crystallographic twofold axis of symmetry parallel to $-a^*$. This local symmetry axis passes through the point $x=0{\cdot}22$, y_1, $z=0{\cdot}005$ where, relative to the choice of origin used here, y_1 is close to zero. (From the $P2_1$ space group symmetry there will be another local symmetry axis parallel to a^* passing through the point $(-0{\cdot}22,\ y_1+\frac{1}{2},\ -0{\cdot}005)$. Referred to these non-crystallographic axes, *A* and *B* are a pair of related sites, as are *D* and *E*, while *C* lies on an axis.

The 'combined' difference Fourier synthesis with coefficients defined by (19) is shown in Fig. 7(*b*). In both Fig. 7(*a*) and Fig. 7(*b*) the two iodine atoms I(1) and I(2) show up clearly above background at positions which are consistent with the interpretation of the (100) Patterson projections (Fig. 4) and with data for the other projections. The iodine peaks straddle the line $y=\frac{1}{2}$ and are equidistant from it, and the coordinates of the iodine atoms deduced from the Fourier maps are consistent with the non-crystallographic axis of symmetry. In the 'combined' synthesis the iodine peaks

Table 2. *Comparison of difference Patterson syntheses shown in Fig.* 6

(*a*) 'Anomalous differences' synthesis		(*b*) 'Isomorphous replacement' synthesis		(*c*) 'Sum' function		(*d*) 'Combined' difference Patterson function	
Peak height	Highest background	Peak	Background	Peak	Background	Peak	Background
10·6	5·4	670	161	849	171	851	149
6·9	5·1	496	153	657	167	648	134
6·9	4·8	441	151	506	142	539	122
5·7	4·7	187	123	207	135	229	120
5·3	4·7	126	120	197	129	173	120

	(*a*)	(*b*)	(*c*)	(*d*)
Average peak height	7·1	384	483	488
Average background	1·52	42·9	44·7	40·2
Peak to background ratio	4·7	9·0	10·8	12·1
Relative improvement	0·52	1·0	1·21	1·35

have approximately double the height observed in the normal difference Fourier synthesis, while the average background density is only 19% higher. A more detailed comparison is given in Table 3. It is a striking feature of the 'combined' synthesis that 8 of the 11 highest 'background' peaks occur in positions which could be related by the non-crystallographic axis. The possibility that these peaks may be due to an effect such as partial substitution at secondary sites is being investigated. The average increase in the height of the

Table 3. *Comparison of difference Fourier syntheses shown in Fig. 7*

(*a*) Normal difference Fourier synthesis		(*b*) 'Combined' difference Fourier synthesis	
Iodine peaks	Highest background peaks*	Iodine peaks	Highest background peaks*
110	60	220	83
105	59	198	72
	58		66
	58		57

	(*a*)	(*b*)
Average peak height	108	209
Average background	14·3	17·0
Peak to background ratio	7·5	12·3
Relative improvement	1·0	1·63

* 'Related' background peaks omitted (see discussion).

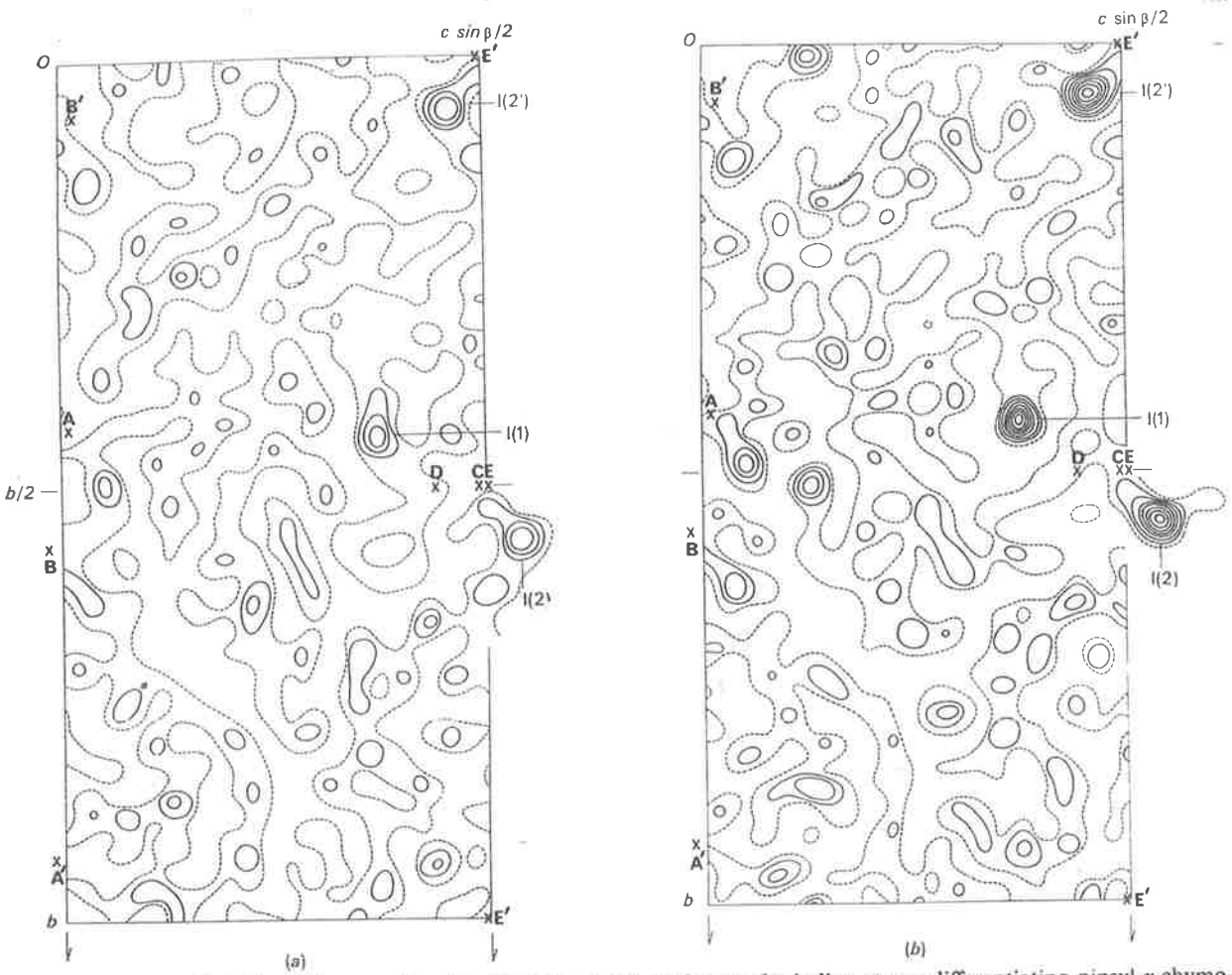


Fig. 7. Comparison of (0*kl*) difference Fourier syntheses used to locate the iodine atoms differentiating pipsyl-α-chymotrypsin and benzenesulphonyl-α-chymotrypsin. The protein phases were determined from data observed for native α-chymotrypsin and $[PtCl_4]^{2-}$-substituted α-chymotrypsin. In the phase determination, the coordinates of the $[PtCl_4]^{2-}$ groups used to calculate structure factors are marked by the crosses labelled *A*, *B*, *C*, *D*, *E*. I(1) and I(2) are the peaks of the difference syntheses indicating the positions of the iodine atoms. (*a*) Normal difference synthesis. (*b*) 'Combined' difference synthesis based on the same protein phases used for (*a*), but making use of the observed anomalous scattering of the iodine atoms in the pipsyl-α-chymotrypsin.

8 'related' peaks is 78% which considerably exceeds the average background increase of 19%.

Having shown that the phases obtained from the $[PtCl_4]^{2-}$ derivative were sufficiently good to reveal the positions of the iodine atoms unambiguously, the converse was next attempted. In this case the observed structure factors for the native and $[PtCl_4]^{2-}$ substituted crystals were substituted for F and F_H in (16), and the phases φ were those determined from the pipsyl and benzenesulphonyl derivatives. The resultant synthesis is shown in Fig. 8(*a*) and can be compared with the 'combined' synthesis [Fig. 8(*b*)] which has coefficients defined by (19). In both syntheses the $[PtCl_4]^{2-}$ groups show up clearly at their expected positions;

Table 4. *Comparison of difference Fourier syntheses shown in Fig.* 8

(*a*) Normal difference Fourier synthesis		(*b*) 'Combined' difference Fourier synthesis	
$PtCl_4]^{2-}$ peaks	Highest background peaks	$[PtCl_4]^{2-}$ peaks	Highest background peaks
142	64	228	73
111	52	184	69
134*	51	177*	64

	(*a*)	(*b*)
Average peak height	129	196
Average background	15·0	16·9
Peak to background ratio	8·6	11·6
Relative improvement	1·0	1·35

* Composite peak.

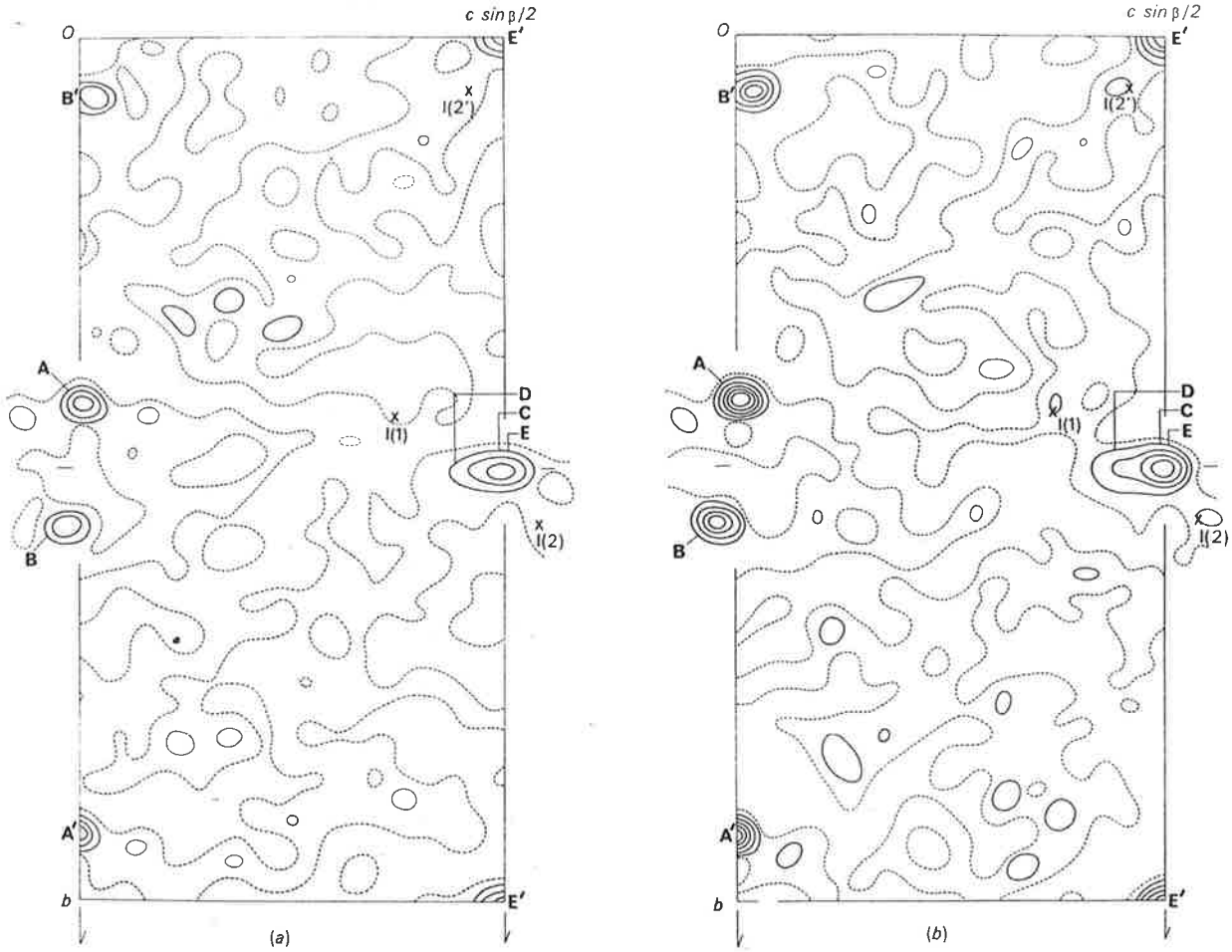


Fig. 8. Comparison of $(0kl)$ difference Fourier syntheses used to locate the chloroplatinite groups in $[PtCl_4]^{2-}$-substituted α-chymotrypsin. The protein phases were determined from the pipsyl- and benzene sulphonyl-α-chymotrypsin data. I(1) and I(2) indicate the coordinates of the iodine atoms used to calculate structure factors for the phase determination. *A*, *B*, *C*, *D* and *E* indicate the sites containing $[PtCl_4]^{2-}$ as revealed by the difference syntheses. These sites are thought to have a relative occupancy of about 100:100:80:50:50. (*a*) Normal difference synthesis. (*b*) 'Combined' difference synthesis.

however, in the 'combined' synthesis the peak-to-background contrast is markedly improved. Table 4 gives a detailed comparison of the two results.

Apart from the usefulness of the 'combined' Fourier synthesis in increasing peak-to-background contrast, it also provides a powerful method of verifying that the indexing of the Friedel related reflexions F_H+ and F_H- is consistent from one heavy atom derivative to the next. If a 'combined' synthesis is calculated with F_H+ and F_H- in the wrong sense, then the heavy atom peaks are markedly reduced rather than being enhanced.

It may be noted that the clarity with which the heavy atom groups show up in Fig. 8(*a*) and (*b*) clearly indicates that atoms as light as iodine can be useful in helping to determine the phases of protein reflexions.

(*c*) *The determination of absolute configuration*

By including anomalous scattering measurements in the determination of protein phases, as was done here, it is possible to determine the absolute configuration of the heavy atoms (and hence of the protein structure) in a simple and straightforward way.

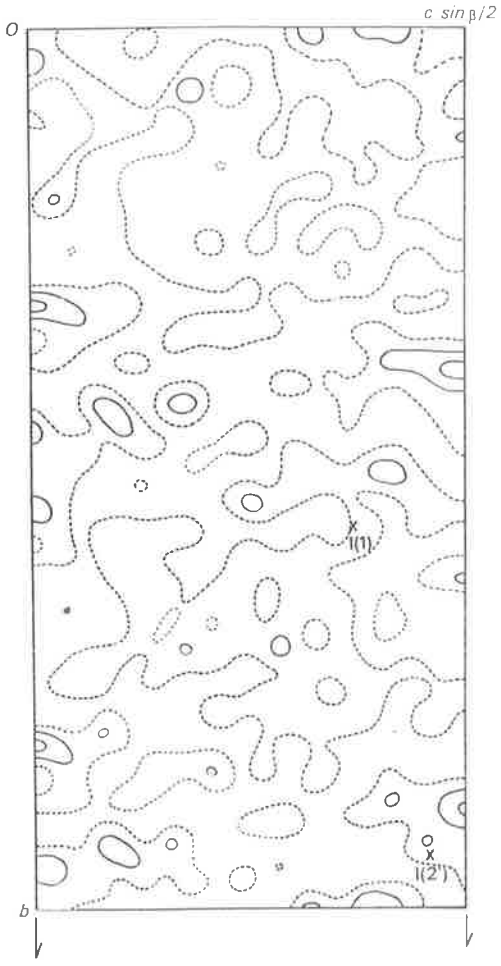


Fig. 9. Alternative difference Fourier synthesis to Fig. 8(*a*) resulting from the choice of the opposite hand for the coordinates of the iodine atoms I(1) and I(2). Contours at the same interval as in Fig. 8.

If the wrong hand is assumed for the coordinates of the heavy atoms, then the combination of anomalous scattering and isomorphous replacement measurements leads to an unambiguous but incorrect value for the protein phase. (The correct and incorrect phases differ by an angle of 2ψ where ψ is the phase of the heavy atom vector *f*.) A Fourier or difference Fourier synthesis calculated with the use of these phases would lead to an 'uninterpretable' result, whereas the choice of the correct hand of the heavy atom coordinates would be expected to lead to a Fourier synthesis with reasonable features (Blow & Rossmann, 1962). Fig. 9 illustrates a difference Fourier synthesis which is in all respects similar to that shown in Fig. 8(*a*) except that the phases used are those obtained after changing the hand of the iodine coordinates (*i.e.* replacing y by $-y$). Clearly Fig. 8(*a*) is the one with 'reasonable features', indicating that the iodine coordinates used for the phase determination and the $[PtCl_4]^{2-}$ positions revealed by this map are the ones with the correct absolute configuration.

I am indebted to Dr D. M. Blow and Dr P. B. Sigler for their continued interest and help throughout this work. I should like to express my thanks to them, and to Miss J. Collard for her able assistance and for preparing the diagrams. The referee made some very useful comments.

References

Bernal, J. D., Fankuchen, I. & Perutz, M. F. (1938). *Nature, Lond.* **141**, 523.

Blow, D. M. (1957). Ph. D. Thesis. Univ. of Cambridge.

Blow, D. M. (1958). *Proc. Roy. Soc.* A, **247**, 302.

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794.

Blow, D. M. & Rossmann, M. G. (1961). *Acta Cryst.* **14**, 1195.

Blow, D. M. & Rossmann, M. G. (1962). *Acta Cryst.* **15**, 1060.

Blow, D. M., Rossmann, M. G. & Jeffery, B. A. (1964). *J. Mol. Biol.* **8**, 65.

Bokhoven, C., Schoone, J. C. & Bijvoet, J. M. (1951). *Acta Cryst.* **4**, 275.

Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *Acta Cryst.* **14**, 1188.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. Roy. Soc.* A, **225**, 287.

Kartha, G. & Parthasarathy, R. (1965). *Acta Cryst.* **18**, 745.

Matthews, B. W. (1966). *Acta Cryst.* **20**, 82.

North, A. C. T. (1965). *Acta Cryst.* **18**, 212.

Ramachandran, G. N. & Raman, S. (1956). *Curr. Sci.* **25**, 348.

Rossmann, M. G. (1960). *Acta Cryst.* **13**, 221.

Rossmann, M. G. (1961). *Acta Cryst.* **14**, 383.

Sigler, P. B., Jeffery, B. A., Matthews, B. W. & Blow, D. M. (1966). *J. Mol. Biol.* In the press.

Sigler, P. B., Skinner, H. C. W., Coulter, C. L., Kallos, J., Braxton, H. & Davies, D. R. (1964). *Proc. Nat. Acad. Sci. Wash.* **51**, 1146.

Stryer, L., Kendrew, J. C. & Watson, H. C. (1964). *J. Mol. Biol.* **8**, 96.

J. Mol. Biol. (1968) **33**, 491–497

Solvent Content of Protein Crystals

B. W. Matthews

Laboratory of Molecular Biology
National Institute of Arthritis and Metabolic Diseases
National Institutes of Health, Bethesda, Md. 20014, U.S.A.

(*Received 1 January 1968*)

An analysis has been made, from the data which are currently available, of the solvent content of 116 different crystal forms of globular proteins. The fraction of the crystal volume occupied by solvent is most commonly near 43%, but has been observed to have values from about 27 to 65%. In many cases this range will be sufficiently restrictive to enable the probable number of molecules in the crystallographic asymmetric unit to be determined directly from the molecular weight of the protein and the space group and unit cell dimensions of the crystal.

For protein crystals the determination of n, the number of molecules in the crystallographic asymmetric unit, is complicated by the presence in the crystals of a variable amount of solvent. It is therefore common to supplement the measurement of the unit cell dimensions of the "wet" crystals with measurements of the cell parameters of the "air-dried" crystals, or with various density measurements in order to obtain an unequivocal value for n. However, in some cases cell dimensions cannot be obtained from air-dried crystals, and if only small crystals are available an accurate determination of the density of the wet crystals is difficult. On the other hand it is well known that for most protein crystals the number of molecules per asymmetric unit is one, although crystals are also found with $n = 1/4, 1/2, 2, \ldots$ etc. (e.g. see Crick & Kendrew, 1957). In any event, to decide between possible alternatives, it is usually necessary only to know the solvent content of the crystals between rather coarse limits. Crick & Kendrew (1957) point out that for most protein crystals the solvent content is between 40 and 60%, but there are exceptions to this general rule. An analysis has been made of all the data currently available in order to find to what extent the over-all range of values observed is sufficiently restrictive to be of use in determining n in further studies. The relevant data from 116 distinct crystal forms which have been reported for a variety of globular proteins are summarized in Figure 1 by plotting the volume of the asymmetric unit, as obtained directly from the X-ray diffraction measurements, against the molecular weight of protein contained in the asymmetric unit. It is convenient to define V_M as the ratio of these two quantities, i.e. V_M is the crystal volume per unit of protein molecular weight. It will be shown that V_M bears a simple relationship to the fractional volume of solvent in the crystal. From Figure 1 it is seen that the range of values assumed by V_M is essentially independent of the volume of the asymmetric unit. The over-all distribution of values observed for V_M is shown in Figure 2, with lowest extremes being 1·68 Å^3/dalton for high potential iron protein (Kraut, Strahs & Freer, *A.C.A. Abstr.* Meeting, August 1967) and 1·72 Å^3/dalton for excelsin (Drenth & Wiebenga, 1955), and highest 3·53 Å^3/dalton and

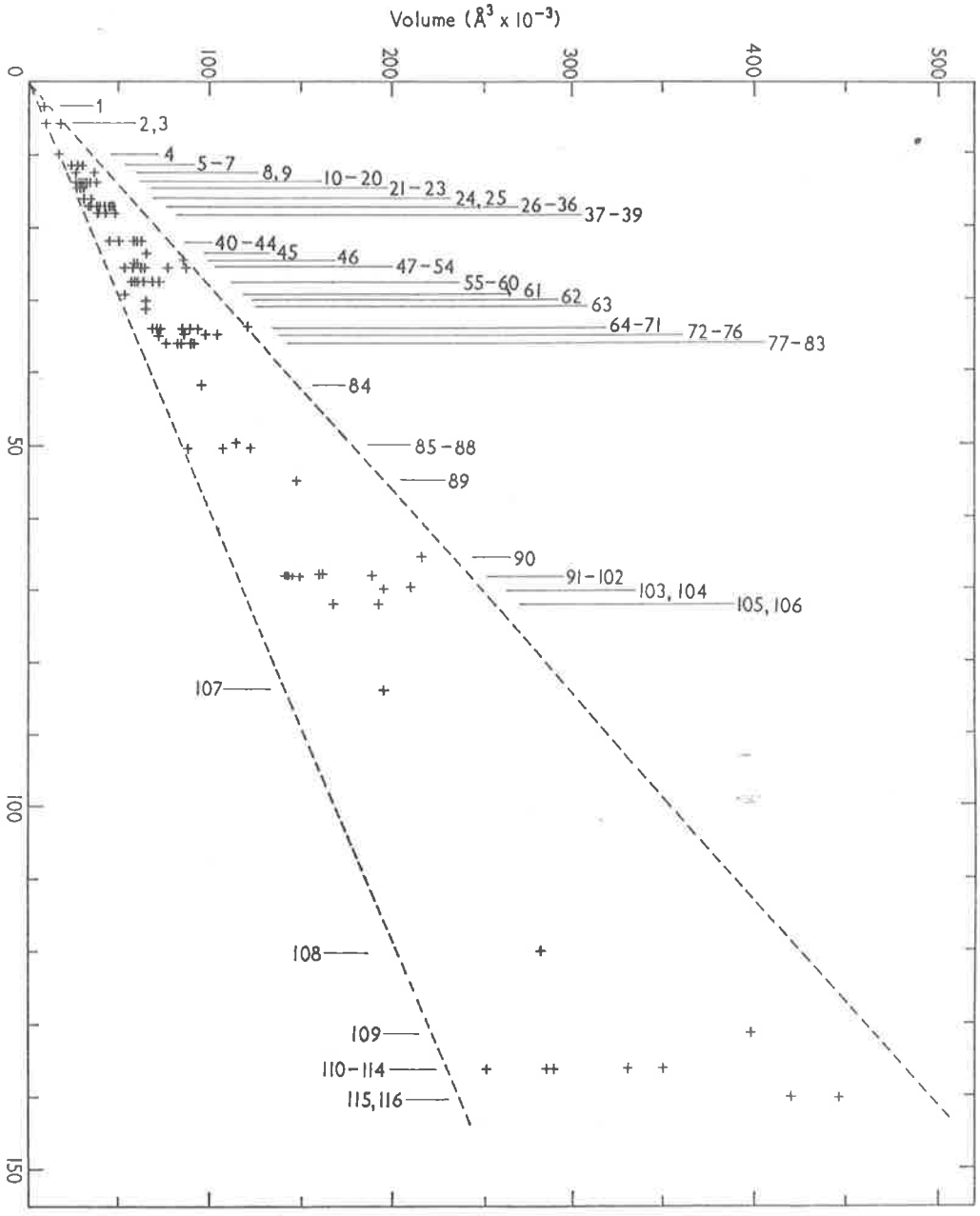

Volume (Å^3 x 10^{-3})
0
100
200
300
400
500
50
100
150
Molecular weight (x 10^{-3})
1
2,3
4
5–7
8,9
10–20
21–23
24,25
26–36
37–39
40–44
45
46
47–54
55–60
61
62
63
64–71
72–76
77–83
84
85–88
89
90
91–102
103, 104
105, 106
107
108
109
110–114
115,116

Fig. 1

FIG. 1. Plot to illustrate the relative variation of the volume of protein crystals. The dashed lines indicate the upper and lower limits observed for V_M. Data are included for the following crystalline proteins which are listed in order corresponding to the numbering in the Figure. Except where stated otherwise, the references may be obtained from the summary of King (1963).

1. Glucagon
2. Ferredoxin (Sieker & Jensen, 1965)
3. Insulin, cubic (Harding, Crowfoot-Hodgkin, Kennedy, O'Connor & Weitzmann, 1966)
4. High potential iron protein (Kraut, Strahs & Freer, *A.C.A. Abstr.* Meeting, Aug. 1967)
5. Insulin, rhombohedral (Harding *et al.*, 1966)
6. Insulin B
7. Insulin A
8. Horse ferricytochrome C
9. Ferricytochrome C (Dickerson, Kopka, Weinzierl, Eisenberg & Margoliash, 1967)
10. Ribonuclease IX
11. Ribonuclease S, Y (Wyckoff *et al.*, 1967)
12. Ribonuclease VII
13. Ribonuclease II (Kartha, Bello & Harker, 1967)
14. Ribonuclease VIII
15. Ribonuclease V
16. Ribonuclease XI
17. Ribonuclease I
18. Ribonuclease XIV
19. Ribonuclease VI
20. Ribonuclease S, W (Wyckoff *et al.*, 1967)
21. Lysozyme, triclinic
22. Lysozyme, orthorhombic
23. Lysozyme, tetragonal (Blake *et al.*, 1965)
24. Erythrocruorin, trigonal (Huber, Formanek, Braun, Braunitzer & Hoppe, 1964)
25. Erythrocruorin, hexagonal (Huber *et al.*, 1964)
26. Myoglobin J
27. Myoglobin A (Kendrew, Dickerson, Strandberg, Hart & Davies, 1960)
28. Myoglobin G
29. Myoglobin B
30. Myoglobin D
31. Myoglobin I
32. Myoglobin F
33. Myoglobin K
34. Myoglobin C2
35. Myoglobin C1
36. Myoglobin H
37. β-Lactoglobulin Y
38. β-Lactoglobulin A
39. β-Lactoglobulin Z
40. Papain S
41. Papain D
42. Papain A
43. Papain C
44. Papain B
45. Apo ferritin B
46. Chymotrypsinogen B, B (Matthews, 1968)
47. Chymotrypsinogen A, D
48. CS_2-chymotrypsinogen A (Kraut, Sieker, High & Freer, 1962)
49. DIP-chymotrypsin, tetragonal (Corey, Battfay, Brueckner & Mark, 1965)
50. γ-Chymotrypsin (Sigler *et al.*, 1964)
51. Chymotrypsinogen A, G (Matthews, 1968)
52. Chymotrypsinogen A, F (Kraut *et al.*, 1962)
53. Chymotrypsinogen A, E (Kraut *et al.*, 1962)
54. Chymotrypsinogen A, B
55. Ribonuclease S, Z (Wyckoff *et al.*, 1967)
56. Ribonuclease X
57. Ribonuclease XIII
58. Ribonuclease IV
59. Ribonuclease III
60. Ribonuclease XII
61. Lysozyme, monoclinic
62. Carbonic anhydrase (Fridborg *et al.*, 1967)
63. Ferritin B
64. Haemoglobin 10
65. Pig haemoglobin 1
66. Carboxypeptidase (Ludwig *et al.*, 1967)
67. Reduced horse haemoglobin (Perutz, Bolton, Diamond, Muirhead & Watson, 1964)
68. Haemoglobin 6
69. Horse oxyhaemoglobin
70. Myoglobin C3
71. Ox haemoglobin, cubic
72. Insulin, monoclinic (Harding *et al.*, 1966)
73. Pepsin, monoclinic (Borisov, Melik-Adamjan, Suckever & Andreeva, 1966)
74. Pepsin, hexagonal
75. Dogfish lactic dehydrogenase 1 (Rossmann, Jeffery, Main & Warren, 1967)
76. Dogfish lactic dehydrogenase 2 (Rossmann *et al.*, 1967)
77. β-Lactoglobulin U (Aschaffenburg, Green, & Simmons, 1965)
78. β-Lactoglobulin R
79. β-Lactoglobulin X
80. β-Lactoglobulin T
81. β-Lactoglobulin S (Aschaffenburg *et al.*, 1965)
82. β-Lactoglobulin W (Aschaffenburg *et al.*, 1965)
83. β-Lactoglobulin P
84. Alcohol dehydrogenase, orthorhombic (Brändén, 1965)
85. Excelsin
86. α-Chymotrypsin (Matthews, Sigler, Henderson & Blow, 1967)
87. Chymotrypsinogen B, C (Matthews, 1968)
88. Chymotrypsinogen A, C
89. Fc γ-G-immunoglobulin (Poljak, Dintzis & Goldstein, 1967)
90. Human mercaptalbumin Hg dimer
91. Haemoglobin 8
92. Haemoglobin H (Perutz & Mazzarella, 1963)
93. Reduced human haemoglobin (Muirhead, Cox, Mazzarella & Perutz, 1967)
94. Ox haemoglobin D (Dunnill, Green & Simmons, 1966)
95. Ox haemoglobin A (Dunnill *et al.*, 1966)
96. Haemoglobin 7
97. Pig haemoglobin II
98. Rabbit haemoglobin 1
99. Ox haemoglobin C (Dunnill *et al.*, 1966)
100. Haemoglobin 2
101. Haemoglobin 4
102. Haemoglobin 5
103. Horse serum albumin
104. Glyceraldehyde phosphate dehydrogenase, PCMB (Watson & Banaszak, 1964)
105. β-Lactoglobulin N (Aschaffenburg *et al.*, 1965)
106. β-Lactoglobulin Q (Aschaffenburg *et al.*, 1965)
107. Alcohol dehydrogenase, monoclinic (Brändén, 1965)
108. Hemerythrin
109. Human serum decanol albumin
110. Lamprey oxyhaemoglobin (Greer, Perutz & Rumen, 1966)
111. Reduced lamprey haemoglobin (Greer *et al.*, 1966)
112. Sickle-cell haemoglobin
113. Haemoglobin 3
114. Haemoglobin 11
115. Lobster glyceraldehyde phosphate dehydrogenase (Watson & Banaszak, 1964)
116. Pig lactic dehydrogenase (Rossmann *et al.*, 1967)

3·43 Å³/dalton, respectively, for cubic ox haemoglobin (North, 1959) and tetragonal chymotrypsinogen B, type B (Matthews, 1968). It will be noted that although relatively few examples are found toward the limits of the range, the frequency distribution within the range is not a symmetric one. The most commonly observed values of V_M are near 2·15 Å³/dalton, whereas the median is at $V_M = 2{\cdot}61$ Å³/dalton. The rather sharp cut-off at the lower end of the range can presumably be considered as corresponding to the closest packing possible for the roughly spherical protein molecules, while the tail of the distribution toward higher V_M values must correspond to progressively looser packing of the molecules, but still with sufficient intermolecular contacts to stabilize the crystal lattice.

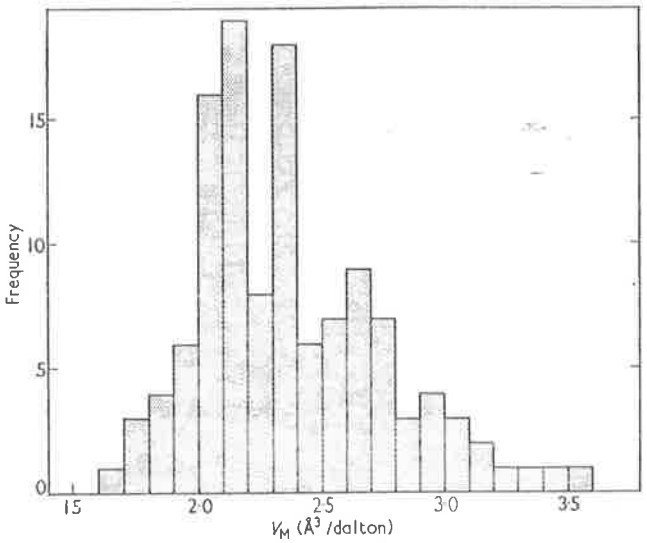


FIG. 2. Frequency distribution of values observed for V_M.
In addition to the data from Fig. 1, the following observations are included: ferritin A, V_M = 2·08 Å³/dalton; chicken lactic dehydrogenase, V_M = 2·84 Å³/dalton (Rossmann *et al.*, 1967); tobacco seed globulin, V_M = 1·76 Å³/dalton; edestin, V_M = 1·75 Å³.

Although the extreme values of V_M vary by a factor of approximately two, the range is still narrow enough to be useful in preliminary studies of most protein crystals. From the knowledge of the unit cell of the crystal, and of the molecular weight of the protein, the values of V_M corresponding to potential values of n may be obtained. In the majority of cases only one V_M value will lie within the acceptable range, so that n is obtained directly. In the event that there are two or more alternatives for n, it will be necessary to resolve this by some other method. While it must also be anticipated that as more protein crystals are studied, examples will be found with V_M lying outside the range quoted above, there is no reason to believe that the limits quoted will need substantial modification, at least in the molecular weight region below about 70,000 for which a reasonably large number of observations have been made. Nevertheless, borderline cases should be regarded with caution.

To express the range of values of V_M in terms of the percentage of solvent contained in the crystals, it may readily be shown that $V_{prot.}$, the fraction of the crystal volume occupied by protein, is given by

$$V_{prot.} = \frac{1{\cdot}66\,\bar{v}}{V_M}$$

where $\bar{v}$ is the partial specific volume of the protein in the crystal, and V_M (as defined above) is the crystal volume in Å^3 per unit of protein molecular weight. For most proteins $\bar{v}$ is approximately 0·74 cc/g, so that unless there is reason to believe that the protein in question has an abnormally low or high partial specific volume, or that the partial specific volume has a different value in the crystal from that in dilute solution, we have the simple approximation that

$$V_{prot.} = \frac{1{\cdot}23}{V_M}.$$

By difference, the fractional volume occupied by the solvent is given by

$$V_{solv.} = 1 - \frac{1{\cdot}66\bar{v}}{V_M} \simeq 1 - \frac{1{\cdot}23}{V_M}.$$

On the basis of this approximation, the protein crystals already examined have a solvent content ranging from about 27 to 65%, with values near 43% occurring most frequently.

Plots similar to Figures 1 and 2 were also made using cell dimensions from air-dried crystals. In this case, as has been noted before (e.g. see Crick, 1957) there is a smaller range of values of V_M—from 1·26 Å^3/dalton and 1·28 Å^3/dalton for edestin and excelsin (Drenth & Wiebenga, 1955), to 1·77 Å^3/dalton and 2·03 Å^3/dalton for orthorhombic insulin R (Einstein & Low, 1962) and hexagonal pepsin (Perutz, 1949). It should be borne in mind that in this case the range of values is based on only 34 sets of cell dimensions, some of which are unavoidably of low accuracy, so that the limits of the range should be regarded with caution.

It must be emphasized that the results quoted here are intended only as a guide in preliminary investigations of protein crystals, and are not intended to be a substitute for measurements of crystal density or solvent content where these can be made. In particular, from the few data which are available, there appears to be a tendency for molecules of higher molecular weight to form crystals containing a relatively higher fractional volume of solvent. Further data will be needed in order to determine the range of values for V_M which might be expected for such proteins. Very recently Rossmann & Labaw (1967) and Longley (1967) have studied by electron microscopy and X-ray diffraction a trigonal modification of crystalline catalase (mol. wt about 250,000). They conclude that there are six catalase molecules in a trigonal cell of dimensions $a = 178 \pm 3$, $c = 241 \pm 4$ Å. This very open arrangement corresponds to $V_M = 4{\cdot}42$ Å^3/dalton, i.e. about 72% solvent, a value higher than has been observed for any smaller crystalline protein. Two other recent studies of different modifications of crystalline catalase (Glauser & Rossmann, 1966; Burgner & McGandy, *A.C.A. Abstr.* Meeting, January 1967) lead to V_M values of 2·52 and 2·86 Å^3/dalton, both of which are above average, although within normal limits for smaller proteins. It must be emphasized strongly, however, that the results quoted here cannot be expected to apply in all other cases. For example, the muscle protein tropomyosin can be crystallized in a three-dimensional lattice, but these crystals have the remarkably high content of about 95% solvent (C. Cohen & D. L. D. Caspar, personal communication).

Relatively few X-ray diffraction studies of crystals of spherical viruses have been reported, and these are not included in the present summary. Although a complete survey was not made of the recent literature, the following examples, which include all X-ray studies listed by King (1963), are probably fairly representative: Southern bean mosaic virus, mol. wt $= 6{\cdot}63 \times 10^6$, $V_M = 2{\cdot}68$ Å^3/dalton (Magdoff, 1960). Polio virus type 1, mol. wt $= 6{\cdot}7 \times 10^6$, $V_M = 3{\cdot}2$ Å^3/dalton (Finch & Klug, 1959). Tomato bushy stunt virus, mol. wt $= 9 \times 10^6$, $V_M = 3{\cdot}2$ Å^3/dalton (Caspar, 1956). Broad bean mottle virus, mol. wt $= 5 \times 10^6$, $V_M = 3{\cdot}4$ Å^3/dalton (Finch, Leberman & Berger, 1967). These have a mean value of $V_M = 3{\cdot}1$ Å^3/dalton, and all are considerably above the average for small protein molecules, but not outside the upper limits which have been observed. In addition, several reports have been made of the crystal structure of turnip yellow mosaic virus (Klug, Longley & Leberman, 1966, and references quoted therein), but in this case the number of virus particles contained in the unit cell does not seem to be well defined. Klug *et al.* (1966) propose that there are usually eight virus, particles per unit cell arranged in a diamond-type lattice, but that in some preparations this number may be increased up to a maximum of sixteen particles per cell to give a double diamond lattice. Assuming a molecular weight of $5{\cdot}5 \times 10^6$ for the virus, the corresponding values of V_M would range from 7·6 to 3·8 Å^3/dalton. In this case it is also possible to obtain isomorphous crystals of the "top component", i.e. of the protein shell of the virus with the core of nucleic acid removed. Clearly if one assumed a molecular weight for the top component particles equal to that of the protein shell, i.e. about 63% of the molecular weight of the intact virus, even higher values for V_M would be obtained.

The majority of the crystals described were obtained by salting out with concentrated solutions of ammonium sulfate or of phosphate. For those crystals grown from solutions containing alcohol the mean value of V_M is 2·3 Å^3/dalton, which is not significantly different from the over-all mean value of 2·37 Å^3/dalton. In other words, there is no indication that the different crystallization systems which have been used have any systematic effect on the solvent content of the crystals. Also there does not appear to be any correlation between the degree of symmetry of the crystals and the amount of solvent contained in them; neither is there any obvious relation between the solvent content and the "polarity ratio" of the protein (Fisher, 1964).

REFERENCES

Aschaffenburg, R., Green, D. W. & Simmons, R. M. (1965). *J. Mol. Biol.* **13**, 194.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Nature*, **206**, 757.

Borisov, V. V., Melik-Adamjan, V. R., Suckever, N. E. & Andreeva, N. S. (1966). *Acta Cryst.* **21**, A157.

Brändén, C. I. (1965). *Arch. Biochem. Biophys.* **112**, 215.

Caspar, D. L. D. (1956). *Nature*, **177**, 475.

Corey, R. B., Battfay, O., Brueckner, D. A. & Mark, F. G. (1965). *Biochim. biophys. Acta*, **94**, 535.

Crick, F. H. C. (1957). In *Methods in Enzymology*, ed. by S. P. Colowick & N. O. Kaplan, vol. 4. New York: Academic Press.

Crick, F. H. C. & Kendrew, J. C. (1957). *Advanc. Protein Chem.* **12**, 134.

Dickerson, R. E., Kopka, M. L., Weinzierl, J. V., Eisenberg, D. & Margoliash, E. (1967). *J. Biol. Chem.* **242**, 3015.

Drenth, J. & Wiebenga, E. H. (1955). *Rec. Trav. Chim. Pays-Bas.* **74**, 813.

Dunnill, P., Green, D. W. & Simmons, R. M. (1966). *J. Mol. Biol.* **22**, 135.

Einstein, J. R. & Low, B. W. (1962). *Acta Cryst.* **15**, 32.
Finch, J. T. & Klug, A. (1959). *Nature*, **183**, 1709.
Finch, J. T., Leberman, R. & Berger, J. E. (1967). *J. Mol. Biol.* **27**, 17.
Fisher, H. F. (1964). *Proc. Nat. Acad. Sci., Wash.* **51**, 1285.
Fridborg, K., Kannan, K. K., Liljas, A., Lundin, J., Strandberg, B., Strandberg, R., Tilander, B. & Wirén, G. (1967). *J. Mol. Biol.* **25**, 505.
Glauser, G. & Rossmann, M. G. (1966). *Acta Cryst.* **21**, 175.
Greer, J., Perutz, M. F. & Rumen, N. (1966). *J. Mol. Biol.* **18**, 547.
Harding, M. M., Crowfoot-Hodgkin, D., Kennedy, A. F., O'Connor, A. & Weitzmann, P. D. J. (1966). *J. Mol. Biol.* **16**, 212.
Huber, R., Formanek, H., Braun, V., Braunitzer, G. & Hoppe, W. (1964). *Ber. Bunsenges. Phys. Chemie*, **68**, 818.
Kartha, G., Bello, J. & Harker, D. (1967). *Nature*, **213**, 862.
Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G. & Davies, D. R. (1960). *Nature*, **185**, 422.
King, M. V. (1963). In *Crystal Data*, ed. by J. D. H. Donnay & G. Donnay, *A.C.A. Monograph No. 5*, p. 1263. American Crystallographic Association.
Klug, A., Longley, W. & Leberman, R. (1966). *J. Mol. Biol.* **15**, 315.
Kraut, J., Sieker, L. C., High, D. F. & Freer, S. T. (1962). *Proc. Nat. Acad. Sci., Wash.* **48**, 1417.
Longley, W. (1967). *J. Mol. Biol.* **30**, 323.
Ludwig, M. L., Hartsuck, J. A., Steitz, T. A., Muirhead, H., Coppola, J. C., Reeke, G. N. & Lipscomb, W. N. (1967). *Proc. Nat. Acad. Sci., Wash.* **57**, 511.
Magdoff, B. S. (1960). *Nature*, **185**, 673.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 499.
Matthews, B. W., Sigler, P. B., Henderson, R. & Blow, D. M. (1967). *Nature*, **214**, 652.
Muirhead, H., Cox, J. M., Mazzarella, L. & Perutz, M. F. (1967). *J. Mol. Biol.* **28**, 117.
North, A. C. T. (1959). *Acta Cryst.* **12**, 512.
Perutz, M. F. (1949). *Research*, **2**, 52.
Perutz, M. F., Bolton, W., Diamond, R., Muirhead, H. & Watson, H. C. (1964). *Nature*, **203**, 687.
Perutz, M. F. & Mazzarella, L. (1963). *Nature*, **199**, 639.
Poljak, R. J., Dintzis, H. M. & Goldstein, D. J. (1967). *J. Mol. Biol.* **24**, 351.
Rossmann, M. G., Jeffery, B. A., Main, P. & Warren, S. (1967). *Proc. Nat. Acad. Sci., Wash.* **57**, 515.
Rossmann, M. G. & Labaw, L. W. (1967). *J. Mol. Biol.* **29**, 315.
Sieker, L. C. & Jensen, L. H. (1965). *Biochem. Biophys. Res. Comm.* **20**, 33.
Sigler, P. B., Skinner, H. C. W., Coulter, C. L., Kallos, J., Braxton, H. & Davies, D. R. (1964). *Proc. Nat. Acad. Sci., Wash.* **51**, 1146.
Watson, H. C. & Banaszak, L. J. (1964). *Nature*, **204**, 918.
Wyckoff, H. W., Hardman, K. D., Allewell, N. M., Inagami, T., Tsernoglou, D., Johnson, L. N. & Richards, F. M. (1967). *J. Biol. Chem.* **242**, 3749.

A computer controlled film scanner for x ray crystallography

B W Matthews, C E Klopfenstein and P M Colman
Institute of Molecular Biology and Department of Chemistry, University of Oregon, Eugene, Oregon, USA

MS *received 28 September 1971, in revised form 6 December 1971*

Abstract A description is given of a computer controlled rotating drum film scanner suitable for the measurement of x ray diffraction photographs. The advantages of such a densitometer for protein crystallography are discussed. By measuring integrated densities 'on line' and taking advantage of the high positional accuracy of the scanner, a very satisfactory combination of speed, accuracy and flexibility can be achieved. It is found that commonly used x ray films are significantly nonlinear even at optical densities less than 1·2 but by using a parabolic modification of the integrated densities the useable density range can be extended to optical densities of at least 2·5.

1 Introduction

During the last few years there has been a revival of interest in the photographic method as a viable technique for measuring diffraction data, particularly for crystal structure analyses of macromolecules. In parallel with this revival has been the development of a number of automatic integrating microdensitometers (Drenth *et al.* 1964, Abrahamsson 1966, Arndt *et al.* 1968, Xuong 1969).

Arndt *et al.* list the requirements which should be fulfilled by a densitometer for macromolecular crystallography: it must measure integrated densities at a rate of not less than one per second, the results must be in a form suitable for further computer processing, and the instrument must be versatile. One might also add the overriding requirement that the instrument be accurate.

The instruments in widest use at the present time are of two quite different types; firstly computer-linked cathode ray tube instruments, for example as described by Arndt *et al.* (1968); secondly rotating drum scanners (Abrahamsson 1966, Xuong 1969, Nockolds and Kretsinger 1970). The advantages of the cathode ray scanner are its speed and flexibility. On the other hand a drum scanner has higher positional accuracy, and can measure optical densities accurately over a much wider range than the cathode ray instrument which becomes increasingly inaccurate above optical densities of 1·2 because of stray light from the halo which surrounds the light spot on the film (Arndt *et al.* 1968). A number of alternative modes of obtaining integrated densities from drum scanners have been

described (Abrahamsson 1966, Xuong 1969, Werner 1969, Nockolds and Kretsinger 1970). In these schemes the film is first scanned sequentially and the individual densities stored on a large capacity intermediate storage device such as a disc or magnetic tape which is processed subsequently to obtain the integrated intensities. Such a mode of operation has several disadvantages. For example it is necessary to measure and store many more densities than are used in the determination of the integrated reflection densities, and for fine intervals of sampling the demands on the capacity of the intermediate storage device may become prohibitive. In addition, if the intermediate storage device is a magnetic tape, and subsequent data processing is done on a large computer the extensive time necessary for reading the tape will lead to inefficient data processing. Also, with an off line scanner it is in general not possible to check sample data during the measurement of a film and thereby to optimize the measurement strategy.

We wish to describe a computer controlled drum scanner in which the computer actively controls the position of the scanner and in which integrated densities are obtained 'on line'. The instrument is fast, flexible and accurate, and we believe it to combine the best features of computer controlled cathode ray scanners, and 'passive' drum scanners.

2 General description

The instrument consists of a Photoscan model B digital microdensitometer interfaced to a Varian 620i computer.† The scanner has been described by Nockolds and Kretsinger (1970).

The film is mounted on a drum 360 mm in circumference which rotates at 4 rev s^{-1}. An optical head mounted on a lead screw moves the source outside and the detector inside the drum along a path parallel to the axis of rotation. The lead screw is driven by a stepping motor such that each step corresponds to 12·5 μm of film. Interface electronics allow moving up to 225 steps in either direction by a single command from the computer. Start pulses for the 8 bit ADC are obtained from an encoder every 50, 100 or 200 μm of rotation. Two conversions are packed into one word and data converted to optical densities are entered into the computer. The actual optical density is referenced to transmission through air, and the optical density range is set by a switch to be 0–2 or 0–3.

To support the densitometer, several routines which could be called by Fortran programs were prepared. These include routines to define an initial axial (i.e. X) position of the densitometer as origin, and to move the densitometer to any X value relative to this origin. Another routine is used to read a row of densities between desired Y (i.e. circumferential) values, and to pack four density values per computer word. The final routine unpacks the density data prior to evaluating the integrated reflection intensity.

In its present configuration the Varian Computer has 8000 words of core storage which is sufficient to allow the programs for intensity measurement to be written in Fortran. We have found this to be sufficient for conventional precession photographs, although additional core store would be desirable for processing oscillation photographs, or 'screenless' precession photographs (Xuong *et al.* 1968).

3 Programs for intensity measurement

We have written computer programs to measure integrated densities from conventional x ray precession photographs in which the reflections lie on a regular lattice. A typical photograph of the protease thermolysin is shown in figure 1. The photograph contains about 6000 reflections which are well resolved in the vertical direction (spacing 1·4 mm), but almost overlap horizontally (spacing 0·9 mm). Before mounting the film on the densitometer three small fiducial holes are punched in the film, two at the extremities of the horizontal axis and one at the extremity of the vertical axis.

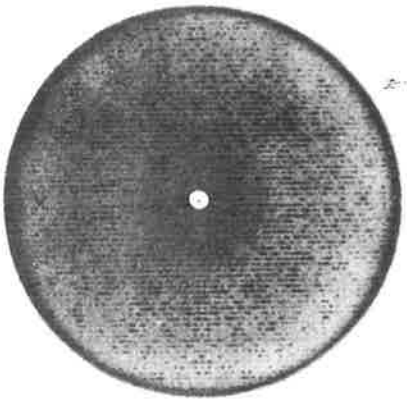


Figure 1 Precession photograph of the ($h0l$) zone of thermolysin. Space group $P6_122$, $a = 94{\cdot}2$ Å, $c = 131{\cdot}4$ Å. Precession angle 20°, crystal to film distance 75 mm, crystal size approximately 0·35 mm in each direction. Photograph taken with an Elliott rotating anode generator, 200 μm focal spot, exposure time 40 h at 40 kV, 40 mA

We have found it convenient to carry out the measurement of a given film in three stages. In the first stage the positions of the three fiducial holes are found by scanning three regions, each about 1 cm square, and each containing one fiducial hole. From the positions of the three holes a transformation is obtained which serves to define the approximate position of any selected reflection on the film.

In the second stage a number of reference reflections, say twelve, are selected. Then, using the approximate transformation, the scanner is sent to each reference reflection in turn where a small region of density values covering the reflection is measured, and hence the exact centre of optical density found (cf. Arndt *et al.* 1968). A least-squares fit to the twelve reference reflections is then carried out to find the 'best' overall transformation which, as will be shown below, may be used to generate the coordinates of any reflection on the film with an accuracy such that no further refinement is necessary. The least-squares analysis is also used to obtain the 'best' lattice parameters, or unit cell dimensions for the film being measured. The program has provisions for checking the correctness of the refined transformation by recentring selected reflections, and by listing optical density values in the neighbourhood of any reflection. The latter facility is very useful in selecting optimum regions for the measurement of peak and background densities.

The third program measures the integrated background-corrected densities. The region on the film within which reflections are to be measured is specified by an inner and outer radius, and by index limits (cf. Arndt *et al.* 1968). The integrated peak density and background density for each reflection are determined by summing over appropriately chosen rectangles. For example with the film shown in figure 1, each peak integration was obtained by summing over a raster of 9×7 points centred on the reflection, and each background was obtained from two regions of 9×3 points above and below the reflection, midway between it and its neighbours. The individual density measurements were made at intervals

† Both the scanner and interface were obtained from Optronics International, Inc, 18 Adams Street, Burlington, Mass., 01803, USA.

of 100 μm. To achieve maximum efficiency the program scans and stores vertical 'strips' of density, chosen to contain as many reflections as possible commensurate with the free core storage of the computer. The integrated densities for these reflections are then obtained directly by using the accurate transformation to predict the peak and background positions for each reflection. Since no further refinement is carried out this is an extremely fast operation and is limited by the data acquisition speed of the scanner. Provision is included for ignoring strong or weak reflections. By using the accurate transformation one can obtain directly the central density for a reflection without checking all possible densities in the 'peak' region. We invariably find the 'central' density to be equal to or very close to the peak density. If this central density is greater than a preset value (e.g. 2·5 OD) no measurement is made. Alternatively if the central density does not exceed the local background by a preset amount, the reflection is ignored. The latter check is useful for omitting weak reflections from second and subsequent films in a film pack without the need to specify these reflections explicitly as measured on a stronger film. The reflection indices and integrated densities are written out on a high speed paper tape punch, or, for preference, on magnetic tape. Additional data may be listed from time to time on the teletype under sense switch control as a check on the correct functioning of the program.

The total time taken by the two alignment programs is about 4–6 min, most of which is for input and output of data. The third program then measures integrated background-subtracted densities at about 2 reflections per second. This rate is limited by the time taken to transfer the desired 'strips' of density from the film to the computer, and could be substantially increased, for example by increasing the rotation speed of the scanner. Also in our present mode of operation the length of the 'strip' is limited by the core storage of the computer. We estimate that an additional 2000 words of core storage would result in a data acquisition rate of about 5 reflections per second. The overall measuring speed is slower than could be achieved by operating the scanner in an off line mode, but in general the off line mode will be much more expensive. We believe the saving in computation cost to be a significant advantage of the present method.

4 Accuracy of the scanner

Some tests of the accuracy of density measurement with the Optronics scanner have been given by Nockolds and Kretsinger (1970). We tested the positional and density reproducibility of the scanner under computer control by moving the scanner back and forth from some arbitrary origin to the same strong reflection and repeatedly measuring individual density values across this reflection. Since the density varied rapidly across the spot, any positional variation of the scanner would cause large fluctuations in the repeated density measurements. This and all other tests were carried out with a sampling interval and light spot size of 100 μm, and the instrument set to read optical densities in the range 0–3·0.

Results of a typical test in which the densities were measured ten times are summarized in table 1. The standard deviations quoted are each the average of ten sets of measurements within the range of density indicated.

For the upper part of the range these values are comparable with those found by Nockolds and Kretsinger (1970), but at lower optical densities they are better by a factor of two. For comparison, the cathode ray instrument described by Arndt *et al.* (1968) has similar accuracy (standard deviation 0·006) up to an optical density of 1·2, but beyond this value there is an increasing departure from linearity amounting to 20% at optical density 2·0. The results quoted in table 1 also indicate

Table 1

Optical density range	Standard deviation
0–1·0	0·007
1·0–2·0	0·013
2·0–2·5	0·017

that the positional reproducibility of the scanner is very satisfactory, with an estimated error of less than about 3 μm.

Other tests of the positional and density accuracy of drum scanners have been carried by Xuong (1969) and Nockolds and Kretsinger (1970) using reference grids and standard density scales. Rather than repeating such measurements we have based further testing on various measurements of precession photographs. In particular it is necessary to justify the use of the scheme described above for the measurement of integrated densities, i.e. to show that a single transformation derived from a selected set of reference reflections, can be used to predict the position of all other reflections on the film with sufficient accuracy that no other recentring is necessary. We find that for a typical high angle precession photograph the RMS discrepancy between the refined centres of density of a set of twelve reference reflections and the 'best' transformation derived from them is about 20 μm. This transformation may then be checked by comparing the 'predicted' centres of other reflections with their individually refined centres of optical density. In a typical test, averaging over 32 different reflections distributed across a film we found that the RMS 'correction' was 22 μm, and the greatest shift 41 μm. Since optical densities are measured at intervals of 100 μm, there is clearly nothing to be gained by refining the positions of individual reflections. For films as in figure 1, where the spots are barely resolved, it is often impossible to meaningfully refine the position of a weak reflection next to a very strong reflection, and in data acquisition schemes where individual refinement is used, special precautions are necessary to handle such cases. When an overall transformation is used, as suggested here, the predicted position of a weak reflection will be as accurate as that for a strong reflection, and no special tests, other than those discussed above for possible rejection of weak reflections need be applied.

The ultimate test of the performance of a film scanner for x ray diffraction must be in terms of the accuracy of measurement of integrated densities. Estimates of this accuracy can be obtained by comparing observed integrated densities for reflections which, by symmetry, are expected to be equal in intensity, or alternatively, comparisons can be made between reflections recorded on different films. Comparisons of reflections recorded on successive films in a film pack provide a sensitive test for systematic errors in the observed densities. Arndt *et al.* (1968) define a symmetry averaging reliability factor

$$R_{\mathrm{sym}}=\frac{\Sigma \mid I_1(\mathbf{h})-\bar{I}(\mathbf{h}) \mid}{\Sigma I_1(\mathbf{h})}$$

where $\bar{I}(\mathbf{h})$ is the mean observed density of two or four symmetry related reflections, and the $\mathbf{h}$ summation is taken over all measured reflections. Similarly a reliability index for film scaling is defined to be

$$R_{\mathrm{sca}}=\frac{\Sigma \mid \bar{I}_1(\mathbf{h})-k\,\bar{I}_2(\mathbf{h}) \mid}{\Sigma \bar{I}_1(\mathbf{h})}$$

where $\bar{I}_1(\mathbf{h})$ is the symmetry averaged intensity of a reflection on the stronger film and $\bar{I}_2(\mathbf{h})$ is the symmetry averaged intensity of a reflection on the weaker film, k is the scale factor and the summations are over reflections common to both films.

For a typical thermolysin photograph with a pack of two films R_{sym} is taken over about 8000 reflections, and $\bar{I}(\mathbf{h})$ is an average of four symmetry related reflections. R_{sca} is typically taken over about 500 reflections. Reliability factors, averaged over a number of film packs are given in table 2, and are compared with values reported elsewhere. On the first film all reflections with central (i.e. peak) densities up to an optical density of 2·5 were measured. For the second film the upper limit was the same, but very weak reflections with a central density of less than 0·06 above the surrounding background density were ignored. Also, in measuring the second film we generally save time by reducing the 'outer radius' of the film, thus ignoring the weak high angle reflections, all of which will have already been measured on the first film. Before calculating

Table 2

R_{sym}	R_{sca}	Maximum density	Source
5·1%	4·4%	1·2	Arndt *et al.* (1968)
6·0%	4·0%†	2·0	Xuong (1969)
—	7·6%‡	2·3	Nockolds and Kretsinger (1970)
4·4%	3·4%	2·5	This work (average film)
3·6%	2·9%	2·5	This work (best film)

† $R_{sca}=[\Sigma(I_1-kI_2)^2/\Sigma I_1^2]^{1/2}$
‡ Reflections not symmetry averaged.

R_{sca} the observed densities were modified by a parabolic correction, as described in the following section, to compensate for the nonlinearity between intensity and density which is appreciable at the higher densities.

It should be emphasized that the respective reliability factors quoted in table 2 will in part reflect the nature and quality of the photographs used. It is clear from table 2 that the accuracy obtainable with an automatic densitometer can be quite high. It is also clear that by using a data measurement scheme as proposed here results are obtained which are at least as accurate as found by other photographic methods.

5 Density and intensity

In measuring data photographically it is desirable that the characteristic relation between incident x ray intensity and photographic density be linear over as wide a range as possible. A large useable range will reduce the number of films necessary to record all intensities within the linear region, and will reduce cumulative errors introduced through film to film scaling factors.

Morimoto and Uyeda (1963) determined the characteristic curve of a number of commercially available x ray films and found, for example, that for Ilford Industrial G and for Kodak No-Screen x ray films the relation between intensity and density is *approximately* linear up to densities of about 1·2. Beyond this value there is an obvious increasing discrepancy between intensity and density. It may be noted, however, that the data of Morimoto and Uyeda suggest that even in the optical density range 0–1·0, which is generally considered to be the 'linear region', there is a *systematic* difference between intensity and density. In other words, the nonlinearity between intensity and density which is obvious at high densities is also present although less pronounced at lower densities. This may be verified by fitting the data of Morimoto and Uyeda to a parabolic curve of the form

$$I=aD+bD^2 \qquad (1)$$

where I is the incident intensity of x rays, D the observed density on the film, and a and b are constants. It is found that over the range of densities investigated, i.e. 0–2·0, for both Ilford Industrial G and No-Screen Kodak film, parabolic relations between density and exposure fit the experimental optical density data within a maximum error of 0·01, i.e. within the accuracy with which the observed densities are quoted. This agreement is significantly better than can be obtained by fitting a straight line ($b=0$) to the observed data, even when the higher densities are excluded. For example, for both films quoted above, the observed values for intensity plotted against density in the optical density range 0–1·2 depart from linearity by up to 0·02. The significance of the nonlinearity at lower densities may be illustrated by noting that for Ilford G and Kodak No-Screen films, the nonlinearity term bD^2 at observed densities of 1·0 is approximately 11% and 8% respectively.

It is clear from these results that in order to extend the useful density range for the films quoted it is essential to take account of the nonlinearity term. Xuong (1969) partially accomplished this by biasing the film scanner logarithmic amplifier circuit, and achieved a linear response over the optical density range 0·6–2·0. A similar method would be to numerically modify the observed densities with an empirical correction determined from standard intensity scales developed under the same conditions as used for the precession photographs. Possible errors in these methods might arise from variations in the development conditions, in the amount of background fogging, or from variations in different batches of film.

Another method which may be used to correct for nonlinearity of the films, and so to extend the useable density range, is to modify directly the background subtracted integrated densities. This approach is an extension of the method developed by R Henderson and J Barrington-Leigh to compensate for nonlinearity in a cathode ray densitometer (Arndt *et al.* 1968) due to the 'Wooster effect' (Wooster 1964) discussed in more detail below (Henderson 1969). With a drum scanner the main source of error is lack of film linearity, rather than intrinsic densitometry errors, and we have found that in such cases the method of Henderson and Barrington-Leigh may be used to enable measurement of diffraction spectra with peak densities up to at least 2·5. The method has the great advantage that it is 'self-calibrating', i.e. the magnitude of the nonlinearity correction is determined automatically for each film pack.

It is not self evident to predict the error that will be introduced in an integrated background subtracted density measurement as a consequence of the parabolic relation between density and intensity for individual density measurements. In the first place the effect of subtracting the background density should be taken into account, and in the second place allowance should be made for the fact that the peak integration is made over a range of individual densities.

We will consider first, for a single density measurement, the effect of subtracting the background. If the peak density is denoted D_p and the background density D_b, then applying the parabolic modification (1) we have

$$I_p=aD_p+bD_p^2 \qquad (2)$$

and, assuming the parabolic exponents to be the same for the background,

$$I_b = aD_b + bD_b^2. \tag{3}$$

Subtracting (3) from (2), the background corrected intensity is given by

$$I = (I_p - I_b) = a(D_p - D_b) + b(D_p^2 - D_b^2). \tag{4}$$

If, on the other hand the parabolic modification were to be applied after subtracting the background from thepeak density, then the resultant intensity would be

$$\begin{aligned} I' &= a(D_p - D_b) + b(D_p - D_b)^2 \\ &= a(D_p - D_b) + b(D_p^2 + D_b^2 - 2D_pD_b) \end{aligned} \tag{5}$$

For I' to equal I we require either that

$$-D_b^2 = D_b^2 - 2D_pD_b \tag{6}$$

or that both D_b^2 and $(D_b^2 - 2D_pD_b)$ be small compared with D_p^2. For equation (6) to hold it is required that $D_b = D_p$, i.e. that the peak and background densities are equal. This will be approximately true for all the weak reflections. The alternate requirement for I to equal I' will be satisfied if D_b is small compared with D_p, i.e. if the background density is much less than the peak density. This will be true for the most intense reflections. For films similar to that shown in figure 1, the alternate requirements for I' to equal I are approximately satisfied for all reflections except the few innermost ones which have a medium density superimposed on a relatively high background.

We now consider whether it is justifiable to apply a parabolic modification not to the individual density measurements, but to their sum, i.e. to an integrated density measurement. For simplicity it will be assumed that the background density is sufficiently small to be ignored. We will also assume that all reflections on a given film have the same shape, i.e. it is assumed that if the film response were perfectly linear, then all reflections would have the same density distribution. This will be approximately true for precession photographs such as that shown in figure 1 except for spot doubling due to crystal misalignment etc. Let the density at some point i in a 'standard reflection' (h_0, k_0, l_0) be D_i, and the integrated density be $D(\mathbf{h}_0)$ where

$$D(\mathbf{h}_0) = \sum_i D_i \tag{7}$$

If all reflections are of the same shape then the density at the ith point of another reflection (hkl) of integrated density c times that of $D(\mathbf{h}_0)$ will be, by definition cD_i. If the relation between intensity and photographic density for an individual density cD_i is parabolic, and given by equation (1), then the integrated intensity for this reflection is given by

$$I(\mathbf{h}) = \sum_i [acD_i + b(cD_i)^2]. \tag{8}$$

If, on the other hand, the individual densities are summed before a parabolic modification is applied then one would have

$$I'(\mathbf{h}) = A \sum_i cD_i + B\left(\sum_i cD_i\right)^2 \tag{9}$$

where A and B are the parabolic exponents for integrated density, analogous to a and b in equation (1), but not necessarily equal to them. To justify the application of the parabolic correction (9) to the integrated densities it is necessary to show that values of A and B can be found such that $I(\mathbf{h}) = I'(\mathbf{h})$ for all reflections on the film, i.e. for all possible values of c. Equating coefficients of c and c^2 in equations (8) and (9), the conditions to be met are that $A = a$, which is trivial, and that

$$B = b \sum_i D_i^2 / \left(\sum_i D_i\right)^2 \tag{10}$$

which is a constant characteristic of the density distribution of the 'standard reflection'. In summary, the relation between x ray diffraction intensity and integrated background-subtracted density can be considered as parabolic if all reflections are of the same shape, and if either the background density is low or the reflection density considerably exceeds that of the background.

Assuming these conditions to be met we write

$$I(\mathbf{h}) = AD(\mathbf{h}) + BD^2(\mathbf{h}) \tag{11}$$

where $I(\mathbf{h})$ is the intensity of the reflection, $D(\mathbf{h})$ the integrated background subtracted density, and A and B are constants. Without loss of generality the proportionality constant A may be put equal to unity. If k is the film scaling factor between successive films in a film pack, say the ith and jth, then

$$I_i(\mathbf{h}) = kI_j(\mathbf{h}) \tag{12}$$

for all reflections.

In the method of Henderson and Barrington-Leigh, k and B are determined from the criterion that the sum

$$\sum \frac{1}{k^2} [I_i(\mathbf{h}) - kI_j(\mathbf{h})]^2 \tag{13}$$

taken over all reflections measured on both the ith and jth films should be a minimum. Henderson and Barrington-Leigh showed that to a good approximation the values of B and k which minimize (13) are given by the following equations, where $B = B'k/(k-1)$

$$B' = \frac{\Sigma I_i^2 \Sigma I_i^2 I_j - \Sigma I_j^3 \Sigma I_i I_j}{\Sigma I_i^4 \Sigma I_i I_j - \Sigma I_i^3 \Sigma I_i^2 I_j} \tag{14}$$

$$k = \frac{\Sigma I_i^4 \Sigma I_i^2 - \Sigma I_i^3 \Sigma I_i^3}{\Sigma I_i^4 \Sigma I_i I_j - \Sigma I_i^3 \Sigma I_i^2 I_j}. \tag{15}$$

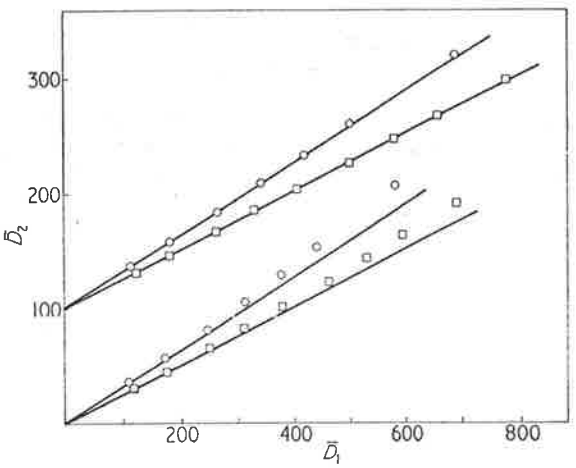


Figure 2 Average integrated density for reflections measured on the second film in a film pack plotted against the average integrated density observed for reflections on the first film. In the upper curves (origin translated 100 units) the observed data have been modified by a parabolic nonlinearity correction as discussed in the text. □, Kodak No-Screen x ray film; ○, Ilford Industrial G film. The straight lines are of slope k where k was determined from equation (15). Each plot was obtained from a bout 2000 individual intensity measurements

To illustrate the effect of the parabolic modification we have plotted in figure 2 the average of the integrated densities measured from successive films in a film pack, both before and after modification. The improvement and the increased density range achieved through the parabolic modification are apparent.

Another method of illustrating the nonlinearity effect is by plotting the variation of the ratio $D_1(\mathbf{h})/D_2(\mathbf{h})$ as a function of the density $D_1(\mathbf{h})$. It can be shown from (11) and (12) that neglecting second order terms

$$\frac{D_1}{D_2}=k[1-B(D_1-D_2)]. \tag{16}$$

Hence, if the nonlinearity is not too pronounced we have

$$\frac{D_1}{D_2}=k-B(k-1)\,D_1. \tag{17}$$

In other words a plot of D_1/D_2 against D_1 should be approximately a straight line with intercept equal to the film to film scale factor k, and slope equal to $-B(k-1)$, where B is the nonlinearity term in equation (11). Such plots, made by grouping the integrated densities into zones, are shown in figure 3. Similar plots have been given in figure 4 of Werner's (1970) paper, and in figure 7 of the paper of Nockolds and Kretsinger (1970). In each case the ratio D_1/D_2 decreases with D_1, and the dependence appears to be approximately linear. The significance of these results can now be adequately explained in terms of equation (17). It may be noted that the decrease in D_1/D_2, observed at relatively low values of D_1 (peak density less than unity) tends to confirm that the films tested are significantly nonlinear even at these low density values. Although plots similar to those shown in figure 3 may be useful in summarizing the data we do not suggest that they be used in preference to numerical calculation using equations (14) and (15) for determining accurate values of B and k. The plots have the disadvantage that for small values of D_2, the ratio D_1/D_2 has a large standard error, which may result in a scatter of the observed points in the region of the extrapolation.

As was indicated in figures 2 and 3, we have used both Ilford Industrial G and Kodak No-Screen x ray film for intensity measurement. The Ilford film has less chemical fog, is less grainy, and tends to give 'cleaner' pictures. On the other hand the Kodak film is more nearly linear in its response to x rays, and its higher absorption reduces the number of films required to cover a desired density range. We find that if a parabolic density modification is used, virtually all intensities from a typical protein crystal may be measured from a pack of two Kodak films. In terms of the agreement factors R_{sym} and R_{sca} there seems to be little to choose between the Kodak and Ilford films.

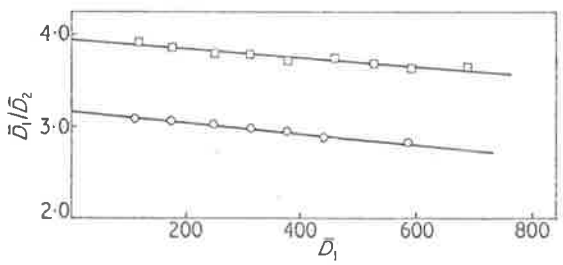


Figure 3 Ratio of the average integrated reflection density for two films plotted as a function of the density on the stronger film. □, Kodak film; ○, Ilford film. The straight lines have intercept k and slope $-B(k-1)$ where B and k were determined for the respective films from equations (14) and (15)

6 Light spot size

To obtain the most accurate intensities from a given x ray film a reasonably small illuminating light spot must be used. However the optimum spot size desirable for a given situation is not obvious. The use of a larger spot will tend to average out film graininess and give a statistically more accurate density. Wooster (1964) suggests that because of the graininess of x ray films it is not worth examining an area less than 100 μm × 100 μm. On the other hand if the spot is made so big that large changes in optical density occur within the sampling area, then the measured film density no longer equals the average optical density for that area (Wooster 1964). This error, referred to previously as the Wooster effect occurs to some degree for all densitometers which measure optical density by means of the transmission of light. Optical density is defined to be the logarithm of the ratio of the incident to transmitted light intensity, and since the logarithm of the average of the transmitted intensity is not the same as the average of the logarithm, an error will be introduced unless the optical density is uniform within the area of the sampling light spot.

For the film shown in figure 1 a typical reflection is contained in a box 0·8 mm × 0·6 mm. The change in density between points 100 μm apart on the sides of a reflection of peak optical density 2·0 is at most about 0·6, e.g. from 1·0 to 1·6. From the formula of Wooster (1964) it follows that a single density measurement with a sampling area of 100 μm × 100 μm would differ from the true mean density by about 3%. Points 200 μm apart within a similar strong reflection may differ in optical density by up to 1·0, which would lead to a maximum Wooster error of about 6% for a single density measurement with spot size 200 μm. These errors are the worst that might be expected. Obviously in determining an integrated intensity many of the contributing densities will have a much smaller Wooster error.

To investigate in a practical case the effect of the light spot size on the integrated intensities we performed the following test. The film shown in figure 1, and the corresponding weaker film were densitometered three times. In each case the individual density measurements were made at intervals of 100 μm, a 9 × 7 raster of points being used for the peak integration, and two 9 × 3 rasters for the background, as described earlier. The three sets of measurements differed only in that the size of the light spot transmitted through the film was set respectively at 50, 100 and 200 μm. The results of the test are given in table 3.

Table 3

Light spot size (μm)	50	100	200
Film to film scale factor k	3·20	3·26	3·23
Parabolic exponent B ($\times 10^6$)	336	402	647
R_{sym} (%)	5·8	4·6	4·1
R_{sca} (%)	4·6	4·2	4·0

Before discussing these results it may be noted that the Wooster effect will always lead to an underestimate of the true density, and will affect strong reflections relatively more than weak ones. It will therefore cause a nonlinearity between true reflection intensity and measured integrated photographic density, similar to that caused by a nonlinear response of the

film to x rays, as discussed above. In fact, it was in an attempt to reduce error due to the Wooster effect in a cathode ray film scanner that Henderson and Barrington-Leigh first developed the parabolic approximation method.

Our purpose in measuring the same film with a different light spot size was to allow differentiation between photographic nonlinearity, and nonlinearity due to the Wooster effect, assuming in both cases the departure from linearity to be essentially parabolic. In table 3 it will be noted that the film to film scale factor k is essentially constant, but the parabolic exponent B increases, particularly for the 200 μm case. A reflection of peak optical density 2·0 has an integrated density of about 400 units. For such a strong reflection the parabolic correction for the 50, 100 and 200 μm cases would be 1·3%, 1·6% and 2·6%, respectively. Since the first two corrections are not very different we may conclude that they both arise mainly from photographic nonlinearity, and that up to a spot size of 100 μm the Wooster effect is essentially negligible for the film tested. On the other hand in going to a 200 μm light spot it appears that for a reasonably strong reflection approximately 1·3% of the total parabolic correction is due to photographic nonlinearity, and that an additional 1·3% correction is necessary to compensate for the Wooster effect. Plots of $\bar{D}_1$ against $\bar{D}_2$ similar to those shown in figure 2, but using data measured with a 200 μm light spot, show that after parabolic modification the points lie almost exactly on a straight line, suggesting that a combination of errors due to both photographic nonlinearity and the Wooster effect can be compensated for by a parabolic modification of the integrated intensities.

In discussing the reliability indices in table 3 it may be noted that R_{sca}, the agreement index between averaged reflections measured on stronger and weaker films is a better measure of accuracy than R_{sym} which measures agreement between reflections of essentially equal intensity on one film, and is not sensitive to any systematic error which is a function of intensity. R_{sym} and R_{sca} are clearly worst for the 50 μm case, presumably because of the larger statistical fluctuations in the individual measurements, and because the spacing between these measurements (i.e. 100 μm) is greater than the light spot diameter. With the 200 μm light spot there will be considerable overlap of the density measurements but such overlap, *per se*, should not introduce errors. The reduction in R_{sym} from 4·6 to 4·1% in going from 100 to 200 μm presumably reflects an improvement in internal consistency within each film due to the better statistical averaging in the 200 μm case. On the other hand R_{sca} remains essentially unchanged in going from 100 to 200 μm, suggesting that the accuracy of the final integrated intensities, after applying the appropriate parabolic modification, is about the same in both cases. In practice we prefer to use the 100 μm light spot which, as has been demonstrated, essentially eliminates possible errors due to the Wooster effect.

Acknowledgments

The parabolic correction method used in this paper was first developed by R Henderson and J Barrington-Leigh. We are particularly grateful to Drs Henderson and D M Blow for sending us details of the method, for acquainting us with the paper of Wooster (1964), and for other helpful comments. We thank D Nelson, K Boekelheide, L Dotson, J Koskinen and S Dixon for their technical assistance. This work was supported in part by the PHS grant GM 15423 and by a Health Science Advancement Award to the University of Oregon (FR06027) from the General Research Support Branch, Bureau of Health Professions Education and Manpower Training, National Institutes of Health.

References

Abrahamsson S 1966 *J. Sci. Instrum.* **43** 931

Arndt U W, Crowther R A and Mallett J F W 1968 *J. Phys.* E: *Sci. Instrum.* **1** 510–6

Drenth J, Kloosterman D, van der Woude J, Croon H C and van Zwet L C M 1965 *J. Sci. Instrum.* **42** 222

Henderson R 1969 *PhD Dissertation, University of Cambridge*

Morimoto H and Uyeda R 1963 *Acta Crystallogr.* **16** 1107

Nockolds C E and Kretsinger R H 1970 *J. Phys.* E: *Sci. Instrum.* **3** 842–6

Werner P E 1969 *Ark. Kemi* **31** 505

Werner P E 1970 *Acta Crystallogr.* **A26** 489

Wooster W A 1964 *Acta Crystallogr.* **17** 878

Xuong N 1969 *J. Phys.* E: *Sci. Instrum.* **2** 485–9

Xuong N, Kraut J, Seely O, Freer S T, and Wright C S 1968 *Acta Crystallogr.* **B24** 289

J. Mol. Biol. (1974) **82**, 513–526

Determination of Molecular Weight from Protein Crystals

B. W. MATTHEWS

Institute of Molecular Biology and Department of Physics
University of Oregon
Eugene, Oreg. 97403, U.S.A.

(*Received 16 April 1973, and in revised form 20 September 1973*)

Different methods for the determination of molecular weight from protein crystals are summarized, and in some cases extended. The use of crystal density measurements is discussed, and several approaches based on such measurements, including density extrapolation (Colman & Matthews, 1971) and crosslinking (Cornick *et al.*, 1973), are described and illustrated by density measurements of crystals of γ-chymotrypsin and β-amylase. The importance of allowing for "bound water" in molecular weight estimation is illustrated by examples, and a summary is given of information which may be gained from densities of crystals soaked in heavy water.

The determination of molecular weight from measurement of the loss of weight of crystals on drying is discussed and illustrated using North's (1959) data.

Also the method of molecular weight estimation from "crystal volume and protein content" (Love, 1957; Heidner *et al.*, 1971) is reviewed, and a new method based on "crystal mass and protein content" is proposed.

1. Introduction

The determination of molecular weights from protein crystals is of interest for two reasons. First, the crystallographic technique can be quite accurate ($\pm 4\%$ in a typical case), and has the advantage over conventional methods that the results are not influenced by dissociation or aggregation of the protein. Second, the symmetry and subunit arrangement of an oligomeric protein can often be determined directly from the molecular weight per unit cell, taken together with the basic crystallographic parameters (e.g. see Matthews & Bernhard, 1973).

This paper summarizes, and in some cases extends, the current theory and practice of molecular weight determination from protein crystals. It will be assumed that the basic crystallographic parameters, i.e. the unit cell parameters and the space group, are known.

2. Crystal Density

(a) *General formulation*

In this section the theory of the determination of protein molecular weight from measurements of the density of single crystals is summarized, and the following related topics are discussed.

(1) The dependence of crystal density on solvent density (Perutz, 1946; Colman & Matthews, 1971).

(2) Allowance for "bound water" in molecular weight estimation.

(3) The use of crosslinking reagents in determining molecular weight (Cornick *et al.*, 1973).

(4) The substitution of H_2O by ^2H_2O in protein crystals.

For a protein, the determination of molecular weight from the crystallographic data is complicated by the fact that the crystal contains not only the protein molecules, but also the intervening solvent (e.g. see Matthews, 1968*a*). In a recent survey of 226 crystal forms of globular proteins (Matthews, 1974) it was found that the non-protein fraction of the unit cell volume varied from 30% to 78%, although for proteins of similar molecular weight the range was somewhat less. Furthermore, the non-protein part of the unit cell is not homogeneous, but, as shown by Adair & Adair (1936) and by Perutz (1946), consists of two parts: a "free solvent" region accessible to external ions and in equilibrium with the external supernatant, and a "bound water" region, not accessible to salts, which can be visualized as a hydration shell surrounding the protein.

The density D_c of the protein crystal can be written as the sum of three parts:

$$D_c = D_p(V_p/V) + D_w(V_w/V) + D_s(V_s/V), \tag{1}$$

where D_p, D_w and D_s are the densities of the protein, bound water and free solvent, (V_p/V), (V_w/V) and (V_s/V) are the fractional volumes occupied by these three components, and V is the unit cell volume (e.g. see Colman & Matthews, 1971).

If the ratio of the bound water volume to the protein volume is denoted f, i.e.

$$f = V_w/V_p, \tag{2}$$

then

$$V_s = V - V_p - V_w = V - V_p - fV_p. \tag{3}$$

Substituting equation (3) into equation (1) we have

$$D_c = D_s - (V_p/V)\,[(1+ f)D_s - fD_w - D_p]. \tag{4}$$

Hence M_p, the molecular weight of protein per asymmetric unit, is given by

$$M_p = \frac{NV_pD_p}{n} = \frac{NVD_p\,(D_c - D_s)}{n\,[D_p - D_s - f(D_s - D_w)]}, \tag{5}$$

where N is Avogadro's number, and n is the number of asymmetric units per unit cell.

Making the simplifying assumption that the density of the protein is equal to the reciprocal of its partial specific volume $\bar{v}_p$, and also using the relation that w, the weight fraction of bound water to protein, is given by

$$w = f\bar{v}_pD_w, \tag{6}$$

then from equation (5) the protein molecular weight per asymmetric unit is given by

$$M_p = \frac{NV\,(D_c - D_s)}{n\,\{1 - \bar{v}_pD_s - w[(D_s - D_w)/D_w]\}}. \tag{7}$$

This is the general expression for the determination of molecular weight from protein crystals. The unit cell volume V is determined from the X-ray diffraction

measurements, and the crystal density D_c is usually measured in a calibrated density gradient (Low & Richards, 1952*a*). The partial specific volume of the protein is most easily estimated from its amino acid composition, if known, or measured pycnometrically in dilute solution (Cohn & Edsall, 1965; McMeekin & Marshall, 1952).

Occasionally protein crystals are grown from water, or from solutions of very low ionic strength, for which the density is equal to that of water, in which case equation (7) reduces to

$$M_p = \frac{NV\,(D_c - D_w)}{n\,(1 - \bar{v}_p D_w)}. \tag{8}$$

In this special case the molecular weight of the protein can be obtained with quite high precision ($\pm 5\%$ or less) from a knowledge of the unit cell volume, the number of molecules in the unit cell, the partial specific volume of the protein and the density of the crystals. However, most protein crystals are grown from concentrated salt solutions and dissolve quite quickly in water, so that equation (8) is not applicable.

It should be emphasized that the derivation of equation (8) from equation (7), and the assumption made above that protein density is reciprocally related to partial specific volume, both imply that the density of the bound water is equal to that of bulk water. In other words, it is assumed that electrostriction (e.g. see Cohn & Edsall, 1965) and the influence of hydrophobic groups (e.g. see Caspar, 1966) have a negligible effect on the density of the bound water. This assumption has some justification in the available crystallographic literature. If, for example, D_w were in fact equal to 1·05 g/cm^3 (see Caspar, 1966) rather than 0·998 g/cm^3, the density of bulk water, then for a typical protein crystal in water the apparent molecular weight calculated using equation (8) would be overestimated by about 5%. A survey of such calculations, using the density data available for proteins such as haemoglobin, α-chymotrypsin, γ-chymotrypsin and thermolysin, suggests no such systematic error. Furthermore, TenEyck (1970) determined, by sedimentation equilibrium experiments, the effect of pressure on the hydration of chymotrypsinogen, and concluded that in this instance the partial molar volume of water changed on hydration by no more than 2·9% in absolute value, the change probably being an increase of about 1%. The survey, and TenEyck's results, also tend to confirm the validity of using partial specific volumes determined pycnometrically or from amino acid composition for crystallographic molecular weight determination, or, conversely, for using crystallographic measurements to determine the partial specific volume of proteins of known molecular weight.

(b) *Experimental*

Measurements of crystal density were made in a manner similar to that described by Colman & Matthews (1971). Using a Pasteur pipette, the crystal is transferred in a droplet of its mother liquor into the water-saturated density gradient (Low & Richards, 1952*a*). Then, using a Hamilton microsyringe, the excess liquid is drawn off, and the crystal gently shaken off the syringe tip. This technique was found to give much more reproducible results than the method of removing excess liquor from the crystal before introducing it into the density column. On occasion, a single crystal was re-wet and re-dried to obtain several density measurements.

In the following sections several aspects of molecular weight determination are discussed, and illustrated by experimental data. The density extrapolation technique is illustrated using the data of Colman & Matthews (1971), and the need to allow for

bound water is demonstrated by calculations based on crystal densities quoted in the literature. The experimental data for crosslinked crystals, and for crystals soaked in deuterium oxide, are original.

(c) *Density extrapolation*

Following the early studies of Perutz (1946), Colman & Matthews (1971) pointed out that it is possible to determine the molecular weights of protein crystals, even in the presence of concentrated salt solutions, by making use of the fact that if the solvent density is varied, then from equation (1), the crystal density will also vary, and that a plot of D_c against D_s will yield a straight line of slope (V_s/V), and of intercept D_0, the density which the crystal would have if all the solvent could be replaced with water. Application of the method to γ-chymotrypsin and to β-amylase (Colman & Matthews, 1971) is illustrated in Figure 1. The observed points lie on straight lines, which may be extrapolated to find the hypothetical crystal density D_0 of the crystal in water. Substituting D_0 for D_c in equation (8), the protein molecular weight may be obtained with an accuracy of about $\pm 4\%$ in a typical case. Details of the density extrapolation technique have been given by Colman & Matthews (1971) and Matthews (1974), and need not be repeated here, except to point out that the slope of the density plot gives the volume fraction of free solvent in the unit cell, and the ratio of the volume of bound water to volume of protein is given by:

$$f = D_p w/D_w = (D_p - D_e)/(D_e - D_w), \tag{9}$$

where D_e is the equilibrium density at which $D_c = D_s$ (Adair & Adair, 1936; Colman & Matthews, 1971).

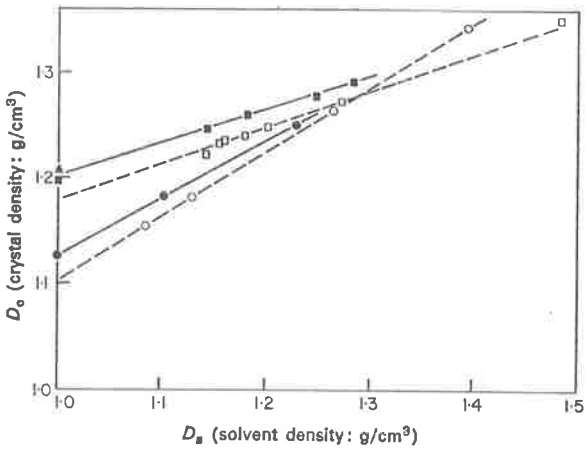


FIG. 1. Plots of crystal density as a function of supernatant density showing the effect of crosslinking with glutaraldehyde. (--□--□--, —■—■—, γ-chymotrypsin; --○--○--, —●—●—, β-amylase). The open symbols are for the normal crystals (data from Colman & Matthews, 1971), and the solid symbols are for the crosslinked crystals. (▲), The density of crosslinked γ-chymotrypsin crystals in water, uncorrected for the expansion of the crystal lattice.

(d) *Allowance for bound water*

Often, crystal densities are measured at only one solvent density, typically that of the supernatant from which the crystals were grown. Consideration of equation (7)

shows that the molecular weight could be found from such measurements if the value of w were known.

In theory, w may be obtained directly by increasing the supernatant density until the crystals remain just suspended (eqn 9), although in practice this is not often done. Note that as the crystal density tends toward the solvent density, then from equation (9), both the numerator and denominator of equation (7) tend to zero, and the molecular weight becomes undefined. In practice, crystal densities measured only for dense supernatant solutions such as saturated ammonium sulfate will tend to give inaccurate molecular weights, because of the increased experimental error in $(D_c - D_s)$ and the increased uncertainty introduced by the $w(D_s - D_w)$ term.

The supernatant density at which the denominator in equation (7) becomes equal to zero is given by:

$$D_s = D_c = D_w(1 + w)/(w + \bar{v}_p D_w), \tag{10}$$

which will equal about 1·25 g/cm^3 in a typical case. As this condition is approached, small errors in either $\bar{v}_p$ or w will introduce large errors in the calculated molecular weight, in addition to the experimental uncertainties described above.

It has been shown in studies of several crystals (e.g. Adair & Adair, 1936; Perutz, 1946; Colman & Matthews, 1971) and by a variety of other techniques (e.g. see Fisher, 1965; Kuntz *et al.*, 1969), that for most proteins the value of w lies in the range 0·15 to 0·35 g water/g protein. Thus, in estimating molecular weight from a pair of crystal and solvent densities, where w is unknown, it would seem preferable to use equation (7), substituting a reasonable value such as 0·25 for w, rather than ignoring the bound water, as is sometimes done.

The consequences of ignoring bound water in the estimation of molecular weights are illustrated in Table 1, in which equation (7) was used to calculate molecular weights from published density data for several protein crystals in a variety of solvents. In one case the bound water was ignored (i.e. w was put equal to zero) and in the other case w was put equal to 0·25 g water/g protein. No example is included in the Table for which the solvent density is close to unity, as in such cases the bound water correction is superfluous.

In each case the omission of the bound water term causes a substantial difference in the molecular weight, the effect becoming more important for higher solvent densities. With the exception of glyceraldehyde phosphate dehydrogenase, the use of $w = 0{\cdot}25$ g water/g protein results in a reasonably accurate value for the calculated molecular weight, suggesting that this is a suitable value to use in practice (although clearly not in preference to a value determined experimentally, as described above). The solvent density for the crystals of glyceraldehyde phosphate dehydrogenase approaches the critical value given in equation (10), and it is likely, therefore, that the apparent error in the calculated molecular weight might be attributed in part to relatively small errors either in the assumed partial specific volume or fraction of bound water.

(e) *Crosslinking*

Recently Cornick *et al.* (1973) have proposed a very simple method of molecular weight estimation, using bifunctional reagents such as glutaraldehyde. Crystals of α-chymotrypsin, soaked for 8 to 24 hours in a 0·25% solution of glutaraldehyde in amine-free phosphate were found to be stable when transferred to water, allowing the crystal density and unit cell dimensions to be measured under salt-free

TABLE 1

Examples of the calculation of molecular weight

Protein	Molecular weight	$\bar{v}_p$ (cm^3/g)	Crystal supernatant	ρ_s (g/cm^3)	ρ_c (g/cm^3)	Calculated† molecular weight ($w = 0$)	Calculated‡ molecular weight ($w = 0{\cdot}25$)	Reference
γ-Chymotrypsin	25,260	0·736	50% saturated $(NH_4)_2SO_4$	1·155	1·232	18,300	24,700	Colman & Matthews (1971)
γ-Chymotrypsin	25,260	0·736	Cs_2SO_4	2·007	1·525	36,000	23,500	Colman & Matthews (1971)
Chymotrypsinogen B								
type B	24,850	0·733	Phosphate	1·092	1·166	19,000	21,300	Matthews (1968*b*)
type C	24,850	0·733	Phosphate, alcohol	1·111	1·226	21,200	25,000	Matthews (1968*b*)
Hemoglobin	66,700	0·749	1·6 M-$(NH_4)_2SO_4$	1·115	1·194	50,300	63,900	Perutz (1946)
Malate dehydrogenase	74,000	0·742	65% saturated $(NH_4)_2SO_4$	1·167	1·227	48,000	70,000	Banaszak *et al.* (1971)
Glyceraldehyde phosphate dehydrogenase	140,000	0·74	3·0 M-$(NH_4)_2SO_4$	1·192	1·243	109,000	184,000	Watson & Banaszak (1964)
β-Amylase	197,000	0·74	30% saturated $(NH_4)_2SO_4$	1·085	1·154	184,000	206,000	Colman & Matthews (1971)
β-Amylase	197,000	0·74	Cs_2SO_4	1·710	1·536	344,000	206,000	Colman & Matthews (1971)

† Calculated mol. wt, ignoring bound water.
‡ Calculated mol. wt, assuming $w = 0{\cdot}25$ gwater/g protein).

conditions. From such measurements Cornick *et al.* (1973) found the molecular weight of α-chymotrypsin to be within 3% of the value determined from the amino acid sequence.

On the other hand, in unpublished experiments Matthews & Colman observed that crosslinking crystals of γ-chymotrypsin caused a significant increase in their density and, therefore, in the apparent molecular weight of the protein. In view of the fact that the simplicity and experimental advantages of the crosslinking technique make it very attractive for molecular weight estimation, and likely to be used in the future, it seemed to be worthwhile to investigate more thoroughly the effect of crosslinking on crystal density.

It was found that crystals of γ-chymotrypsin required overnight soaking in a 2% glutaraldehyde solution in 55% saturated ammonium sulfate, pH 5·6, to prevent dissolution on transfer to water. Such crystals cracked when placed into water, but gave diffraction patterns similar, at least to low resolution, to those of the native protein crystals (Matthews, 1974). Over several hours the diffraction pattern gradually deteriorated, being essentially lost after two days. Crystals of γ-chymotrypsin soaked in 1% glutaraldehyde solutions exhibited more severe cracking, and partly dissolved when placed in water.

The densities of crystals crosslinked with 2% glutaraldehyde, plotted as a function of supernatant density, are shown in Figure 1. As with the non-crosslinked native crystals (Colman & Matthews, 1971), the density relation is linear. When extrapolated to a solvent density of unity, the crystal density is in good agreement with that observed for the crosslinked crystals placed in water.

Very similar results were obtained for crystals of β-amylase, also crosslinked overnight in a 2% solution of glutaraldehyde in 40% saturated ammonium sulfate, pH 4·0. It is clear from the density plots shown in Figure 1, that at least under the conditions used for these experiments, crosslinking caused a significant increase in crystal density and, therefore, in apparent protein molecular weight, amounting to 14% for γ-chymotrypsin and 25% for β-amylase. This is obviously much more than would be expected for simple crosslinking of lysine residues, and suggests that a substantial amount of polymerization of the glutaraldehyde must occur, or that more than one glutaraldehyde molecule reacts with each lysine ε-amino group (Quiocho & Richards, 1964; Richards & Knowles, 1968; Korn *et al.*, 1972). It will also to be noted that for both density plots the slope decreases slightly, from 34% to 31% for γ-chymotrypsin, and from 61% to 54% for β-amylase, presumably reflecting a corresponding decrease in the free solvent volume due to displacement by the glutaraldehyde.

In the case of β-amylase it proved possible to stabilize the crystals with overnight crosslinking in 1% and 0·5% solutions of glutaraldehyde, although not with 0·25%. The densities of such crystals, after transfer to water, were 1·112 g/cm^3 and 1·101 g/cm^3, respectively, i.e. less than the more heavily crosslinked crystals (1·127 g/cm^3). The density of the least crosslinked crystals is, within experimental error, equal to that of the extrapolated density for the untreated crystals. Taken together with the results of Cornick *et al.* (1973), these measurements suggest that if protein crystals can be stabilized by soaking in, say, 0·25 to 0·5% glutaraldehyde solutions, then the error in molecular weight will probably be small (i.e. at most a few per cent), but that molecular weights determined from more heavily crosslinked crystals must be regarded with caution. In practice, an estimate of the approximate increase in

crystal density due to the crosslinking agent could be obtained by determining the difference in density between the native and crosslinked crystals under solvent conditions where both can be measured; for example, under the conditions of crystal growth.

During experiments with the crosslinked γ-chymotrypsin crystals the density in water was observed to decrease slowly with time. This could have been due either to a gradual expansion of the crystal lattice, or to slow diffusion of residual salts out of the crystal. The experiments of Wyckoff *et al.* (1967), showing that ammonium sulfate diffused into and out of crystals of ribonuclease S with a half-time of about 90 seconds, suggested that slow diffusion was unlikely. This was confirmed by measuring the change in density of crosslinked crystals of γ-chymotrypsin on transfer from 55% saturated to fully saturated solutions of ammonium sulfate. Under these conditions the cell dimensions do not change, so that any density changes are due only to the change in salt concentration within the crystal. The results, illustrated in Figure 2, show that in fact the diffusion of salts into and out of the crystals is rapid, with an approximate half-time of about two minutes for the crystals used in these experiments, which were about 0·4 mm in each direction. It also follows from these experiments that for complete solvent exchange, crosslinked crystals should be transferred to water at least 30 minutes before their density is measured.

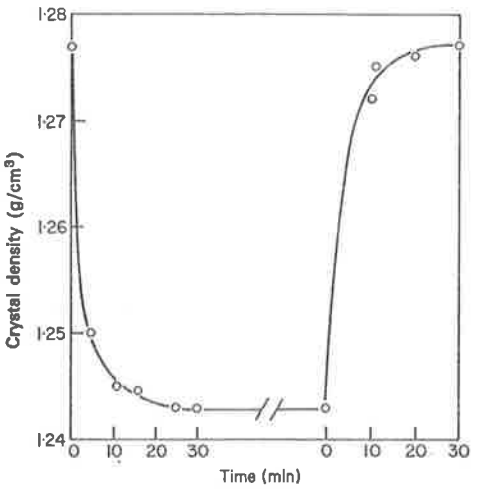


FIG. 2. Density of crosslinked crystals of γ-chymotrypsin measured after transfer from saturated ammonium sulfate to 55% saturated ammonium sulfate, and *vice versa*.

The cell dimensions of crosslinked γ-chymotrypsin crystals were measured directly as a function of time, by taking a series of precession photographs of the same crystal, and the unit cell volume was found to increase by about 6% during one day (Table 2).

For crystals of crosslinked β-amylase, no significant change in cell dimensions was observed on transfer to water, although in this case the poor quality of the diffraction patterns of both the native and crosslinked crystals made accurate measurements impossible.

Whether or not the slow change in cell parameters observed for γ-chymotrypsin is likely to be the rule rather than the exception for crosslinked protein crystals in

TABLE 2

Cell dimensions of crystals of γ-chymotrypsin

Crystal condition	*a*,*b* (Å)	*c* (Å)	Fractional change in cell volume (%)
Native (55% saturated ammonium sulfate)	69·6 (1)†	97·7 (1)	−1·0
Crosslinked (55% saturated ammonium sulfate)	69·9 (1)	97·8 (1)	0
Crosslinked (45 min in water)	71·0 (2)	98·5 (2)	3·9
Crosslinked (90 min in water)	71·3 (2)	99·0 (2)	5·3
Crosslinked (24 h in water)	71·5 (2)	99·1 (2)	6·0

† Numbers in parentheses give the estimated standard deviation in the last place.

water remains to be determined, but it would seem advisable to measure the crystal density and the cell dimensions at approximately the same time.

(f) *Use of heavy water*

In this section a brief summary is given of information that may be gained from density measurement of protein crystals transferred from H_2O to ^2H_2O.

On making such a transfer it would be expected that the deuterium would rapidly exchange with hydrogen in the free solvent and in the bound water regions, and also with most, if not all, of the hydrogen-bonding hydrogens within the protein (Schoenborn, 1971). The occasional water molecules trapped within some protein molecules would also presumably be exchanged. If primes are used to denote the variables that take up new values during this exchange, then one can denote the ^2H_2O density as D_w', the crystal density as D_c', the protein density as D_p', and so on.

In spite of these changes, the volume of free solvent will remain unchanged, so that from equation (1) the slope of a plot of crystal density *versus* solvent density for a crystal in ^2H_2O solutions should be the same as that in H_2O. Therefore, if the slope has been determined from density measurements in water solutions it would in principle be possible to extrapolate from a single density measurement in a ^2H_2O solution to find the density of the protein crystal in pure ^2H_2O.

Writing D_0' for the density of the protein crystal in pure ^2H_2O, it can be shown from equation (7), as it applies in ^2H_2O and H_2O, that the partial specific volume of the protein (in water), i.e. the one quantity that has not been previously determined by crystal density measurements, is given by:

$$\bar{v}_p = (1/D_p) = \frac{(D_w' - D_w) - (D_0' - D_0) + \delta(D_0 - D_w)}{D_w'D_0 - D_wD_0'}, \qquad (11)$$

where δ is the fractional increase in the protein density in the presence of ^2H_2O, i.e.

$$\delta = (D_p' - D_p)/D_p = (\bar{v}_p - \bar{v}_p')/\bar{v}_p' = (M_p' - M_p)/M_p. \qquad (12)$$

An approximate estimate of the magnitude of δ can be obtained by noting that the average number of hydrogen atoms per amino acid residue potentially exchanged by deuterium will be about two, i.e. one in the backbone imino group of every residue,

one to two for most of the polar side chains except lysine 3 and arginine 6, but none for the hydrophobic side chains. Since the average molecular weight per amino acid for a typical protein is about 100, δ will have a value of about 2%.

To test equation (11) in a real case, density measurements were made for crystals of thermolysin both in water and in ^2H_2O. Thermolysin is particularly suitable for such a test, since the crystals are stable in the presence of very low salt concentrations (Matthews *et al.*, 1972; Colman *et al.*, 1972), so that a density extrapolation is unnecessary. Before the density measurements the crystals were soaked overnight in pure H_2O or ^2H_2O, and the supernatant was replaced several times to allow complete exchange. Such crystals had indistinguishable cell dimensions and diffraction patterns, indicating that no change in conformation of the protein had occurred.

The respective densities of the thermolysin crystals were $D_0 = 1{\cdot}176 \pm 0{\cdot}003$ g/cm^3 in H_2O and $D_0' = 1{\cdot}230 \pm 0{\cdot}003$ g/cm^3 in ^2H_2O. Substituting $D_w' = 1{\cdot}105$ g/cm^3, and $D_w = 0{\cdot}998$ g/cm^3, equation (11) leads to a partial specific volume of $0{\cdot}74 \pm 0{\cdot}07$ cm^3/g if δ is assumed to be zero, and $0{\cdot}78 \pm 0{\cdot}07$ cm^3/g if δ is put equal to 0·02. The agreement with the expected value for $\bar{v}_p$ of 0·73 cm^3/g (Ohta *et al.*, 1966) is reasonably good, though it is obvious from the large standard deviations that ^2H_2O substitution cannot be recommended as a general method for the determination of partial specific volume. The large standard deviation is due to the fact that in both the numerator and denominator of equation (11) one is taking the difference between quantities that are not very different in their absolute magnitudes.

In principle it would be possible to avoid changes in the partial specific volume of the protein and, therefore, put δ in equation (11) equal to zero, by using water containing the heavy isotope of oxygen, rather than that of hydrogen, but since the uncertainty in δ is not the main source of error, this refinement would be of limited benefit.

The measurements presented above are of interest in that they provide independent evidence supporting the view that in protein crystals not only is the free solvent exchanged, but also the bound water, since if this were not so it would be expected that the increase in crystal density on going from H_2O to ^2H_2O would be about 0·035 g/cm^3 rather than the observed value of 0·054 g/cm^3.

The bound water is not to be envisaged as a static hydration shell, but rather as a region of rapidly exchanging water molecules, from which the salt ions are excluded (e.g. see Kuntz *et al.*, 1969). This was confirmed by transferring crystals of thermolysin from H_2O to ^2H_2O and measuring their density as a function of time. The results were very similar to those shown in Figure 2, and indicated a diffusion half-time of less than two minutes. It is not possible by the methods described here to distinguish between the time for replacement of the bound water and that of the bulk water in the crystal, except to say that the replacement of both must be substantially complete within three to four minutes.

3. Crystal Water Content

As discussed above, the essential difficulty in estimating molecular weights from protein crystals lies in determining the relative fractions in the crystal of protein and solvent. North (1959) proposed that the necessary information might be obtained from a knowledge of the water content of the crystal, determined by measuring the loss of weight of crystals on drying. There has been renewed interest in the method

following the recent development of a more accurate method for determining the water content (Berthou *et al.*, 1972). These authors have discussed various methods, such as vacuum desiccation, or heating to 110°C, by which the crystals may be dried. Although such methods give consistent results, it has long been known (Haurowitz, 1950) that crystals dried in this manner retain between 5 and 10% of the protein weight in water. This unknown amount of water will of course lead to an uncertainty in the molecular weight. It is also necessary to allow for the fact that the crystal contains not just protein and solvent, but also bound water, as discussed in previous sections (e.g. see North, 1959).

In order to derive general expressions for the calculation of molecular weight from crystal dehydration measurements, the following symbols will be used.

d: the fractional loss of weight of a crystal on drying (in g water/g wet crystal).

s: the fractional mass of salt contained in the protein-free mother liquor (in g salt/g mother liquor). The value of s can be measured by noting the fractional loss of weight of the solvent on drying, or can be calculated from the molarity of the constituents of the mother liquor.

u: the fractional mass of water retained in the crystal on drying (in g water/g protein). The value of u is not readily accessible to experimental measurement, but is generally assumed to be in the range 5 to 10% (Haurowitz, 1950). The term u is not to be confused with w, which denotes the bound water, i.e. the fractional mass of water in the crystal (in g water/g protein) that is inaccessible to salt, and can often be estimated using equation (9).

From consideration of the mass of one unit cell, and the loss of weight on drying, it is straightforward to show that the protein molecular weight is given by:

$$M_p = \frac{NVD_c(1 - s - d)}{n(1 + u - s - sw)}. \tag{13}$$

Although equation (13) was not stated explicitly, the general method of estimating molecular weight from the crystal density, and from s and d, was first proposed by North (1959). It may be noted that in this formulation it is necessary to know or to assume the value of w, whereas no use is made of the solvent density.

In North's case, the molecular weight of hemoglobin in a cubic crystal form was calculated using the value $w = 0{\cdot}3$ g/g, determined by Perutz (1946) for monoclinic crystals of hemoglobin, and u was assumed to be 0·075 g/g. Substituting North's values, $D_c = 1{\cdot}240$ g/cm^3, $d = 0{\cdot}367$ g/g and $s = 0{\cdot}398$ g/g, into equation (13), the molecular weight of half a hemoglobin molecule is calculated to be 37,700. As is apparent from equation (13), the consequences of any error in w become more serious for crystals grown from concentrated salt solutions (i.e. s large), but in a typical case fairly large uncertainties in w can be tolerated. For example, if w is decreased from 0·3 to 0·2 g/g, the calculated molecular weight increases by only 7%. Possible errors in u will have a more serious effect. For example, if u were changed from 5 to 10%, the apparent molecular weight would decrease by 9%.

In the special case of crystals grown from salt-free or low-salt solutions, i.e. $s = 0$, equation (13) reduces to:

$$M_p = \frac{NVD_c(1 - d)}{n(1 + u)}. \tag{14}$$

In this case knowledge of w is not required; neither, it may be pointed out, is the partial specific volume of the protein. Apart from the experimental errors in the determination of D_c and d, the only uncertainty in the molecular weight comes from $(1 + u)$, which can be put equal to 1·075 with a probable error of not more than a few per cent.

A method of taking advantage of the obvious simplification in molecular weight determination for salt-free crystals would be to use the crosslinking technique in a manner analogous to that described above for crystal density measurements. In this case, crystals would be stabilized by crosslinking, transferred to water to remove all salt, and then d determined by drying these "salt free" crystals. In contrast to the crosslinking–density technique described previously, knowledge of the partial specific volume of the protein would not be required but, on the other hand, a possible uncertainty in the calculated molecular weight would accrue from the $(1 + u)$ term in equation (14), and from the presence of the crosslinking agent.

Another approach to molecular weight determination by crystal drying, applicable to all protein crystals whether grown from concentrated or dilute salt solutions, makes use of the solvent density D_s. It is possible to derive by straightforward, although somewhat tedious algebra, the following expression for the protein molecular weight:

$$M_p = \frac{NV[D_c(1 - d) - sD_s(D_c - D_w)/(D_s - D_w)]}{n[1 + u - sD_s(1 - \bar{v}_p D_w)/(D_s - D_w)]}. \qquad (15)$$

In contrast to equation (13), knowledge of w is not required. Again using North's (1959) data for cubic hemoglobin as an example, assuming $u = 7{\cdot}5\%$ and using the solvent density $D_s = 1{\cdot}240$ g/cm^3 (measured for the crystal supernatant described by North), and a partial specific volume of 0·749 cm^3/g (Svedberg & Fahraeus, 1926), the half-molecular weight of hemoglobin is calculated to be 37,600, in excellent agreement with the value of 37,700 obtained from equation (13). By substituting the molecular weight determined from equation (15) back into equation (13), and solving for w, one obtains 29·4% as the weight fraction of bound water, in excellent agreement with Perutz's (1946) value. It is not obvious why the molecular weights calculated from both equations (13) and (15) are somewhat higher than the expected value of about 33,300, although it may be noted that if the fractional water retained in the crystal on drying were assumed to be $u = 15\%$ rather than 7·5%, the discrepancy would be eliminated.

Another expression that may be useful if only small crystals are available, and density measurements difficult, requires a knowledge of w, and of D_s, but not of the crystal density; it is

$$M_p = \frac{NVD_sD_w(1 - s - d)}{n[(d + wd - w)(1 + u)D_w + (D_w\bar{v}_p + u + w + wu)(1 - s - d)D_s]}. \qquad (16)$$

In summary, molecular weight determinations from measurements of the loss of weight of crystals on drying can be moderately accurate in favourable cases, but in general will be limited by the uncertainty in the amount of water that remains in the crystal after drying, and by the experimental difficulty in determining precisely the fractional loss of mass on drying the crystals.

4. Crystal Volume and Protein Content

Love (1957) proposed a very simple method of molecular weight determination based on measurement of the mass of protein in a crystal of known volume. The method requires fairly large crystals, which are sufficiently well-formed that the volumes of individual crystals can be measured under a suitable microscope (Love, 1957; Low & Richards, 1952*b*). The protein crystals are dissolved in a known volume of solvent, and the total protein content measured either from the optical density of the solution, or by the method of Lowry *et al.* (1951).

If m_v is the mass of protein per unit volume of the crystal, then the protein molecular weight is given by:

$$M_p = \frac{NVm}{n} \tag{17}$$

(see Love, 1957).

This method of molecular weight determination, which was independently proposed by Heidner *et al.* (1971), has the great advantage that it requires no knowledge of crystal or solvent density, of crystal water content, or of the partial specific volume of the protein. On the other hand, accurate estimation of the volumes of individual crystals may not be feasible.

5. Crystal Mass and Protein Content

A method related to that described in the previous section, and apparently not proposed before, makes use of the mass of protein per unit *mass* of crystal. A known mass of crystals is dissolved in a small volume of a suitable solvent, and the protein content determined from the optical density, or by Lowry's method (Lowry *et al.*, 1951). From m_m, the mass of protein per unit mass of wet crystals, the protein molecular weight may be obtained using the formula

$$M_p = \frac{NVm_mD_c}{n} . \tag{18}$$

Other than the fact that the proposed method requires a knowledge of the crystal density, it has all the advantages of the "crystal volume and protein content" method. In addition, the volume determination of individual crystals is obviated, so that an aggregate of small or irregular crystals could be used.

Furthermore, the crystal density need not be known as accurately as for those methods of molecular weight determination based on crystal density, for in equation (18), the crystal density is used directly, whereas in equations (7) and (8) a *difference* between two densities is used. Also the crystal mass enters equation (18) directly, which is preferable to the difference in mass, required when the fractional loss of weight on drying is used (eqns 14, 15 and 16).

The main disadvantage of the proposed method would be in the determination of crystal mass, but with the microbalance used by Berthou *et al.* (1972) the masses of fairly large individual crystals can apparently be measured with ease to an accuracy of better than 1%. Also Berthou *et al.* (1972) found that the loss of weight on drying masses of microcrystals agreed with that of single crystals within about 3%, suggesting that if care were taken to remove as much of the mother liquor as possible, molecular weight determinations based on the mass of microcrystals need not be susceptible to large errors.

I thank Dr P. M. Colman, W. R. Kester and P. S. F. Lo for making some of the density measurements. Also I am grateful to Drs H. W. Wyckoff and R. Henderson for a discussion concerning the use of isotopes other than deuterium, and Dr P. B. Sigler for stimulating discussions, and for sending a preprint of his paper.

I also thank Dr D. C. Phillips for a number of helpful editorial comments, which led to a substantial expansion in the scope of this report.

This study was supported in part by grants from the National Science Foundation (GB30823X) and the National Institutes of Health (GM20066, GM15423 and FR06027), and by the award of an Alfred P. Sloan Research Fellowship and a Public Service Career Development Award (GM70585) from the Institute of General Medical Sciences.

REFERENCES

Adair, G. S. & Adair, M. E. (1936). *Proc. Roy. Soc. Ser. B*, **120**, 422–446.

Banaszak, L. J., Tsernoglou, D. & Sade, M. (1971). In *Probes of Structure and Function of Macromolecules and Membranes* (Chance, B., Yonetani, T. & Mildvan, A. S., eds), vol. 2, pp. 71–80, Academic Press, New York.

Berthou, J., Cesbron, F. & Laurent, A. (1972). *J. Mol. Biol.* **71**, 809–813.

Caspar, D. L. D. (1966). CIBA Symposium on *Principles of Biomolecular Organization* (Wolstenholme, G. E. W. & O'Connor, M., eds), pp. 7–34, Little & Brown, Boston.

Cohn, E. J. & Edsall, J. T. (1965). *Proteins, Amino Acids and Peptides*, Hafner Publishing Co., New York. Originally published 1943, Reinhold Publishing Corp.

Colman, P. M. & Matthews, B. W. (1971). *J. Mol. Biol.* **60**, 163–168.

Colman, P. M., Jansonius, J. N. & Matthews, B. W. (1972). *J. Mol. Biol.* **70**, 701–724.

Cornick, G., Sigler, P. B. & Ginsberg, H. S. (1973). *J. Mol. Biol.* **73**, 533–538.

Fisher, H. (1965). *Biochim. Biophys. Acta*, **109**, 544–550.

Haurowitz, F. (1950). *Chemistry and Biology of Proteins*, p. 91, Academic Press, New York.

Heidner, E. G., Weber, B. H. & Eisenberg, D. (1971). *Science*, **171**, 677–680.

Korn, A. H., Feairheller, S. H. & Filachione, E. M. (1972). *J. Mol. Biol.* **65**, 525–529.

Kuntz, I. D., Jr, Brassfield, T. S., Law, G. D. & Purcell, G. V. (1969). *Science*, **163**, 1329–1331.

Love, W. E. (1957). *Biochim. Biophys. Acta*, **23**, 465–471.

Low, B. W. & Richards, F. M. (1952*a*). *J. Amer. Chem. Soc.* **74**, 1660–1666.

Low, B. W. & Richards, F. M. (1952*b*). *Nature (London)*, **170**, 412–413.

Lowry, O. H., Rosebrough, N. J., Farr, A. L. & Randall, R. J. (1951). *J. Biol. Chem.* **193**, 265–275.

Matthews, B. W. (1968*a*). *J. Mol. Biol.* **33**, 491–497.

Matthews, B. W. (1968*b*). *J. Mol. Biol.* **33**, 499–501.

Matthews, B. W. (1974). In *The Proteins*, 3rd edn (Neurath, H. & Hill, R. L., eds), Academic Press, New York. In the press.

Matthews, B. W. & Bernhard, S. A. (1973). *Annu. Rev. Biophys. Bioeng.* **2**, 257–317.

Matthews, B. W., Jansonius, J. N., Colman, P. M., Schoenborn, B. P. & Dupourque, D. (1972). *Nature New Biol.* **238**, 37–41.

McMeekin, T. L. & Marshall, K. (1952). *Science*, **116**, 142–143.

North, A. C. T. (1959). *Acta Crystallogr.* **12**, 512–514.

Ohta, Y., Ogura, Y. & Wada, A. (1966). *J. Biol. Chem.* **241**, 5919–5925.

Perutz, M. F. (1946). *Trans. Faraday Soc.* **42B**, 187–195.

Quiocho, F. A. & Richards, F. M. (1964). *Proc. Nat. Acad. Sci., U.S.A.* **52**, 833–839.

Richards, F. M. & Knowles, J. R. (1968). *J. Mol. Biol.* **37**, 231–233.

Schoenborn, B. P. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 569–575.

Svedberg, T. & Fahraeus, R. (1926). *J. Amer. Chem. Soc.* **48**, 430–438.

TenEyck, L. F. (1970). Ph.D. Dissertation, Princeton University.

Watson, H. C. & Banaszak, L. J. (1964). *Nature (London)*, **204**, 918–920.

Wyckoff, H. W., Doscher, M., Tsernoglou, D., Inagami, T., Johnson, L. N., Hardman, K. D., Allewell, N. M., Kelly, D. M. & Richards, F. M. (1967). *J. Mol. Biol.* **27**, 563–578.

LINKOAHO, M. V. (1971). *Philos. Mag.* **23**, 191–198.
STEVENS, E. D. (1974). *Acta Cryst.* A**30**, 184–189.
STEWART, A. T. & BROCKHOUSE, B. N. (1958). *Rev. Mod. Phys.* **30**, 250–255.
WAKOH, S. & KUBO, Y. (1980). *J. Phys. F*, **10**, 2707–2715.
WAKOH, S. & YAMASHITA, J. (1971). *J. Phys. Soc. Jpn*, **30**, 422–427.
WEISS, R. J. & DEMARCO, J. J. (1965). *Phys. Rev. A*, **140**, 1223–1225.
ZACHARIASEN, W. H. (1967). *Acta Cryst.* **23**, 558–564.

Acta Cryst. (1981). A**37**, 701–710

An Oscillation Data Collection System for High-Resolution Protein Crystallography

BY M. F. SCHMID, L. H. WEAVER, M. A. HOLMES, M. G. GRÜTTER, D. H. OHLENDORF, R. A. REYNOLDS, S. J. REMINGTON* AND B. W. MATTHEWS†

Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oregon, USA

(*Received* 13 *November* 1980; *accepted* 25 *March* 1981)

Abstract

An oscillation data collection system for protein crystallography is described. The system consists of a modified Enraf–Nonius precession camera with cylindrical cassette, and stepping motor driven by a flexible microprocessor control system. The X-ray source is graphite-monochromatized radiation from an Elliott GX-21 rotating-anode generator run at 5·5 kW on a focal spot measuring 0·5 × 5·0 mm. The potential advantages of using a relatively large focal spot in conjunction with a graphite monochromator are discussed. Conditions for optimum collimation and X-ray intensity are considered, and it is shown that appropriately designed collimators with adjustable apertures can have substantial advantages over commercially available pinhole collimators. The oscillation films are processed by a procedure based on that of Rossmann [*J. Appl. Cryst.* (1979), **12**, 225–238]. Determination of the initial alignment of the film is facilitated by a pair of reference pins incorporated in the cylindrical cassette. These pins ensure that the position of the film in the cassette is known, and avoid the need for fiducial marks. The crystal alignment and film measurement technique is fully automatic, requiring no prior input other than the approximate starting orientation of the crystal, the approximate unit-cell dimensions, and the angular oscillation range. An alternative method for the determination of crystal orientation is proposed which has been found to be somewhat superior to that of Rossmann, especially for smaller unit cells.

1. Introduction

The use of screenless X-ray photography for the rapid, efficient measurement of diffraction data from protein crystals is well established, and an increasing number of laboratories have developed, and are using, such techniques (*e.g.* Xuong & Freer, 1971; Arndt, Champness, Phizackerley & Wonacott, 1973; Schwager, Bartels & Jones, 1975; Rossmann, 1979; Wilson & Yeates, 1979; and the collection of papers in Arndt & Wonacott, 1977).

In this paper we wish to describe an oscillation data collection system we have developed which includes a number of features we have found to be of value, and may be of interest to other laboratories.

These features include graphite monochromatization to reduce background and enhance crystal life, the use of a broad-focal-spot X-ray generator to reduce exposure times by a factor of four, improved collimation geometry, and cylindrical cassettes for high-resolution data collection. The film scanning program, adapted from that of Rossmann (1979), permits the accurate determination of crystal alignment from the oscillation photograph itself, with no prior input other than the approximate orientation of the crystal. We have adopted an alternative strategy for refinement of the crystal orientation which we have found to work better than the 'convolution' procedure developed by Rossmann for virus crystal photography.

The basic hardware consists of two Enraf–Nonius precession cameras, modified for oscillation photography, each controlled by a microprocessor. The cameras are mounted on an Elliott GX-21 rotating-anode generator with Charles Supper graphite monochromators. Films are scanned by an Optronics P3000 rotating-drum scanner interfaced to a Varian V76

* Present address: Max-Planck-Institut für Biochemie, D-8033 Martinsried bei München, Federal Republic of Germany.
† To whom correspondence should be addressed.

0567-7394/81/050701-10$01.00 © 1981 International Union of Crystallography

computer. The V76 computer is used simply to write optical densities on magnetic tape. Typically, nine 7 × 5 in (171·5 × 122·5 mm) films, scanned with a 100 μm raster, can be accommodated on one 2400 ft, 800 bpi tape. Data processing and data reduction are performed on a VAX 11/780 computer.

2. The oscillation camera

The oscillation–rotation cameras consist of modified Enraf–Nonius precession cameras with a stepping motor driven by a microprocessor controller, patterned after those used in Dr Robert Huber's laboratory (Huber, personal communication; Arndt & Wonacott, 1977, p. 30).

The possibility of precession and oscillation geometry on the same camera permits the use of precession as well as still photographs during the initial alignment of the crystal. Also, with some space groups we have found it useful to take a precession photograph of one zone (*e.g.* the centrosymmetric zone) of the same crystal which has been used for oscillation photography. On the other hand, the use of a modified precession camera does not have the convenience of a carousel with a number of pre-loaded cassettes as used in the commercially available oscillation camera systems. Cylindrical cassettes with a radius of 80 mm permit data collection to Bragg spacings of 1·6 Å on 175 × 125 mm sheets of film. A cylindrical cassette has obvious advantages over a flat cassette for high-resolution data collection, and has simpler geometry than that of *V*-shaped cassettes (Arndt & Wonacott, 1977, p. 27). The spot shape is more constant and the obliquity corrections are simpler for a cylindrical cassette than for a *V*-shaped one.

In commercially available cassettes, fiducial marks are made on the film either by allowing light to enter narrow holes, which often produces an ill-defined reference mark, or by using an ancillary X-ray source to direct X-rays through the fiducial holes onto the film. The latter method is not only inconvenient, but does not always work successfully. It is a common and very frustrating experience to find that the fiducial marks on the film are sufficiently out of register with the actual diffraction pattern that one cannot locate reflections at their expected positions on the film (*e.g.* Arndt & Wonacott, 1967, pp. 25–26; Wilson & Yeates, 1979; Rossmann, 1979). A method of avoiding these difficulties, which we have found to work well, is to incorporate a pair of metal pins in the film cassette itself in exactly the same relative positions that they occupy on the film scanner. The scanner mounting holes are punched in the films *before* they are loaded in the cassette. Then, when the film is subsequently mounted on the film scanner, it is automatically aligned in precisely the same way that it was in the cassette during X-ray photography. Since the film scanner starts at an arbitrary position when it scans across the film, it is necessary to have a reference mark on the film to define the origin in the horizontal direction. The spot made by the direct beam is very convenient for this purpose. In principle, one can dispense with the direct-beam mark if the background on the film is dark enough for the film scanner to detect the shadow on the left and right sides of the film made by the edge of the cassette. Either of these steps in optical density can also serve to define the horizontal origin.

There is another advantage in knowing the precise orientation of the film in the cassette. To a close approximation, a rotation of the film in the cassette is equivalent to a rotation of the crystal about the X-ray beam. Therefore a rotation of the film could be incorrectly attributed to an apparent rotation of the crystal, which in turn will lead to errors in the Lorentz–polarization correction and in other geometrical factors.

The stepping motor for the oscillation or rotation motion is driven by a microprocessor controller which we have found to be both flexible and very reliable. Advantages of current microprocessor technology have been described by Reeke (1977). The controller has a repertoire of commands whereby the user can drive the crystal to a specified angle, and specify the oscillation range, speed of rotation, number of cycles, and so on. Should the X-ray generator turn off during an exposure, the microprocessor enters a 'wait' mode, giving the user the option of resuming the exposure when the generator is returned to power.

3. Collimation

Conditions for optimum collimation have been discussed at length by Huxley (1953) and by Arndt & Sweet (1977). In their discussion it is assumed that the protein crystal is smaller than the exit pinhole of the collimator, but in practice many laboratories have found it advantageous to use crystals much larger than the collimator diameter, both to maximize the intensity of reflections on the film and also to permit multiple exposures of different parts of a single crystal. As shown below, some commercially available collimators are not well suited to this geometry, and there are advantages in using appropriately designed adjustable collimators.

A schematic diagram of the collimation system with a graphite monochromator is shown in Fig. 1. The horizontal and vertical dimensions of the X-ray focal spot are denoted f_x and f_y. With a monochromator present, it is, in general, necessary to place an aperture as close to the exit of the monochromator as possible in order to reduce the divergence of the monochromated beam (see the following section). In practice, this

aperture can be a simple pinhole made by piercing a lead sheet with a needle, and mounted on an X–Y translation stage to facilitate alignment. For generality, we denote the horizontal and vertical dimensions of the aperture as f'_x and f'_y. The source aperture of the collimator is usually large (about 1 mm) and plays no part in the collimation.

As discussed by Arndt & Sweet (1977), when the protein crystal is smaller than the specimen aperture ($a_x < a'_x$), simple geometry shows that the width of the spot on the film is

$$w_x = a_x + (a_x + f'_x)\frac{d}{s}. \quad (1)$$

If, however, the crystal is larger than the specimen aperture ($a_x > a'_x$), then the spot width will be

$$w_x = a'_x + (a'_x + f'_x)\frac{d'}{s'}. \quad (2)$$

For Enraf–Nonius collimators, for example, the specimen aperture is about 30 mm from the crystal, which results in a considerable increase in spot size. Substituting some typical numbers (s = 120 mm, d = 80 mm, $f'_x = 0{\cdot}45$ mm) a large crystal with collimator aperture $a'_x = 0{\cdot}4$ mm will produce a spot of width 1·44 mm, whereas a crystal measuring 0·4 mm a side will produce a spot on the film measuring 0·97 mm. The larger spot size is undesirable, not only because the diffracted X-rays are spread over a larger area, but also because spot overlap limits resolution.

The obvious solution is to have the specimen aperture as close to the crystal as possible. In addition, it is desirable to be able to adjust the vertical and horizontal dimensions of the specimen aperture independently. If, for example, one had a crystal with a long reciprocal axis in one direction, but not the other, then one might wish to use a narrow aperture in this direction in order to resolve the spots, but retain a wider aperture in the other direction in order to increase the total scattered intensity, and the resultant accuracy of intensity measurement.

Both these objectives can be met by using collimators similar to those proposed by Love, Hendrickson, Herriott, Lattmann & McCorkle (1965). As shown in Fig. 2, the adjustable apertures are obtained with polished flat-faced set screws. The length of the collimator is such that it ends as close to the crystal as possible (about 1 mm). Also the adjustable apertures are set as far forward in the collimator as possible. We have found that the guard apertures are not strictly necessary, presumably because the defining apertures are quite flat, and act as their own guards. This means that one only needs to use one vertical aperture and one horizontal aperture. Obviously, the two sets of adjustable screws closest to the protein crystal are the ones used, bringing the defining aperture even closer to the crystal. In practice, the adjustable collimators are easier to make than the standard ones, since the difficult step of drilling the pinhole is avoided. The alignment of the adjustable apertures is best accomplished by mounting the collimator with a light behind, and viewing it from the front with a low-power microscope. The collimator mount and microscope of a Weissenberg camera are ideal for the purpose. Centering of the aperture is checked by rotation. Since the collimators are easy to make it is convenient to have a selection available with commonly used aperture sizes.

It remains to select the actual aperture size, effective focal-spot size, and the distance s' in Fig. 1. Much of the treatment of pinhole collimation given by Arndt & Sweet (1977) is germane, and can be consulted. In practice, the optimum collimator aperture size is 0·3–0·5 mm, smaller values being used only when mitigated by cell edges longer than 150 Å or so. It has to be remembered that if the monochromator crystal is vertical, then the vertical divergence of the X-ray beam is from the focal spot and not from the monochromator. (There is a small additional divergence α,

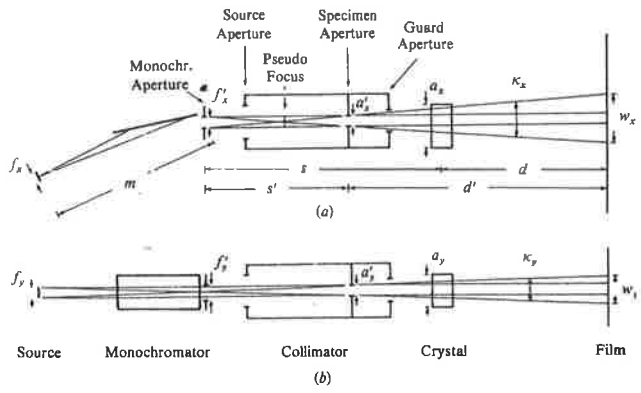


Fig. 1. Schematic illustration of collimation geometry when a monochromator is present and when the dimensions of the crystal exceed those of the specimen aperture of the collimator. The monochromator is mounted in the vertical plane. (*a*) Plan. (*b*) Elevation.

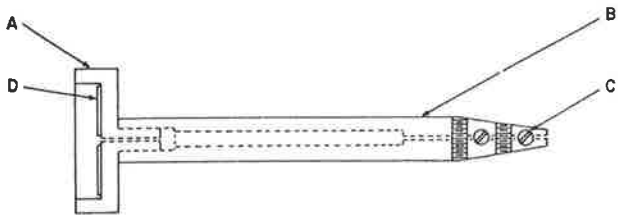


Fig. 2. Adjustable X-ray collimator patterned after that described by Love *et al.* (1965). The components are as follows. *A* Brass cup. *B* Brass tube with optional taper. The taper is usually not necessary for oscillation photography. *C* Flat-faced, polished, (4-80) brass screws, to provide adjustable collimation. Usually the front pairs of screws can be used as both defining and guard apertures (see text). *D* lead shield. A full set of drawings for the collimator is available on request from the authors.

discussed in the following section, owing to the mosaicity of the graphite crystal.) Therefore, the vertical crossfire (Fig. 1*b*) is

$$\kappa_y = \frac{(f_y + a'_y)}{s' + m} + \alpha, \tag{3}$$

whereas the horizontal crossfire (Fig. 1*a*) is

$$\kappa_x = \frac{f'_x + a'_x}{s'}. \tag{4}$$

For a 0·5 mm focal spot, 0·4 mm collimator aperture, $(s' + m)$ equal to its minimum practical value of about 220 mm, and $\alpha = 0{\cdot}06°$ (see next section), the vertical crossfire has the reasonably low value of 0·29°, as is desirable to minimize the spindle rotation required to sweep through each Bragg reflection of the protein crystal. Decreasing the focal-spot size (*e.g.* to 0·2 mm) will decrease the vertical crossfire but not the horizontal. As discussed in the next section, use of a small focal spot may unnecessarily reduce useful beam intensity.

Up to a point, the diameter of the pinhole (f'_x and f'_y) can be increased in order to maximize beam intensity, commensurate with resolving reflections on the film. Increasing f'_x will increase intensity, but will also increase the horizontal dimension of the spots on the film (equation 2). In the vertical direction, f'_y does not define the useful X-ray beam (Fig. 1*b*). Ideally, f'_y should be just large enough not to restrict the limiting cone of X-rays between the focal spot and the specimen aperture (Fig. 1*b*). Decreasing f'_y below this value will reduce beam intensity but will not reduce the crossfire, which is determined by f_y and a'_y (equation 3). In principle, f'_y could be eliminated (*i.e.* f'_x could be a vertical slit) but the aperture f'_y does prevent unnecessary radiation exiting from the monochromator. In practice we have found it convenient to define f'_x and f'_y by a pinhole about 0·45 mm in diameter.

Where possible, the crystal-to-monochromator distance s (or s') should be kept to a minimum, in order to maximize intensity. Only when required by the need to increase resolution (equation 2) should the camera be moved away from the monochromator exit pinhole.

4. Monochromatic radiation

The potential advantages of strictly monochromatic radiation for screenless X-ray photography are well known. The unwanted white-radiation background, as well as radiation streaks, are completely eliminated. Also, use of monochromatic radiation is expected to enhance useful crystal life, since the crystal is only exposed to radiation which potentially contributes to the measured diffraction pattern. Our experience with a number of protein crystals confirms this expectation. We have found that crystals can often be exposed ten times longer in a monochromatic beam than in a comparable nickel-filtered beam, before seeing equivalent radiation damage. [Also, for reasons not well understood, but reported informally by others, we find that radiation damage seems to be less for short, intense exposures than equivalent long, weak (or interrupted) exposures.]

The reputed disadvantages of monochromated radiation for oscillation photography are a reduction in intensity, relative to a nickel-filtered source, and limitations on resolution owing to 'fuzziness' of the diffracted spots (*e.g.* Arndt & Sweet, 1977). We have found that by use of an appropriate X-ray source, and by optimizing the collimation geometry, these disadvantages can be eliminated. In this section we discuss the use of the monochromator, *per se*, and how the need to collimate the monochromatic beam affects the choice of focal-spot size.

As a preliminary comparison, we recorded a series of oscillation photographs with a conventional rotating-anode generator (2·0 × 0·2 mm focal spot, 1·6 kW) with three different beam geometries: (1) nickel-filtered radiation with pinhole collimation; (2) a double mirror system (Harrison, 1968); (3) graphite-monochromatized radiation as described below. The resultant films were not measured, so that the following two conclusions are subjective. (1) There is no advantage in using a mirror–mirror system unless one is forced to do so by a cell dimension exceeding about 150 Å. This is in accord with the findings of Harrison, reported by Arndt & Sweet (1977). (2) A graphite monochromator with appropriate collimation produces by far the 'cleanest' pictures, but requires two to three times the exposure relative to the nickel-filtered beam. Resolution of cell edges up to 150 Å was not a problem. As discussed below, use of a larger focal-spot size can dramatically reduce the exposure time.

The reflection of X-rays from a vertical monochromator crystal is illustrated in Fig. 3(*a*). It is expected that a monochromator crystal will produce a pseudo-image of the X-ray source equidistant from the monochromator (*e.g.* see Fig. 5.13 of Arndt & Sweet, 1977) but, as Fig. 3(*a*) shows, such an image is not perfect. In addition, X-rays incident on the monochromator may penetrate 0·3 mm or so before being reflected, which will also blurr the focus. Photographs taken at increasing distances from the monochromator exit port do not reveal a sharp image of the focal spot. In addition, the reflected X-rays diverge continuously in the vertical plane as one moves away from the monochromator. This suggests that maximum intensity will be obtained by placing the monochromator crystal as close to the X-ray source as possible (as is common practice), and (other considerations aside) placing the camera as close as possible to the exit port of the monochromator housing.

When used in the manner described, the monochromator will produce an extended source of X-rays. If the monochromator crystal is mounted vertically, the width of this effective source will be about 0·6 mm or so, plus additional parasitic scattering, and the horizontal divergence of the reflected beam will originate from the monochromator rather than the focal spot (Fig. 3*a*).

In the vertical direction, the beam diverges from the focal spot (Fig. 1*b*). In addition, as pointed out to us by Dr H. Wyckoff, there is an additional divergence, α, owing to the mosaicity of the graphite crystal (Fig. 3*b*). Straightforward geometry shows that

$$\alpha = \frac{2\omega d_1 \sin\theta}{d_1 + d_2}, \qquad (5)$$

where d_1 is the focal-spot-to-monochromator distance, d_2 the monochromator-to-specimen-aperture distance, ω the mosaic spread and θ the Bragg angle for the graphite monochromator. In general, α will be about 0·06° or less.

Because the monochromator crystal provides such a broad source, it will give 'fuzzy' spots on the film unless either the distance from monochromator to protein crystal is increased, or additional collimation is introduced, or a combination of these is used, as discussed in the previous section. We now ask what can be done to maximize the reflected intensity from the monochromator itself. One obvious step, already mentioned, is to place the monochromator crystal as close as possible to the X-ray focal spot. The second important factor is focal-spot size. The objective is to have each point on the used area of the monochromator crystal reflect as many X-ray quanta as possible in directions which will reach the protein crystal. In the horizontal direction the two factors which need to be considered are the pseudo-focusing of the monochromatic beam (Fig. 3*a*), and the limitations imposed by the mosaic spread of the monochromator crystal. As shown in Fig. 1(*a*), the pseudo-focus lies between the monochromator aperture and the specimen aperture, and the size of the pseudo-focus can be usefully increased until it fills the limiting cone between the two apertures. Since the two apertures are usually of comparable size, the maximum useful width of the pseudo-focus, which in turn is approximately equal to the size of the focal spot itself, is equal to f'_x (or a'_x), *i.e.* about 0·45 mm in practice. The effect of the mosaic spread of the graphite monochromator is to limit the width of the focal spot which can be 'seen' by a given point on the monochromator. If a monochromator mosaic spread of 0·4° is assumed, the limiting focal-spot width for a rotating-anode generator with a typical focal-spot-to-monochromator distance of 80 mm and a vanishingly small monochromator aperture is 0·56 mm. As the monochromator aperture f'_x is increased in size, the focal spot could, in principle, be increased by the same amount, although there is clearly no gain in intensity when the pseudo-focus exceeds the limits described above (*i.e.* about 0·45 mm). In other words, it is usually the collimation of the X-ray beam rather than the mosaic spread of the monochromator which determines the maximum useful size of the focal spot. Where the X-ray generator design allows, it is clearly advantageous to increase the focal-spot size to equal this value, since a larger focal spot can provide a greater total flux than a small spot.

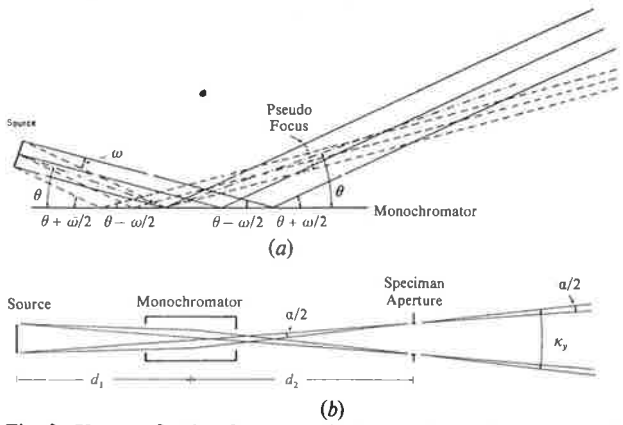


Fig. 3. X-ray reflection from a vertical monochromator crystal. (*a*) Plan. The angle ω is the mosaic spread of the monochromator. (*b*) Elevation. As explained in the text, α is the additional divergence caused by the mosaic spread of the monochromator. For simplicity, only the rays which define the maximum divergence are drawn.

In the vertical direction (Figs. 1*b*, 3*b*), the beam diverges primarily from the focal spot, and the monochromator can be effectively ignored. Here, useful intensity is proportional to the height f_y of the focal spot, and the upper limit on f_y is determined by the desired vertical crossfire (equation 3). As discussed previously, a focal spot of 0·5 mm will result in a vertical crossfire of about 0·29° or less, which is quite acceptable.

These arguments suggest that, with typical collimation geometry, the optimum focal-spot size for use with monochromatized radiation is about 0·5 × 0·5 mm, as seen from the X-ray port. This contrasts with normal pinhole collimation, where diffracted intensity, defined as maximum film darkening, is claimed to depend strictly on source brightness rather than total flux (Arndt & Sweet, 1977). In practice we have found that an Elliott GX-21 generator running at 6 kW on a 0·5 × 5 mm focal spot reduces exposure times by a factor of four relative to a generator running at 1·6 kW on a 0·2 × 2 mm focal spot. This factor of four more than offsets the factor of two to three increase in exposure which is normally considered to be necessary to obtain comparable intensities with monochromatized radiation relative to β-filtered radiation (Arndt & Sweet, 1977). The above argument suggests that a monochromatized beam from a high-brilliance

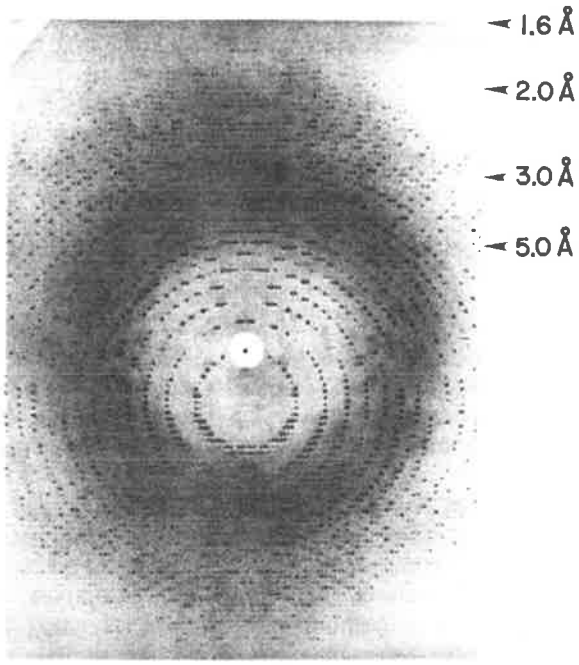
sealed X-ray tube run at 1·5 kW with a focal spot of 0·4 × 8 mm ought to have an intensity comparable with a rotating-anode source run at 1·6 kW, although the sealed tube will lose intensity because of its shallow take-off angle.

An oscillation photograph for thermolysin, showing the resolution of spots along a horizontal c axis of 131·4 Å and the rectangular spot shape produced by an adustable collimator, is shown in Fig. 4. The crystal used for this photograph was a hexagonal prism 1·4 mm in the horizontal direction, and 0·45 mm in diameter.

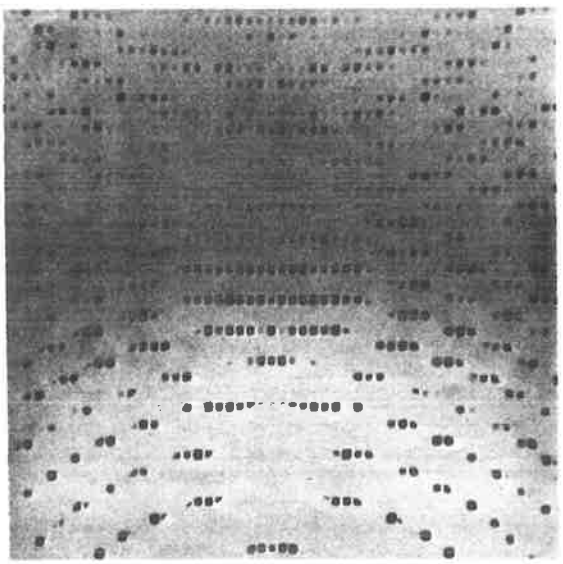
5. Intensity measurement

The system we have developed for intensity measurement is based on that of Rossmann (1979). One of the advantages of this approach is that the alignment of the crystal is determined automatically from the same film that is used for intensity measurement. The method was developed initially for virus photography, where only a single exposure could be taken of each crystal (Schutt & Winkler, 1977; Rossmann, 1979), and contrasts with the alternative approach in which additional still or other alignment photographs are used to determine the exact orientation of the crystal (Xuong & Freer, 1971; Arndt *et al.*, 1973; Nyborg & Wonacott, 1977; Wilson & Yeates, 1979). Clearly, the reliance on additional photographs for determination of crystal alignment has several disadvantages (Schwager, Bartels & Jones, 1975). One way of determining crystal alignment from an oscillation photograph is to identify manually a number of the partially recorded reflections and to base the alignment on these reflections (Schwager *et al.*, 1975; Jones, Bartels & Schwager, 1977). This method reduces the amount of computing, but requires manual intervention for each film. The Rossmann (1979) procedure is entirely automatic, although it does require additional computing. As described below, we have found it desirable to modify Rossmann's procedure somewhat for use with protein (as opposed to virus) crystals. In the following discussion we discuss only those aspects of our procedure which differ from that described by Rossmann.

The first step is to find the transformation between the scanner coordinate frame and the camera coordinate system, *i.e.* the Q matrix of Rossmann (1979). As described previously, this is greatly simplified by the use of alignment pins incorporated within the cassette. The accurate determination of the Q matrix uses whole reflections, and differs slightly from Rossmann in that the Q matrix is constrained to be orthogonal. In other words we preserve the 90° angle and the relative lengths of the vertical and horizontal axes of the scanner coordinate frame, rather than trying to use the Q matrix to allow for errors in the crystal-to-film distance and anisotropic shrinking of the film. We refine separately the crystal-to-film distance and a parameter correcting for any vertical misalignment



(*a*)



(*b*)

Fig. 4. (*a*) Oscillation photograph for thermolysin. The arrows indicate the resolution in ångströms in the central region of the film. Unit-cell dimensions $a = b = 94·2$, $c = 131·4$ Å. Crystal 0·45 × 0·45 × 1·4 mm. Collimator aperture 0·30 mm (horizontal), 0·40 mm (vertical). Crystal-to-film distance 80 mm. Focal spot approximately 0·5 × 5·0 mm; 39 kV, 140 mA, exposure time 6 h. Oscillation range 1·0°. Because of the cylindrical cassette, the reflections lie on parallel vertical lines (layer lines). (*b*) Enlargement showing the central region of the film from about 25 to 2·3 Å resolution.

between the cassette cylinder axis and the spindle axis (this being the most difficult alignment adjustment to make on the cassettes).

Our most significant difference from the procedure of Rossmann is in the refinement of the orientation matrix A relating the crystal alignment to the camera coordinate frame. The crystal orientation is defined by rotations about three axes, the vertical X axis, the horizontal Y (spindle) axis, and the horizontal Z (X-ray beam) axis. Rotations about the Z axis cause a bodily rotation of the X-ray pattern on the film, and are easily determined from the positions of the whole spots on the film. Rotations about Y, the spindle axis, cause additional reflections to appear at one end of the crystal oscillation range, and other reflections to be 'lost' at the other extreme of the oscillation range. This effect is most noticeable along the equator (vertical axis) of the film. Rotations about the vertical axis X cause some reflections, most noticeably near the meridian (horizontal axis), to begin to be lost, but others will begin to appear.

Rossmann's approach is to consider small rotational increments of the crystal about the X and Y axes, assuming that the correct oscillation range and mosaic spread of the crystal are known. For each increment, any new reflection which would come into the reflecting position is noted, and the measured intensity of all such reflections is summed. Similarly, any reflections 'lost' are also noted. If, for a given increment, more intensity is gained than lost, then one continues to increment in that direction until the optimum orientation is reached, at which point the intensity gained minus intensity lost is a maximum (*i.e.* any rotation causes no significant intensity gain, but does cause a loss of intensity). Rossmann describes the method as a 'convolution' technique, meaning that one fits an expected 'window' of reflections to the intensity distribution which is actually observed on the film. Rossmann's procedure depends on the change in the intensity distribution at the leading edge and the trailing edge of the window (*i.e.* at the beginning and the ending of each lune of reflections on the film). As described below, we have found it desirable to base the orientation refinement on, essentially, the center of gravity of the intensity distribution within each lune. Our approach involves the following steps (see also the caption to Fig. 5).:

1. All reflections which occur in 0·1° oscillation ranges about the spindle axis are computed. Since, for the purposes of this computation, we choose 0° mosaicity, a reflection will not appear in more than one range.

2. The reflections in each range are measured, and the average intensity for that range is computed (Fig. 5*a*).

3. The center of gravity of intensity for the overall intensity distribution is taken as the new spindle setting angle φ_Y, and $\delta\varphi_Y$ is defined as the difference between this value and the previously assumed value of φ_Y.

For the X-axis refinement, the orientation angle φ_X is incremented in steps of 0·1°, and, for each value of φ_X, φ_Y is swept through its oscillation range and all reflections occurring within this range are summed and averaged (Fig. 5*b*). We have found it desirable to exclude reflections in the 90° sectors at the top and

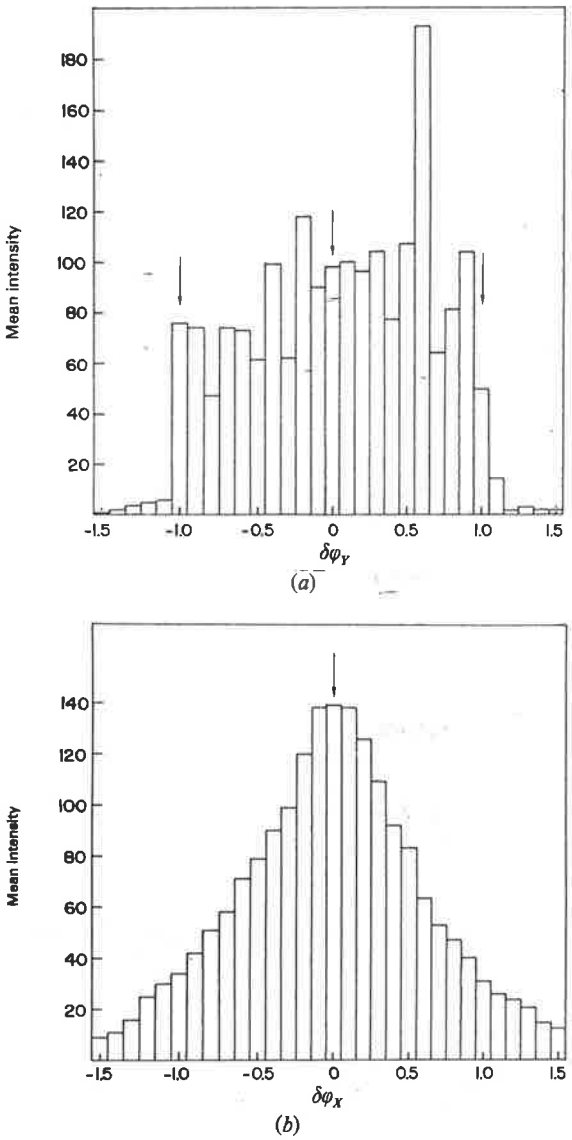


Fig. 5. Crystal orientation profiles for a typical T4 phage lysozyme oscillation film. Cell dimensions $a = b = 61{\cdot}2$, $c = 96{\cdot}8$ Å; oscillation range 2°; resolution for refinement 3·0 Å. (*a*) In the Y-axis refinement the crystal orientation is incremented in steps of 0·1° and the average reflection intensity for each increment is plotted. The theoretical limits of the oscillation range and the center of gravity of the intensity distribution are indicated. (*b*) For X-axis refinement, φ_X is incremented in steps of 0·1°, and, for each step, φ_Y is swept through the oscillation range of 2°. The plotted intensity is the average for all reflections within the 2° range, and the optimum value of φ_X is shown.

bottom of the film in order to eliminate reflections near the equator which would appear in several X-angle steps. (Conversely, in the Y-axis refinement we exclude reflections in the 90° sectors to the left and right of the film.) As with the Y-axis refinement, the optimum value of φ_X is taken to be the center of gravity of the intensity distribution (Fig. 5*b*). Cycles of refinement about X and Y are alternated, and refinement usually achieves a movement of less than 0·01° in $\delta\theta_X$ and $\delta\theta_Y$ per cycle (our criterion for completion) in less than five cycles.

There are several advantages to this method. Both Rossmann's procedure and our modification depend on all reflections having 'average' intensity. In practice, of course, this means a large enough sample of reflections must be used that the random variations in both the number of reflections and their intensities in different regions of reciprocal space are smoothed out. Our modification makes use of virtually all reflections (at least for Y refinement) in computing the weighted average of the intensity *vs* orientation, whereas Rossmann uses only those which are gained or lost on rotations as small as 0·01°. This may be satisfactory for a virus crystal, but, at least in our experience, gives less-consistent results for typical protein crystals, especially those with smaller unit cells.

Another major advantage is that the proposed method does not assume a knowledge of the mosaic spread in advance, whereas Rossmann's approach is rather sensitive to the choice of mosaicity. This can be seen from the representative profile shown in Fig. 5(*a*). If, for instance, too large a mosaicity is chosen, then the 'search window' will be wider than the actual profile, and all reflections gained or lost on rotation of the crystal will be outside the oscillation range and in the noise region. On the other hand, too small a mosaic spread will place the 'search window' completely inside the actual profile, so that again the optimum position will not be well determined.

A comparison of the refinement techniques for two different proteins is given in Table 1. The angles $\delta\theta_X$ and $\delta\theta_Y$ are the refined orientation angles relative to the approximate starting values. In the table, $\delta\theta_X$ and $\delta\theta_Y$ are calculated for a series of contiguous oscillations of the same crystal. If the crystal has not slipped during the series of exposures, the values of $\delta\theta_X$ and $\delta\theta_Y$ should be the same for each exposure, and the variations in $\delta\theta_X$ and $\delta\theta_Y$ give an estimate of the accuracy of the determination of the crystal orientation angles. For simplicity of comparison, the values of $\delta\theta_X$ and $\delta\theta_Y$ quoted in Table 1 are not the actual calculated values, but are the differences in $\delta\theta_X$ and $\delta\theta_Y$ from their mean value for the series of films. Two different situations are illustrated in Table 1. The first example is nine contiguous films of a platinum derivative of goose lysozyme (Grütter, Rine & Matthews, 1979), with space group $P2_1$ and cell dimensions $a = 38·3$, $b = 65·7$, $c = 45·2$ Å, $\beta = 116°$. The second example is four contiguous films of a bacteriochlorophyll protein (Matthews, Fenna, Bolognesi, Schmid & Olson, 1979), cell dimensions $a = b = 112·4$, $c = 98·4$ Å, space group $P6_3$. For the larger bacteriochlorophyll protein cell, the proposed refinement procedure gives results which are somewhat more consistent than Rossmann's procedure, although there is little to choose between the two techniques. However, in the case of the smaller goose lysozyme cell, which has far fewer reflections on each film pack, the proposed procedure is clearly superior. Here, the r.m.s. variation in both $\delta\theta_X$ and $\delta\theta_Y$ is 35% lower for the proposed method than for Rossmann's method. It appears that Rossmann's method works best when there is a large number of reflections on the film, and, therefore, a reasonable number of reflections close to the edge of each lune. It is these 'edge' reflections which are critically important in Rossmann's orientation procedure. In contrast, the proposed method uses all the reflections in each lune, and the accuracy does not seem to depend very much on the number of reflections on the film. In Table 2 we give the variation in the crystal orientation angles for several different proteins. The results suggest that the typical uncertainty in the calculated orientation angles φ_X and φ_Y is about 0·05°. As proposed by Schutt & Winkler (1977) and Rossmann (1979), it is also possible to obtain higher precision by 'post-refinement' techniques, although this is not so important when one can add together partial reflections recorded from the same crystal on successive film packs. Where a

Table 1. *Comparison of crystal orientation methods*

The values of $\delta\theta_X$ and $\delta\theta_Y$ give the refined values of the crystal orientation angles relative to assumed, approximate starting values of θ_X and θ_Y, determined from a series of contiguous oscillations of the same crystal. For ease of comparison, the quoted values of $\delta\theta_X$ and $\delta\theta_Y$ are the differences from the average, rather than the raw calculated values.

Crystal	Rossmann procedure		Proposed procedure	
Goose lysozyme	$\delta\theta_X$ (°)	$\delta\theta_Y$ (°)	$\delta\theta_X$ (°)	$\delta\theta_Y$ (°)
Film 1	−0·031	−0·171	0·033	0·008
2	0·054	0·186	0·022	0·027
3	−0·116	0·109	−0·072	0·069
4	−0·016	0·009	0·010	−0·091
5	0·014	0·069	0·037	0·095
6	−0·026	0·039	0·004	−0·024
7	0·049	0·049	0·021	0·097
8	0·024	−0·096	−0·026	−0·051
9	0·049	0·174	−0·024	−0·011
R.m.s. value	0·051	0·117	0·033	0·073
Bacterio-chlorophyll protein	$\delta\theta_X$ (°)	$\delta\theta_Y$ (°)	$\delta\theta_X$ (°)	$\delta\theta_Y$ (°)
Film 1	−0·008	0·037	−0·039	−0·041
2	0·042	0·051	0·021	0·041
3	−0·008	−0·109	−0·004	−0·061
4	−0·028	0·021	0·023	0·062
R.m.s. value	0·026	0·064	0·025	0·052

Table 2. *Accuracy of crystal orientation*

Crystal	Number of contiguous films	R.m.s. $\delta\theta_X$ (°)	R.m.s. $\delta\theta_Y$ (°)
Goose lysozyme	9	0·033	0·073
Bacteriochlorophyll protein: Run 1	4	0·025	0·052
Bacteriochlorophyll protein: Run 2	8	0·038	0·042
T4 phage lysozyme	5	0·023	0·032
Group-specific protease	20	0·054	0·082
Thermolysin: Run 1	7	0·094	0·036
Thermolysin: Run 2	9	0·022	0·023

contiguous set of films is available, as in Tables 1 and 2, it is possible, in principle, to increase the precision by using the average of the orientation angles determined for the set, rather than the individual values. We tested the potential advantage of such averaging by processing a contiguous set of phage lysozyme films in two ways. In one case the individually determined orientation parameters for each film were used, and in the second case a constant overall average orientation was used for all the films. When the respective data sets were merged together, the respective R values were essentially identical. In the latter case the merged data set contained about 1% more reflections because, with constant orientation parameters, each film exactly matches the adjacent films in the set, whereas, with individually determined orientation parameters, a few partially recorded reflections on adjacent films are 'lost'. On balance, there seems to be little to be gained by averaging the crystal orientation parameters and reprocessing all the films.

The proposed method for determining the crystal orientation will routinely converge to the correct solution if the assumed orientation of the crystal is correct within a degree or so. The range of convergence has not been tested fully, but we have, on occasion, seen correct convergence where the assumed orientation was in error by 3°. In cases where the alignment of the crystal is in doubt, it is desirable to carry out an initial refinement with the low-angle data, *e.g.* to 5 Å resolution, before including the high-resolution data.

As discussed by Rossmann (1979) and others (Diamond, 1969; Ford, 1974), it is desirable to use profile fitting to improve the accuracy of the weak reflections, to estimate standard deviations, and also to provide criteria for the rejection of doubtful measurements. Our procedure follows that described by Rossmann. Intensities of fully recorded reflections are estimated by comparison with a profile which varies across the film. Partially recorded reflections are measured by summing optical densities within designated peak and background areas. Partial reflections on adjacent films are summed, but are usually ignored if the reflection extends over more than two exposures. Variation in spot shape is taken into account by means of the profile shape, which can change smoothly across the film. Because of the cylindrical geometry, spot-shape variation is less than for a flat or V-shaped cassette. Also, the obliquity correction varies least for cylindrical geometry, in which case it depends only on the distance from the equator. The necessary obliquity corrections are given by Wonacott (1977), although two errors need to be corrected. In equation (7.32), $A(\beta)^2$ should be replaced by $A(\beta)$, and in equation (7.33), $A(\beta)^n$ should be replaced by $A(\beta)$.

Along with other investigators we have found that the measured intensities of partially recorded reflections tend to be systematically higher than the fully recorded reflections, especially for the weak data (Irwin, Nyborg, Reid & Blow, 1976; Wilson & Yeates, 1979). It is important, when measuring weak data, not to discard reflections for which the measured intensity is negative. Such negative values should be kept and averaged with measurements of the reflection from other films, otherwise the weak reflections will be systematically overestimated. Since partial reflections occur on more films than do whole ones, they will, on average, be overestimated by more than the whole ones. The inclusion of the 'negative' reflections reduces, but does not eliminate the systematic overestimation of the partial reflections. The overestimation is given by R_{bias} (*cf.* Irwin *et al.*, 1976), where

$$R_{\text{bias}} = 100\frac{\sum(\langle I_F\rangle - I_P)}{\sum\langle I_F\rangle}. \qquad (6)$$

$\langle I_F\rangle$ is the average intensity of the fully-recorded reflections and I_P the intensity determined by summing partially recorded reflections. In a test with the 1·8 Å resolution data set for T4 phage lysozyme, the value of R_{bias} when all negative intensity measurements were excluded was −7·6%. Including the negative measurements reduced the bias value to −3·3%.

Some representative data processing statistics are given in Table 3. The R values depend on several factors such as the resolution and the redundancy of the data, but they do give an indication of the overall quality of the data. In most cases the crystals used were about 0·4 to 0·5 mm in cross section, although the goose lysozyme crystals were thinner (0·4 × 0·4 × 0·15 mm). The lower R value for the goose lysozyme crystals may be due in part to the lower absorption in this case. In contrast, the phage lysozyme crystals were equilibrated with electron-dense phosphate solutions which increased the absorption. The relatively high R value for the bacteriochlorophyll protein is due in part to the large number of films and the high redundancy (approximately threefold) of the data. As an independent test of the quality of the oscillation data we compared the thermolysin data, measured to 1·6 Å,

Table 3. *Data processing statistics*

Protein	Resolution (Å)	Films	Crystals	Total reflections	Unique reflections	R^*_{sym}	R^*_{merge} (Including partials)	R^*_{merge} (Deleting partials)
Thermolysin	1·6	48	5	95 806	32 660	5·0	8·8	7·1
T4 phage lysozyme	1·9	28	4	32 410	13 466	6·7	11·6	10·2
Mutant T4 lysozyme, N14	2·1	19	2	19 126	10 502	–	9·5	7·9
Bacteriochlorophyll protein	1·9	76	9	126 269	46 764	4·6	11·6	10·0
Goose lysozyme	2·0	32	2	18 739	10 959	–	5·2	–
Cro repressor	2·4	15	2	18 940	11 987	–	7·6	6·8

* $R = 100 \sum |I - \bar{I}| / \sum I$. R_{sym} is the R value for symmetry-related intensities recorded on the same film, and the value quoted is the average over all films in the data set. R_{merge} gives the agreement between intensities measured on different films.

with a previous data set collected to 2·3 Å resolution by precession photography (Colman, Jasonius & Matthews, 1972). For the 11 824 reflections common to the two sets of data the R value on intensities was 6·6%. Similarly, the R value for 6500 reflections common to a 2·4 Å resolution precession data set and a 1·8 Å resolution oscillation data set for T4 phage lysozyme was 10·8%. For the bacteriochlorophyll protein, 15 000 common reflections to 2·8 Å resolution had an R value of 9·8%. Recently, the oscillation data collection system has been used to determine the structure of the cro repressor protein from bacteriophage λ to 2·8 Å resolution (Anderson, Ohlendorf, Takeda & Matthews, 1981).

We are particularly grateful to Dr Robert Huber for copies of his modifications to the Enraf–Nonius precession camera and cylindrical cassette, on which the system described here is based. Also, we are very grateful to Dr Michael Rossmann for his film measurement program, and numerous discussions on its use. We also thank Dr Philip Evans for his film-merging program, Dr Harold Wyckoff for a discussion on monochromatization and collimation, and the Machine and Electronics Shops at the University of Oregon for designing and constructing the oscillation system.

This work was supported in part by grants from the National Science Foundation (PCM-8014311) and the National Institutes of Health (GM 20066; GM 21967); by a PHS Postdoctoral Fellowship (GM 07221) to MFS and a Damon Runyon–Walter Winchell Cancer Fund Postdoctoral Fellowship (DRG-303-F) to DHO; and by a grant from the M.J. Murdock Charitable Trust.

References

ANDERSON, W. F., OHLENDORF, D. H., TAKEDA, Y. & MATTHEWS, B. W. (1981). *Nature (London)*, **290**, 754–758.

ARNDT, U. W., CHAMPNESS, J. N., PHIZACKERLEY, R. P. & WONACOTT, A. J. (1973). *J. Appl. Cryst.* **6**, 457–463.

ARNDT, U. W. & SWEET, R. M. (1977). In *The Rotation Method in Crystallography*, edited by U. W. ARNDT & A. J. WONACOTT, pp. 43–62. Amsterdam: North-Holland.

ARNDT, W. & WONACOTT, A. J. (1977). *The Rotation Method in Crystallography*. Amsterdam: North Holland.

COLMAN, P. M., JANSONIUS, J. N. & MATTHEWS, B. W. (1972). *J. Mol. Biol.* **70**, 701–724.

DIAMOND, R. (1969). *Acta Cryst.* A**25**, 43–55.

FORD, G. C. (1974). *J. Appl. Cryst.* **7**, 555–564.

GRÜTTER, M. G., RINE, K. L. & MATTHEWS, B. W. (1979). *J. Mol. Biol.* **135**, 1029–1032.

HARRISON, S. C. (1968). *J. Appl. Cryst.* **1**, 84–90.

HUXLEY, H. E. (1953). *Acta Cryst.* **6**, 457–465.

IRWIN, M. J., NYBORG, J., REID, B. R. & BLOW, D. M. (1976). *J. Mol. Biol.* **105**, 577–586.

JONES, A., BARTELS, K. & SCHWAGER, P. (1977). In *The Rotation Method in Crystallography*, edited by U. W. ARNDT & A. J. WONACOTT, pp. 105–118. Amsterdam: North-Holland.

LOVE, W. E., HENDRICKSON, W. A., HERRIOTT, J. R., LATTMAN, E. E. & MCCORKLE, G. L. (1965). *Rev. Sci. Instrum.* **36**, 1655–1656.

MATTHEWS, B. W., FENNA, R. E., BOLOGNESI, M. C., SCHMID, M. F. & OLSON, J. M. (1979). *J. Mol. Biol.* **131**, 259–285.

NYBORG, J. & WONACOTT, A. J. (1977). In *The Rotation Method in Crystallography*, edited by U. W. ARNDT & A. J. WONACOTT, pp. 139–152. Amsterdam: North-Holland.

REEKE, G. N. JR (1977). In *The Rotation Method in Crystallography*, edited by U. W. ARNDT & A. J. WONACOTT, pp. 33–42. Amsterdam: North-Holland.

ROSSMANN, M. G. (1979). *J. Appl. Cryst.* **12**, 225–238.

SCHUTT, C. & WINKLER, F. K. (1977). In *The Rotation Method in Crystallography*, edited by U. W. ARNDT & A. J. WONACOTT, pp. 173–186. Amsterdam: North-Holland.

SCHWAGER, P., BARTELS, K. & JONES, A. (1975). *J. Appl. Cryst.* **8**, 275–280.

WILSON, K. & YEATES, D. (1979). *Acta Cryst.* A**35**, 146–157.

WONACOTT, A. J. (1977). In *The Rotation Method in Crystallography*, edited by U. W. ARNDT & A. J. WONACOTT, pp. 75–103. Amsterdam: North-Holland.

XUONG, N. H. & FREER, S. T. (1971). *Acta Cryst.* B**27**, 2380–2387.

SECTION G. REVIEWS ON PROTEIN STRUCTURE AND FUNCTION

Ann. Rev. Phys. Chem. 1976. 27:493–523
Copyright © 1976 by Annual Reviews Inc. All rights reserved

X-RAY CRYSTALLOGRAPHIC STUDIES OF PROTEINS

✣2633

B. W. Matthews
Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oregon 97403

INTRODUCTION

Following the dramatic impact of the first protein structure determinations, the field of macromolecular crystallography has expanded with ever-increasing rapidity. In many cases the techniques used are the same as those pioneered by Perutz and Kendrew and their disciples, but the preoccupation has shifted from methods to results. The first glimpses of protein architecture, provided by myoglobin (1) and hemoglobin (2), showed that proteins with related function could have structural similarity. In general, however, the early structure determinations, of lysozyme (3), ribonuclease (4), carboxypeptidase (5), chymotrypsin (6) and so on (Table 1), demonstrated most strikingly the enormous variability of protein structure. Recent structure analyses have provided further examples of diversity in structure, but at the same time have shown the existence of similar patterns of folding within different proteins that may or may not have related function and may or may not have recognizably homologous amino acid sequences. The origin of these recurrent "domains" or "super secondary structures" is the subject of considerable debate.

Not only has the number of known structures continued to grow, but also the quality of a few structure determinations has been increased by improved data collection and refinement techniques to the point that relatively precise statements can be made about protein conformation and function.

In a review of this type it is impossible to discuss all the protein structures known at the present time. Rather, the intention is to describe those areas of protein crystallography within which significant advances have been made during the last few years. Table 1 lists those protein structures for which the polypeptide chain has been traced, and includes selected references to the more recent literature. Also a number of relatively recent reviews are available (7–13).

Table 1 Protein structure determinations[a]

Year	Protein	Species	Molecular weight[b]	Resolution (Å)	Reference[c]
1960	Myoglobin	Sperm whale	17,800	2.0, 1.4	1, 14–16
1965	Lysozyme	Hen egg white	14,600	2.0	3, 17–19
1967	Carboxypeptidase A	Bovine	34,600	2.8, 2.0	5, 20–21
	α-Chymotrypsin	Bovine	25,000(2)	2.0	6, 22
	Ribonuclease A	Bovine	13,700	2.0	4
	Ribonuclease S	Bovine	13,700	3.5, 2.0	23, 24
1968	Hemoglobin, oxy	Horse	66,500($\frac{1}{2}$)	2.8, 2.5	25–29
	Papain	Papaya	23,000	2.8	30, 31
1969	γ-Chymotrypsin	Bovine	25,000	2.7	32, 33
	Hemoglobin	Chironomous	17,000	2.8, 2.5	34, 35
	Insulin	Porcine	5,800(2)	2.8, 1.5	36–38
	Rubredoxin	*Clostridium pasteurianum*	6,000	3.0, 1.5	39–41
	Subtilisin BPN′	*Bacillus amyloliquefaciens*	27,500	2.5, 2.0	42, 43
1970	Chymotrypsinogen	Bovine	25,000	2.5	44
	Elastase	Porcine	25,900	3.5	45
	Hemoglobin, deoxy	Horse	66,500($\frac{1}{2}$)	2.8	46
	Hemoglobin, deoxy	Human	66,500	3.5	47
	Lactate dehydrogenase	Dogfish	140,000($\frac{1}{4}$)	2.8, 2.0	48, 49
	Trypsin inhibitor	Bovine pancreas	6,500	2.5, 1.5	50, 51

Table 1 (*Continued*)

Year	Protein	Species	Molecular weight[b]	Resolution (Å)	Reference[c]
1971	Carbonic anhydrase C	Human	30,000	2.0	52, 53
	Cytochrome b_5	Calf liver	11,000	2.8, 2.0	54, 55
	Cytochrome *c*, ferri	Horse	12,500	2.8	56, 57
	Cytochrome *c*, ferro	Tuna	12,500	2.5, 2.0	57, 58
	Flavodoxin	*Clostridium pasteurianum*	16,000	3.25, 1.9	59, 60
	Hemoglobin	*Glycera dibranchiata*	18,200	2.5	61, 62
	Hemoglobin	Lamprey	18,000	2.0	63
	High potential iron protein	Chromatium	9,650	2.25, 2.0	64–66
	Myogen	Carp	11,000	2.0, 1.85	67–69
	Nuclease	*Staphylococcus aureus*	16,800	2.0, 1.5	70, 71
	Subtilisin Novo	*Bacillus subtilis*	27,500	2.8	72
	Trypsin	Bovine	25,000	2.7	73–75
1972	Concanavalin A	Jack bean	108,000($\frac{1}{4}$)	2.0	76–78
	Concanavalin A	Jack bean	108,000($\frac{1}{4}$)	2.4	79, 80
	Cytochrome *c*, ferro	Bonito	12,500	2.3	81–83
	Ferredoxin	*Peptococcus aerogenes*	6,000	2.5, 2.0	84–86
	Flavodoxin	*Desulfovibrio vulgaris*	16,000	2.5	87
	Insulin	Porcine	5,800(2)	3.1	88
	Malate dehydrogenase	Porcine	72,000	3.0, 2.5	89–91
	Ribonuclease A	Bovine	13,700	2.5	92–93
	Thermolysin	*Bacillus thermoproteolyticus*	34,600	2.3	94–96

Table 1 (*Continued*)

Year	Protein	Species	Molecular weight[b]	Resolution (Å)	Reference[c]
1973	Alcohol dehydrogenase	Horse liver	80,000($\frac{1}{2}$)	2.9, 2.4	97, 98
	Bence–Jones protein Mcg	Human	23,000(2)	3.5, 2.3	99, 100
	α-Chymotrypsin	Bovine	25,000(2)	2.8	101, 102
	Cytochrome *c*, ferro	Tuna	12,020	2.45, 2.0	57, 103
	Cytochrome c_2	*Rhodospirillum rubrum*	12,480	2.0	104
	Cytochrome c_{550}	*Micrococcus denitrificans*	14,000	4.0, 2.5	57, 105
	Fab fragment	Human	55,000	2.8, 2.0	106, 107
	Glyceraldehyde-3-phosphate dehydrogenase	Lobster	143,000	3.0, 2.9	108, 109
	Lactate dehydrogenase, ternary complex	Dogfish	140,000	3.0	110, 111
	Trypsin-pancreatic trypsin inhibitor	Bovine	30,000	2.8, 1.9	112–114
1974	Adenylate kinase	Porcine	22,000	3.0	115
	Bence–Jones protein REI	Human	11,000(2)	2.8, 2.0	116, 117
	Fab Fragment, McPC 603	Mouse	55,000	3.0	118, 119
	Flavodoxin, oxidized	*Clostridium MP*	16,000	1.9	60, 120
	Insulin	Porcine	5,800(2)	1.8	121
	Lysozyme	Human	14,500	2.5	122
	Lysozyme	Bacteriophage T4	18,700	2.5	123
	Phosphoglycerate kinase	Horse muscle	38,000	3.0	124
	Phosphoglycerate kinase	Yeast	45,000	3.5	125
	Phosphoglycerate mutase	Yeast	110,700($\frac{1}{2}$)	3.5	126
	Prealbumin	Human	54,000($\frac{1}{2}$)	2.5	127
	Rhodanese	Bovine liver	32,000	3.9, 3.0	128, 129
	Trypsin-soybean trypsin inhibitor	Porcine, soybean	43,500	2.6	130, 131

Table 1 (*Continued*)

Year	Protein	Species	Molecular weight[b]	Resolution (Å)	Reference[c]
1975	Bacteriochlorophyll protein	*Chlorobium limicola*	127,000(1/3)	2.8	132
	Bence–Jones fragment Au	Human	11,000	2.5	133
	Bence–Jones fragment Rhe	Human	12,500(2)	3.0	134
	Carbonic Anhydrase B	Human	30,000	2.2	135
	Carboxypeptidase B	Bovine	34,000	2.8	136, 137
	Glucagon	Porcine	3,600	3.0	138
	Hexokinase	Yeast	51,000	2.7	139, 140
	Lysozyme, triclinic	Hen egg white	14,600	1.5	141
	Pepsin	Porcine	35,000	2.7	142
	Phospholipase A_2	Porcine	14,500	3.0	143
	Protease SGPB	*Streptomyces griseus*	18,400	2.8	144
	Superoxide dismutase	Bovine	16,000(2)	3.0	145
	Thioredoxin S_2	*Eschericia coli*	11,700(2)	2.8	146
	Triose phosphate isomerase	Chicken muscle	52,000	2.5	147
	Trypsin	Bovine	25,000	1.8	148–150

[a] Additional unpublished structure determinations, indicated in personal communications, include the following: seal whale myoglobin (H. Scouloudi); tyrosyl-tRNA synthetase (D. M. Blow); wheat germ agglutinin (C. S. Wright); methemerethrin (L. H. Jensen); actinidin (E. N. Baker); fungal acid protease (D. R. Davies); sickle cell hemoglobin (W. E. Love); neurotoxin (B. W. Low); phosphorylase *a* (R. J. Fletterick); α-lactalbumin (D. C. Phillips); leghemoglobin (B. K. Vainshtein); Fc antibody fragment (R. Huber); neurotoxin (D. Tsernoglou); dihydrofolate reductase (J. Kraut); *B. stearothermophilus* GPD (A. Wonacott).

[b] The number of molecular weight units, if different from unity, is indicated in parentheses.

[c] References are given to the original structure determination and to selected recent literature.

METHODS

Structure Determination

The method of isomorphous replacement, first applied to proteins by Green, Ingram & Perutz (151), has been employed in the great majority of protein structure determinations and is likely to remain preeminent for the forseeable future.

Nevertheless, it has recently been demonstrated that it is possible to use a known structure to derive subsequent related structures without the use of isomorphous replacement (152). For example, Schmid & Herriott (136, 137) have used the known structure of carboxypeptidase A (20, 21) to derive the atomic structure of carboxypoptidase B, even though the respective enzymes have a number of differences in their amino acid sequences. Also Fehlhammer, Bode, and Schwager (149, 150) used the structure of trypsin as seen in a trypsin-trypsin inhibitor complex to derive the structure of crystalline free trypsin.

The basic idea is to first orient the "known" structure in the "unknown" unit cell, and then, by a series of refinement cycles, to proceed from the known to the unknown structure. The orientation procedure is nontrivial and has been discussed extensively elsewhere (e.g. see 153). Recent experience suggests that the orientational parameters may be better determined by calculating the "rotation function" in Patterson space rather than in reciprocal space (e.g. see 149). The method is restricted to proteins that have similar three-dimensional structures, although it is not clear at this time how "similar" they need be for the method to be successful. The method can be very helpful in cases where isomorphous replacements cannot be readily obtained, and could in principle be used to derive a series of rather closely related protein structures, as, for example, has already been demonstrated with the immunoglobulins (133).

Data Collection

RAPID DATA ACQUISITION Most of the structure determinations listed in Table 1 have employed an automatic diffractometer to measure the X-ray diffraction data; the reflections usually are recorded one at a time. Also, several structural studies have been pursued effectively using conventional Buerger precession photography, coupled with an automated film-measuring device (6, 67, 76, 94, 124, 132).

Recently there has been considerable progress in the development of more efficient modes of data collection, on the one hand using photographic methods, and on the other hand using position-sensitive X-ray area detectors.

In the photographic case, the crystal is usually oscillated slowly through a series of small contiguous angular ranges during which the reflections are recorded on a series of films. The measurement of the films is nontrivial, but the overall efficiency of data collection can be very high (154–156). The method has great potential for the measurement of diffraction data from macromolecules with large unit cells, including virus crystals, and has also been used with striking success in Huber's laboratory for the rapid acquisition of high (beyond 2 Å) resolution data (149). Related techniques have also been used for medium-resolution studies of proteins in a number of laboratories (e.g. see 44, 113).

Two different types of position-sensitive area detector have been proposed for rapid data acquisition. In one scheme the diffracted X-rays fall on a scintillation screen and the resultant light flash is amplified and recorded either on film or on a TV screen (157–159). In the second method, developed by Xuong and collaborators, the X-rays are detected by a multi-wire proportional counter, and the counts accumulated in a large-core computer memory (160). Although none of the position-sensitive detectors have, to date, been used to complete a full structure determination, this seems to be only a matter of time. Xuong's system, in particular, has been used for extensive data collection at low and intermediate resolution and appears to have great promise, although measurements made with the present instrument have a rather high background which limits the accuracy of the weaker high-resolution data (N. Xuong and L. H. Jensen, personal communication).

SYNCHROTRON RADIATION There has been considerable interest in the potential use of synchrotron radiation for diffraction studies (161, 162). Although very intense beams of X-rays can be obtained, the useful intensity for "conventional" crystallography is at present only about one- to tenfold greater than from a rotating anode generator, which is not a sufficient gain to offset the inconvenience of working with a synchrotron facility. It will require intensity gains of at least 100-fold before synchrotron radiation has a major impact in "conventional" structure determination. On the other hand there are special circumstances where the unique properties of synchrotron radiation lead to important advantages. For example, in low-angle diffraction studies the strict collimation of the beam leads to an improvement in resolution and an increase of 50-fold or so in intensity (162). Also the ability to vary continuously the wavelength of the radiation allows one to make diffraction measurements close to absorption edges of any metals bound to the protein, thus maximizing the small intensity changes arising from anomalous scattering. Such enhancement of anomalous scattering is likely to be helpful in locating the bound metals, and, to some extent, in solving protein structures, but will not obviate the need for isomorphous heavy atom derivatives in macromolecular crystallography.

LOW-TEMPERATURE CRYSTALLOGRAPHY There has been renewed interest recently in the feasibility of data collection at low temperature. Potential advantages of this technique include the reduction of radiation damage, a possible increase in resolution, and the slowing-down of enzymatically catalysed reactions to the point that productive enzyme-substrate complexes can be studied directly by X-ray techniques. The diffraction data used for the structure determination of, for example, carboxypeptidase A and staphylococcal nuclease were collected at crystal temperatures between 0°C and 4°C (20, 21, 70). At lower temperatures crystal disorder due to formation of ice crystals will normally occur, leading to gross deterioration of the diffraction pattern (163). Hass (164, 165) investigated the feasibility of preserving crystal perfection by soaking glutaraldehyde-cross-linked crystals of lysozyme and lactate dehydrogenase in glycerol or sucrose, and then rapidly cooling them in liquid nitrogen. Such crystals showed a tenfold decrease in the rate of radiation damage, but the quality of the diffraction data appear to be somewhat inferior to that measured at room temperature.

More recently Petsko (166) has shown that crystals of a number of different proteins could be successfully cooled to −70°C or so by coupling the cooling cycle with a gradual replacement of the normal mother liquor with salt-free aqueous/organic liquids. Protein denaturation in such mixtures is minimized by selecting conditions such that the dielectric constant is near that of water at 20°C (167). For a variety of crystals tested, Petsko found that at temperatures near −70°C, radiation damage was negligible. In most cases cooling did not result in an improvement in resolution; however, in two cases where the crystals already diffracted unusually well at room temperature (to 1.1–1.3 Å), an improvement was noted. This is consistent with the notion that the "temperature factor" of most protein crystals is primarily due to crystal disorder, and that it is only for unusually well-ordered crystals that true thermal motion is a significant component of the overall thermal factor. Further developments in this area are awaited with interest.

Structure Refinement

Many factors influence the accuracy of a protein structure determination. Further, it is not easy to assess the probable errors in a given structural study.

In general, a low-resolution (up to 4 Å) electron density map will reveal little more than the overall shape of the molecule, although there are exceptions such as the globins (1, 2) and, more recently, the hemerythrins (168, 169), where the high proportion of helix allowed the approximate course of the polypeptide chain to be followed.

At about 3.5 Å resolution it may be possible in favorable cases to trace the probable course of the polypeptide chain (23, 45, 47, 126), although details of backbone and side-chain conformations cannot be seen.

Beyond 3.0 Å resolution it becomes possible to distinguish backbone and side-chain geometry and, knowing the amino acid sequence, to assign approximate atomic coordinates. The accuracy of such preliminary coordinates will depend not only on the resolution, but also on the size and perfection of the crystals, the number and quality of the isomorphous heavy-atom derivatives, and the location of the atom in question within the protein molecule. From a typical 2.5 Å medium-resolution electron density map, preliminary coordinates for most of the structure can be obtained with a probable error of about 0.5 Å, but there may be a few regions on the surface of the molecule where the likely error can be several angstroms. As we describe in the following section, by using high-resolution data (beyond 2 Å), considerable improvement in accuracy can be realized.

Several years ago Jensen and colleagues achieved an important breakthrough in demonstrating that it was possible to refine the structure of a protein by difference Fourier and least-squares techniques (41, 170). In the case of rubredoxin, using data to 1.5 Å resolution, it was possible to reduce the crystallographic residual R to 12.6% and the coordinate error to less than 0.2 Å. More recently, a number of other proteins have been refined and in some instances the estimated standard deviation of the atomic coordinates reduced to less than 0.1 Å (18, 22, 28, 51, 66, 69, 113, 117). In refining a protein model it is necessary to ensure that it retains acceptable stereochemistry (171–173); improved methods of applying constraints are now available

(174, 175). One of the current anomalies in protein refinement is that the discrepancy between observed and calculated structure amplitudes for a "fully refined", stereochemically constrained structure is usually 20–25%, even though the expected errors in the observed data account for perhaps a third of this residual.

In addition to the above refinement of atomic coordinates to optimize the agreement between the observed and model structure amplitudes, there has also been interest in the development of methods to improve the electron density map itself, prior to using it to obtain a set of starting coordinates. Sayre has proposed a method of phase refinement and extension, to be carried out in reciprocal space, and applied it to rubredoxin (176, 177). Alternatively, it has been proposed that an equivalent result can be achieved more economically by real space modification of the electron density map (178, 179), in a manner similar to that originally proposed by Hoppe & Gassmann (180). [Experience with other refinement procedures has been reviewed elsewhere (13).] In both the real and reciprocal space refinement procedures, it is necessary to have available a set of approximate starting phase angles to a resolution of about 2.5 Å, and also structure amplitudes to higher resolution (e.g. 1.5 Å). These are essentially the same starting conditions for a "conventional" refinement in which the coordinates of an approximate model are refined directly. At this time it is not clear which refinement procedure is likely to be optimal in a given situation. Sayre's method can be useful, but is expensive in terms of computing time. The density modification procedure also has its place; however the claim of Collins et al (179) that efficiencies achieved by density modification coupled with an interactive graphics system represent "an order of magnitude improvement in the state of the art" is debatable. As stated by Jensen and co-workers (181), "The new methods contribute to improving the interpretability of the electron density maps while classical refinement results both in improved interpretability *and in a set of atomic coordinates*" (my italics).

In addition to the *crystallographic* refinement described above, a number of protein structures have also been subject to energy minimization (e.g. see 16, 19, 182).

CRYSTAL AND SOLUTION

The possibility that the conformation of a macromolecule in a crystal might, in one way or another, be different from that in solution has been a matter of concern since the inception of protein crystallography. Over the years several lines of evidence have suggested that the process of crystal formation does not significantly modify the structure of a protein. Recent work has tended to strengthen this conclusion and has provided explanations for some apparent anomalies.

It is an important characteristic of protein crystals that usually about 50% of the crystal volume is occupied by solvent. Even the most tightly packed crystals contain at least 30% solvent, while at the other extreme crystals of some larger proteins contain up to 78% solvent (13, 183). The solvent content of 226 different protein crystals is summarized in Figure 1. There are always channels of solvent extending through the crystal, so that the molecular environment within the crystal

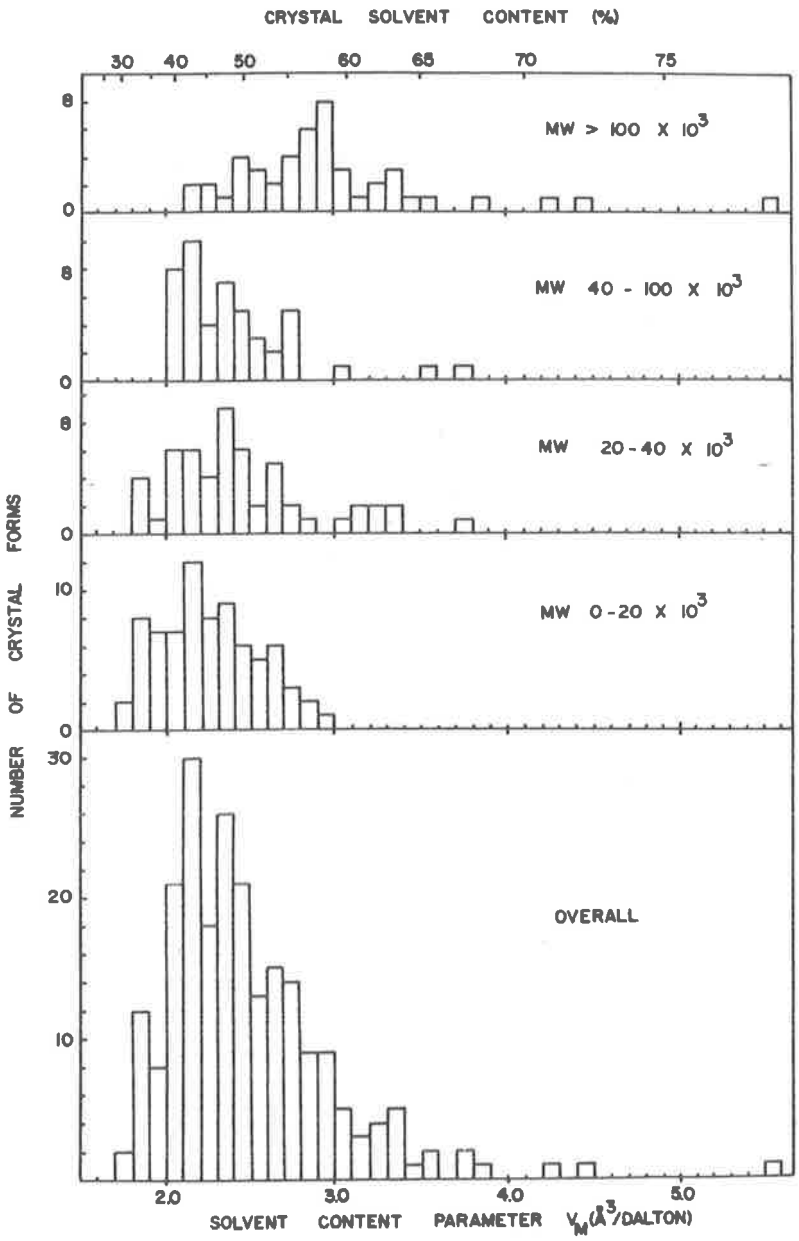


Figure 1 Solvent content parameter V_M and solvent content of 226 different crystal forms of globular proteins. (Adapted from reference 13.)

is in fact rather similar to that in solution. A particularly striking example is provided by the crystals of the trimeric bacteriochlorophyll protein from *Chlorobium limicola* where there are channels 62 Å in diameter extending throughout the crystal (Figure 2) (132). Presumably even small proteins could be diffused into such crystals.

In general, only a small fraction of the surface of a given protein molecule participates in intermolecular contacts within the crystal, and the crystal stabilization energy is only about 6 kcal mole^{-1} (95, 184), so that on general thermodynamic grounds one would not expect crystallization to perturb substantially the conformation of a protein.

This has in fact been shown directly by comparing the independently determined tertiary structures of the same or related proteins crystallized in different space groups in the presence of radically different solvents. For example, the structures of subtilisin BPN', crystallized from 2.1 M ammonium sulfate, and of subtilisin Novo, crystallized from 55% acetone, were determined independently, and the rms difference between 1895 atoms in the respective structures was found to be 1.55 Å (42, 72, 185). A number of other examples of this type can be cited (13). A recent and more accurate comparison has been made between the crystal structure of free bovine trypsin and the structure of trypsin complexed with bovine pancreatic trypsin inhibitor, both of which have been refined in Huber's laboratory (113, 114, 149, 150). In this case the coordinates of the main chain atoms of free trypsin and of the trypsin component of the trypsin-trypsin inhibitor complex differ, on the average, by only 0.26 Å, and the backbone dihedral angles differ by 8.0°. Thus, even at the unusually high precision of these two structure determinations, with estimated coordinate standard deviations of about 0.1 Å, no significant structural changes in the backbones of the respective molecules are seen. Bode & Schwager (150) also comment on the close similarity between trypsin and the refined structure of α-chymotrypsin (22), notwithstanding the differences in amino acid sequences of these homologous serine proteases. Accurate conformational data of this type provide further evidence that protein structure as seen in the crystal is likely to be very close to that in solution.

Evidence of a different type comes from the observation that a number of enzymes are catalytically active in the crystalline phase, albeit with reduced efficiency in some cases. Examples include ribonuclease S (186), ribonuclease A (187), carboxypeptidase A (188), and γ- and α-chymotrypsin (189). There are examples of anomalous behavior, but these now appear to have been resolved. For example, Johansen & Vallee (190, 191) found that carboxypeptidase A specifically labelled at Tyr 248 with diazotized arsanilic acid was red in solution, but yellow in the crystal form, indicating that in solution the azophenol moiety complexes with the active-site Zn^{2+}, but not in the crystals, in apparent conflict with the mobility of Tyr 248 observed in crystalline carboxypeptidase A (20, 21). Subsequently Quiocho et al (192) pointed out that the form of the enzyme used by Johansen & Vallee (CPA_γ) has seven residues less than that used for the crystallographic analysis (CPA_α), and crystallizes somewhat differently. Furthermore, modified CPA_α is red, both in the crystals and when dissolved, indicating that in both cases Tyr 248 is free to approach the zinc ion. Thus it seems that the apparent difference in behavior of CPA_γ in

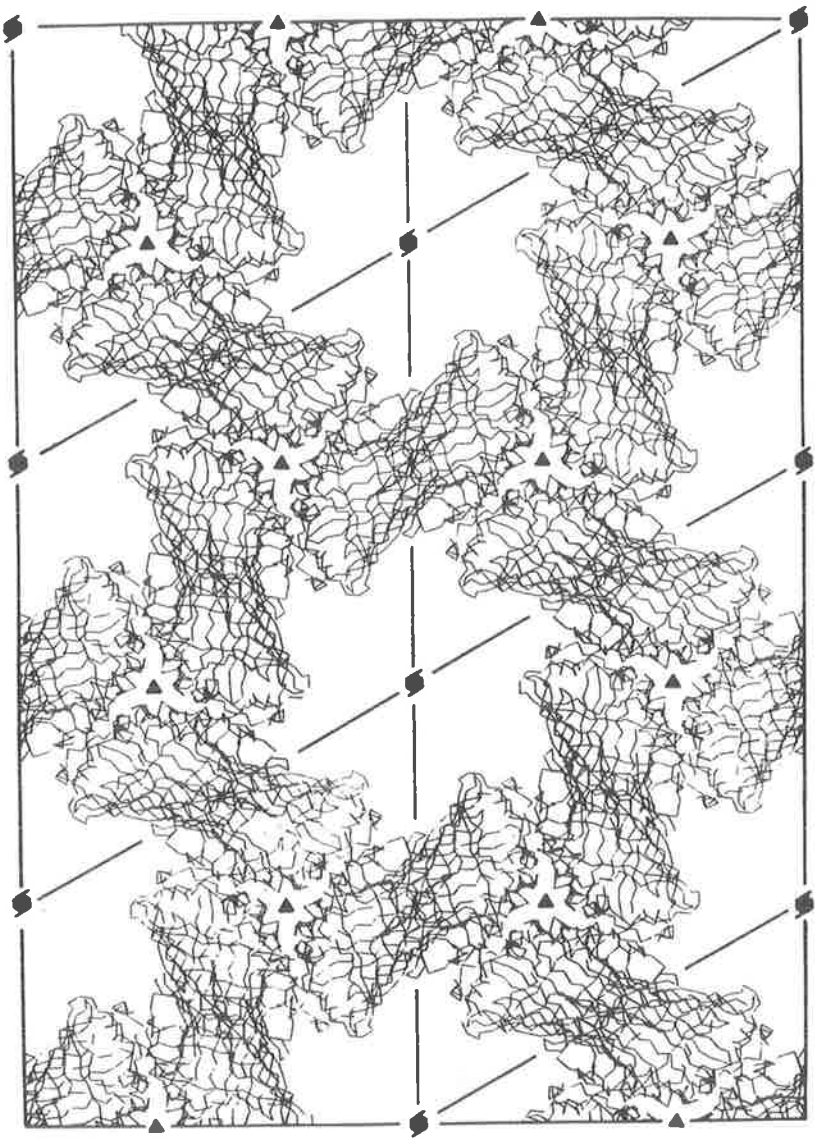


Figure 2 Molecular packing in hexagonal crystals of a bacteriochlorophyll protein (132). The protein molecules, for which only the backbone is shown, are projected parallel to the hexagonal axis. Cell dimensions are $a = b = 112.4$ Å, $c = 98.4$ Å. For clarity, the symbols for the 2_1 screw symmetry elements at the midpoints of the cell edges and at the centres of each unit cell have been omitted.

the crystals and in solution arises from the crystal-packing interactions that restrict the movement of Tyr 248, and not from an intrinsic change in conformation of the enzyme induced by the crystallization process (193), although Johansen & Vallee dispute this conclusion (194).

A second example of apparently anomalous behavior in the crystal, relative to that in solution, is provided by concanavalin A. The same crystal form of this lectin, a phytohemagglutanin that binds mono, oligo- and polysaccharides, has been studied extensively by the Rockefeller group (76–78) and by Hardman and Ainsworth (79, 80). Studies on the binding of β-(*o*-iodophenyl)-D-glucopyranoside (β-IPG) (195) and of *myo*-inositol (196) to the crystalline lectin suggested that the saccharide binding site was in a deep cavity at least 20 Å from the double Ca^{2+}–Mn^{2+} metal binding site, although it was also shown on the one hand that *myo*-inositol does not inhibit hemagglutination (76) and on the other hand that the binding of β-IPG was probably due to the iodophenyl moiety rather than the saccharide portion of the inhibitor (197). These results were in conflict with the finding of Kalb & Levitzki (198) that the integrity of the metal binding site was necessary for the binding of carbohydrates, suggesting that the carbohydrate probably bound near the metals. Furthermore, Brewer et al (199) showed rather convincingly by NMR that the sugar-binding site was about 10 Å from the manganese ion.

Very recently this paradox has been resolved. Hardman & Ainsworth (200) have co-crystallized, in a new crystal form, the complex of methyl-α-D-mannopyranoside with concanavalin A and have shown in a 6 Å resolution analysis that the sugar-binding site is in fact 10 to 14 Å from the manganese atom. Becker et al (201), in an independent study at 3.5 Å resolution on cross-linked crystals, have located the bound saccharide 13 Å from the manganese ion. They have also shown that the structural changes induced by removal of the essential metal ions are restricted to this region.

In the case of both carboxypeptidase A and of concanavalin A, apparent differences in behavior between crystal and solution have turned out to be due to steric hindrance resulting from intermolecular contacts in the crystal. In carboxypeptidase A_γ, such contacts prevent the free motion of Tyr 248, while in the crystal form of concanavalin A used for the original structure determinations, access to the biologically relevant carbohydrate binding site is prevented by the presence of a neighboring molecule. This behavior is, of course, possible for any crystalline enzyme, especially those that bind bulky substrates, and must be considered as a possibility in any crystallographic study. In total, however, the evidence that macromolecular structures determined crystallographically are biologically relevant is now overwhelming.

FAMILIES OF STRUCTURES

Homologous Proteins

It was one of the dramatic early discoveries of protein crystallography that the tertiary structures of myoglobin and hemoglobin were closely related (1, 2). Since that time many other examples of structural similarity have been reported. This has

not been surprising for proteins with obviously homologous amino acid sequences and related catalytic function. However, examples now exist of pairs of proteins with extensive common structural elements, yet unrelated biological activities and apparently unrelated primary structures. The origin of this structural similarity is, at present, the subject of a great deal of debate.

The most straightforward examples of structural similarity are provided by families of proteins with homologous amino acid sequences and related functions. In every case where clear sequence homology exists between two proteins, their tertiary structures have been shown to be very similar, even in cases where up to 80% of the amino acids are nonidentical. Examples include the globins, where the "myoglobin fold" has been shown to occur not only in sperm whale (1), seal (202), and tuna (203) myoblogin and the α- and β-chains of horse hemoglobin (25, 26), but also in the hemoglobin from sea lamprey (63), a marine annelid worm (62), a larval insect (34), and lupin root nodules (204). Also the structures of the serine proteases α- and γ-chymotrypsin (22, 205), elastase (45), and trypsin (75, 150) have been shown to be obviously similar, as have the structures of respiratory cytochromes *c* from eukaryotic and bacterial sources (57, 105), and also a photosynthetic cytochrome c_2 from *Rhodospirillum rubrum* (104). Within each of these families of proteins, variations in amino acid sequence, or insertions and deletions, cause modifications of the basic structure, but in each case the overall structural similarity remains obvious. A somewhat extreme case is provided by an extracellular protease from *Streptomyces griseus* (144). This microbial serine protease has short active site peptides identical to the corresponding peptides of the mammalian serine proteases, but there are large differences between the respective amino acid sequences. The microbial enzyme has 186 amino acid residues compared with 245 for α-chymotrypsin, and not only are there deletions of ten or more contiguous residues in the microbial enzyme relative to the mammalian one, but also there are large insertions. Even when the sequences are aligned as optimally as possible, less than 20% of the corresponding loci are occupied by the same amino acid. Nevertheless, the recent structure determination of the microbial protease (144) shows that two thirds of the residues are in a conformation similar to that observed for the pancreatic enzymes. Obviously the two structures are derived from a common precursor. They have the same structural core, and differ primarily in the disposition of structural loops, present only in one or the other enzyme, on the surface of the molecule.

There is a structural relation of a different character between the bacterial serine protease subtilisin and α-chymotrypsin. In this case the active sites of the respective serine proteases have been shown to be virtually identical, both with respect to residues involved in substrate binding and in catalysis, with a *rms* difference in the location of corresponding atoms of only 0.8 Å (206). However, the overall folding of the polypeptide backbone of the respective enzymes is completely different, and the amino acid sequences have no detectable homology.

Dehydrogenases

The examples quoted above provide extreme examples of two types of structural relatedness. On the one hand, the mammalian serine proteases, for example, have

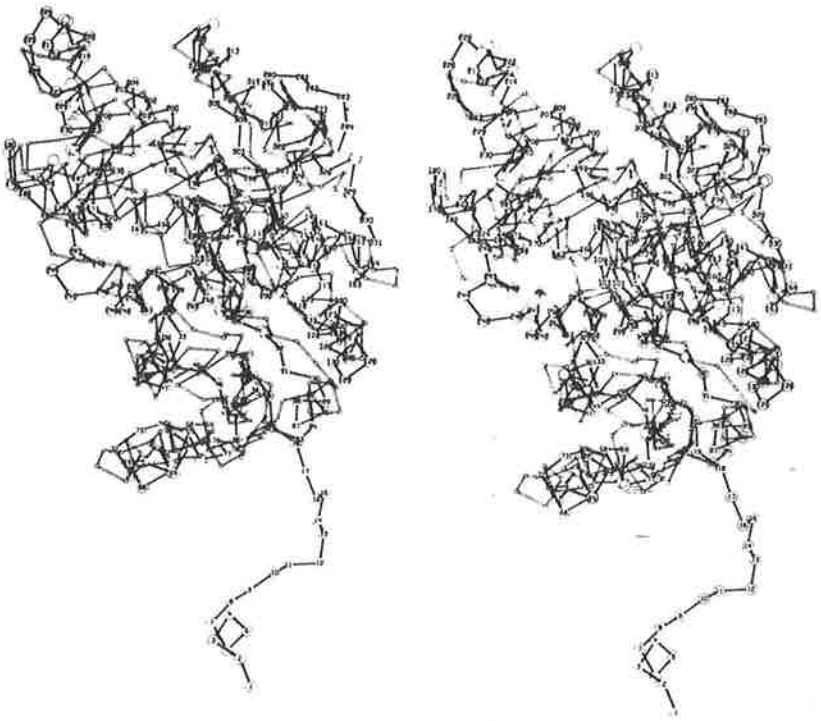


Figure 3 Stereo view comparing the structures of lactate dehydrogenase and soluble malate dehydrogenase. The malate subunits (open bonds) associate to form dimers, whereas the lactate subunits associate to form tetramers that are stabilized in part by the long "tail" present only in the lactate subunits. [From Rao & Rossmann (207) with permission.]

homologous amino acid sequences and similar three dimensional structures, and clearly diverged from a common ancestral precursor. At the other extreme, the mammalian serine proteases and subtilisin have *analogous* active sites, but are otherwise quite different, and provide a clearcut example of convergent structural evolution.

During the last few years an increasing number of related structures have been noted that do not obviously belong to either of the above categories. The best example is provided by the dehydrogenases and, more generally, by those proteins that require nucleotides as a cofactor. Following the structure determinations of dogfish lactate dehydrogenase (LDH) (48) and soluble porcine malate dehydrogenase (s-MDH) (89, 90), it was obvious that these two dehydrogenases had very similar tertiary structures, notwithstanding their different physical properties. Rao &

Rossmann (207) quantitatively compared the two structures and found, with the exception of the first 20 residues of LDH, which are involved in stabilization of the tetramer, that the two structures, including about 250 equivalent C^{α} atoms, differed by about 2.7 Å (Figure 3). In this case there is little doubt that the two dehydrogenases derived from a common precursor, presumably having rather broad dehydrogenase activity.

The subsequent elucidation of the structure of lobster glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (108) and of horse liver alcohol dehydrogenase (LADH) (97) complicated the picture. Here it was observed, in common with LDH and s-MDH, that the polypeptide chain of GAPDH folded to form two rather distinct domains, but only one of these domains was similar to that in LDH and s-MDH. For LADH also, two domains were found, but again only one was similar to that observed in the other dehydrogenases. In each case the common structural domain is that which is responsible for binding the obligatory cofactor nicotinamide adenine dinucleotide (NADH) and is now referred to as the "nucleotide-binding domain". The other half of each dehydrogenase constitutes the so-called "catalytic" domain, which is quite different for LDH, GAPDH, and LADH, and is thought to be responsible for the different dehydrogenase specificity exhibited by the respective enzymes. It is also of interest to note that the nucleotide-binding domain of LDH, S-MDH, and GAPDH is formed by the amino terminal half of the polypeptide chain, whereas for LADH it is the carboxyl half of the molecule (91).

The "idealized" nucleotide binding fold consists of six strands of parallel β-sheet with four helices that are parallel to each other and antiparallel to the individual strands of the β-sheet (Figure 4) so that the domain has an approximate two-fold axis of symmetry between the two central stands of β-sheet. The domain can be imagined as being built up by laying down a single strand of β-sheet, followed by an antiparallel length of helix that returns the polypeptide chain to a position such that the second strand of β-sheet can be laid down, and so on. The nucleotide-binding domain illustrated schematically in Figure 4 occurs in each of the four dehydrogenases. While there are substantial variations in the details of folding observed in the four enzymes (Figure 5), the preservation of the basic elements of the super secondary structure are apparent. Furthermore the location and geometry of cofactor binding have been determined for each enzyme, and found to be essentially the same within the respective domains (49, 90, 91, 98, 109, 208). The observed structural equivalence has been used by Rossmann and coworkers (91, 209, 210) and by Ohlsson et al (208) to align the amino acid sequences of the respective enzymes. Even with the known structural data as a guide, the sequences are not obviously homologous, although two glycine residues were observed to be invariant, and a repeating pattern of hydrophobic residues was noted. Calculation of the minimum base change per codon for the aligned sequences tended to confirm that they had in fact diverged from a common precursor. Not surprisingly, prior attempts to align sequences of dehydrogenases, made without the benefit of structural data, turned out to be misleading. Rossmann and colleagues have proposed that the NAD binding domain observed in the dehydrogenases evolved from a common mono-nucleotide-binding ancestor that itself originated as a three-strand precursor. It is assumed

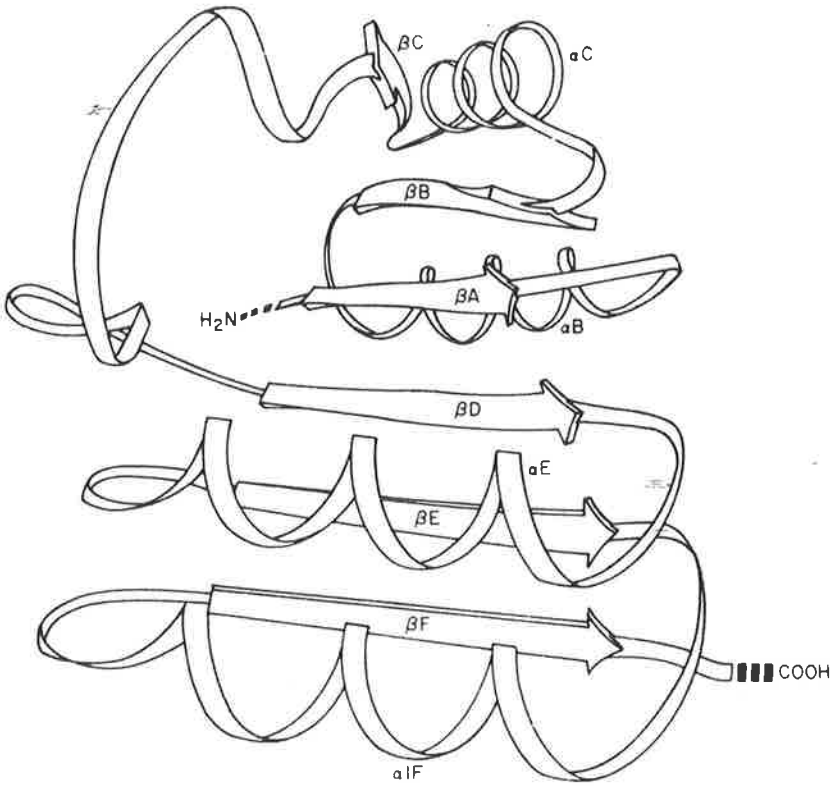


Figure 4 Schematic representation of the structure of the coenzyme-binding domain present in dehydrogenases. [From Ohlsson et al (208) with permission.]

that the precursor first formed a six-strand domain by duplication, and subsequently fused with a variety of other genes to yield the structures seen today.

Homology or Analogy?

In addition to the dehydrogenases, structural studies of a number of other glycolytic enzymes are in progress, and the striking result has emerged that a number of these proteins contain domains more or less similar to the "nucleotide-binding fold". For example, phosphoglycerate kinase (PGK) from both horse (124) and yeast (125) has been shown to include a domain of topology similar to that in the four dehydrogenases. Furthermore, the site of binding of the cofactor ATP is similar to that of the AMP portion of NAD in the dehydrogenases. In the case of porcine adenylate kinase there is a domain that resembles that of the dehydrogenases and PGK, but here two of the β-sheet strands are interchanged, and one is missing (115, 211). In this case the location of the substrate binding has not yet been determined with

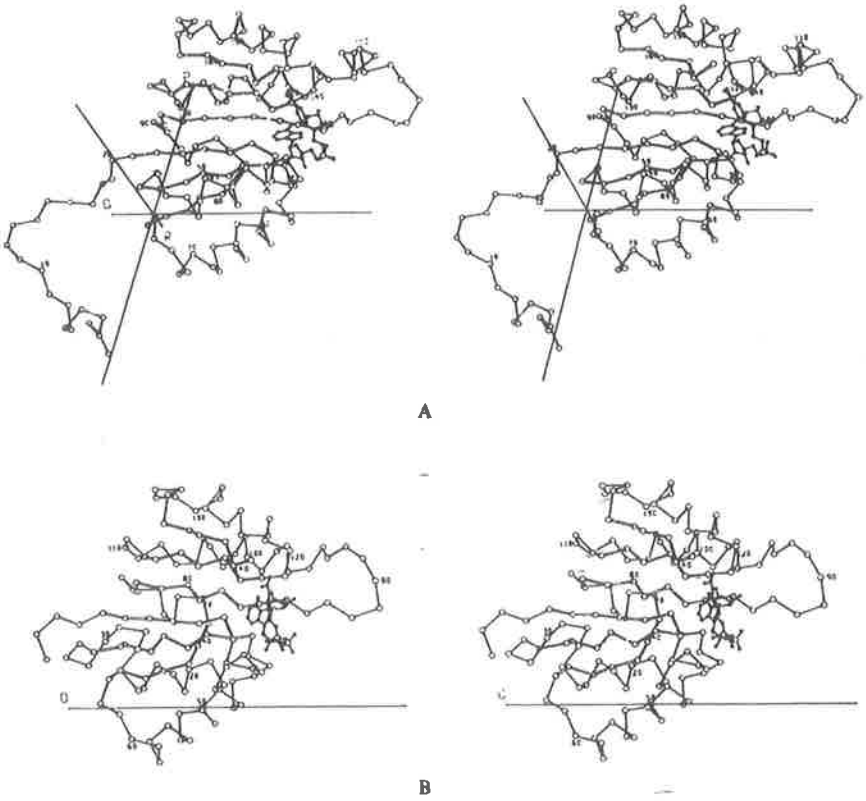
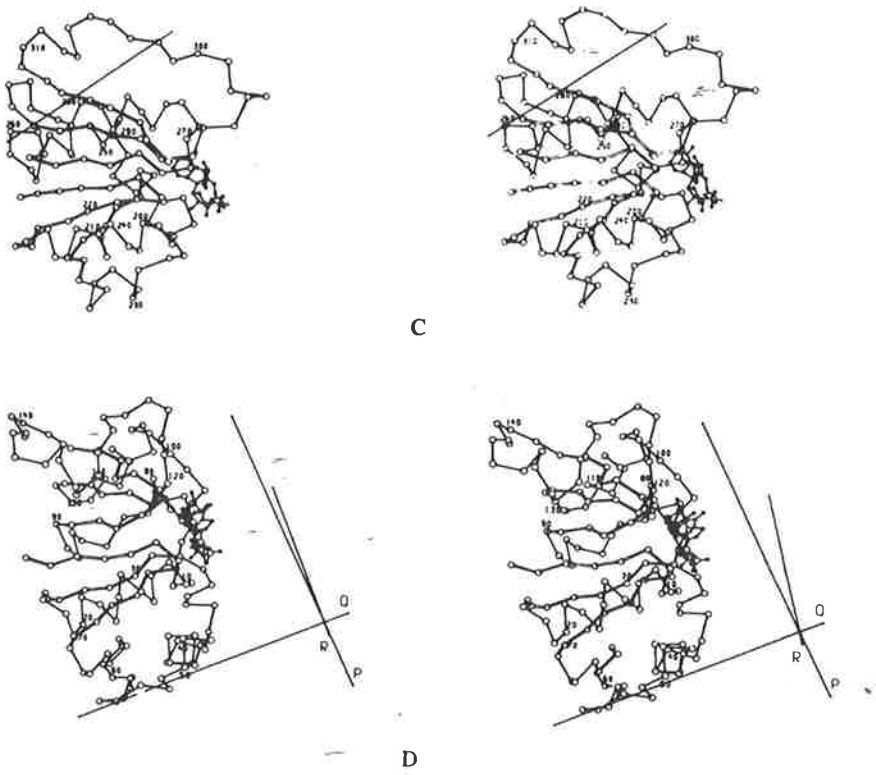


Figure 5 Stereo views of the nucleotide-binding domains in (A) LDH, (B) s-MDH, (C) GAPDH, and (D) LADH. [From Rossmann et al (91) with permission.]

certainty. For another phosphorylating enzyme, yeast hexokinase, the region of the structure that binds AMP includes six strands of β-sheet and superficially resembles the dehydrogenase domain, but here there are greater differences than was the case for adenylate kinase. In particular, two of the six β-sheet strands are antiparallel to the other four, and two of the LDH helices are either absent or differently located in hexokinase (139, 140). In yet another glycolytic enzyme, triose phosphate isomerase, the structure is built up of alternating β-sheet strands and helices and includes "super secondary" structures similar to the NAD-binding domain in LDH (147). Finally, structural domains resembling that first seen in LDH occur not only in other dehydrogenases and in other enzymes of the glycolytic pathway, but also in flavodoxin, subtilisin, and rhodanese (91, 207, 210).

While it is tempting to speculate that the presence of a similar structural element in different proteins indicates a common evolutionary precursor, the very fact that a region of β-sheet flanked by helices occurs so commonly suggests that such a



structural element may be a readily formed stable entity that could have evolved independently on more than one occasion.

Rossmann et al (209) and Ohlsson et al (208) have emphasized that in very distantly related proteins, derived from a common precursor, essentially no sequence homology need remain, and in such cases it is essential to use structural relatedness to identify corresponding amino acids in the respective sequences. There are, however, difficulties with this approach, for in estimating evolutionary distances from changes in amino acid sequence (e.g. see 212, 213) one should compare those amino acids that are "neutral", and free to change, whereas in comparing structures that are similar, one may be selecting residues that are "essential" for the preservation of the structural domain (209). Therefore any pair of similar domains, when compared, will tend to show some "relatedness" between their respective sequences whether or not they derived from a common precursor. Furthermore, when two amino acid sequences derived from a common precursor differ substantially, it

becomes probable that "unchanged" residues can occur through genetic change followed by subsequent reversion (212, 213).

Although originally favoring independent evolution of flavodoxin and the dehydrogenases (209), Rossmann and coworkers now consider divergent evolution to be more probable, and have constructed an evolutionary tree relating not only flavodoxin and the dehydrogenases, but also subtilisin and adenylate kinase (210). Schulz & Schirmer (211) also lean toward an evolutionary relation between the dehydrogenases, kinases, flavodoxin, and subtilisin, and Blake & Evans (124) and Bryant et al (125) favor a relation between the dehydrogenases and phosphoglycerate kinase. On the other hand, Steitz et al (140) have argued that yeast hexokinase is not evolutionarily related to either adenylate kinase or to phosphoglycerate kinase or for that matter to any of the glycolytic enzymes whose structures are known.

As suggested by Blake (214), it is by no means impossible that within this large group of apparently closely and more distantly related enzymes, there are both convergent and divergent relationships. It is clearly of importance to distinguish these relationships, and one can only hope that further structural studies will provide some unequivocal answers.

Other Related Structures

In addition to the "nucleotide-binding fold" several other examples of structural relatedness have been recognized recently. The first and perhaps the most striking of these is a "β-sheet barrel" consisting of seven strands, antiparallel except for the first and last strand, which was first observed as the basic structural element of immunoglobulin domains (99, 106, 118, 119, 133, 134, 215, 216) (Figure 6), and now found to occur in bovine superoxide dismutase (145, 217) (Figure 7). Approximately 50 residues in the respective domains can be superimposed within about 2.0–2.5 Å, depending on the type of immunoglobulin domain being compared. In the dismutase structure the polypeptide connecting two pairs of β-sheet strands extends from the cylinder as long loops and helps form the copper and zinc binding sites. The analogous polypeptide connections in the immunoglobulin variable domain extend to form two of the three bends that form the hypervariable antigen binding site.

Richardson et al (217) point out that the probability of generating the same topology observed in the immunoglobulins and superoxide dismutase by chance is very small, although if one allows for the fact that the types of folding commonly observed in other proteins are likely to be favored, then the probability is much increased (although still small).

Another case where evolutionary relatedness is possible, although not proven, is for cytochrome b_5 and the globins, which have been shown by Rossmann & Argos (218) to have topological similarities. In this case it is not clear to what extent the heme group, around which the polypeptide chain is wrapped, may tend to influence or restrict the conformation of the folded protein.

Finally, it has recently been observed (218) that portions of the structures of the lysozymes from bacteriophage T4 (123) and from hen egg white (HEW) (3) are similar (Figure 8), notwithstanding the fact that the respective sequences have no detectable homology. By allowing extensive deletions in the respective polypeptides,

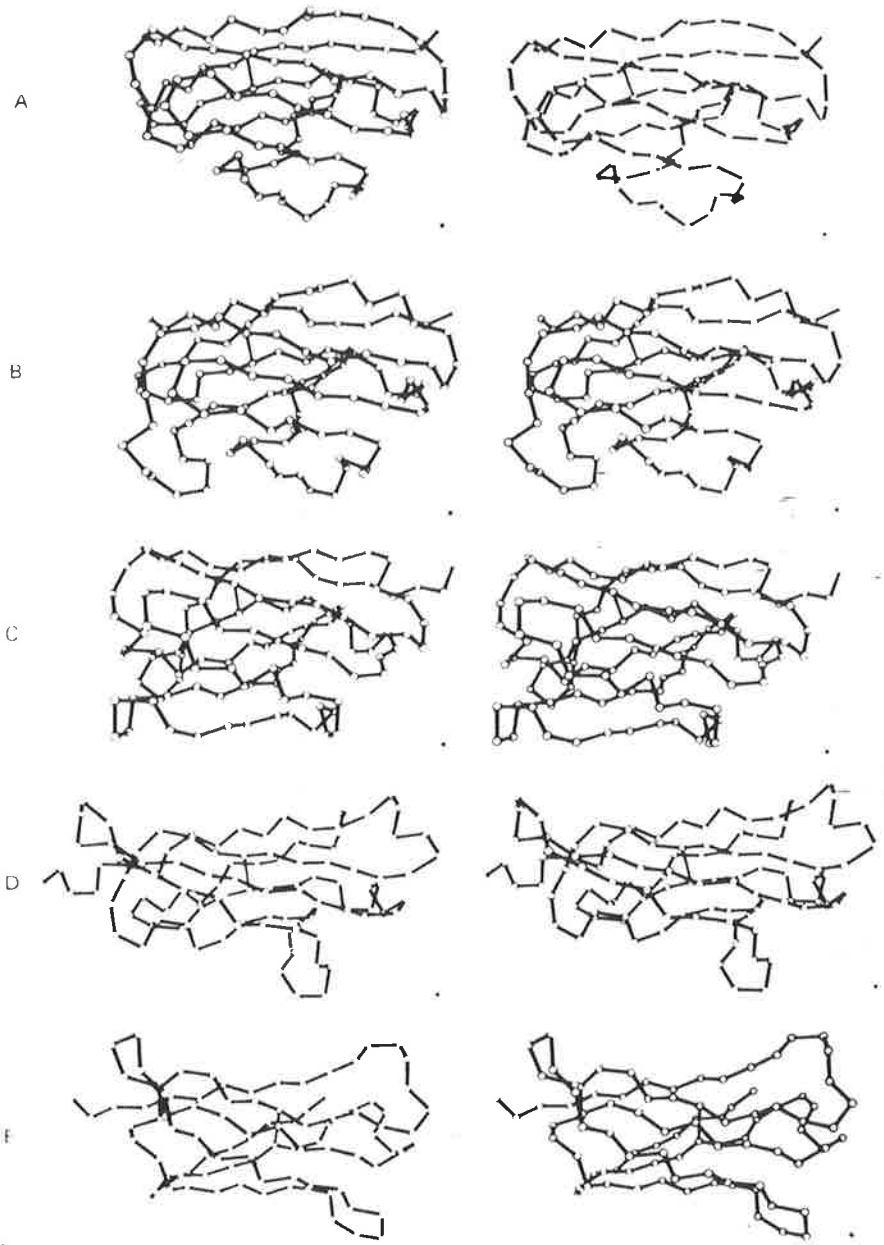


Figure 6 Stereo representation of the α-carbon backbones of five different immunoglobulin domains. (A) REI is a human V_K domain and (B), (C), (D), and (E) are V_L, V_H, C_L, and C_H1 domains from the mouse Fab McPC 603. [From Davies, Padlan & Segal (215), with permission.]

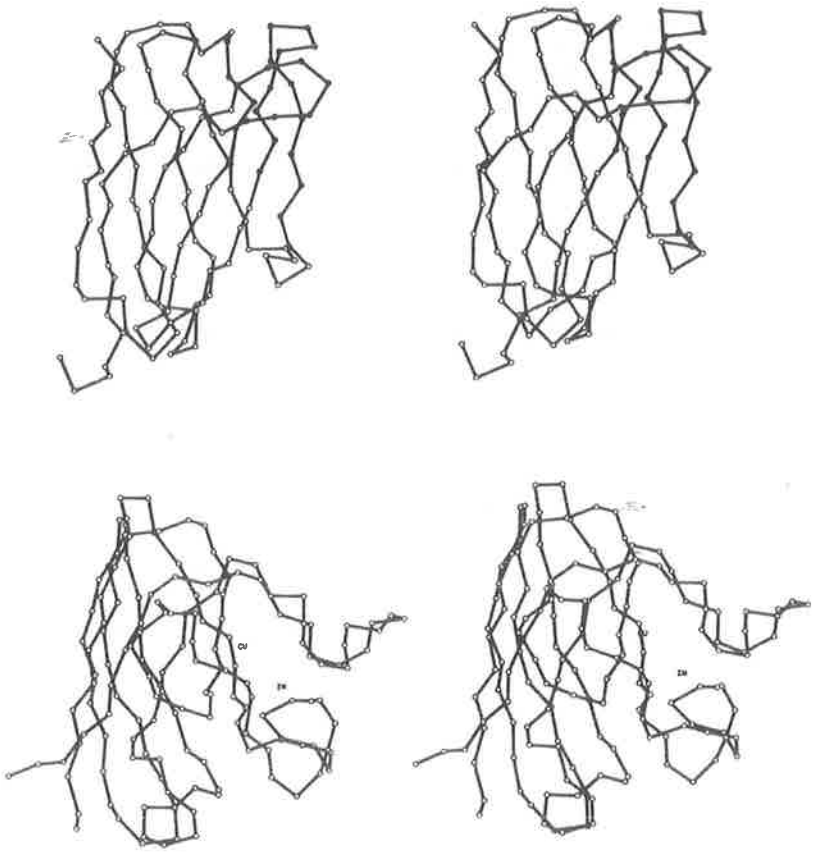


Figure 7 Stereo drawings illustrating (*a*) the backbone of an immunoglobulin domain (Figure 6C) and (*b*) the subunit of Cu, Zn superoxide dismutase. The hypervariable residues of the immunoglobulin domains are indicated by solid circles. [From Richardson et al (217), with permission.]

Rossmann & Argos (218) found that 78 residues could be made "equivalent", with an average C_α–C_α distance of 4.1 Å. The region of most striking agreement occurs between portions of the amino terminal halves of the two lysozymes (Figure 8), and includes part of the respective substrate binding sites. Using the sequence alignment proposed by Rossmann & Argos, the catalytically essential Glu 35 of HEW coincides with the "essential" Glu 11 of phage lysozyme (17, 123). On the other hand, the overall minimum base change per codon is 1.53, a value expected for random agreement.

The two "essential" acids described above occur within segments of 33 residues in the respective lysozymes, which can be well aligned without invoking any insertions or deletions. We have compared the structure and sequence of these respective

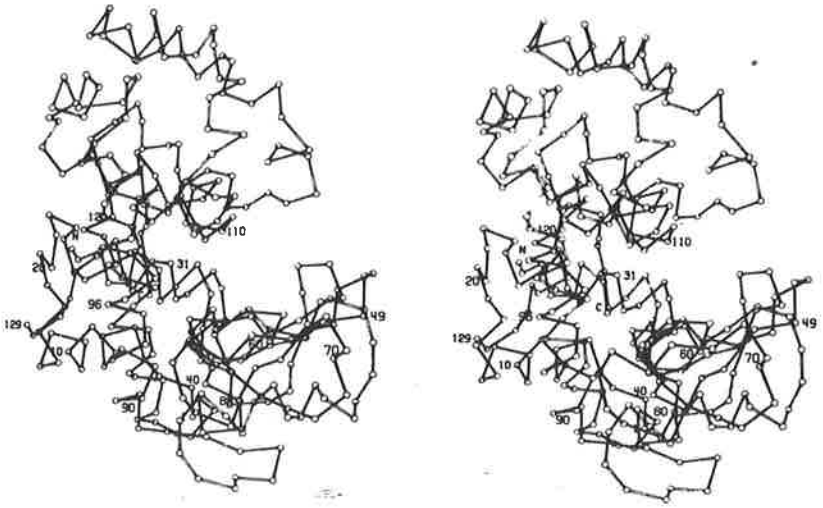


Figure 8 Superposition of the backbones of hen egg white lysozyme (open lines) and phage lysozyme (solid lines). [From Rossmann & Argos (218) with permission.]

segments of polypeptide in the two lysozymes by translating one segment relative to the other, in increments of a single residue, and optimizing, by rotation and translation, the fit between corresponding alpha-carbons. The results of this comparison, shown in Figure 9, reveal a well-defined minimum, corresponding to the "best" structural agreement. However, on this basis, Glu 11 of phage lysozyme is aligned with Ser 36 of HEW lysozyme and not with Glu 35. Furthermore, the minimum base change per codon does not have a minimum value for either the optimum structural alignment (Glu 11-Ser 36), or the optimal catalytic alignment (Glu 11-Glu 35). This test highlights both the strengths and the limitations of structural data in defining possible evolutionary relatedness. On the one hand the structural data clearly indicate that the two enzymes include a common "lysosyme fold," but the structural data do not permit the respective sequences to be aligned in an unequivocal manner, let alone allow one to use the sequence data to differentiate between convergent and divergent evolution for these two lysozymes.

PROTEIN FOLDING

The subject of protein folding is too broad to be more than touched upon here; reference can, however, be made to the recent reviews of Wu et al (213) and of Anfinsen & Scheraga (219).

Although it is now generally accepted that the structure of the native protein is the one with lowest free energy, and that protein folding is probably dominated by short-range interactions, and although coordinates for thirty or so proteins are now available (Table 1), the prediction of the three-dimensional structure of a

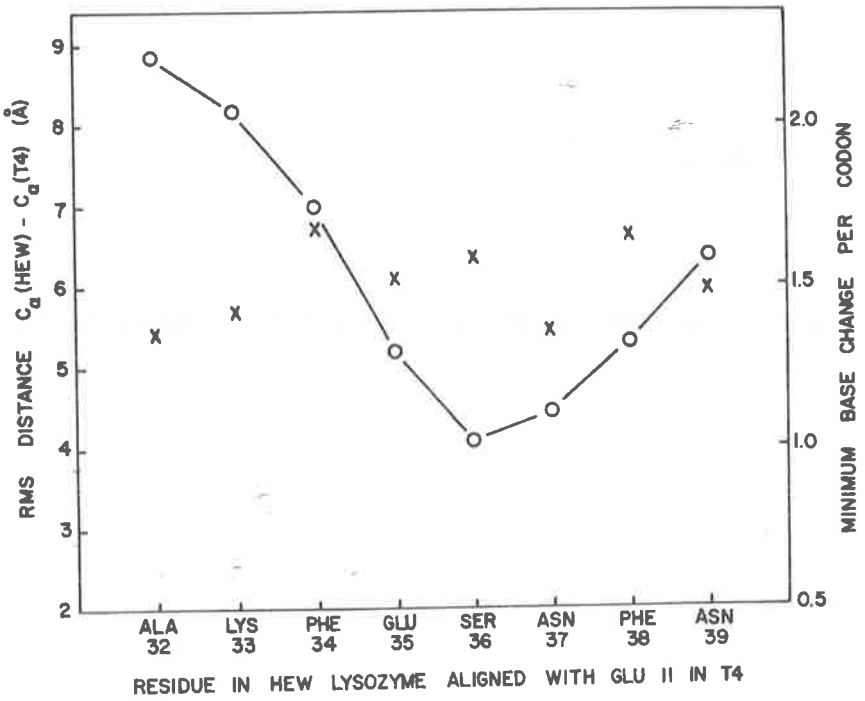


Figure 9 Structural and sequence comparison of a 33-residue segment of T4 phage lysozyme with successive overlapping 33-residue segments of hen egg white lysozyme. The abscissa gives the number of the residue in hen egg white lysozyme which is aligned with Glu 11 in phage lysozyme. For each alignment the rms distance between corresponding α-carbon atoms, after refinement, is indicated by a circle, and the minimum base change per codon indicated by a cross. (Unpublished results of S. J. Remington, W. F. Anderson, and B. W. Matthews).

protein, given its amino acid sequence, remains one of the major problems in physical chemistry.

Following the early observations that the helix content of a protein seemed to be correlated with its amino acid composition (220, 221), a number of investigators have attempted to use the known protein structures to develop empirical rules for the location of regular secondary structure, such as α-helices, β-sheet, and bends (222–248).

Recently, two tests were carried out in which the different empirical rules that have been proposed were used to predict the locations of secondary structure in adenylate kinase (249) and in the lysozyme from bacteriophage T4 (250). In each case the structure of the protein was not known to the investigators making the predictions, and, of course, the proteins were not included among the known structures from which the prediction schemes were devised.

The success of a given prediction can be measured in a number of ways, one of which is by calculation of the correlation between prediction and observation. The correlation coefficient C gives an immediate indication of how much better a given prediction is than a random one, $C = 1$ indicating perfect agreement, $C = 0$ indicating a prediction no better than a random one, and $C = -1$ indicating total disagreement between prediction and observation. The correlation coefficients for adenylate kinase and for phage lysozyme are given in Table 2.

Table 2 Correlation between predicted and observed secondary structure

Method	Adenylate kinase			T4 phage lysozyme		
	C_α	C_β	C_t	C_α	C_β	C_t
Barry and Friedman (249)	0.44	—	—	0.14	0.22	—
Burgess et al (240)	0.25	—	—	0.38	−0.08	0.19
Chou and Fasman (242, 243)	0.51	0.50	0.58	0.42	0.09	0.12
Finkelstein and co-workers (250)	—	—	—	0.27	−0.08	0.21
Guzzo (222)	—	—	—	0.34	—	—
Kotelchuck and Scheraga (229)	0.01	—	—	—	—	—
Leberman (233)	—	—	—	0.32	—	—
Lewis et al (230)	0.24	—	—	—	—	—
Lewis et al (234)	—	—	0.62	—	—	—
Lim (244–247)	0.46	0.58	—	0.21	−0.08	—
Nagano (238, 241)	0.47	0.41	0.58	—	—	—
Nagano and Hasegawa (238, 241)	—	—	—	0.20	0.04	0.01
Prothero (223)	—	—	—	0.19	—	—
Ptitsyn and Finkelstein (232)	0.61	0.58	—	0.23	−0.06	—
Robson and Pain (235)	0.35	—	—	—	—	—
Schellman (248)	—	—	—	0.41	0.34	—

As can be seen, the agreement between prediction and observation is moderately good in some instances, but poor in others. No predictive method is consistently superior to all the others. It is quite clear that even the most successful predictive methods leave a great deal to be desired, and can give very misleading results in some cases. Furthermore the predictive schemes based on many known protein structures are not dramatically more successful than those based on a few, and it cannot be anticipated that the accuracy of current predictive methods will be substantially increased simply by the inclusion of more data from additional protein structure determinations (250).

ACKNOWLEDGMENTS

I am most grateful to many colleagues for sending details of their work in advance of publication. Figures 3, 5, and 8 are reproduced with permission of Dr. M. G. Rossmann, Figure 4 with permission of Dr. C. I. Brändén, and Figures 7 and 8 with permission of Dr. D. R. Davies. Support during the writing of the review was

provided by NSF Grant BMS74-18407, NIH Grants GM20066, GM21967, GM15423, and GM70585 (Career Development Award, NIGMS), and the Alfred P. Sloan Foundation.

Literature Cited

1. Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R. 1960. *Nature* 185 : 422–27
2. Perutz, M. F., Rossmann, M. G., Cullis, A., Muirhead, H., Will, G., North, A. C. T. 1960. *Nature* 185 : 416–22
3. Blake, C. C. F. et al 1965. *Nature* 206 : 757–63
4. Kartha, G., Bello, J., Harker, D. 1967. *Nature* 213 : 862–65
5. Ludwig, M. L. et al 1967. *Proc. Natl. Acad. Sci. USA* 57 : 511–14
6. Matthews, B. W., Sigler, P. B., Henderson, R., Blow, D. M. 1967. *Nature* 214 : 652–56
7. Eisenberg, D. 1970. *Enzymes* 1 : 1–89
8. Dickerson, R. E. 1972. *Ann. Rev. Biochem.* 41 : 815–42
9. Blake, C. C. F. 1972. *Prog. Biophys.* 25 : 83–130
10. Vainshtein, B. K., Borisov, V. V. 1973. *Usp. Biol. Khim.* 14 : 91–145
11. Blundell, T. L., Johnson, L. N. 1973. *MTP Bienn. Rev. Sci. Phys. Chem. II, Chem. Crystallogr.*
12. Matthews, B. W., Bernhard, S. A. 1973. *Ann. Rev. Biophys. Bioeng.* 2 : 257–312
13. Matthews, B. W. *The Proteins*, ed. H. Neurath, R. C. Hill, Vol. III. New York: Academic. 3rd ed. In press
14. Kendrew, J. C. 1963. *Science* 139 : 1259–66
15. Watson, H. C. 1969. *Prog. Stereochem.* 4 : 299–333
16. Levitt, M., Lifson, S. 1969. *J. Mol. Biol.* 46 : 269–79
17. Imoto, T., Johnson, L. N., North, A. C. T., Phillips, D. C., Rupley, J. A. 1972. *Enzymes* 7 : 665–868
18. Diamond, R. 1974. *J. Mol. Biol.* 82 : 371–91
19. Levitt, M. 1974. *J. Mol. Biol.* 82 : 393–420
20. Quiocho, F. A., Lipscomb, W. N. 1971. *Adv. Protein Chem.* 25 : 1–78
21. Hartsuck, J. A., Lipscomb, W. N. 1971. *Enzymes* 3 : 1–56
22. Birktoft, J. J., Blow, D. M. 1972. *J. Mol. Biol.* 68 : 187–240
23. Wyckoff, H. W., Hardman, K. D., Allewell, N. M., Inagami, T., Johnson, L. N., Richards, F. M. 1967. *J. Biol. Chem.* 242 : 3984–88
24. Richards, F. M., Wyckoff, H. W. 1971. *Enzymes* 4 : 647–806
25. Perutz, M. F., Muirhead, H., Cox, J. M., Goaman, L. C. G., Mathews, F. S., McGandy, E. L., Webb, L. E., 1968. *Nature* 219 : 29–32
26. Perutz, M. F., Muirhead, H., Cox, J. M., Goaman, L. C. G. 1968. *Nature* 219 : 131–39
27. Perutz, M. F. 1972. *Nature* 237 : 495–99
28. Fermi, G. 1975. *J. Mol. Biol.* 97 : 237–56
29. Ten Eyck, L. F., Arnone, A. 1976. *J. Mol. Biol.* 100 : 3–11
30. Drenth, J., Jansonius, J. N., Koekoek, R., Swen, H. M., Wolthers, B. G. 1968. *Nature* 218 : 929–32
31. Drenth, J., Jansonius, J. N. Koekoek, R., Wolthers, B. G. 1971. *Adv. Protein Chem.* 25 : 79–115
32. Davies, D. R., Cohen, G. H., Silverton, E. W., Braxton, H. P., Matthews, B. W. 1969. *Acta Crystallogr. Sec. A* 25 : S182
33. Segal, D. M., Powers, J. C., Cohen, G. H., Davies, D. R., Wilcox, P. E. 1971. *Biochemistry* 10 : 3728–37
34. Huber, R., Epp, O., Formanek, H. 1969. *Naturwissenschaften* 56 : 362–67
35. Huber, R., Epp, O., Steigemann, W., Formanek, H. 1971. *Eur. J. Biochem.* 19 : 42–50
36. Adams, M. J., Blundell, T. L., Dodson, E. J., Dodson, G. G., Vijayan, M., Baker, E. N., Harding, M. M., Hodgkin, D. C., Rimmer, B., Sheat, S. 1969. *Nature* 224 : 491–95
37. Blundell, T. L., Dodson, G., Hodgkin, D., Mercola, D. 1972. *Adv. Protein Chem.* 26 : 279–402
38. Dodson, E. J., Isaacs, N. W., Rollett, J. S. 1976. *Acta Crystallogr. Sec. A* 32 : 311–15
39. Herriott, J. R., Sieker, L. C., Jensen, L. H. 1969. *Acta Crystallogr. Sec. A* 25 : S186
40. Herriott, J. R., Sieker, L. C., Jensen, L. H., Louenberg, W. 1970. *J. Mol. Biol.* 50 : 391–406
41. Watenpaugh, K. D., Sieker, L. C., Herriott, J. R., Jensen, L. H. 1973. *Acta Crystallogr. Sec. B* 29 : 943–56
42. Wright, C. S., Alden, R. A., Kraut, J. 1969. *Nature* 221 : 235–42
43. Paulos, T. L., Alden, R. A., Freer, S. T., Birktoft, J. J., Kraut, J. 1976. *J. Biol. Chem.* In press
44. Freer, S. T., Robertus, J. D., Wright, H. T., Xuong, Ng. H. 1970. *Biochemistry* 9 : 1997–2009

45. Shotton, D. M., Watson, H. C. 1970. *Philos. Trans. R. Soc. London Ser. B* 257 : 111–18
46. Bolton, W., Perutz, M. F. 1970. *Nature* 228 : 551–52
47. Muirhead, H., Greer, J. 1970. *Nature* 228 : 516–19
48. Adams, M. J., Ford, G. C., Koekoek, R., Lentz, P. J., McPherson, A., Rossmann, M. G., Smiley, I. E., Schevitz, R. W., Wonacott, A. J. 1970. *Nature* 227 : 1098–103
49. Holbrook, J. J., Liljas, A., Steindel, S. J., Rossmann, M. G. 1976. *Enzymes* 11 : 191–292
50. Huber, R., Kukla, D., Rühlmann, A., Epp, O., Formanek, H. 1970. *Naturwissenschaften* 57 : 389–92
51. Deisenhofer, J., Steigemann, W. 1975. *Acta Crystallogr. Sec. B* 31 : 238–50
52. Kannan, K. K. et al 1971. *Cold Spring Harbor Symp. Quant. Biol.* 36 : 221–31
53. Notstrand, B., Vaara, I., Kannan, K. K. 1975. *Isozymes I. Molecular Structure*, pp. 575–99. San Francisco: Academic
54. Mathews, F. S., Argos, P., Levine, M. 1971. See Ref. 52, pp. 387–95
55. Argos, P., Mathews, F. S. 1975. *J. Biol. Chem.* 250 : 747–51
56. Dickerson, R. E., Takano, T., Eisenberg, D., Kallai, O. B., Samson, L., Cooper, A., Margoliash, E. 1971. *J. Biol. Chem.* 246 : 1511–35
57. Dickerson, R. E., Timkovich, R. 1976. *Enzymes* 11 : 397–547
58. Takano, T., Kallai, O. B., Swanson, R., Dickerson, R. E. 1971. *J. Biol. Chem.* 248 : 5234–55
59. Ludwig, M. L., Anderson, R. D., Apgar, P. A., Burnett, R. M., LeQuesne, M. E., Mayhew, S. G. 1971. See Ref. 52, pp. 369–80
60. Mayhew, S. G., Ludwig, M. L. 1975. *Enzymes* 12 : 57–118
61. Love, W. E., Klock, P. A., Lattman, E. E., Padlan, E. A., Ward, K. B. Jr., Hendrickson, W. 1971. See Ref. 52, pp. 349–57
62. Padlan, E. A., Love, W. E. 1974. *J. Biol. Chem.* 249 : 4067–78
63. Hendrickson, W. A., Love, W. E. 1971. *Nature New Biol.* 232 : 197–203
64. Carter, C. W. Jr., Freer, S. T., Xuong, Ng. H., Alden, R. A., Kraut, J. 1971. See Ref. 52, pp. 381–85
65. Carter, C. W. Jr., Kraut, J., Freer, S. T., Xuong, N.-H., Alden, R. A., Bartsch, R. G. 1974. *J. Biol. Chem.* 249 : 4212–25
66. Freer, S. T., Alden, R. A., Carter, C. W. Jr., Kraut, J. 1975. *J. Biol. Chem.* 250 : 46–54
67. Kretsinger, R. H., Nockolds, C. E., Coffee, C. J., Bradshaw, R. A. 1971. See Ref. 52, pp. 217–20
68. Kretsinger, R. H., Nockolds, C. E. 1973. *J. Biol. Chem.* 248 : 3313–26
69. Moews, P. C., Kretsinger, R. H. 1975. *J. Mol. Biol.* 91 : 201–28
70. Arnone, A., Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. E. Jr., Richardson, D. C., Richarson, J. S., Yonath, A. 1971. *J. Biol. Chem.* 246 : 2302–16
71. Collins, D. M., Cotton, F. A., Hazen, E. E. Jr., Legg, M. J. 1976. *Proc. Harry Steenbock Symp., 4th*, ed. M. Sundaralingam, S. T. Rao. In press
72. Drenth, J., Hol, W. G. J., Jansonius, J. N., Koekoek, R. 1971. See Ref. 52, pp. 107–16
73. Stroud, R. M., Kay, L. M., Dickerson, R. E. 1971. See Ref. 52, pp. 125–40
74. Stroud, R. M., Kay, L. M., Dickerson, R. E. 1974. *J. Mol. Biol.* 83 : 185–208
75. Kreiger, M., Kay, L. M., Stroud, R. M. 1974. *J. Mol. Biol.* 83 : 209–30
76. Edelman, G. M., Cunningham, B. A., Reeke, G. N. Jr., Becker, J. W., Waxdal, M. J., Wang, J. L. 1972. *Proc. Natl. Acad. Sci. USA* 69 : 2580–84
77. Becker, J. W., Reeke, G. N. Jr., Wang, J. L., Cunningham, B. A., Edelman, G. M. 1975. *J. Biol. Chem.* 250 : 1513–24
78. Reeke, G. N. Jr., Becker, J. W., Edelman, G. M. 1975. *J. Biol. Chem.* 250 : 1525–47
79. Hardman, K. D., Ainsworth, C. F. 1972. *Biochemistry* 11 : 4910–19
80. Hardman, K. D. 1973. *Adv. Exp. Med. Biol.* 40 : 103–23
81. Kakudo, M., Ashida, T., Tanaka, N., Tsukihara, T., Yamane, T. 1972. *Acta Crystallogr. Sec. A* 28 : S34
82. Ashida, T., Tanaka, N., Yamane, T., Tsukihara, T., Kakudo, M. 1973. *J. Biochem. (Tokyo)* 73 : 463–65
83. Tanaka, T., Yamane, T., Tsukihara, T., Kakudo, M. 1975. *J. Biochem. (Tokyo)* 77 : 147–62
84. Sieker, L. C., Adman, E., Jensen, L. H. 1972. *Nature* 235 : 40–42
85. Adman, E. T., Sieker, L. C., Jensen, L. H. 1973. *J. Biol. Chem.* 248 : 3987–96
86. Jensen, L. H. 1974. *Ann. Rev. Biochem.* 43 : 461–74
87. Watenpaugh, K. D., Sieker, L. C., Jensen, L. H., Legall, J., Dubourdieu, M. 1972. *Proc. Natl. Acad. Sci. USA* 69 : 3185–88
88. Sakabe, N., Sakabe, K., Katayama, C. 1972. *Acta Crystallogr. Sec. A* 28 : S34
89. Hill, E., Tsernoglou, D., Webb, L., Banaszak, L. J. 1972. *J. Mol. Biol.* 72 : 577
90. Webb, L. E., Hill, E. J., Banaszak, L. J. 1973. *Biochemistry* 12 : 5101–9

91. Rossmann, M. G., Liljas, A., Brändén, C.-I., Banaszak, L. J. 1975. *Enzymes* 11 : 62–94
92. Carlisle, C. H., Gorinsky, B. A., Mazumdar, S. K., Palmer, R. A., Yeates, D. G. R. 1972. *Acta Crystallogr. Sec. A* 28 : S36
93. Carlisle, C. H., Palmer, R. A. Mazumdar, S. K., Gorinsky, B. A., Yeates, D. G. R. 1974. *J. Mol. Biol.* 85 : 1–18
94. Matthews, B. W., Jansonius, J. N., Colman, P. M., Schoenborn, B. P., Dupourque, D. 1972. *Nature* 238 : 37–41
95. Matthews, B. W., Weaver, L. H., Kester, W. R. 1974. *J. Biol. Chem.* 249 : 8030–44
96. Weaver, L. H., Kester, W. R., Ten Eyck, L. F., Matthews, B. W. 1976. *Experientia Supplementum.* 26 : 31–39
97. Brändén, C.-I., Eklund, H., Nordström, B., Boiwe, T., Söderlund, G., Zeppezauer, E., Ohlsson, I., Åkeson, Å. 1973. *Proc. Natl. Acad. Sci. USA* 70 : 2439–42
98. Eklund, H., Nordström, B., Zeppezauer, E. Söderlund, G., Ohlsson, I., Boiwe, T., Söderberg, B.-O., Tapia, O., Brändén, C.-I., Åkeson, Å. 1976. *J. Mol. Biol.* 102 : 27–59
99. Schiffer, M., Girling, R. L., Ely, K. R., Edmundson, A. B. 1973. *Biochemistry* 12 : 4620–31
100. Edmundson, A. B., Ely, K. R., Abola, E. E., Schiffer, M., Panagiotopoulos, N. 1975. *Biochemistry* **14** : 3953–61
101. Tulinsky, A., Vandlen, R. L., Morimoto, C. N., Mani, N. V., Wright, L. H. 1973. *Biochemistry* 12 : 4185–92
102. Mauridis, A., Tulinsky, A., Liebman, M. N. 1974. *Biochemistry* 13 : 3661–66
103. Takano, T., Kallai, O. B., Swanson, R., Dickerson, R. E. 1973. *J. Biol. Chem.* 248 : 5234–55
104. Salemme, F. R., Freer, S. T., Xuong, Ng. H., Alden, R. A., Kraut, J. 1973. *J. Biol. Chem.* 248 : 3910–21
105. Timkovich, R., Dickerson, R. E. 1973. *J. Mol. Biol.* 79 : 39–56
106. Poljak, R. J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizackerley, R. P., Saul, F. 1973. *Proc. Natl. Acad Sci. USA* 70 : 3305–10
107. Poljak, R. J. 1975. *Adv. Immunol.* 21 : 1–33
108. Buehner, M., Ford, G. C., Moras, D., Olsen, K. W., Rossmann, M. G. 1973. *Proc. Natl. Acad. Sci. USA* 70 : 3052–54
109. Moras, D., Olsen, K. W., Sabesan, M. N., Buehner, M., Ford, G. C., Rossmann, M. G. 1976. *J. Biol. Chem.* 250 : 9137–62
110. Adams, M. J., Buehner, M., Chandrasekhar, K., Ford, G. C., Hackert, M. L., Liljas, A., Rossmann, M. G., Smiley, I. E., Allison, W. S., Everse, J., Kaplan, N. O., Taylor, S. S. 1973. *Proc. Natl. Acad. Sci. USA* 70 : 1968–72
111. White, J. L., Hackert, M. L., Buehner, M., Adams, M. J., Ford, G. C., Lentz, P. J., Jr., Smiley, I. E., Steundel, S. J., Rossmann, M. G. 1976. *J. Mol. Biol.* 102 : 759–79
112. Rühlmann, A., Kukla, D., Schwager, P., Bartels, K., Huber, R. 1973. *J. Mol. Biol.* 77 : 417–36
113. Huber, R., Kukla, D., Bode, W., Schwager, P., Bartels, K., Deisenhofer, J., Steigemann, W. 1974. *J. Mol. Biol.* 89 : 73–101
114. Huber, R., Bode, W., Kukla, D., Kohl, V., Ryan, C. A. 1975. *Biophys. Struct. Mech.* 1 : 189–201
115. Schulz, G. E., Elzinga, M., Marx, F., Schirmer, R. H. 1974. *Nature* 250 : 120–23
116. Epp, O. et al 1974. *Eur. J. Biochem.* 45 : 513–24
117. Epp, O., Lattman, E. E., Schiffer, M., Huber, R., Palm, W. 1975. *Biochemistry* 14 : 4943–52
118. Segal, D. M. et al 1974. *Proc. Natl. Acad. Sci. USA* 71 : 4298–302
119. Davies, D. R., Padlan, E. A., Segal, D. M. 1975. *Ann. Rev. Biochemistry* 44 : 639–67
120. Burnett, R. M. et al 1974. *J. Biol. Chem.* 249 : 4383–92
121. Peking Insulin Structure Research Group. 1974. *Sci. Sin.* 17 : 752–78
122. Banyard, S. H., Blake, C. C. F., Swan, I. D. A. 1974. *Lysozyme*, ed. E. F. Osserman, R. E. Canfield, S. Beychok, pp. 71–79. Oxford: Pergamon
123. Matthews, B. W., Remington, S. J. 1974. *Proc. Natl. Acad. Sci. USA* 71 : 4178–82
124. Blake, C. C. F., Evans, P. R. 1974. *J. Mol. Biol.* 84 : 585–601
125. Bryant, T. N., Watson, H. C., Wendell, P. L. 1974. *Nature* 247 : 14–17
126. Campbell, J. W., Watson, H. C., Hodgson, G. I. 1974. *Nature* 250 : 301–3
127. Blake, C. C. F., Geisow, M. J., Swan, I. D. A., Rerat, C., Rerat, B. 1974. *J. Mol. Biol.* 88 : 1–12
128. Smit, J. D. G., Ploegman, J. H., Pierrot, M., Kalk, K. H., Jansonius, J. N., Drenth, J. 1974. *Israel J. Chem.* 12 : 287–304
129. Bergsma, J., Hol, W. G. J., Jansonius, J. N., Kalk, K. H., Ploegman, J. H., Smit, J. D. G. 1975. *J. Mol. Biol.*

98 : 637–43
130. Blow, D. M., Janin, J., Sweet, R. M. 1974. *Nature* 249 : 54–57
131. Sweet, R. M., Wright, H. T., Janin, J., Chothia, C. H., Blow, D. M. 1974. *Biochemistry* 13 : 4212–28
132. Fenna, R. E., Matthews, B. W. 1975. *Nature* 258 : 573–77
133. Fehlhammer, H., Schiffer, M., Epp, O., Colman, P. M., Lattman, E. E., Schwager, P., Steigemann, W., Schramm, H. J. 1975. *Biophys. Struct. Mech.* 1 : 139–146
134. Wang, B. C., Yoo, C. S., Sax, M. 1975. *Acta Crystallogr. Sec. A* 31 : S30
135. Kannan, K. K. et al 1975. *Proc. Natl. Acad. Sci. USA* 72 : 51–55
136. Schmid, M. F., Herriott, J. R. 1975. *Acta Crystallogr. Sec. A* 31 : S30
137. Schmid, M. F., Herriott, J. R. 1976. *J. Mol. Biol.* 103 : 175–90
138. Sasaki, K., Dockerill, S., Adamiak, D. A., Tickle, I. J., Blundell, T. 1975. *Nature* 257 : 751–57
139. Fletterick, R. J., Bates, D. J., Steitz, T. A. 1975. *Proc. Natl. Acad. Sci. USA* 72 : 38–42
140. Steitz, T. A., Fletterick, R. J., Anderson, W. F., Anderson, C. 1976. *J. Mol. Biol.* 104 : 197–222
141. Kurachi, K., Sieker, L. C., Jensen, L. H. 1975. *J. Biol. Chem.* 250 : 7663–67
142. Andreeva, N. S. *Abst. Soviet–France Symp. Protein Struct., Pouschino USSR Sept. 1975*
143. Drenth, J., Enzing, C., Kalk, K. H., Vessies, J. 1975. *Acta Crystallogr. A* 31. S36
144. Delbaere, L. T. J., Hutcheon, W. L. B., James, M. N. G., Thiessen, W. E. 1975. *Nature* 257 : 758–63
145. Richardson, J. S., Thomas, K. A., Rubin, B. H., Richardson, D. C. 1975. *Proc. Natl. Acad. Sci. USA* 72 : 1349–53
146. Holmgren, A., Södeberg, B.-O., Eklund, H., Brändén, C.-I. 1975. *Proc. Natl. Acad. Sci. USA* 72 : 2305–9
147. Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D., Waley, S. G. 1975. *Nature* 255 : 609–14
148. Bode, W., Schwager, P. 1975. *FEBS Lett.* 56 : 139–43
149. Fehlhammer, H., Bode, W. 1976. *J. Mol. Biol.* 98 : 683–92
150. Bode, W., Schwager, P. 1976. *J. Mol. Biol.* 98 : 693–717
151. Green, D. W., Ingram, V. M., Perutz, M. F. 1954. *Proc. R. Soc. London Ser. A* 225 : 287–307
152. Tollin, P. 1969. *J. Mol. Biol.* 45 : 481–90
153. Rossman, M. G., ed. *The Molecular Replacement Method.* 1972. New York: Gordon & Breach
154. Milledge, H. J. 1963. *Acta Crystallogr.* 16 : 72
155. Xuong, Ng.-H., Kraut, J., Seely, O., Freer, S. T., Wright, C. 1968. *Acta Crystallogr. Sec. B* 24 : 289–90
156. Arndt, U. W., Wonacott, A. J., eds. 1976. *The Rotation Method in Crystallography.* ASP Biol. Med. Press B. V. In press
157. Arndt, U. W., Gilmore, D. J. 1975. *Adv. Electron. Electron Physics.* In press
158. Minor, T. C., Milch, J. R., Reynolds, G. T. 1974. *J. Appl. Crystallogr.* 7 : 323–30
159. Hashizume, H., Kohra, K., Oguro, Y. 1975. *J. Appl. Crystallogr.* 8 : 249–54
160. Cork, C., Fehr, D., Hamlin, R., Vernon, W., Xuong, N. H., Perez-Mendez, V. 1973. *J. Appl. Crystallogr.* 7 : 319–23
161. Rosenbaum, G., Holmes, K. C., Witz, J. 1971. *Nature* 230 : 434–37
162. Holmes, K. C. 1974. *Endeavour* 23 : 50–66
163. Low, B. W., Chen, C. C. H., Berger, J. E., Singman, L., Pletcher, J. F. 1966. *Proc. Natl. Acad. Sci. USA* 56 : 1746–50
164. Haas, D. J. 1968. *Acta Crystallogr. Sec. B* 24 : 604
165. Haas, D. J., Rossmann, M. G. 1970. *Acta Crystallogr. Sec. B* 26 : 998–1004
166. Petsko, G. A. 1975. *J. Mol. Biol.* 96 : 381–92
167. Dozou, P. 1973. *Mol. Cell Biochem.* 1 : 15–27
168. Hendrickson, W. A., Klippenstein, G. L., Ward, K. B. 1975. *Proc. Natl. Acad. Sci. USA* 72 : 2160–64
169. Stenkamp, R. E., Sieker, L. C., Jensen, L. H., Loehr, J. S. 1976. *J. Mol. Biol.* In press
170. Jensen, L. H. 1974. *Ann. Rev. Biophys. Bioeng.* 3 : 81–93
171. Diamond, R. 1966. *Acta Crystallogr.* 21 : 253–66
172. Warme, P. K., Gō, N., Scheraga, H. A. 1972. *J. Comp. Phys.* 9 : 303–17
173. Hermans, J., Jr., McQueen, J. E., Jr. 1974. *Acta Crystallogr. Sec. A* 30 : 730–39
174. Dodson, E. J., Isaacs, N. W., Rollett, J. S. 1976. *Acta Crystallogr.* A32 : 311–15
175. Ten Eyck, L. F., Weaver, L. H., Matthews, B. W. 1976. *Acta Crystallogr.* A32 : 349–50
176. Sayre, D. 1972. *Acta Crystallogr. Sec. A* 28 : 210–12
177. Sayre, D. 1974. *Acta Crystallogr. Sec. A* 30 : 180–84

178. Barrett, A. N., Zwick, M. 1971. *Acta Crystallogr.* 27 : 6–11
179. Collins, D. M., Cotton, F. A., Hazen, E. E. Jr., Meyer, E. F. Jr., Morimoto, C. N. 1975. *Science* 190 : 1047–53
180. Hoppe, W., Gassmann, J. 1968. *Acta Crystallogr. Sec. B* 24 : 97–107
181. Adman, E. T., Watenpaugh, K. D., Sieker, L. C., Jensen, L. H. 1975. *Science*, Letters to the Editor, 14 March
182. Warme, P. K., Scheraga, H. A. 1975. *Biochemistry* 14 : 3509–17
183. Matthews, B. W. 1968. *J. Mol. Biol.* 33 : 491–97
184. Perutz, M. F. 1965. *J. Mol. Biol.* 13 : 646–68
185. Robertus, J. D., Alden, R. A., Kraut, J. 1971. *Biochem. Biophys. Res. Commun.* 42 : 334–7
186. Doscher, M., Richards, F. M. 1963. *J. Biol. Chem.* 238 : 2399–406
187. Bello, J., Nowoswiat, E. F. 1965. *Biochim. Biophys. Acta.* 105 : 325–32
188. Quiocho, F. A., Richards, F. M. 1966. *Biochemistry* 5 : 4062–76
189. Sigler, P. B., Jeffery, B. A., Matthews, B. W., Blow, D. M. 1966. *J. Mol. Biol.* 15 : 175–92
190. Johansen, J. T., Vallee, B. L. 1971. *Proc. Natl. Acad. Sci. USA* 68 : 2532–35
191. Johansen, J. T., Livingston, D. M., Vallee, B. L. 1972. *Biochemistry* 11 : 2585–88
192. Quiocho, F. A., McMurray, C. H., Lipscomb, W. N. 1972. *Proc. Natl. Acad. Sci. USA* 69 : 2850–54
193. Lipscomb, W. N. 1973. *Proc. Natl. Acad. Sci. USA* 70 : 3797–801
194. Johansen, J. T., Vallee, B. L. 1975. *Biochemistry* 14 : 649–60
195. Becker, J. W., Reeke, G. N. Jr., Edelman, G. M. 1971. *J. Biol. Chem.* 246 : 6123–25
196. Hardman, K. D., Ainsworth, C. F. 1972. *Nature New Biol.* 237 : 54–55
197. Hardman, K. D., Ainsworth, C. F. 1973. *Biochemistry* 12 : 4442–48
198. Kalb, A. J., Levitzki, A. 1968. *Biochem. J.* 109 : 669–72
199. Brewer, C. F., Marcus, D. M., Grollman, A. P., Sternlicht, H. 1974. *J. Biol. Chem.* 249 : 4614–16
200. Hardman, K. D., Ainsworth, C. F. 1976. *Biochemistry* 15 : 1120–28
201. Becker, J. W., Reeke, G. N. Jr., Cunningham, B. A., Edelman, G. M. 1976. *Nature* 259 : 406 9
202. Scouloudi, H. 1969. *J. Mol. Biol.* 40 : 353–77
203. Lattman, E. E., Nockolds, C. E., Kretsinger, R. H., Love, W. E. 1971. *J. Mol. Biol.* 60 : 271–77
204. Vainshtein, B. K., Arutyunyan, E. H., Kuranova, I. P., Borisov, V. V., Sosfenov, N. I., Pavlovsky, A. G., Grebendo, A. I., Konareva, N. V. 1975. *Nature* 254 : 163–64
205. Cohen, G. H., Matthews, B. W., Davies, D. R. 1970. *Acta Crystallogr. B* 26 : 1062–69
206. Robertus, J. D., Alden, R. A., Birktoft, J. J., Kraut, J., Powers, J. C., Wilcox, P. E. 1972. *Biochemistry* 11 : 2439–49
207. Rao, S. T., Rossmann, M. G. 1973. *J. Mol. Biol.* 76 : 241–56
208. Ohlsson, I., Nordström, B., Brändén, C.-I. 1974. *J. Mol. Biol.* 89 : 339–54
209. Rossmann, M. G., Moras, D., Olsen, K. W. 1974. *Nature* 250 : 194–99
210. Eventoff, W., Rossmann, M. G. 1975. *Chemical Rubber Company Critical Reviews in Biochemistry* ed. G. Fasman, pp. 111–40. Cleveland: Chemical Rubber Co. Press
211. Schulz, G. E., Schirmer, R. H. 1974. *Nature* 250 : 142–44
212. Margoliash, E., Fitch, W. M., Dickerson, R. E. 1971. *Biochemical Evolution and the Origin of Life*, ed. E. Schoffeniels, pp. 52–95. Amsterdam: North Holland
213. Wu, T. T., Fitch, W. M., Margoliash, E. 1974. *Ann. Rev. Biochem.* 43 : 539–66
214. Blake, C. C. F. 1974. *Nature* 250 : 284–85
215. Davies, D. R., Padlan, E. A., Segal, D. M. 1975. *Contemporary Topics in Molecular Immunology*, ed. F. P. Inman, W. J. Mandy, 4 : 127–55. New York: Plenum
216. Padlan, E. A., Davies, D. R. 1975. *Proc. Natl. Acad. Sci. USA* 72 : 819–23
217. Richardson, J. S., Richardson, D. C., Thomas, K. A., Silverton, E. W., Davies, D. R. 1976. *J. Mol. Biol.* In press
218. Rossmann, M. G., Argos, P. 1976. *J. Mol. Biol.* In press
219. Anfinsen, C. B., Scheraga, H. A. 1975. *Adv. Protein Chem.* 29 : 205–300
220. Szent-Györgyi, A. G., Cohen, C. 1957. *Science* 126 : 697
221. Davies, D. R. 1964. *J. Mol. Biol.* 9 : 605–9
222. Guzzo, A. V. 1965. *Biophys. J.* 5 : 809–22
223. Prothero, J. W. 1966. *Biophys. J.* 6 : 367–70
224. Cook, D. A. 1967. *J. Mol. Biol.* 29 : 167–71
225. Periti, P. F., Quagliarotti, G., Liquori, A. M. 1967. *J. Mol. Biol.* 24 : 313–22

226. Schiffer, M., Edmundson, A. B. 1968. *Biophys. J.* 8 : 29–39
227. Dunnill, P. 1968. *Biophys. J.* 8 : 865–75
228. Low, B. W., Lovell, F. M., Rudko, A. D. 1968. *Proc. Natl. Acad. Sci. USA* 60 : 1519–26
229. Kotelchuck, D., Scheraga, H. A. 1969. *Proc. Natl. Acad. Sci. USA* 62 : 14–21
230. Lewis, P. N., Gō, N., Gō, M., Kotelchuck, D., Scheraga, H. A. 1970. *Proc. Natl. Acad. Sci. USA* 65 : 810–15
231. Ptitsyn, O. B., Finkelstein, A. V. 1970. *Dokl. Akad. Nauk. SSSR* 195 : 221–24
232. Ptitsyn, O. B., Finkelstein, A. V. 1970. *Biofizika USSR* 15 : 757–67
233. Leberman, R. 1971. *J. Mol. Biol.* 55 : 23–30
234. Lewis, P. N., Momany, F. A., Scheraga, H. A. 1971. *Proc. Natl. Acad. Sci. USA* 68 : 2293–97
235. Robson, B., Pain, R. H. 1971. *J. Mol. Biol.* 58 : 237–59
236. Wu, T. T., Kabat, E. A. 1971. *Proc. Natl. Acad. Sci. USA* 68 : 1501–6
237. Finkelstein, A. V., Ptitsyn, O. B. 1971. *J. Mol. Biol.* 52 : 613–24
238. Nagano, K. 1973. *J. Mol. Biol.* 75 : 401–21
239. Kabat, E. A., Wu, T. T. 1973. *Proc. Natl. Acad. Sci. USA* 70 : 1473–77
240. Burgess, A. W., Ponnuswamy, P. K., Scheraga, H. A. 1974. *Isr. J. Chem.* 12 : 483–94
241. Nagano, K. 1974. *J. Mol. Biol.* 84 : 337–72
242. Chou, P. Y., Fasman, G. D. 1974. *Biochemistry* 13 : 211–22
243. Chou, P. Y., Fasman, G. D. 1974. *Biochemistry* 13 : 222–45
244. Lim, V. I. 1974. *Biofizika* 19 : 366–78
245. Ibid, pp. 562–75
246. Lim, V. I. 1974. *J. Mol. Biol.* 88 : 857–72
247. Ibid, pp. 873–94
248. Schellman, C. G. 1975. In *Chemical Rubber Company Critical Reviews in Biochemistry*, ed. G. D. Fasman. Cleveland: Chemical Rubber Co. Press
249. Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, O. B., Kabat, E. A., Wu, T. T., Levitt, M., Robson, B., Nugano, K. 1974. *Nature* 250 : 140–42
250. Matthews, B. W. 1975. *Biochim. Biophys. Acta* 405 : 442–51

REPRINTED FROM
THE PROTEINS, 3rd Edition, VOL. III
© 1977
ACADEMIC PRESS, INC.
NEW YORK SAN FRANCISCO LONDON

4

X-Ray Structure of Proteins

B. W. MATTHEWS

I. Introduction 404
A. Historical Résumé 404
B. Characteristics of the X-Ray Technique 404
C. Scope of This Review 410
II. X-Ray Crystallography 411
A. Introduction 411
B. Structure Analysis of Single Crystals 411
C. Isomorphous Replacement and Anomalous Scattering . . 417
D. The X-Ray Image 445
E. Comparison of Different Structures 456
F. Novel Methods for Structure Determination 461
G. Neutron Diffraction 466
III. Protein Crystals 468
A. Solvent Content 468
B. Crystallization 477
C. Preliminary Information 479
D. Structure in the Crystal and in Solution 496
E. Heavy Atom Derivatives 499
F. Data Collection 502
IV. Protein Structures 509
A. Heme Proteins 509
B. Nonheme Electron Transport Proteins 515
C. Proteolytic Enzymes 518
D. Carbonic Anhydrase 536
E. Insulin 537
F. Lysozyme 538
G. Nucleases 539
H. Dehydrogenases 543
I. Calcium-Binding Carp Myogen 548
J. Concanavalin A 549
K. Other Proteins 553

V. Conclusions 553
A. Protein Conformation 553
B. Protein Evolution 568
C. The Future. 570
References 572

I. INTRODUCTION

A. Historical Résumé

As long ago as 1934, Bernal and Crowfoot (1934), having taken the first X-ray diffraction photographs of wet pepsin crystals, immediately realized the potential significance of their findings—that "the arrangement of atoms inside the protein molecule is of a perfectly definite kind" and that at least in principle the X-ray technique could be used to determine that atomic arrangement. Furthermore, Bernal (1939) very quickly pointed out that a possible method of solving protein structures might be "by some physical artifice, such as the introduction of a heavy atom."

As is now well known, it took nearly thirty years of effort to bring these predictions to fruition with the determination of the structure of myoglobin at essentially atomic resolution (Kendrew *et al.*, 1960) and of hemoglobin at 5.5 Å resolution (Perutz *et al.*, 1960). [Fascinating accounts of the early days of protein crystallography have been given by Bragg (1965), Hodgkin and Riley (1968), and Perutz (1970a).] Five years later the atomic structure of lysozyme was determined by Phillips and colleagues (Blake *et al.*, 1965), followed in 1967 by reports of the structures of ribonuclease A (Kartha *et al.*, 1967), α-chymotrypsin (Matthews *et al.*, 1967), ribonuclease S (Wyckoff *et al.*, 1967b), and carboxypeptidase A (Ludwig *et al.*, 1967). Since 1967 "new structures" have appeared in the literature with increasing frequency, as summarized in Table I. It is encouraging to note that not only is the number of new structures per year increasing, but also that the list includes larger and technically more demanding proteins. Furthermore, the average time per structure determination is decreasing, as is the amount of protein required.

B. Characteristics of the X-Ray Technique

The first requirement for an X-ray structure analysis is that the protein be in the form of a crystal. It follows that the structure determined

will be that of the macromolecule in the crystal. Often the conditions necessary to crystallize proteins, e.g., the use of concentrated salt solutions or organic solvents, are far from the normal physiological environment of the protein; also the pH may be very different. In Section III,D we will discuss an extensive body of evidence suggesting that the crystallization conditions notwithstanding, in a number of specific instances the crystal structure is, by several criteria, very similar to the structure in solution. Nevertheless, this is a limitation of the X-ray technique which must be considered for each individual case.

One of the most characteristic features of the X-ray method is that it provides information about the static or, more strictly, the time- and space-averaged structure, the time being typically measured in hours, and the space enclosing thousands of molecules. It may be possible to determine the structure of a protein before or after it changes conformation, or perhaps to chemically "freeze" it in some stable intermediate state, but it is not possible with X-rays to directly observe transient intermediate states or to determine rates of conformational change.

Typically, crystals of a protein suitable for X-ray crystallography take at least a week to grow, often much longer, and rarely less than a day. Therefore, crystallization of unstable proteins or protein–ligand complexes may not be possible. On the other hand, protein crystals are very "open" since they contain channels through which solvents and small molecules may diffuse freely (see Section III,A). In many cases it has been possible to carry out chemical reactions within the pregrown crystals, and so to determine the structure of an enzyme–ligand complex with a half-life measured in hours. New data acquisition techniques, using area detectors currently being developed, may reduce this time by at least an order of magnitude.

We conclude this section by commenting on the accuracy of the X-ray technique. In favorable cases it may be possible to determine the coordinates of every nonhydrogen atom in a protein crystal, including bound solvent molecules, insofar as their position is essentially static. However, in contrast to X-ray diffraction studies of crystals of "small molecules," in which postional uncertainties of a few thousandths of an angstrom are commonplace, the accuracy of the coordinates of a protein will at best be measured in tenths of angstroms, and for some parts of the structure may be an order of magnitude larger. This inherent limitation arises from the fact that the molecules in a protein crystal are not exactly aligned but have an intrinsic disordering which is much greater than, for example, in a crystal of urea. The degree of disordering varies from protein to protein and from one crystal form to another. It may be influenced by the purity

TABLE I

Determination of Protein Structures[a]

Year	Protein	Molecular weight[b]	Nominal resolution of first interpretable map (Å)	Solvent content parameter[c] V_M (Å³/dalton)	Reference
1960	Myoglobin	17,800	2.0	1.98	Kendrew *et al.* (1960)
1965	Lysozyme	14,600	2.0	2.03	Blake *et al.* (1965)
1967	Carboxypeptidase A	34,600	2.8	2.12	Ludwig *et al.* (1967)
	α-Chymotrypsin	25,000 (2)	2.0	2.12	Matthews *et al.* (1967)
	Ribonuclease A	13,700	2.0	2.15	Kartha *et al.* (1967)
	Ribonuclease S	13,700	3.5	2.81	Wyckoff *et al.* (1967b)
1968	Hemoglobin, horse oxy	66,500 ($\frac{1}{2}$)	2.8	2.61	Perutz *et al.* (1968a,b)
	Papain	23,000	2.8	2.59	Drenth *et al.* (1968)
1969	γ-Chymotrypsin	25,000	2.7	2.35	Davies *et al.* (1969)
	Hemoglobin, chironomous	17,000	2.8	1.90	Huber *et al.* (1969)
	Insulin	5,800 (2)	2.8	1.96	Adams *et al.* (1969a)
	Rubredoxin	6,000	3.0	2.18	Herriott *et al.* (1969)
	Subtilisin BPN′	27,500	2.5	2.15	Wright *et al.* (1969)
1970	Chymotrypsinogen	25,000	2.5	2.56	Freer *et al.* (1970)
	Elastase	25,900	3.5	2.46	Shotton and Watson (1970)
	Hemoglobin, horse deoxy	66,500 ($\frac{1}{2}$)	2.8	2.26	Bolton and Perutz (1970)
	Hemoglobin, human deoxy	66,500	3.5	2.07	Muirhead and Greer (1970)
	Lactate dehydrogenase	140,000 ($\frac{1}{4}$)	2.8	2.97	Adams *et al.* (1970a)
	Trypsin inhibitor	6,500	2.5	1.85	Huber *et al.* (1970)

1971	Carbonic anhydrase	30,000	2.0	1.86	Kannan *et al.* (1971)
	Cytochrome b_5	11,000	2.8	2.02	Mathews *et al.* (1971b)
	Cytochrome *c*, horse ferri	12,500	2.8	2.84	Dickerson *et al.* (1971)
	Cytochrome *c*, tuna ferro	12,500	2.48	2.26	Takano *et al.* (1971)
	Flavodoxin	16,000	3.25	2.60	Ludwig *et al.* (1971)
	Hemoglobin, glycera	18,200	2.5	2.14	Love *et al.* (1971)
	Hemoglobin, lamprey	18,000	2.0	1.86	Hendrickson and Love (1971)
	High potential iron protein	9,650	2.25	1.75	Carter *et al.* (1971)
	Myogen, carp	11,000	2.0	2.17	Kretsinger *et al.* (1971b)
	Nuclease, staphylococcal	16,800	2.0	2.19	Arnone *et al.* (1971)
	Subtilisin Novo	27,500	2.8	2.06	Drenth *et al.* (1971a)
	Trypsin	24,000	2.7	2.27	Stroud *et al.* (1971)
1972	Concanavalin A	108,000 ($\frac{1}{4}$)	2.0	2.26	Edelman *et al.* (1972)
	Concanavalin A	108,000 ($\frac{1}{4}$)	2.4	2.26	Hardman and Ainsworth (1972)
	Cytochrome *c*, bonito ferro	12,500	2.3	1.85	Kakudo *et al.* (1972)
	Ferredoxin	6,000	2.5	1.90	Sieker *et al.* (1972)
	Flavodoxin	16,000	2.5	2.9	Watenpaugh *et al.* (1972)
	Insulin	5,800 (2)	3.1	1.96	Sakabe *et al.* (1972)
	Malate dehydrogenase	72,000	3.0	2.48	Hill *et al.* (1972)
	Ribonuclease A	13,700	2.5	2.15	Carlisle *et al.* (1972)
	Thermolysin	34,600	2.3	2.43	Matthews *et al.* (1972a,b)

(*Continued*)

TABLE I (*Continued*)

Year	Protein	Molecular weight[b]	Nominal resolution of first interpretable map (Å)	Solvent content parameter[c] V_M (Å³/dalton)	Reference
1973	Alcohol dehydrogenase	80,000 ($\frac{1}{2}$)	2.9	2.39	Brändén *et al.* (1973)
	Bence Jones protein	23,000 (2)	3.5	3.05	Schiffer *et al.* (1973)
	α-Chymotrypsin	25,000 (2)	2.8	2.12	Tulinsky *et al.* (1973)
	Cytochrome *c*, tuna reduced	12,020	2.45	2.33	Takano *et al.* (1973)
	Cytochrome c_2	12,480	2.0	2.05	Salemme *et al.* (1973)
	Cytochrome c_{550}	14,000	4.0	2.0	Timkovitch and Dickerson (1973)
	Fab fragment	55,000	2.8	2.29	Poljak *et al.* (1973)
	Glyceraldehyde-3-phosphate dehydrogenase	143,000	3.0	2.92	Buehner *et al.* (1973)
	Lactate dehydrogenase, ternary complex	140,000	3.0	2.78	Adams *et al.* (1973a)
	Trypsin-pancreatic trypsin inhibitor	30,000	2.8	3.2	Rühlmann *et al.* (1973)
1974	Adenyl kinase	22,000	3.0	2.18	Schulz *et al.* (1974a)
	Bence Jones protein	11,000 (2)	2.8	3.7	Epp *et al.* (1974)
	Fab fragment	55,000	3.1	4.2	Segal *et al.* (1974)
	Flavodoxin, oxidized	16,000	1.9	2.41	Burnett *et al.* (1974)

	Insulin	5,800 (2)	1.8[d]	1.96	Peking Insulin Structure Group (1974)
	Lysozyme, human	14,500	2.5	1.98	Banyard *et al.* (1974)
	Lysozyme, phage T4	18,700	2.5	2.77	Matthews and Remington (1974)
	Phosphoglycerate kinase, horse	38,000	3.0	2.56	Blake and Evans (1974)
	Phosphoglycerate kinase, yeast	45,000	3.5	2.56	Bryant *et al.* (1974)
	Phosphoglycerate mutase	110,700 ($\frac{1}{2}$)	3.5	2.66	Campbell *et al.* (1974)
	Prealbumin	54,000 ($\frac{1}{2}$)	2.5	2.23	Blake *et al.* (1974)
	Rhodanese	37,000	3.9	2.16	Smit *et al.* (1974)
	Trypsin-soybean trypsin inhibitor	43,500	2.6	3.17	Blow *et al.* (1974)
1975[e]	Carbonic anhydrase B	30,000	2.2	1.86	Kannan *et al.* (1975)
	Hexokinase	51,000	2.7	2.83	Fletterick *et al.* (1975)
	Superoxide dimutase	16,000 (2)	3.0	2.21	Richardson *et al.* (1975)
	Triosephosphate isomerase	49,000	2.5	2.48	Banner *et al.* (1975)

[a] This compilation gives only the reference to the original structure determination. Citations of more recent work are included in the text.
[b] The number of molecular weight units in the crystallographic asymmetric unit, if different from unity, is indicated in parentheses.
[c] See Section III,A.
[d] A 2.5 Å resolution electron density map was obtained in 1971.
[e] Incomplete listing. Additional references are given by Matthews (1976).

(both chemical and conformational) of the protein and by its molecular weight (large proteins tending to be more disordered), but seems to be mainly dependent upon the nature of the contacts which an individual molecule makes with its neighbors in the crystal. Crystals with a relatively high solvent content, and therefore presumably fewer intermolecular contacts, tend to give diffraction patterns indicative of greater random disorder than tightly packed crystals. Whatever the reason for its occurrence, the degree of disorder will determine the ultimate accuracy or resolution of the structure determination.

Sometimes a crystal will diffract to "high resolution" but for various reasons will be studied at "low resolution," often as a stepping-stone on the way to an ultimate high resolution analysis. We will return to the question of resolution later, but it is important at the outset to realize that a low resolution (e.g., 5 Å) crystallographic study cannot give the same detail as is potentially available in a higher resolution (e.g., 2–3 Å) analysis. It might also be added that it is the responsibility of the crystallographer not to overinterpret his electron density maps, at whatever the resolution.

C. Scope of This Review

This chapter is intended to review the present status of the X-ray diffraction technique, as it applies to single crystals of macromolecules, and to discuss the more important results which have been obtained.

Other reviews which may be consulted for additional background information include the classic exposition by Crick and Kendrew (1957), the chapter by Dickerson (1964) in the second edition of "The Proteins," the review by Eisenberg (1970), and more technical discussions by Holmes and Blow (1965), Phillips (1966), and North and Phillips (1969). The results of protein crystallography have also been reviewed within the last several years by Perutz (1969a), Dickerson and Geis (1969), Blow and Steitz (1970), Dickerson (1972), Matthews and Bernhard (1973), Liljas and Rossmann (1974), and Matthews (1976).

For detailed information on the techniques of X-ray diffraction, books by Stout and Jensen (1968) and by Lipson and Cochran (1966) are highly recommended, although they are directed more toward small molecule than macromolecular crystallography.

II. X-RAY CRYSTALLOGRAPHY

A. Introduction

The science of X-ray crystallography was born when Friedrich *et al.* (1912) succeeded in diffracting X-rays from crystals of copper sulfate, thereby demonstrating that the structure of crystals is periodic and that X-rays are wavelike in nature. Bragg (1913) simplified the interpretation of the diffraction spectra by treating each spectrum as a "reflection" from sets of planes through the crystal. Each set of planes can be uniquely defined by the three "Miller indices" (hkl), and the amplitude of diffraction denoted $F(hkl)$. In fact, the X-rays are scattered by the electrons distributed through the crystal, and it can be shown that the electron density of the crystal $\rho(xyz)$ is related to the "structure amplitudes" $F(hkl)$ by the Fourier transformation

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l F(hkl) \exp i\phi(hkl) \exp\{-2\pi i(hx + ky + lz)\} \quad (1)$$

where $\phi(hkl)$ is the phase angle of the X-ray wave diffracted from the (hkl) planes relative to the phase of the undiffracted wave, and (xyz) are coordinates in the crystal relative to axes defined by unit vectors parallel to the edges of a "unit cell" of volume V. The summation is taken over all the diffraction spectra $F(hkl)$.

To perform the Fourier summation for a given crystal it is necessary to know the structure amplitudes $F(hkl)$ and the phase shifts $\phi(hkl)$ of each scattered wave. The structure amplitudes are directly observable, but the phase shifts are not, and herein lies the fundamental difficulty in X-ray crystallography, i.e., the desired electron density of the protein $\rho(xyz)$ cannot be calculated directly from the observable scattered spectra $F(hkl)$. In words, Eq. (1) states that the electron density in the unit cell of a protein can be considered as the sum of a series of components, in the same way that a musical note consists of the sum of its constituent harmonics. Each diffraction spectrum $F(hkl) \exp i\phi(hkl)$ can be thought of as defining the amplitude and relative phase of one of the harmonics of the electron density distribution.

B. Structure Analysis of Single Crystals

1. X-Ray Diffraction

It is not the purpose of this chapter to review the theory of X-ray diffraction from single crystals, but rather to concentrate on those areas of

the subject which are relevant to contemporary macromolecular crystallography. Some of the terms which may be confusing to the noncrystallographer are defined below.

A protein crystal is composed of a periodically repeated arrangement of molecules, with the intervening spaces occupied largely by disordered solvent molecules in equilibrium with the crystal supernatant. The *unit cell* is the smallest parallelepiped within the crystal which, by *translations* parallel to its edges, will eventually cover the whole crystal. In some instances the definition is relaxed somewhat to allow choice of orthogonal axes; in this case, one part of the unit cell contents will be related to another part by a translation along a face- or body-diagonal of the unit cell. In cases such as this the cell is described as "face centered" or "body centered," in contrast to the more common primitive cell. The unit cell often has internal symmetry elements which relate different parts of the cell. The smallest *unique* volume within the crystal is termed the *asymmetric unit*, and will have a volume equal to an integral fraction of the unit cell volume. By use of the symmetry elements present within the unit cell, the asymmetric unit can be made to cover the volume of the unit cell. For crystals, only certain combinations of symmetry elements are possible, and these are systematically enumerated in the *International Tables for X-Ray Crystallography* (1952). Often the asymmetric unit of a protein crystal will contain one molecule of the protein, but this need not be the case. The asymmetric unit may contain more than one molecule or it may contain only a fraction of the molecule. In the latter case, as will be discussed later in more detail, information about the symmetry of the molecule will be obtained (see Section III,C,3).

2. *The Real Lattice and the Reciprocal Lattice*

The diffraction of X-rays by a crystal can perhaps be best visualized by comparison with the diffraction of light by a diffraction grating.

A single narrow slit will diffract a beam of light passing through it, but since very little light is transmitted, the diffraction pattern will be very weak. Similarly, an isolated protein molecule would diffract a beam of X-rays, but the intensity of diffraction would be immeasurably small.

If many slits are placed side by side, distance d apart, then one obtains a diffraction grating and, as is well known, light directed normally onto the slits will be diffracted into orders of diffraction, so that

$$d \sin \theta_n = n\lambda \tag{2}$$

where λ is the wavelength and θ_n the angle of diffraction for the nth

order. Since many slits diffract and the energy transmitted is concentrated at the angles θ_n, the intensity of diffraction at these angles is much greater than for a single slit. Note that the *angle* of diffraction θ_n is inversely related to the spacing between the slits but is not dependent upon the shape of the individual slits.

A protein crystal has periodicity in three dimensions and is in fact a three-dimensional diffraction grating. A beam of X-rays incident on the crystal will be diffracted into a series of orders, but because of the three-dimensional periodicity, three indices (hkl) are necessary to index a given spectrum. The angles of diffraction, $2\theta_{hkl}$, are given by Bragg's law:

$$2d_{hkl} \sin \theta_{hkl} = \lambda \tag{3}$$

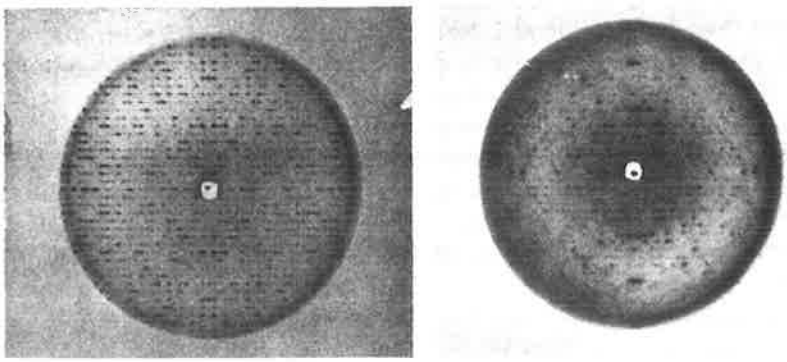
where d_{hkl} is the distance between "Miller" planes passing through the real lattice points. Note again that the angles of diffraction are related inversely to the spacings between the molecules but are not dependent upon the structure of the individual molecules within the unit cell. On the other hand, the amplitudes of diffraction $F(hkl)$ are intimately related to the three-dimensional structure of the individual molecules.

It is convenient to describe the diffraction pattern of a crystal in terms of a *reciprocal lattice*. The *real lattice* is defined by the three base vectors **a**, **b**, and **c** parallel to the edges of the unit cell of the protein crystal. The *reciprocal lattice* is defined by three base vectors **a***, **b***, and **c***. If the real lattice is orthogonal, then the reciprocal lattice is also orthogonal, and $\mathbf{a} = 1/\mathbf{a}^*$, etc. The reciprocal lattice is no more than a very convenient mathematical construction to help visualize diffraction from an object periodic in three dimensions. Reference should be made to standard crystallographic texts, such as that by Stout and Jensen (1968), for ways in which it can be applied.

The lattice points of the reciprocal lattice are identified by indices (hkl) and it can be shown that the point (hkl) is of distance $1/d_{hkl}$ from the origin of the reciprocal lattice, along the line normal to the (hkl) Miller plane.

A typical X-ray diffraction pattern showing the ($h0l$) plane of the reciprocal lattice is shown in Fig. 1a. In this case the crystal is of γ-chymotrypsin, mounted with its c axis horizontal and its a axis vertical. In the real lattice, c is larger than a ($a = 69.6$ Å, $c = 97.7$ Å), so that in the diffraction pattern the horizontal spacing between reflections (proportional to c^*) is less than the vertical spacing (proportional to a^*).

To illustrate the various features of the diffraction pattern, we have



(a) (b)

Fig. 1 Buerger precession photographs of the ($h0l$) reciprocal lattice plane of crystalline γ-chymotrypsin. The h axis is vertical, and the l axis horizontal. Precession angle $\mu = 12°$. (a) Crystals cross-linked with glutaraldehyde; supernatant, 90% saturated ammonium sulfate. (b) Cross-linked crystals soaked in water for several hours. (The diffraction pattern deteriorates with time. After two days virtually no reflections remain.)

included for comparison (Fig. 1b) the diffraction pattern of another crystal of γ-chymotrypsin, grown in the same way as the first, i.e., from a 55% saturated ammonium sulfate solution, but then cross-linked with a solution of 2% glutaraldehyde, and subsequently transferred into pure water. Normally, γ-chymotrypsin crystals quickly dissolve in water, but this is prevented by the glutaraldehyde cross-links. The crystals of cross-linked γ-chymotrypsin do not obviously alter when equilibrated with water, but inspection of Figs. 1a and 1b shows that changes have occurred at the molecular level. The first crystal (Fig. 1a) was photographed in the presence of 90% saturated ammonium sulfate and was in fact also cross-linked, so that the differences are due not to the glutaraldehyde, but rather to the transfer from ammonium sulfate to water. Although the patterns are obviously different, there are many similarities. The spacings between the reflections are approximately the same, although careful measurement shows that the reciprocal lattice points for the cross-linked crystal in water are slightly closer together, indicating that the real lattice has expanded ($a_* = 71.5$ Å, $c = 99.1$ Å). Although certain reflections have different intensities in the two photographs, a great deal of similarity remains, indicating that the γ-chymotrypsin molecules in the two crystals still have fairly similar conformations.

The central region of the photographs, i.e., the "low resolution" diffraction pattern, contains information about the large scale structure of

the molecules and the intervening solvent. For one crystal the solvent is water, which is less electron dense than the protein, whereas in the other case the 90% saturated ammonium sulfate is of higher electron density than the protein. This drastic change in "contrast" between protein and solvent in the crystal causes the large changes in intensity visible in the central region of the photographs. As one proceeds from the center of the photograph toward the outer edge, the resolution increases. In other words, one is observing diffraction from planes through the crystal which are successively closer together. For this photograph, the limiting resolution, i.e., the limiting interplanar spacing, is about 3.7 Å. In Fig. 1a the diffraction spectra extend to the edge of the photographs, showing that the structure of all the γ-chymotrypsin molecules is very similar and that the molecules are aligned throughout the crystal with fidelity. On the other hand, the "high resolution" (i.e., small d spacings) diffraction spectra for the cross-linked γ-chymotrypsin crystal in water are very weak, suggesting that at a resolution of 4 Å either the structures of the individual molecules have been perturbed, or the intact molecules have been disordered somewhat throughout the crystal, or a combination of both. The diffuse regions in Fig. 1b are due in part to random disorder, but suggest that localized regions of "pseudo order" may occur.

Many protein crystals give diffraction patterns measurable to a resolution in the range of 2–3 Å; a few, particularly those of lower molecular weight, give patterns extending to spacings of 1.5 Å or less. On the other hand, some proteins give diffraction patterns poorer than that illustrated in Fig. 1b. Obviously, it is the extent of the diffraction pattern observable for a given crystal which determines the resolution with which the structure of that protein may ultimately be defined.

3. *Crystallographic Symmetry*

The presence of symmetry elements in a protein crystal will often be revealed by corresponding symmetry in the reciprocal lattice, although the relation is not unique, since the reciprocal lattice always exhibits at least a center of inversion.

In Fig. 1a the intensities in the top half of the photograph are equal to those in the bottom half. Similarly, reflections on the left are equal to those on the right. The left-right symmetry could indicate several possible symmetry elements in the crystal: for example, a vertical mirror plane, a vertical twofold axis, or a vertical 2_1 screw axis. (The screw symmetry element P_q defines a rotation of $2\pi/P$ followed by a translation of q/P of the unit cell edge.) Proteins consist of L-amino acids and have a unique handedness, so that the mirror plane can be

excluded. Also, the diffraction pattern allows a distinction to be made between the existence of a twofold axis and a screw axis in the crystal, parallel to *a*. Along the vertical axis of the photograph (Fig. 1a), every second reflection has zero intensity. (The same is also true for the horizontal axis.) These *systematic absences* are characteristic of a 2_1 screw axis rather than a twofold axis. From additional photographs showing other parts of the reciprocal lattice, additional symmetry information is obtained and, at least for crystals of γ-chymotrypsin, all symmetry elements, i.e., the *space group*, may be uniquely defined.

4. *Structure Determination*

We have described how the electron density within the unit cell can be visualized as the sum of many components or waves. Each reflection or reciprocal lattice point (hkl) corresponds to one wave passing through the unit cell. The contribution of the (hkl) reflection to the electron density is given by

$$\rho_{hkl}(xyz) = (1/V)F(hkl) \exp i\phi(hkl) \exp\{-2\pi i(hx + ky + lz)\} \quad (4)$$

where $\phi(hkl)$ is the phase angle of the wave relative to the origin of the cell.

Note that if h, k, and l are small, the wavelength is large, but if the indices are large, the wavelength of the electron density component is small. Thus the low angle diffraction pattern contributes to the broad features of the electron density, but the high angle reflections contribute the fine detail, as was illustrated with Figs. 1a and 1b.

Summing over all the observed diffraction spectra, the electron density is given by

$$\rho(xyz) = \frac{1}{V}\sum_h\sum_k\sum_l \mathbf{F}(hkl) \exp\{-2\pi i(hx + ky + lz)\} \quad (5)$$

where the relation between the *structure factor* $\mathbf{F}(hkl)$ and the structure amplitude $F(hkl)$ is

$$\mathbf{F}(hkl) = F(hkl) \exp i\phi(hkl) \quad (6)$$

As we have mentioned before, the fundamental difficulty of X-ray crystallography is that calculation of the electron density requires the knowledge of the structure amplitudes, which are directly observable, and the structure factor phase angles, which are not.

The inverse relation to Eq. (5), i.e., the equation which expresses the structure factor in terms of the electron density, can be written as

$$\mathbf{F}(hkl) = \int_x\int_y\int_z \rho(xyz) \exp 2\pi i(hx + ky + lz)\, dx\, dy\, dz \quad (7)$$

where the integral is over one unit cell. Alternatively, the electron density can be considered not as a continuous function, but as a sum of the electron densities of all the atoms in the unit cell. In this form, Eq. (7) reduces to

$$\mathbf{F}(hkl) = \sum_{j=1}^{N} f_j(hkl) \exp 2\pi i(hx_j + ky_j + lz_j) \tag{8}$$

where the summation is over the N atoms in the unit cell, the jth atom having coordinates $(x_j y_j z_j)$ and strength of scattering f_j in the direction of the (hkl) reflection. Roughly, f_j is proportional to the number of electrons associated with the jth atom. In words, Eq. (8) states that the (hkl) X-ray wave diffracted from a crystal can be considered as the sum of a series of waves, one for each atom in the unit cell, with proper allowance being made for the different path lengths for each of the wavelets.

Equation (8) implies that if the coordinates of all the atoms in a crystal structure are known, then the amplitudes of X-ray diffraction expected from that crystal are calculable. In other words, if the structure is known, then its diffraction pattern may be calculated. This provides one method of checking whether a proposed structure is correct, i.e., it should, within experimental error, allow prediction of the observed diffraction spectra. This comparison of observed and calculated structure amplitudes is used routinely in small molecule crystallography and fiber diffraction to assess the accuracy of the structure determination, but has been less meaningful when applied to globular proteins. Often the criteria for the correctness of a proposed protein structure have been that it be consistent with an electron density map obtained, say, by using the isomorphous replacement method, and that it be stereochemically plausible. One of the most encouraging recent developments in protein crystallography has been the demonstration by Jensen and colleagues (Watenpaugh *et al.*, 1973a) that the structure of at least one protein can be "refined" to give excellent agreement between the observed and calculated diffraction data (see Section II,D,3).

C. Isomorphous Replacement and Anomalous Scattering

1. Historical Background

The atomic structures of the proteins listed in Table I have all been determined by application of the method of isomorphous replacement, first applied to protein crystals by Green *et al.* (1954). The idea that the phases of the X-ray reflections from a protein crystal could be

determined by the introduction of heavy atoms into the crystal is not new, being suggested by J. D. Bernal in the 1930's (see, for example, Bragg, 1965; Hodgkin and Riley, 1968). Nevertheless, as is now well known, it took many years for the idea to evolve into the working technique which is currently of such eminence in crystallographic studies of proteins.

Isomorphous replacement was used as early as 1927 by Cork (1927) in studying the alums. Bokhoven *et al.* (1951) subsequently extended the method to the study of noncentrosymmetric reflections using what would now be termed the method of single isomorphous replacement. They also suggested that by using a double isomorphous replacement, a unique phase determination would be obtained. The details of the double (or multiple) isomorphous replacement were worked out by Harker (1956). Another contribution of great practical value was the method introduced by Blow and Crick (1959) for the treatment of experimental errors in the isomorphous replacement technique.

In addition to the determination of protein phases by isomorphous replacement, it is possible to obtain useful phase information by utilizing the anomalous (i.e., out-of-phase) scattering of the substituent atoms (Bijvoet, 1954; Okaya and Pepinsky, 1960; Ramachandran and Raman, 1956). Blow (1958) was one of the first to show that anomalous scattering effects could be used to help in the phase determination of macromolecules.

2. *Theory of Isomorphous Replacement*

Let the scattering factor for the (hkl) reflection from a native protein crystal be denoted $\mathbf{F}_P(\mathbf{h})$, and suppose that a small number of heavy atoms is introduced at specific sites in each unit cell of the crystal. Let the structure factor of the modified crystal be $\mathbf{F}_{PH}(\mathbf{h})$. Then, by vector addition, we have

$$\mathbf{F}_{PH}(\mathbf{h}) = \mathbf{F}_P(\mathbf{h}) + \mathbf{F}_H(\mathbf{h}) \tag{9}$$

where $\mathbf{F}_H(\mathbf{h})$ denotes the structure factor of the heavy atoms. In subsequent discussion, where there is no ambiguity, we will omit the reflection index (**h**). We will assume for the moment that the positions and occupancy of the sites of heavy atom substitution are known, so that from Eq. (8) the heavy atom scattering $\mathbf{F}_H$ may be calculated. The amplitudes F_P and F_{PH} are experimentally obtainable, and the essential problem is to combine these with $\mathbf{F}_H$ to yield $\phi(\mathbf{h})$, the phase of the native protein structure factor. A convenient graphical solution, introduced by Harker (1956), is illustrated in Fig. 2a. The known vector $\mathbf{F}_H$ is drawn as shown, and circles of radii equal to F_P and F_{PH} are drawn,

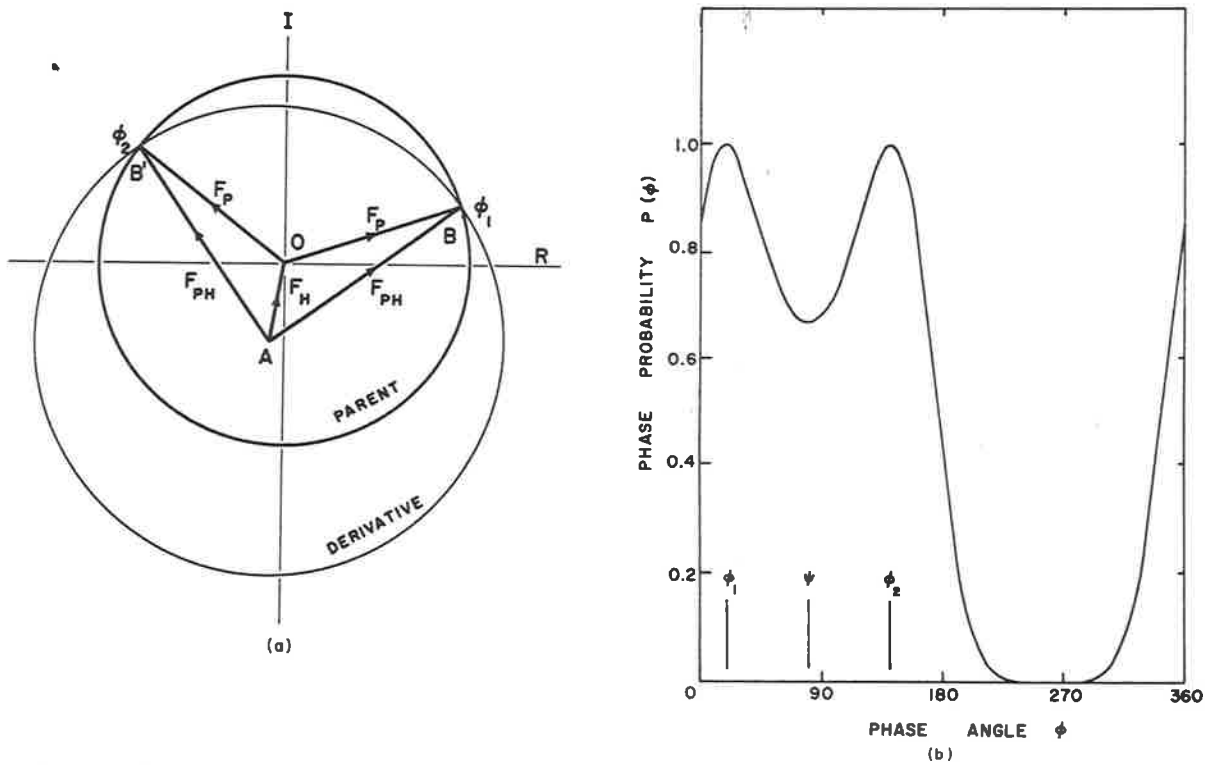


Fig. 2 (a) Harker construction for a single isomorphous replacement. ϕ_1 and ϕ_2 are the "most probable" phases for F_P. (b) Phase probability distribution for a single isomorphous replacement. This and subsequent probabilities are unnormalized. Taken from Matthews (1969).

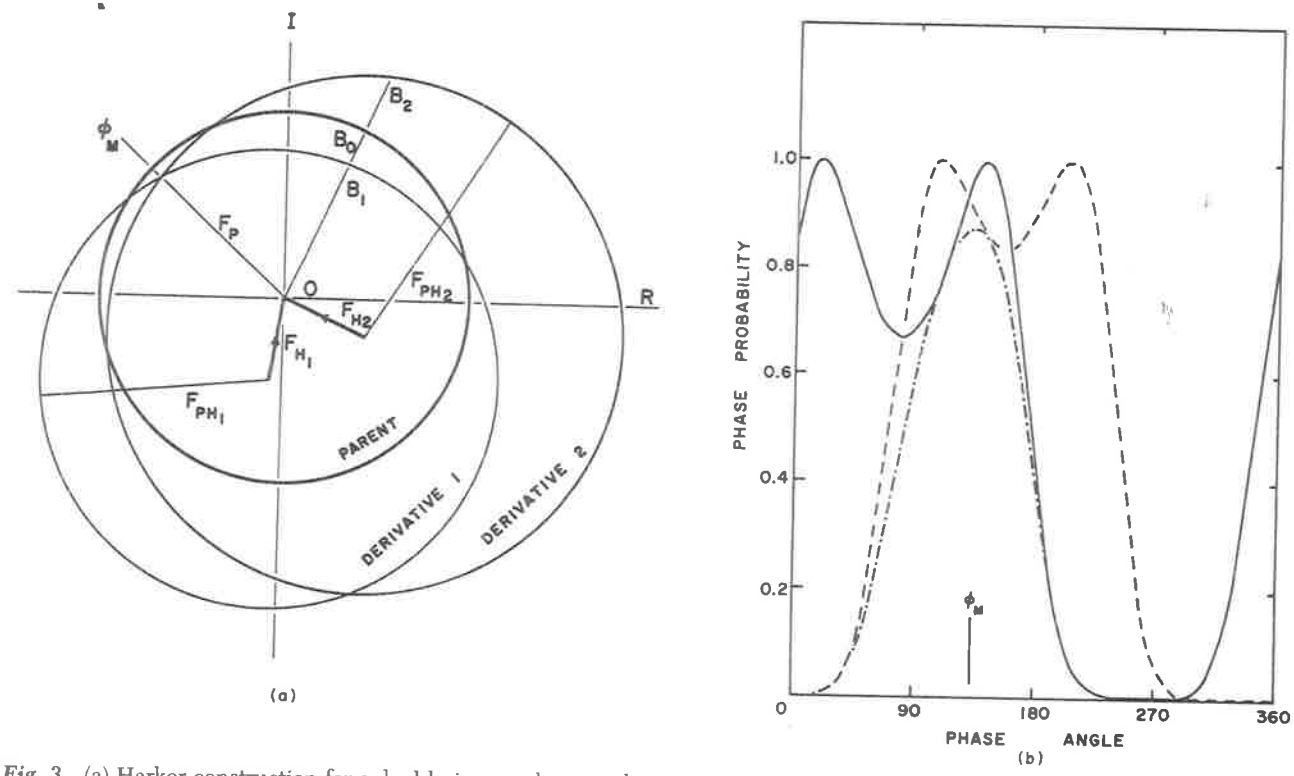


Fig. 3 (a) Harker construction for a double isomorphous replacement. ϕ_M is the "most probable" phase for F_P. (b) Phase probability distribution corresponding to the double isomorphous replacement shown in (a). The curve for derivative 1 is solid, for derivative 2 is broken, and for the combined distribution is drawn as a chain line. Taken from Matthews (1969).

respectively, about O and A. The points of intersection B and B′ of the two circles give possible phase angles ϕ_1 and ϕ_2, both of which satisfy Eq. (9). To resolve the ambiguity and select the correct phase angle for ϕ, three alternatives have been employed: (i) multiple isomorphous replacement, (ii) anomalous scattering, and (iii) use of direct methods. The first method of deciding between the possible solutions ϕ_1 and ϕ_2 in Fig. 2a is by the use of a second isomorphous substitution. If the sites of heavy atom substitution are different in the two isomorphs, then in general, different heavy atom structure factors $\mathbf{F}_{H1}$ and $\mathbf{F}_{H2}$ will be obtained, and the information from them may be graphically combined as illustrated in Fig. 3a. In the absence of all experimental errors, and assuming that the introduction of the heavy metals causes no perturbation of the protein structure, then the three circles in Fig. 2a would intersect at a point, giving a unique solution for ϕ, the phase of $\mathbf{F}_P$. In practice, however, neither of these conditions is likely to hold, so that the circles will not exactly coincide at one point. Nevertheless, it is clear from the figure that the correct phase angle is probably near ϕ_2 rather than ϕ_1. Another complication, related to the experimental precision with which F_P and F_{PH} can be measured, concerns the reliability of the phase information obtained from different isomorphs. For a given reflection, the magnitude of F_H (Fig. 2a) might be very small, so that the circles about O and A are almost concentric and their points of intersection ill defined. In fact, they might not intersect at all. On the other hand, for the same reflection, a second isomorph might have large F_H, with solutions ϕ_1 and ϕ_2 well defined. Clearly, in this case more weight should be given to the information from the second isomorph.

A general method of dealing with cases of this sort was proposed by Blow and Crick (1959) and has become widely adopted. Descriptions of the method have been given by Dickerson *et al.* (1961), Cullis *et al.* (1961a), and Matthews (1969). In essence, the Blow–Crick approach is to assume that no phase angle is ever known with certainty; rather, any phase angle ϕ for the native protein has a finite probability of being correct.

Consider the vector diagram shown in Fig. 4 in which $\mathbf{F}_H$ is known and we wish to determine the probability $P(\phi)$ that the phase angle ϕ is the correct phase of $\mathbf{F}_P$. Strictly, one should allow for experimental errors in $\mathbf{F}_H$, F_P, and F_{PH}, and evaluate the probability that the vector $\mathbf{F}_P$ occupies all possible positions in the Argand diagram. However, the analysis is simplified by making the reasonable assumption that all errors may be treated as if they occur in the measurement of F_{PH}. The probability of ϕ being correct is then given by

$$P(\phi) = N \exp(-\epsilon^2/2E^2) \tag{10}$$

where ϵ is the amount by which the phase triangle (Fig. 4) fails to close for this value of ϕ, E is the rms error in the measurements, and N is a normalizing factor so that the sum of all probabilities is unity, i.e.,

$$\int_0^{2\pi} P(\phi)\, d\phi = 1 \tag{11}$$

If we define $\mathbf{F}_C$ to be the vector sum of $\mathbf{F}_H$ and $F_P \exp(i\phi)$, then

$$\epsilon = F_C - F_{PH} \tag{12}$$

and

$$F_C = \{F_P^2 + F_H^2 + 2F_P(A_H \cos \phi + B_H \sin \phi)\}^{1/2} \tag{13}$$

where

$$\mathbf{F}_H = F_H \exp i\psi = A_H + iB_H \tag{14}$$

The unnormalized probability distribution corresponding to Figs. 2a and 3a is shown in Figs. 2b and 3b. Separate probability distributions $P_1(\phi)$, $P_2(\phi)$, . . . may be obtained for each isomorphous derivative and are combined together to give the overall phase probability distribution for a given reflection as

$$P(\phi) = \Pi\, P_i(\phi) = N \exp\{-(\Sigma\, \epsilon_i^2/2E_i^2)\} \tag{15}$$

The joint probability corresponding to the graphical construction in Fig. 3a is given in Fig. 3b, and in this case, the most probable phase is that which most nearly satisfies the observed data for the two isomorphs.

The main objections to the Blow–Crick treatment are that it assumes no error in F_P and that the results are not in a form suitable for

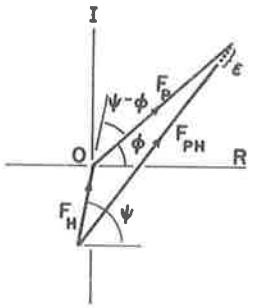


Fig. 4 Vector diagram illustrating the lack of closure, ϵ, of an isomorphous replacement phase triangle. Taken from Matthews (1969).

easy incorporation of phase information from a subsequent heavy atom derivative. Cullis *et al.* (1961a) have proposed a modification of Blow and Crick's method to empirically allow for the fact that all the observations may have errors, but their method has certain computational disadvantages. Raiz and Andreeva (1970) have proposed a general method of phase determination which includes, as special cases, the Blow–Crick formalism and that of Cullis *et al.* (1961a). Hendrickson and Lattman (1970) have shown that the phase probability $P(\phi)$ can be expressed in the form

$$P(\phi) = N \exp(A \cos \phi + B \sin \phi + C \cos 2\phi + D \sin 2\phi) \quad (16)$$

where A, B, C, and D are constants which encode the phase information (see also Rossmann and Blow, 1961). Hendrickson and Lattman (1970) point out that simply by adding the appropriate constants it is possible to combine together phase information from different sources such as isomorphous replacement, anomalous scattering, direct methods of phase determination, partial structure phase information, and information from noncrystallographic symmetry. Nevertheless, the Blow–Crick method remains the most popular at the present time and has been proved adequate in many successful structure determinations.

Whatever method is employed to evaluate the phase probability distribution, it still remains to select the Fourier coefficient for inclusion in the calculation of the electron density map of the protein. Blow and Crick (1959) point out that although the "most probable" electron density map, i.e., the one using the most probable phases, is most likely to be correct, it may also contain spurious features due to inclusion of phase angles which, although most probable, are still likely to be seriously in error. Blow and Crick suggest that rather than using the phase at the maximum of the probability distribution, it may be preferable to use a compromise solution given by the centroid of the probability distribution. They show the corresponding synthesis to be the "best Fourier," defined as *that Fourier transform which is expected to have a minimum mean square difference from the Fourier transform of the true F's when averaged over the whole unit cell.* Comparisons of the "best" and "most probable" Fouriers have been made by Dickerson *et al.* (1961) and by Cullis *et al.* (1961b). The respective electron density maps were found to be very similar, although in both cases the "best" map was the one preferred.

The centroid of the phase probability distribution can be defined as a vector of length mF_P and phase angle ϕ_B, where ϕ_B is the "best" phase. Taking the mean of all possible values of the protein structure

factor $F_p \exp(i\phi)$, weighted by the probability $P(\phi)$ that ϕ is correct, we obtain

$$m \cos \phi_B = \frac{\int P(\phi) \cos \phi \, d\phi}{\int P(\phi) \, d\phi} \tag{17a}$$

and

$$m \sin \phi_B = \frac{\int P(\phi) \sin \phi \, d\phi}{\int P(\phi) \, d\phi} \tag{17b}$$

The quantity m, which acts as the weighting factor for F_P, is called the "figure of merit" of the phase determination. Dickerson *et al.* (1961) have shown that the mean square error in the electron density of a "best" electron density map is given by

$$\Delta\rho^2 = \frac{2}{V^2} \sum_{0}^{\infty} \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} F_P^2(\mathbf{h})[1 - m^2(\mathbf{h})] \tag{18a}$$

Cohen *et al.* (1970) tested this equation with electron density maps of α- and γ-chymotrypsin and found it to give reasonably accurate values for the errors in the electron density maps.

Henderson and Moffat (1971) have discussed errors in difference Fourier syntheses (see Sections II,C,*4* and II, E,*1*) and concluded that they are best estimated by the equation

$$\overline{\Delta\rho^2} = \frac{1}{V^2} \Sigma\Sigma\Sigma\{(F_{PH} - F_P)^2(2 - m^2) + \delta^2\} \tag{18b}$$

where δ is the rms experimental error in the determination of $(F_{PH} - F_P)$.

3. *Anomalous Scattering*

Although the multiple isomorphous replacement can be very successful for phase determination, it is often difficult to obtain even one or two suitable isomorphs and it is important to extract as much information as possible from those derivatives which are available. One such source of information derives from measurement of small intensity differences in the diffraction patterns caused by "anomalous" or out-of-phase scattering by the inner electrons of any heavy atoms present in the crystal.

The atomic scattering factor of an atom can be expressed as

$$\mathbf{f} = f_0 + \Delta f' + i \, \Delta f'' = f' + if'' \tag{19}$$

where f_0 is the normal scattering factor far from an absorption edge, and $\Delta f'$ and $\Delta f''$ are correction terms which arise from dispersion effects. The quantity $\Delta f'$ usually corresponds to a small reduction in the normal (in-phase) scattering, and $\Delta f''$, normally negative, corresponds to scattering with phase $\pi/2$ ahead of the normal scattering. For atoms such as carbon and nitrogen, $\Delta f'$ and $\Delta f''$ are negligibly small, but for typical "heavy atoms," they may equal 10% to 20% of the total scattering.

Assuming that there are N heavy atoms in the unit cell, the nth atom having coordinates (x_n, y_n, z_n) and atomic scattering factor $f'_n + if''_n$, then the total scattering of these atoms is given by the vector sum of an inphase structure factor

$$\mathbf{F}_H(\mathbf{h}) = A_H(\mathbf{h}) + iB_H(\mathbf{h}) \tag{20}$$

and an out-of-phase structure factor

$$\mathbf{F}''_H(\mathbf{h}) = A''_H(\mathbf{h}) + iB''_H(\mathbf{h}) \tag{21}$$

where it may be shown (e.g., see Matthews, 1969) that

$$A_H(\mathbf{h}) = \sum_{n=1}^{N} f'_n(\mathbf{h}) \cos \phi \tag{22a}$$

$$B_H(\mathbf{h}) = \sum_{n=1}^{N} f'_n(\mathbf{h}) \sin \phi \tag{22b}$$

$$A''_H(\mathbf{h}) = -\sum_{n=1}^{N} f''_n(\mathbf{h}) \sin \phi \tag{22c}$$

and

$$B''_H(\mathbf{h}) = \sum_{n=1}^{N} f''_n(\mathbf{h}) \cos \phi \tag{22d}$$

where

$$\phi = 2\pi(hx + ky + lz) \tag{22e}$$

If anomalous scatterers are present in a protein crystal, the amplitudes of the (hkl) and $(\overline{hkl})$ reflections, which are normally equal, become slightly different, allowing information about the phase of the reflection to be obtained. The origin of the difference between the amplitudes of a heavy atom substituted crystal, denoted F_{PH+} and F_{PH-}, is illustrated in Figs. 5a and 5b. The protein itself contains no heavy atoms so that its out-of-phase scattering is negligible and one need only consider the in-phase scattering, defined by

$$A_P(\mathbf{h}) = A_P(\bar{\mathbf{h}}) \tag{23a}$$

and

$$B_P(\mathbf{h}) = -B_P(\bar{\mathbf{h}}) \tag{23b}$$

If the phase angle of the protein scattering for the (hkl) reflection is ϕ, then that for $(\bar{h}\bar{k}\bar{l})$ is $-\phi$, as illustrated in Fig. 5a. For the heavy atom scatterers, the out-of-phase scattering cannot be ignored and may lead to the situation illustrated in Fig. 5a. On vectorially summing the protein plus heavy atom scattering to obtain F_{PH+} and F_{PH-}, it can be seen that the resultants are unequal. The information contained in Fig. 5a can be expressed more conveniently by reflecting the $(\bar{h}\bar{k}\bar{l})$ diagram through the real axis onto the (hkl) diagram, as in Fig. 5b.

The way in which the anomalous scattering data are used to obtain information about the phase of F_P may be illustrated by a construction analogous to that in Fig. 2a. Given the heavy atom parameters, the components of the heavy atom scattering F_H and F''_H (Fig. 5a) may be calculated and drawn as shown in Fig. 6. If circles of radii equal to the observed amplitudes F_{PH+} and F_{PH-} are drawn with respective centers marked + and − (Fig. 6), then in the absence of all errors the intersections of these circles will give two alternate solutions for the phase angle of the heavy atom derivative (although not, directly, for the unsubstituted protein). Note that the phase ambiguity from the isomorphous replacement method (Fig. 2) is symmetrical about a line paral-

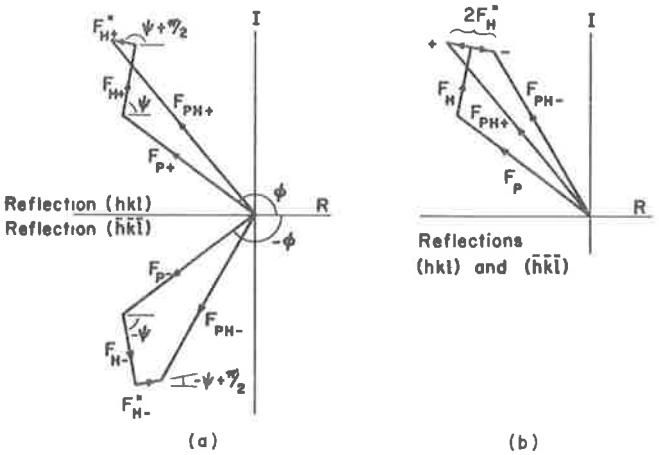


Fig. 5 (a) Vector diagrams illustrating anomalous scattering for the reflections (hkl) and $(\bar{h}\bar{k}\bar{l})$. (b) Combined vector diagram for reflections (hkl) and $(\bar{h}\bar{k}\bar{l})$. Taken from Matthews (1969).

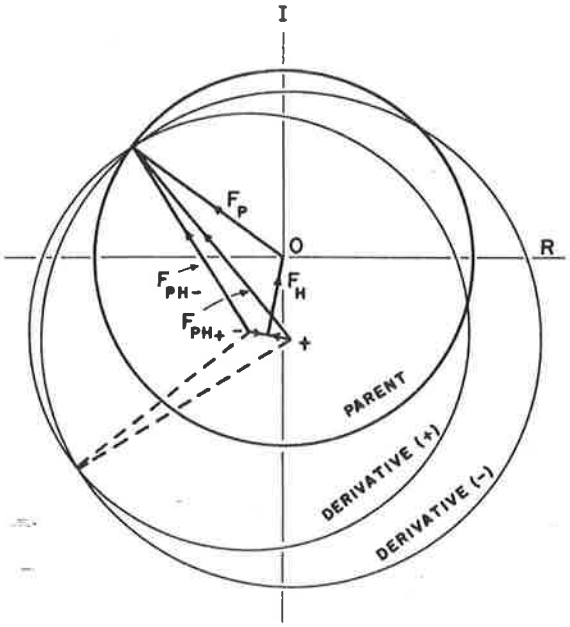


Fig. 6 Harker construction for a single isomorphous replacement with anomalous scattering, in the absence of errors. Taken from Matthews (1969).

lel to the in-phase heavy atom scattering $\mathbf{F}_H$, whereas the phase ambiguity from the anomalous scattering method is symmetrical about a line parallel to $\mathbf{F}''_H$. In general, these lines will be at right angles, or nearly so. In other words, the phase information provided by the anomalous scattering technique is exactly complementary to the phase information from isomorphous replacement. It is this complementarity which makes the combination of isomorphous replacement with anomalous scattering so powerful. Either method, used alone, will give an ambiguous solution for the desired phase angle (assuming only one isomorph is available), but taken together they will always, at least in principle, give a unique solution. As will be discussed in Section II,C,*4*, use can be made of the same complementarity in locating and refining the positions of heavy atom substituents.

Returning to Fig. 6, it was shown that the anomalous scattering data lead directly to alternate phase angles for F_{PH} rather than F_P. However, in the absence of error, F_P is defined simply as the vector from 0 to the intersection of the F_{PH+} and F_{PH-} circles. In practice, the amplitude of the protein scattering, F_p, is observable and can be incorporated into Fig. 6 by drawing a circle about 0 with radius F_P. In the ab-

sence of all error, and assuming perfect isomorphism, the three phase circles will intersect at a point, as in Fig. 6, resolving the ambiguity inherent in the anomalous scattering data taken alone. In practice, however, the three circles may not coincide, and methods based upon that of Blow and Crick (1959) have been developed to combine the conflicting data in the most advantageous way (Blow and Rossmann, 1961; North, 1965; Matthews, 1966a). A typical vector diagram illustrating the treatment of errors in the isomorphous replacement–anomalous scattering technique is shown in Fig. 7, in which ϵ_+ and ϵ_- represent the experimental errors which would have to occur if the phase ϕ were the correct phase for F_P. To satisfy just the anomalous scattering measurements, F_{PH-} and F_{PH+} must meet at a point, or, what amounts to the same thing, $(\epsilon_+ - \epsilon_-)$ must be zero. This requirement leads to a phase probability distribution for the protein phase angle given by

$$P_{\text{ano}}(\phi) = N \exp\left\{ - \frac{1}{2E'^2} [F_{PH-} - F_{PH+} + 2F''_H \cos(\gamma + \omega)]^2 \right\} \qquad (24)$$

where E' is the rms error in $(\epsilon_+ - \epsilon_-)$. The expression on the right can be evaluated using the following relationships (Matthews, 1966a):

$$\sin \gamma = [F_P(B_H \cos \phi - A_H \sin \phi)]/(F_C F_H) \qquad (24a)$$

$$\cos \gamma = [F_H^2 + F_P(A_H \cos \phi + B_H \sin \phi)]/(F_C F_H) \qquad (24b)$$

$$\sin \omega = (A_H B''_H - A''_H B_H)/F''_H F_H \qquad (24c)$$

$$\cos \omega = (A_H A''_H + B_H B''_H)/F''_H F_H \qquad (24d)$$

The phase probability distribution $P_{\text{iso}}(\phi)$ determined from the iso-

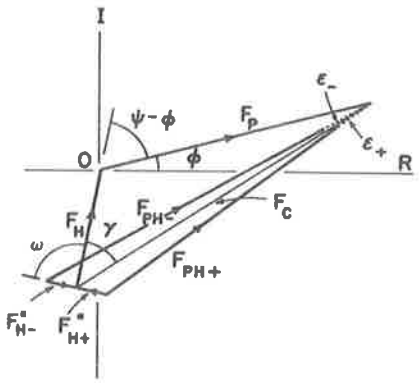


Fig. 7 Vector diagram illustrating lack of closure in the anomalous scattering method. Taken from Matthews (1969).

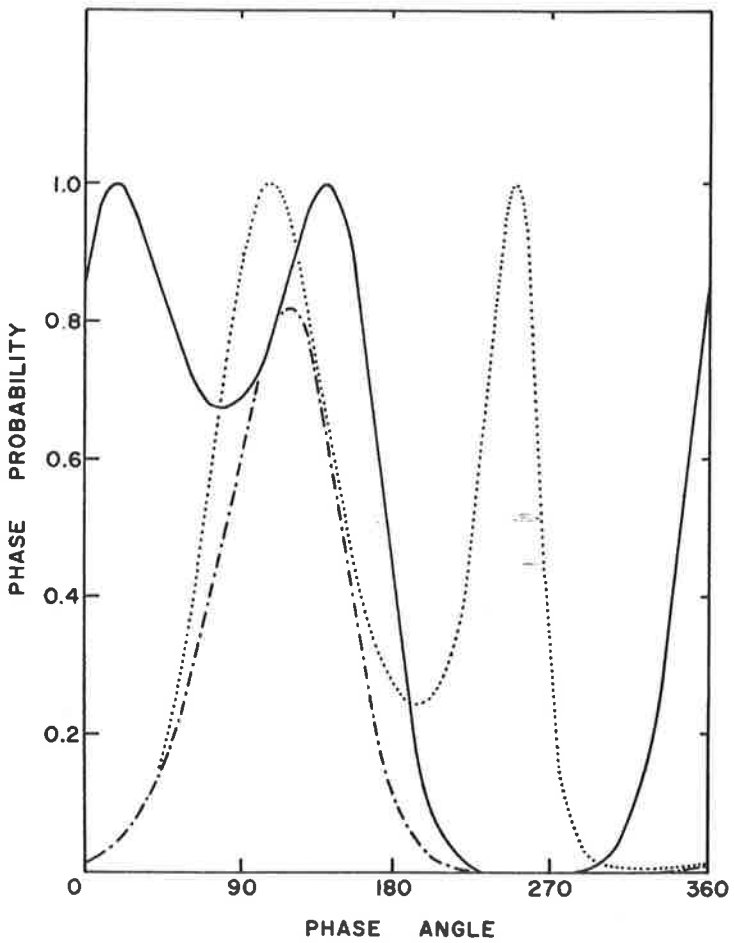


Fig. 8 Combination of isomorphous replacement (solid line) and anomalous scattering (dotted line) phase probabilities to obtain a combined probability distribution (chain line). Taken from Matthews (1969).

morphous replacement data illustrated in Fig. 7 is identical with Eqs. (10), (12), and (13), substituting $(\epsilon_+ + \epsilon_-)/2$ in place of ϵ, and $(F_{PH+} + F_{PH-})/2$ in place of F_{PH}. The overall phase distribution from both isomorphous and anomalous scattering data is given by the product of the two distributions. Examples of the isomorphous replacement and anomalous scattering probability distributions corresponding to Fig. 7 are given in Fig. 8.

Although the intensity differences due to anomalous scattering effects are smaller than those due to isomorphous replacement, there

are several factors which tend to offset this disadvantage (e.g., see Blow, 1958; North, 1965). In particular, the experimental errors may be smaller and the consequences of imperfect isomorphism are much less severe for the anomalous scattering method.

Before using anomalous scattering data for phase determination it is essential to have the coordinates of the heavy atom substituents in their correct absolute configuration, since use of the wrong "hand" will lead to nonsensical phase angles (Blow and Rossmann, 1962; Matthews, 1966b). If two or more heavy atom isomorphs are available, several methods may be employed to find the absolute configuration of the substituents. For example, the mean value of the figure of merit may be used as a measure of the consistency of phase determinations from different isomorphs, and this value should be higher when the correct enantiomorph is chosen. Also, the average "lack of closure" of the phase triangles should be less (e.g., see Lipscomb *et al.*, 1966; Colman *et al.*, 1972a). Other methods of hand-determination have employed various Fourier syntheses (Matthews, 1966b; Drenth *et al.*, 1967; Kraut, 1968). If only one derivative is available and the native protein does not contain anomalous scatterers, the only possible procedure is to compute two electron density maps using phase angles calculated for each of the alternate enantiomorphs. The "correct" heavy atom arrangement should lead to the more physically reasonable map. This method was in fact used in the study of rubredoxin (Herriott *et al.*, 1970).

4. Location of Heavy Atoms

Before any heavy atom isomorph can be used for phase determination, the sites of substitution must be determined from the observed differences in scattering amplitude caused by introduction of the heavy atoms. Although many approaches to this problem have been proposed, the basic method used and proved to be successful in many applications is the calculation of difference Patterson and difference Fourier syntheses.

a. **Difference Patterson Syntheses.** The relation between the structure factor for the protein, $\mathbf{F}_P$, for the substituted crystal, $\mathbf{F}_{PH}$, and for the heavy atom(s), $\mathbf{F}_H$, is illustrated in Fig. 9. Assuming that only one heavy atom derivative is available, then the data available experimentally are the magnitudes F_P and F_{PH}. Ideally, one would like to know the *vectors* $\mathbf{F}_H$ which, by a suitable Fourier transformation (Eq. 5), could be combined to yield directly the heavy atom electron density distribution. If only the amplitudes of the heavy atom scattering

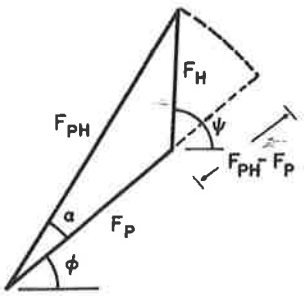


Fig. 9 Vector diagram illustrating, for an arbitrary reflection, the relation between the structure factor for the protein, $\mathbf{F}_P$, for the substituted crystal, $\mathbf{F}_{PH}$, and for the heavy atom(s), $\mathbf{F}_H$.

are known, then one may calculate the well-known synthesis developed by Patterson (1934)

$$P(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l F_H(hkl)^2 \cos 2\pi(hx + ky + lz) \qquad (25)$$

The Patterson function does not yield the atomic sites directly, but gives the distribution of the vectors between pairs of atoms.

From Fig. 9, the amplitude F_H is given by

$$F_H^2 = F_{PH}^2 + F_P^2 - 2F_P F_{PH}^2 \cos \alpha \qquad (26)$$

$$= (F_{PH} - F_P)^2 + 2F_P F_{PH} (1 - \cos \alpha) \qquad (27)$$

In the special case of a centrosymmetric projection, $\alpha = 0$ for all except some of the weak reflections, so that

$$F_H^2 = (F_{PH} - F_P)^2 \qquad (28)$$

Thus, for a protein with one or more centrosymmetric projections, the difference Patterson synthesis with coefficients equal to the amplitudes of heavy atom scattering can be calculated directly. An example is given in Fig. 10a for a single site derivative of thermolysin obtained by replacement of the active site zinc with mercury (Matthews *et al.*, 1972a). Note that although there is heavy atom substitution at only one site, there are many peaks in the Patterson function. In general, if the unit cell contains N atoms, the Patterson function will contain N^2 vector peaks. Obviously, the satisfactory interpretation of the projected Patterson function of a multisite derivative, especially in a high symmetry space group, may be impossible, and in such cases three-dimensional data may be essential.

For general (noncentrosymmetric) reflections, α in Eq. (27) is un-

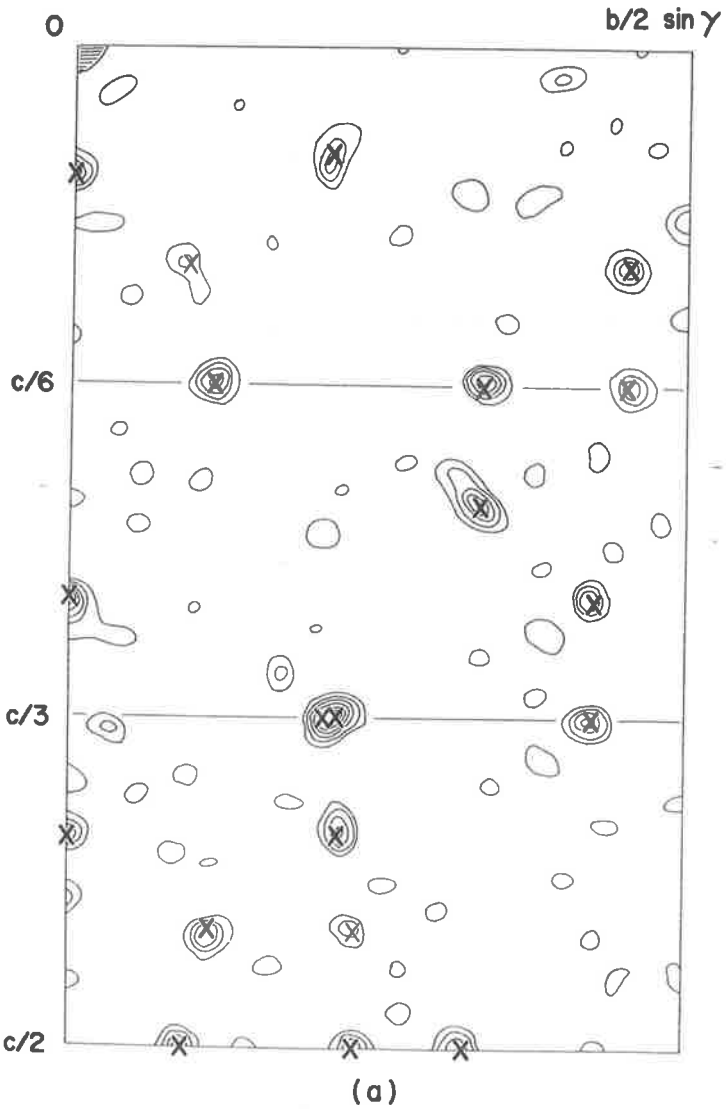
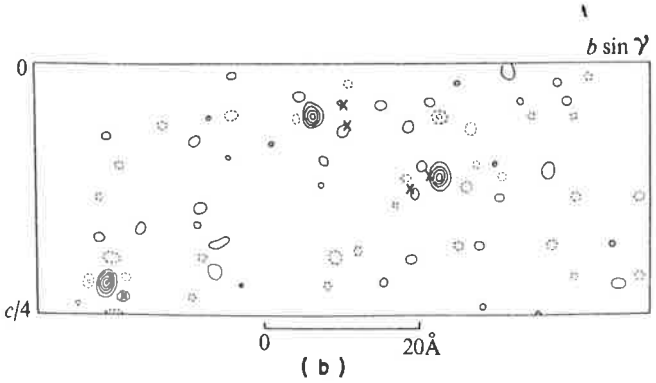


Fig. 10 (a) Difference Patterson projection calculated from the centrosymmetric (0*kl*) diffraction data for thermolysin, using the differences between zinc-free crystals and crystals in which the zinc has been replaced by mercury. Substitution occurs at only one site per molecule, and the expected mercury–mercury vector positions are indicated by light crosses for single-weight peaks and darker crosses for double-weight ones. Resolution 2.5 Å.



(b) Difference Fourier projection using as amplitudes the difference between mercury-thermolysin and native (zinc) thermolysin. Phase angles were determined from a dimercury acetate heavy atom isomorph (the two mercury sites per molecule are indicated by crosses). In this space group, one obtains, in effect, a simultaneous projection of three molecules so that although replacement of zinc by mercury occurs at only one site per molecule, three peaks are obtained in the difference Fourier synthesis. Resolution 2.4 Å. Taken from Matthews *et al.* (1972b).

known, and F_H^2 cannot be found from the observable data (excluding, for the moment, possible anomalous scattering effects). However, Blow (1958) and Rossmann (1960) found that even for noncentrosymmetric data a difference Patterson synthesis with coefficients $(F_{PH} - F_P)^2$ could still be used to determine the heavy atom positions. It can be shown that this synthesis, calculated with noncentrosymmetric reflections, will have peaks at the expected heavy atom–heavy atom vector positions, but reduced to about half their full height and, in addition, there will be some background features (Phillips, 1966; Kartha and Parthasarathy, 1965a). Experience has shown the difference Patterson method, both in projection and with three-dimensional data, to be a very useful aid in determining at least the major sites of occupancy in potential isomorphous derivatives. Once this start has been made, other methods may be employed to locate sites of minor substitution using approximate phase angles derived from this first derivative to analyze other potential isomorphs.

Just as the "isomorphous" difference Patterson synthesis can be used to locate isomorphously substituted heavy atoms, the "anomalous" difference Patterson, with coefficients $(F_{PH+} - F_{PH-})^2$, may be used to locate anomalously scattering groups (Rossmann, 1961). It can be shown that this function will have peaks corresponding to the vectors between the anomalous scattering centers, along with some background "noise" (Kartha and Parthasarathy, 1965a). In practice, the anomalous scattering differences $(F_{PH+} - F_{PH-})$ may be very small

and difficult to measure with precision, so that inaccuracies in an anomalous scattering Patterson will tend to be due to experimental limitations rather than to approximations in the theory. Nevertheless, an anomalous scattering difference Patterson can provide a completely independent check on heavy atom coordinates derived from an isomorphous difference Patterson.

In discussing phase determination using isomorphous replacement and anomalous scattering, we stressed the intrinsic complementarity of these two effects, and Kartha and Parthasarathy (1965a) and Matthews (1966b) have shown how use can also be made of this complementarity in calculating difference Patterson syntheses. On combining the isomorphous replacement and the anomalous scattering data, the desired heavy atom scattering amplitude F_H is given to a good approximation by

$$F_H^2 = F_P^2 + F_{PH}^2 - 2F_PF_{PH}\{1 - [wk(F_{PH+} - F_{PH-})/2F_P]^2\}^{1/2} \quad (29)$$

where $w = 1$ and k is the ratio of real to the anomalous scattering of the heavy atom groups (Matthews, 1966b). Making further approximations, Eq. (29) reduces to

$$F_H^2 = (F_{PH} - F_P)^2 + (k/2)^2 (F_{PH+} - F_{PH-})^2 \quad (30)$$

(Kartha and Parthasarathy, 1965a; Kartha, 1969). Expressions slightly more accurate than Eq. (29) have been given (W. E. Love, personal communication; Singh and Ramaseshan, 1966) but have not been shown to be worth applying in practice. The improvement which can be obtained by combining isomorphous replacement and anomalous scattering data for a noncentrosymmetric projection is shown in Fig. 11. In practice, it seems desirable to allow for the larger relative experimental error in the anomalous differences by the introduction of the empirical weighting factor w in Eq. (29). The value $w = 0.75$ has been found to be satisfactory in several cases (e.g., Matthews, 1966b; Herriott *et al.*, 1970; Mathews *et al.*, 1972).

***b.* Difference Fourier Syntheses.** Once the approximate heavy atom sites of one or more isomorphous derivatives have been determined, approximate phase angles for the protein structure factors may be determined and used to mutually confirm heavy atom sites determined from Patterson functions and to study potential new isomorphs. Given such phase angles, $\phi(\mathbf{h})$, one may calculate a "difference Fourier synthesis," with coefficients

$$m(F_{PH} - F_P) \exp(-i\phi) \quad (31)$$

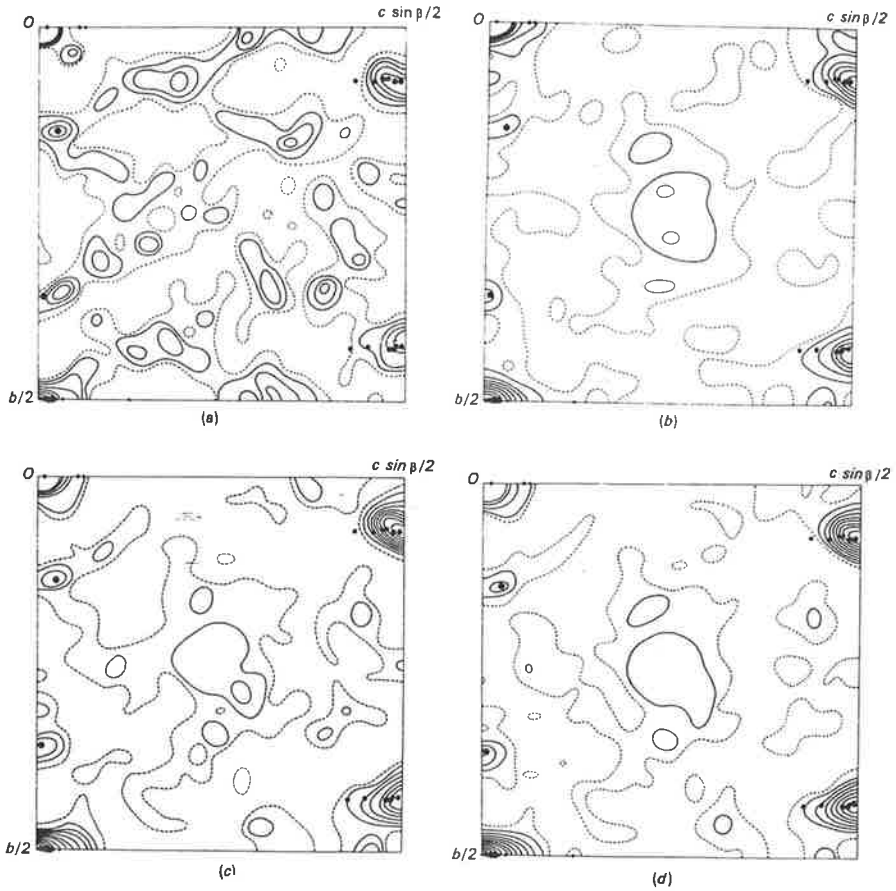


Fig. 11 Comparison of different (0*kl*) difference Patterson syntheses used to locate the heavy atom groups in $[PtCl_4]^{2-}$-substituted α-chymotrypsin. The position and size of the small solid circles indicate the position and approximate height of the expected vector peaks. (a) "Anomalous difference" synthesis. Contours are drawn at one twenty-fifth the interval used for (b), (c), and (d). (b) "Isomorphous replacement" synthesis. (c) "Sum" synthesis [Eq. (30)]. (d) "Combined" synthesis [Eq. (29)]. Taken from Matthews (1966b).

Kraut *et al.*, 1962; Steinrauf, 1963; Stryer *et al.*, 1964). For a centrosymmetric projection, the difference Fourier function will give, to a good approximation, the difference between the electron density of the substituted and native protein. If the two crystal structures are isomorphous, the protein density will cancel, leaving the desired heavy atom density. An example of a difference Fourier for a centrosymmet-

ric projection of thermolysin, showing the substitution of mercury at the active site, is given in Fig. 10b. Note that the difference Fourier function is much easier to interpret than the corresponding difference Patterson function. The difference Fourier synthesis can also be used successfully for noncentric reflections. In this case, the difference coefficient does not equal $\mathbf{F}_H$, as in the centrosymmetric case, but equals, approximately, the projection of $\mathbf{F}_H$ on $\mathbf{F}_P$ (Fig. 9). It can be shown theoretically and by test calculations (Luzzati, 1953; Steinrauf, 1963) that a difference Fourier synthesis calculated with noncentrosymmetric reflections will reveal the true difference density, but with peak heights reduced to about 60% of their true value. The difference Fourier method has been very successful not only in locating heavy atoms but also, following the study of azide-myoglobin by Stryer *et al.* (1964), in studying the binding of substrates, inhibitors, and coenzymes to crystalline proteins.

In the same way that anomalous scattering and isomorphous replacement data may be used to complement each other in the calculation of difference Patterson coefficients [Eq. (29)], they may also be combined to obtain the desired heavy atom vector $\mathbf{F}_H$ (Fig. 9) rather than its projection on $\mathbf{F}_P$ as in the normal isomorphous replacement difference Fourier synthesis (Matthews, 1966b). Given the protein phase angle ϕ, the magnitude of the desired coefficient $mF_H \exp(-i\psi)$ can be obtained from Eq. (29), and its phase ψ is given (Matthews, 1966b) by

$$\psi = \beta + \phi - 180° \tag{32}$$

where

$$\sin \beta = kF_{PH}(F_{PH+} - F_{PH-})/2F_PF_H \tag{32a}$$

and

$$\cos \beta = (F_P^2 + F_H^2 - F_{PH}^2)/2F_PF_H \tag{32b}$$

In locating heavy atoms, difference Fourier syntheses are superior in several respects to difference Pattersons, but in some circumstances they can give misleading results. The Fourier calculation provides a map in which there is only one peak per potential heavy atom site, whereas the difference Patterson will usually contain many interrelated vector peaks, perhaps overlapping and obscured by background density. The success of the Fourier method depends upon the quality of both the phase angles used and the derivative being evaluated, whereas the Patterson method depends only upon the latter. Until one has good reason to believe that the derivative or derivatives

used to calculate phase angles are reliable, such phase angles must be used with caution. In particular, features in difference electron density maps which coincide with the sites of the heavy atoms used for the phase determination must be considered as probable artifacts. In general, both difference Patterson and difference Fourier syntheses should be calculated. The Patterson functions should independently provide coordinates for the respective heavy atoms of each derivative, and these can be checked by using phase angles from one derivative to calculate difference Fourier syntheses of the others, and so on. As derivatives are shown to be reliable, they may be included in subsequent phase angle calculations. Some of the possible dangers inherent in the difference Fourier method are described by Dickerson *et al.* (1967).

One problem which needs to be considered, especially in certain low-symmetry space groups, notably $P2_1$, which occurs rather commonly, is the determination of the relative coordinates of atoms in different derivatives. In space group $P2_1$, the origin of coordinates along the b axis is arbitrary, but once fixed, e.g., by assigning y-coordinates to the heavy atoms in one isomorph, must be used consistently for all subsequent derivatives.

One method of dealing with this problem, suggested by Rossmann (1960), is by calculation of difference Patterson functions between pairs of derivatives, with coefficients of the form $(F_{PH1} - F_{PH2})^2$. This synthesis will give positive peaks for vectors between atoms in the same derivative and negative features for the "correlation" vectors between atoms in different isomorphs.

A second possible correlation Patterson function uses coefficients $(F_{PH1} - F_P)$ $(F_{PH2} - F_P)$ (Steinrauf, 1963; Kartha and Parthasarathy, 1965b). In this case, only the correlation vectors are obtained as positive peaks, and the intraderivative vectors are eliminated, although there is also some spurious background density. Kartha and Parthasarathy (1965b) show that by including anomalous scattering data, this background may be eliminated.

A third method of correlating the origin of different isomorphs, and possibly the most satisfactory, is by difference Fourier syntheses (for example, see Matthews, 1966b). In this method, coordinates are assigned to one derivative and are used to calculate approximate protein phase angles ϕ_1 which are in turn used to calculate a difference synthesis with coefficients $m(F_{PH2} - F_P) \exp(-i\phi_1)$. This map will reveal the heavy atom sites in the second derivative relative to the same origin adopted for the phase calculation. Furthermore, if anomalous scattering data are available, the same calculation may be used to deter-

mine the absolute configuration of the heavy atom coordinates (Matthews, 1966b).

c. **Weighting Functions.** Following the classical paper by Blow and Crick (1959) on the treatment of errors in the isomorphous replacement method, it has become standard procedure to calculate the "best" electron density map of the protein (see Section II,C,2). Also, when calculating difference Fourier syntheses, it is usual to weight each coefficient by the figures of merit $m(\mathbf{h})$ of the phase determinations, as in Eq. (31).

It may also be possible to improve the "signal to noise" of difference Patterson and difference Fourier syntheses by the use of additional weighting functions.

One may define a "best" difference Patterson or difference Fourier synthesis in the Blow–Crick sense. The "best" difference Patterson has coefficients

$$F_P^2 + F_{PH}^2 - 2wF_PF_{PH} \tag{33}$$

and the "best" difference Fourier has coefficients

$$m(wF_{PH} - F_P)\exp(-i\phi) \tag{34}$$

(Woolfson, 1956; Sim, 1960; Srinivasan, 1968; B. W. Matthews, unpublished observations), where w is a weighting factor. For centrosymmetric reflections

$$w = \tanh x_c \tag{35a}$$

and

$$x_c = F_PF_{PH}/(\overline{F_{PH} - F_P)^2} \tag{35b}$$

where the overbar denotes the average value. For noncentrosymmetric reflections

$$w = I_1(2x_a)/I_0(2x_a) \tag{35c}$$

where I_1 and I_0 are Bessel functions with imaginary argument, and

$$x_a = F_PF_{PH}/2(\overline{F_{PH} - F_P)^2} \tag{35d}$$

The denominator in Eqs. (35b) and (35d) is evaluated as a function of scattering angle.

The physical interpretation of these equations is illustrated by considering a typical centric reflection. Most likely, the true amplitude difference between the parent and derivative structure amplitudes is $F_{PH} - F_P$, but there is a probability that the true heavy atom structure

amplitude may be $F_{PH} + F_P$. The "best" difference Fourier or difference Patterson has coefficients which provide a compromise between the two possibilities, weighted according to their respective likelihood of occurring. Note that the weighting factors have the greatest effect on the weak reflections and should be most worth using when the average intensity difference between the parent and the isomorphous derivative is large. On the other hand, the weak reflections are generally the least accurately measured, and a weighting procedure which enhances their relative contribution might be of doubtful value. In unpublished tests of the weighting functions using derivatives of thermolysin and T4 phage lysozyme, the peak-to-background ratio was improved by up to 7% for centrosymmetric difference Fourier projections, but little if any improvement was observed in three-dimensional difference Fourier syntheses and in Patterson functions.

Another type of weighting has been proposed by Moews and Bunn (1971), who suggest that in difference Patterson syntheses, the acentric reflections should be given a weight of 1/0.64 to allow for the fact that the experimentally determined quantity $F_{PH} - F_P$ is less for these reflections, on the average, than the desired amplitude F_H. Although not suggested by Moews and Bunn, an obvious extension of their method would be to apply a similar weighting to the acentric reflections in difference Fourier syntheses containing both centric and acentric reflections.

The possible use of direct methods for the location of heavy atoms is discussed in Section II,F,1.

5. *Refinement of Heavy Atom Parameters*

Once the approximate sites of heavy atom substitution have been located, the heavy atom parameters should be determined as accurately as possible before phase angles are calculated. Two general techniques may be employed: least squares refinement and Fourier methods.

a. **Least Squares Refinement.** The simplest case to consider is refinement of a single derivative using only the centric reflections. In this case, one wishes to determine the heavy atom parameters which minimize the sum of the squares of the residuals

$$S = \sum_{\mathbf{h}} w(kF_{PH} - |F_P \pm F_H|)^2 \tag{36}$$

where k is a scaling factor and w a weighting factor. Minimization may

be achieved either by a trial and error steepest descents process (Hart, 1961) or by conventional least squares equations (Lundberg, 1965). This method can work well for centrosymmetric projections and was used successfully for lysozyme and thermolysin, since a relatively large fraction of all the reflections were centric (Blake *et al.*, 1965; Matthews *et al.*, 1972a). When this method can be used, the average lack of closure error E necessary for the Blow–Crick phase determination is given *approximately* as the rms residual. It should be noted, however, that the value obtained will be an underestimate, since the sign ($\pm$) in Eq. (36) is chosen to make the residual, and therefore E, as small as possible. Because of experimental errors this choice will be incorrect for some reflections.

Refinement of a single isomorphous derivative using acentric data was attempted by Rossmann (1960) using the criterion that the best heavy atom parameters should minimize

$$S = \sum_{h} w[(kF_{PH} - F_P)^2 - F_H^2]^2 \tag{37}$$

The terms on the right suffer from the same errors as in the isomorphous replacement difference Patterson function [Eqs. (27) and (28)], and this method has not been widely adopted.

In principle, the errors inherent in Eq. (37) can be removed by the inclusion of anomalous scattering data which, taken with the isomorphous differences, allow the magnitude of the heavy atom scattering to be determined experimentally, as for the difference Patterson function [Eqs. (29) and (30)]. The difference between the calculated and observed values for F_H may then be minimized. This method was first suggested by Kartha (1965, 1969) and applied to the refinement of the ribonuclease isomorphs. It has also been used for rubredoxin (Herriott *et al.*, 1970), staphylococcal nuclease (Arnone *et al.*, 1971), cytochrome b_5 (Mathews *et al.*, 1972), and flavodoxin (Watenpaugh *et al.*, 1972).

Although the refinement techniques for single isomorphs discussed above have been very successful in a number of cases, the mode of refinement most favored at present utilizes the data from as many derivatives as are available. The technique is based upon that proposed by Dickerson *et al.* (1961). The starting heavy atom parameters are first used to calculate approximate protein phase angles ϕ_B. Each derivative is then refined in turn by minimizing (for the Jth derivative) the sum

$$S_J = \sum_{h} w[k_J F_{PHJ} - (F_P \exp i\phi_B + \mathbf{F}_{HJ})]^2 \tag{38}$$

Note that the residual being minimized is equal to the "lack of closure" vector ϵ in Fig. 4. After each derivative has been refined, the new heavy atom parameters are used to recalculate improved protein phase angles, and the refinement process is repeated until convergence has been achieved, i.e., the protein phase angles converge to essentially fixed values. Since its introduction, this method has become increasingly popular (e.g., see Kraut *et al.*, 1962; Lipscomb *et al.*, 1966; Dickerson *et al.*, 1968). Experience has shown that the process generally works best when the "best" protein phases ϕ_B are used. Different weighting schemes have been employed by different authors (e.g., see Lipscomb *et al.*, 1966; Dickerson *et al.*, 1968). Blow adopted the simple expedient of using the figure of merit of the phase determination as the weighting factor and found this to be satisfactory in the refinement of the α-chymotrypsin isomorphs (Blow and Matthews, 1973).

It is possible to reduce computer time by repeating several cycles of least squares refinement for each derivative before recalculating the phase angles. Nevertheless, convergence can still be very slow, particularly if one heavy atom isomorph dominates the phase determination, or if two or more of the derivatives are so similar that the phase probability distributions are bimodal for many reflections. This problem is due to the fact that the protein phases are determined in such a way as to be most compatible with the heavy atom coordinates of the dominant isomorph and consequently do not exert a strong corrective pressure on these parameters (e.g., see Dickerson *et al.*, 1968). Situations like this can sometimes be dealt with by omitting in turn from the phase determination the isomorph to be refined (Blow and Matthews, 1973). This procedure results in a set of protein phases that are on the whole less accurate than those obtained from all the isomorphs, but the phase angles are not biased toward the coordinates of the heavy atom derivative being refined. Although the phase angles are approximate, they can lead to rapid convergence when the conventional procedure fails to converge correctly.

In each round of least squares refinement, the "lack of closure error" E_J for the Jth derivative may be reestimated for each derivative as the rms value of the lack of closure $\epsilon_J(\phi_M)$ of the phase triangles at the most probable phase ϕ_M. The problem of *ideal* estimation of E_J and of the occupancy factors for the heavy atom binding sites remains unsolved. The E_J values obtained using phase angles from all derivatives except the Jth will be *overestimated* due to errors in the phase angles. On the other hand, bias introduced using phase angles calculated with the help of the Jth derivative will give *underestimates* of E_J and, it

seems likely, corresponding overestimates of the occupancies (Blow and Matthews, 1973).

As an overall check on the accuracy of the estimation of the E's, and therefore on the figures of merit of the phase determination, the overall ratio Q of the lack of closure error to E may be evaluated during the phase determination

$$Q_J = \langle \epsilon_J / E_J \rangle \tag{39}$$

where the mean is taken over all reflections and over each contributor (isomorphous or anomalous) to the overall phase probability distribution. If the E's are correctly estimated, Q should be approximately equal to unity.

To assess the usefulness of a given heavy atom isomorph, several "R factors" have been proposed. The one used most commonly was introduced by Cullis *et al.* (1961a):

$$R_C = \frac{\Sigma \, \| F_{PH} \pm F_P | \pm F_H |}{\Sigma \, | F_{PH} \pm F_P |} \tag{40}$$

where the summations are over the centrosymmetric reflections. A second residual, R_K, proposed by Kraut *et al.* (1962) and useful for noncentric data, has also been widely used:

$$R_K = \Sigma \, |\epsilon(\phi_M)| / \Sigma \, F_{PH} \tag{41}$$

For good derivatives R_C has been observed to be around 40%, and R_K around 10% or less (e.g., see Dickerson *et al.*, 1968). For the centric reflections it is approximately true that

$$R_K = R_C \{ \Sigma \, | F_{PH} - F_P | / \Sigma \, F_P \} \tag{42}$$

but exactly true only if the signs used in Eq. (40) are determined from all the isomorphs and not just from the isomorph being evaluated. The residual R_K suffers from the disadvantage that it tends to zero as the heavy atom scattering becomes smaller. A low value of R_K should therefore not be taken as proof of the quality of a given isomorph.

Another quantity which has been found useful in practice as an indicator of the quality of phase determination is the ratio of the mean lack of closure, E, to the average heavy atom scattering F_H, and the analogous quantities for anomalous scattering. If F_H is less than E, the resultant phase information is of marginal benefit (e.g., see Matthews, 1969).

***b.* Fourier Refinement.** The least squares method of refinement can be programmed to proceed semiautomatically and is mathemati-

cally the most satisfactory, but it has several limitations. For example, it may not refine atoms initially placed outside a certain range of convergence, and it cannot be used to locate "additional" heavy atom binding sites; also the least squares process, of itself, is not constrained by the physical realities of the situation. Thus it is useful to check and to supplement the least squares method by Fourier techniques, using as phase angles those that are available at the current stage of refinement.

A convenient synthesis for this purpose, termed the "residual" or "double difference" synthesis, has been used in a number of laboratories (e.g., see Hoppe, 1959; Blake *et al.*, 1963; Sigler *et al.*, 1966; Matthews, 1969; Blow and Matthews, 1973). It has coefficients equal to

$$m(F_{PH} - F_C) \exp(-i\phi_C) \tag{43}$$

where F_C is the amplitude of the resultant of $\mathbf{F}_P$ and $\mathbf{F}_H$ [Eq. (13)], and ϕ_C is the phase of $\mathbf{F}_C$. These coefficients correspond to the lack of closure vector ϵ in Fig. 4, weighted according to the figure of merit of the current phase determination. This synthesis is expected to reveal the difference between the true heavy atom density, differentiating a native protein and its isomorph, and the approximation to that density in terms of the current heavy atom parameters. Additional heavy atom sites not accounted for are expected to appear as positive peaks, while spurious sites will be indicated by negative regions. In the absence of all error, the "residual" electron density would be identically zero. Since $|F_{PH} - F_C|$ is generally less than $|F_{PH} - F_P|$, the error [Eq. (18a)] in a residual synthesis will be less than in the corresponding difference map.

6. *Practical Limitations*

In practice, it is difficult to obtain satisfactory isomorphous derivatives, and it is therefore of interest to know the minimum number of isomorphs necessary to solve a protein structure. Also, one would like to know how "heavy" a substituent atom needs to be for effective phase determination.

The use of a single isomorph was considered by Kartha (1961) and by Blow and Rossmann (1961). If the heavy atom arrangement is not centrosymmetric, then useful phase information can be obtained without the inclusion of anomalous scattering data. The essence of the single isomorphous replacement method is to use the weighted mean of the two possible phase angles or, what amounts to the same thing, to take the sum of the two phase solutions. In Blow and Rossmann's

method, the "best" single isomorphous phase angle is taken to be the centroid of the single isomorphous phase probability distribution $P_{iso}(\phi)$. In Kartha's method, an algebraic solution is employed which is equivalent to the Blow–Rossmann result, assuming no experimental error (i.e., $E = 0$). The most serious limitation of the single isomorphous replacement method is due to ghost images, which are related to the true electron density by inversion through any partial centers of symmetry in the heavy atom arrangement. Nevertheless, the method can be used to provide a preliminary set of phase angles from a first isomorph which may then be employed to analyze other potential derivatives by the powerful difference Fourier method discussed in Section II,C,5. In such analyses, care must be taken to distinguish real heavy atom peaks from the spurious ghost peaks. Comparison with Patterson functions and refinement of the potential heavy atom positions can be helpful in this regard.

As with the single isomorphous replacement method, useful preliminary phase angles may be obtained from a single set of anomalous scattering measurements (Matthews, 1969), possibly from a heavy atom present in the "native" structure. In the single anomalous scattering method there will be ghost images at the same positions as in the single isomorphous replacement method, but in this case they are negative and would therefore not be confused with the heavy atom binding sites (Mathews, 1969).

The combination of anomalous scattering with isomorphous replacement data for a single heavy atom provides a unique determination of the protein phase angle. Since experimental limitations will introduce errors, it is desirable to include data from additional isomorphs, although one isomorph is in principle sufficient, as has been demonstrated by Jensen and collaborators, who used a single derivative to obtain interpretable electron density maps of both rubredoxin and flavodoxin (Herriott *et al.*, 1970; Watenpaugh *et al.*, 1972). Although these proteins have relatively low molecular weights (6,300 and 16,000), there is no reason why larger proteins could not be solved by the same method, given sufficient experimental accuracy in intensity measurement. Nevertheless, there will be a definite improvement in phase determination when more than one isomorph is used, and most protein structure determinations have employed at least three heavy atom derivatives.

The question of how heavy a substituent atom should be has been considered by Crick and Kendrew (1957) and by Eisenberg (1970). With current experimental techniques it is more often the lack of isomorphism associated with heavy atom binding, rather than lack of

electron density at the heavy atom site, that is the limiting factor. Cases in which excellent isomorphism allowed the use of relatively light substitution include the comparison of pipsyl- with tosyl-α-chymotrypsin, where two iodine atoms were bound to two protein molecules having a combined molecular weight of 50,000 (Matthews *et al.*, 1967); the replacement of calcium by barium in staphylococcal nuclease (36 electrons per 16,800 MW) (Arnone *et al.*, 1971); and the substitution of mercury for zinc in thermolysin (50 electrons per 34,600 MW) (Matthews *et al.*, 1972a, 1974).

Although invariance of the unit cell parameters is necessary for perfect isomorphism, it is not a sufficient condition. It follows from the theory of Crick and Magdoff (1956) that when cell dimension changes do occur, the resultant errors are dependent upon the absolute rather than the relative cell change. This may make it more difficult, at a given resolution, to find satisfactory derivatives for proteins with large unit cells.

D. The X-Ray Image

1. Resolution

Once the phase angles of the protein structure factors have been determined, it is straightforward to calculate, using Eq. (5), the electron density distribution of the protein crystal. However, the number of Fourier components included in building up this three-dimensional image of the protein structure will, for one reason or another, be limited, and as a consequence the image will be imperfect. If the number of F's included in Eq. (5) is limited to those corresponding to minimum spacing d between the Miller planes and there are no experimental errors, then atoms closer together than about $0.71\,d$ cannot be resolved in the resultant three-dimensional image (James, 1948). Such an image is usually described as an electron density map with a resolution of d.

Since the number of reflections included in the electron density map is proportional to d^{-3}, it is apparent that maps with high resolution (small d) require many more data than those with low resolution. Furthermore, the measurement of high resolution data is made more difficult by the fact that the average intensity per reflection is less, and the intensities decrease more rapidly with X-ray exposure than do the low order reflections. Also, at high resolution the effects of nonisomorphism become much more serious (e.g., see Crick and Magdoff, 1956).

North and Phillips (1969) studied the effect of resolution in a real

case by calculating a series of electron density maps for lysozyme at increasing values of d. In going from 5.5 to 2.0 Å resolution, the number of reflections increased from 473 to 8542. North and Phillips found that α-helices were apparent even at $d = 6$ Å and that the main chain could be followed fairly well for $d \leq 4$ Å, although the carbonyl groups were not visible with $d \geq 3.5$ Å. Some side chains were visible in the 4 Å resolution map, but the S—S bridges did not have the characteristic high density observed at higher resolution ($d \leq 3$ Å). In going from $d = 2.5$ Å to $d = 2.0$ Å, the number of reflections was doubled, but the lower resolution map included very nearly as much detail and was only slightly less easy to interpret than the 2 Å map. Presumably the rather marginal improvement observed was due to the increased error in the high resolution phase angles, reflected in their lower mean figure of merit ($\overline{m} = 0.4$).

The general conclusions of North and Phillips, based upon the lysozyme maps, have been largely confirmed by other studies. For example, α-helices were readily visible in low resolution maps of myoglobin, hemoglobin, and a number of other proteins. Wyckoff *et al.* (1967b), Shotton and Watson (1970), and Muirhead and Greer (1970) were able to follow the course of the polypeptide chains of ribonuclease S, elastase, and human deoxyhemoglobin, respectively, from 3.5 Å resolution electron density maps (although, in each case, rather similar structures had been determined previously from studies at higher resolution). A number of interpretable electron density maps have been calculated at resolutions of about 2.8 Å, e.g., carboxypeptidase (Ludwig *et al.*, 1967), papain (Drenth *et al.*, 1968), and horse oxyhemoglobin (Perutz *et al.*, 1968a). North and Phillips (1969) questioned the value of extending the range of reflections included in the image beyond about $d = 3$ Å, although as subsequently demonstrated with carboxypeptidase (Lipscomb *et al.*, 1970), when the diffraction data and the heavy atom derivatives are of sufficiently high quality, a substantial improvement in the detail of the electron density map can be achieved by increasing the resolution to 2 Å or higher.

2. *Interpretation of Electron Density Maps*

In all the early crystallographic studies of proteins, the electron density map at low resolution ($d = 5$–6 Å) was calculated as a check and a stepping-stone on the way to the ultimate high resolution structure determination. Usually, low resolution electron density maps have been represented by solid models cut out so as to enclose all the electron density above an arbitrary threshold level. Such models can

show the general outline of the molecular boundary and the location of helices, but usually the structural information obtained is very limited, at least compared with high resolution studies.

The notable exceptions to this generalization were myoglobin and hemoglobin. The usually high proportion of helices allowed the course of the polypeptide chain to be followed with reasonable certainty and the approximate tertiary structure to be defined (Kendrew *et al.*, 1958; Perutz *et al.*, 1960). Furthermore, the low resolution studies of myoglobin and hemoglobin, taken together, led to the important findings that the tertiary structures of myoglobin and hemoglobin were very similar and that in hemoglobin the heme groups were far apart (in spite of the well-known cooperatively exhibited by hemoglobin on binding oxygen). In contrast, the next low resolution electron density map reported was for chymotrypsinogen (Kraut *et al.*, 1962), which has very little helix, and the map yielded disappointingly little information about the molecular architecture. One might speculate on how the evolution of the field of protein crystallography would have been altered if myoglobin and hemoglobin had turned out to have nonhelical structures or if their low resolution electron density maps had been as unintelligible as those obtained later for proteins with lower helix content.

High resolution electron density maps have been displayed in different ways. Kendrew *et al.* (1960) used colored markers clipped to a forest of rods to obtain a three-dimensional lattice into which a wire model of the structure could be built directly. The large scale (2 in. = 1 Å), the restrictions imposed by the rods, and the numerous lacerations suffered by the builders are factors which have discouraged subsequent use of this technique (e.g., see Barry and North, 1971).

Most high resolution electron density maps are first visualized by drawing contours onto sheets of plexiglass, or onto thinner transparent sheets suitably supported, to allow a number of sections of the electron density to be viewed simultaneously. Most workers prefer a scale of 2 cm = 1 Å, although supplemental "minimaps" have been used to obtain an overall impression of the quality of the map and for purposes of comparison and display.

More recent innovations include the representation of electron density maps and protein structures by means of computer controlled display systems (e.g., see Levinthal, 1966; Barry and North, 1971). Although there is no doubt that such display systems provide a very effective method of displaying, and recording on film, known structural information, it is not yet clear to what extent such systems will facilitate the interpretation of "new" electron density maps or the

refinement of an approximate structure to best fit an electron density map. Attempts to automate the interpretation of electron density maps by a "skeletonization" procedure have met with only limited success (Greer, 1974).

The "interpretation" of an electron density map can have different levels of meaning, depending upon the quality and resolution of the map, whether or not the amino acid sequence is known, and other factors. In general, however, an "interpretable" map is one in which the course of the polypeptide backbone can be followed with reasonable certainty from one end to the other. Experience has shown that prior knowledge of the primary structure of a protein helps considerably in tracing this backbone path, since the known sequence of side chains can provide continual checks that the interpretation is not seriously in error. In several cases it has been stated that it would have been impossible to follow the course of the polypeptide chain through the electron density map without knowledge of the primary structure. At this level of interpretation, it is very useful to have chemical markers to provide starting points and checks. Examples include prosthetic groups (Kendrew *et al.*, 1958), specifically labeled active site residues, amino groups, and tyrosine residues (see Section III,E,2). Also, the course of the polypeptide chain and the nature of the side chains should be consistent with the observed sites of binding of the heavy atom ligands used in the structure determination. Obviously, in the preliminary interpretation of an electron density map, there are other criteria which may be used, such as reasonable stereochemistry, no extensive regions of unexplained electron density, and no isolated buried charged groups (which is not to say that exceptions to these "rules" may not occur). At a later stage, when a detailed model of the structure is being evaluated, more stringent checks may be employed, as will be discussed below.

Generally, models of protein structures derived from electron density maps have been constructed from "Kendrew–Watson" models, consisting of skeletal components with stereochemically idealized bond lengths and angles (available from Cambridge Repetition Engineers, Greens Road, Cambridge, U.K.). With practice, these models are convenient to work with and have the great advantage of "openness," i.e., the interior of the molecule is readily visible and is accessible if adjustments are necessary. On the other hand, a real protein will not have idealized stereochemistry, so that in accurate model building, it may occasionally be necessary to slightly deform the model components to obtain the best fit with the electron density map. Also with the Kendrew–Watson components, the space-filling

properties of the model are not immediately obvious, although with practice one quickly develops an "eye" for bad van der Waals contacts and for probable hydrogen bonding distances. Painting the model components in appropriate colors helps in the visualization of their chemical identity. Rupley and Bruice (1968) have described a method of attaching Courtauld space-filling model components to the Kendrew–Watson skeletal framework. In constructing models from a known set of coordinates, it is often much quicker and more accurate to use the dihedral angles in addition to the individual atomic coordinates (Haas and Lentz, 1969; Yankeelov and Coggins, 1971). Methods of constructing models of the polypeptide backbone have been developed by Haas (1970) and Rubin and Richardson (1972).

With the exception of myoglobin, models of the first proteins to be "solved" were constructed by transferring coordinates from the electron density map to a model-building frame—an extremely tedious and not very accurate process. If unusually well-defined high resolution electron density maps were available, as with carboxypeptidase A (Lipscomb *et al.*, 1968) and rubredoxin (Watenpaugh *et al.*, 1973a), the model-building process could be eliminated, and approximate atomic coordinates could be obtained directly by placing markers into the electron density map. However, even in such cases it is usually desirable to construct a model of the protein to aid in the visualization of the structure.

Several years ago Richards (1968) developed a simple optical device which allowed the electron density map and the model to be compared directly. In Richards' optical comparator, a half-silvered mirror is placed at 45° to the electron density sections, and the model placed in such a position that an observer looking through the mirror at the electron density map sees both the map and the reflection of the model superimposed on the electron density map. The important feature of the comparator is not the use of the 45° angle, but the fact that the electron density map and the model are in mirror image positions. Another arrangement, using Richards' principle but simpler to construct, has the mirror parallel to the electron density sections (Matthews *et al.*, 1972a; Colman *et al.*, 1972a). The parallel layout allows easy alignment and has the advantages that the model is accessible from either side and the mirror being vertical, does not bend. Also, the model can be closer to the mirror which tends to improve the accuracy of model building, especially for larger molecules. The use of the mirror at 45° as devised by Richards (1968) allows photographs simultaneously showing the electron density map and the protein model to be taken normal to the electron density sections (see Fig. 12).

Experience has shown that, even with the best protein electron density maps currently available, it is not possible to determine the primary structure from the X-ray results alone. Nevertheless, it may be possible to identify many of the amino acid residues correctly. At the very least, the electron density map will provide a number of checks on the overall correctness of the sequence of amino acids as determined by chemical methods.

Several attempts at "sequence determination" by X-ray diffraction have been reported. For both myoglobin (Kendrew *et al.*, 1961) and carboxypeptidase A (Lipscomb *et al.*, 1969), parts of the amino acid sequences were known prior to calculation of the electron density maps, and attempts to identify the remainder of the primary structure were 48% correct for myoglobin and 60% for carboxypeptidase A. For thermolysin, the primary structure was determined chemically at about the same time as the electron density map (Titani *et al.*, 1972; Matthews *et al.*, 1972a). In this case, a rather cursory study of the 2.3 Å resolution map led to the correct identification of 53% of the 316 residues (Matthews *et al.*, 1972b; Colman *et al.*, 1972a). For rubredoxin, about 40 of the 54 amino acids (i.e., 74%) were identified correctly from the excellent electron density map and, in addition, "most of the rest did not differ by more than an atom or two from what they have subsequently turned out to be" (Watenpaugh *et al.*, 1973a; Herriott *et al.*, 1973). An example of the usefulness of the X-ray data in checking the chemical results is provided by papain: the electron density map (Drenth *et al.*, 1968, 1971b) showed that the tentative amino acid sequence determined chemically (Light *et al.*, 1964) differed from the correct sequence by three major transpositions and the deletion of thirteen residues. Usually, errors in an amino acid sequence determined by chemical methods will be rather different in character from those produced by X-ray crystallography. In the former case, transpositions and insertions or deletions of several residues at a time may occur, whereas in the X-ray case the most common error will result from confusion between residues which are structurally (not necessarily chemically) similar. For example, threonine, valine, leucine, and isoleucine may be difficult to differentiate, expecially the first two, which have almost identical electron densities. Thus the sequence of the general size and shape of the amino acid side chains can be defined by the X-ray method, but for unambiguous identification of each residue, particularly those which may be functionally important, it is essential to combine the chemically defined primary structure with the X-ray data.

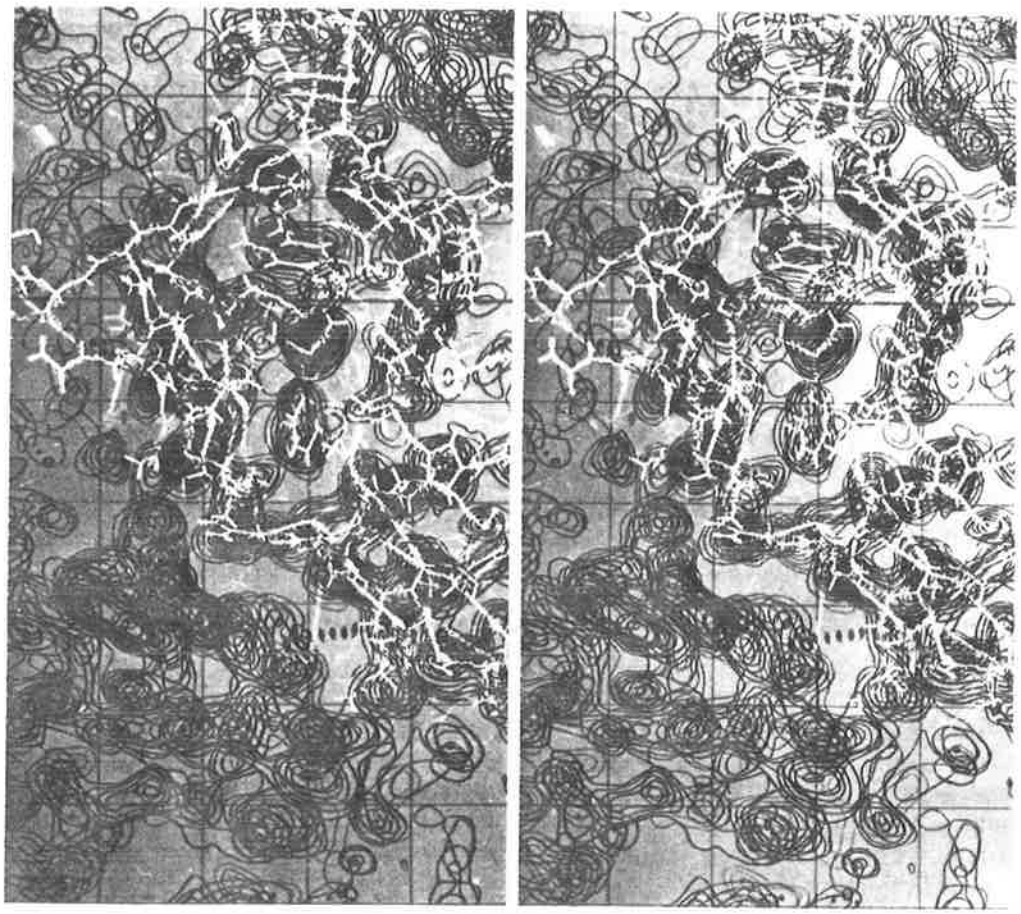
3. *Refinement of Protein Structures*

A set of atomic coordinates, measured either from a carefully constructed skeletal model or from markers placed in an electron density map, will typically have an average error of 0.2–0.5 Å for a 2.0–2.8 Å resolution electron density map. The error for some atoms may be considerably higher. In order to make strong statements about possible catalytic mechanisms and to define as precisely as possible the conformation of the protein, it is clearly desirable to reduce the positional uncertainties to the smallest possible values commensurate with the limitations of the observable data.

A typical crystal of a protein having a molecular weight of 20,000 will give about 13,000 independent X-ray reflections to a resolution of 2 Å. The same protein will contain about 1400 atoms, excluding hydrogen. Assuming that each atom can be defined by three positional parameters and a "thermal" or "disorder" parameter, and that all the atoms are independent, then if all the reflections can be measured, the ratio of observations to parameters is about 2:1. This may be compared with the diffraction data for a "small molecule" crystal which are typically measured to a resolution of $d = 0.77$ Å ($d_{\min}$ for CuKα). In this case, the ratio of observations to parameters is about 20:1.

It seems apparent, at least at first glance, that at a resolution of 2 Å the parameters necessary to define a model of a protein are not strongly overdetermined, especially if a substantial fraction of the diffraction data is experimentally unobserved. On the other hand, a great deal is known about the stereochemistry of the polypeptide backbone and of the various side chains. Although free rotation is possible about some bonds, many of the atoms are in essentially fixed relative positions so that the ratio of the number of observations to the number of *independent* parameters necessary to define the structure is in fact more favorable than might at first seem to be the case. The problem, then, is to refine the initial model parameters to obtain the best possible fit with the observed diffraction amplitudes, but at the same time to avoid violating the known rules of stereochemistry.

Conventional methods of refinement of small molecule structures, such as Fourier and least squares methods (Lipson and Cochran, 1966; Stout and Jensen, 1968), maximize the agreement between the observed diffraction pattern and that calculated for a model of the structure, but usually do so without regard to the stereochemistry of the model. This approach may be satisfactory if the parameters are strongly overdetermined, as for small molecules, but must be applied with caution to proteins. Also, consideration needs to be given to the



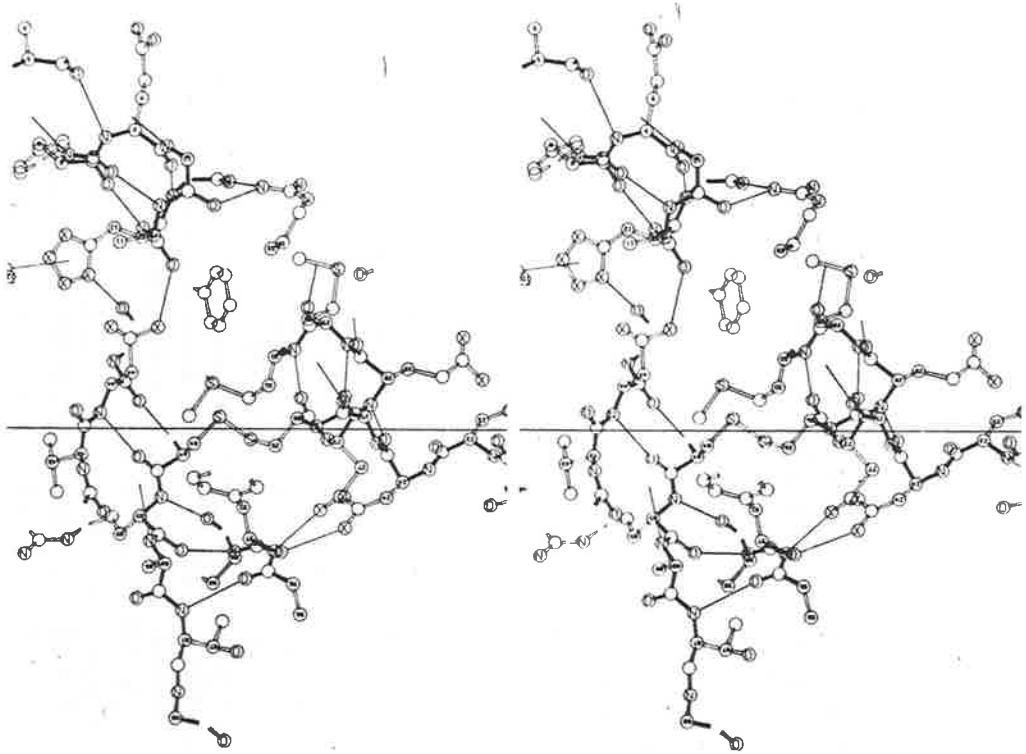


Fig. 12 Stereophotograph of part of the 2 Å resolution electron density map of ribonuclease S, with the image of the atomic model superimposed. Computer-drawn stereopictures of the model are also shown. Taken from Wyckoff *et al.* (1967b). Stereo viewers may be purchased, for example, from Ward's Natural Science Establishment, Inc., Rochester, New York, Model 25 W 2951.

fact that in a typical electron density map of a protein, the individual atoms will not be resolved, so that the refinement of one atom will not be independent of its neighbors.

In early attempts to refine the structure of myoglobin, traditional Fourier refinement was employed. In this method, the assumed (approximate) model is used to calculate structure amplitudes F_c and phase angles ϕ_c [Eq. (8)]. A new electron density map is then calculated with coefficients $F_P \exp(-i\phi_c)$, i.e., combining the observed structure amplitudes with the calculated phases. It is expected that the new map will show the atoms somewhat closer to their correct positions. Also, if atoms present in the real structure are not included in the model, their presence should be revealed in the new map by peaks of positive density. In practice, with myoglobin, additional atoms were located by this method, but the positions of "known" atoms were not improved significantly (Brändén *et al.*, 1963; Watson *et al.*, 1963; Watson, 1969). The overall correctness of the model can in principle be tested by checking the agreement between the *calculated* structure amplitudes F_c and the *observed* amplitudes F_P. The most common measure of agreement, the "R factor" (reliability index), is defined by the equation

$$R = \frac{\Sigma |F_P - F_c|}{\Sigma F_P} \tag{44}$$

For a random (incorrect) noncentrosymmetric structure, R is expected to have a value of 0.59 (Wilson, 1950). A satisfactory starting model for a small molecule crystal typically has a reliability index of less than about 30%. It has therefore always been worrisome to find that R factors calculated for protein models typically are in the range of 40%–50%, e.g., 44% for carboxypeptidase A (Lipscomb *et al.*, 1970), 43% for low resolution lysozyme data (Joynson *et al.*, 1970), 44% for subtilisin BPN′ (Alden *et al.*, 1971), and, increasing with resolution, from about 35% to 50% for α-chymotrypsin (Birktoft and Blow, 1972). With a protein, it is difficult to estimate the contribution to F_c of the solvent and more or less bound water molecules, but it is not obvious that the high values of R can be attributed entirely to these effects.

Recently, Jensen and colleagues found that *difference* Fourier syntheses, which are extensively used in small molecule crystallography, were very effective in refining the structure of rubredoxin (Watenpaugh *et al.*, 1973a). The difference Fourier synthesis has coefficients $(F_P - F_c) \exp(-i\phi_c)$ (Lipson and Cochran, 1966) and is expected to highlight any discrepancies between the model and the real structure. In principle, the refinement of a protein structure by the dif-

ference Fourier method can be done automatically with a computer, but because of the limitations of the data, steps must be taken to avoid cumulative coordinate shifts which result in impossible stereochemistry. Watenpaugh *et al.* (1973a) refined the structure of rubredoxin by visual inspection of a series of difference electron density maps and were thus able to continually check the physical reasonableness of the indicated changes in the structure. After four difference maps, R decreased from 39% to 22%. Jensen and colleagues then used a quasi full-matrix least squares method to further refine the rubredoxin structure, ultimately reducing R to 12.6%. This strikingly low value of R (for proteins) is, of course, only meaningful insofar as the model for which it is calculated is stereochemically reasonable. Watenpaugh *et al.* (1973a) have tested this point in several ways; for example, excluding the terminal residues, the rms deviation in the observed lengths of the C^{α}—C^{β} bonds was 0.20 Å, in good agreement with the standard deviation of 0.19 Å derived from the final cycle of least squares refinement.

Freer *et al.* (1975) reduced manual intervention by alternating automated difference Fourier refinement with adjustment of the model to conform to known bond lengths and angles (e.g., Diamond, 1966; Hermans and McQueen, 1974). This procedure was also successful in the refinement of carp muscle parvalbumin (Moews and Kretsinger, 1975).

Diamond (1971) has proposed a sophisticated method of refinement which minimizes the rms difference between the electron density of a model and the electron density calculated from the isomorphous replacement method. The model is considered as a flexible chain with fixed bond lengths, and refinement is in terms of rotations of dihedral angles. If desired, interbond angles may also be allowed to vary and can be made elastically stiffer than the dihedral angles. Diamond's approach has several desirable features, including the preservation of local stereochemistry and the fact that one region of the electron density map can be considered at a time, but it does not consider possible interactions between nonbonded atoms or changes in bond energy on rotation about a single bond.

Birktoft and Blow (1972) used Diamond's method to refine the α-chymotrypsin structure and found that the method did give improved fit between the model and the electron density map, but tended to introduce improbable stereochemistry in regions in which the electron density was not well defined. These stereochemical inconsistencies were minimized by subsequent "energy refinement" (Levitt and Lifson, 1969). In this technique, contributions to the conformational

energy from covalent bonds, hydrogen bonds, and nonbonded interactions are approximated by simple analytical functions, and the energy is minimized by varying the coordinates. This procedure will force atoms into stereochemically reasonable positions, but not necessarily into those which agree best with the electron density map. A combination of "real space" refinement, using Diamond's procedure, and "energy refinement" was successful in the refinement of hen egg-white lysozyme (Diamond, 1974; Levitt, 1974). Also, Huber and colleagues have used Diamond's procedure to refine the structure of basic pancreatic trypsin and the trypsin–inhibitor complex (Deisenhofer and Steigemann, 1975; Huber *et al.*, 1974).

A novel method of refinement has been proposed by Sayre (1972) and applied with some success to rubredoxin (Sayre, 1974). The technique seems at present to be most suited to small proteins for which high resolution data (1.5 Å) are available.

E. Comparison of Different Structures

It is now very clear that while there are many different globular proteins, many of these have very similar amino acid sequences and three-dimensional structures. Obvious examples include the globins, the cytochromes, and the serine proteases.

It is not the province of this review to discuss the comparison of proteins at the level of their primary structure, although there is an abundance of evidence that proteins that have homologous amino acid sequences also have similar three-dimensional structures (see Section III,D, especially Fig. 20).

It is not uncommon that two proteins are found to have related catalytic activity and/or apparent sequence homology, and the question then arises, "To what extent can the X-ray method be used to compare these structures, and what types of comparison can be made?" The answer to this question is not trivial because comparisons can be made in several different ways. Assuming that both proteins can be crystallized and that a set of "native" diffraction data can be measured for both crystals, the following representative situations might occur:

(a) The two crystal forms are "isomorphous"
(b) The crystal forms are different, but
  (i) both structures are known, at least to low resolution
  (ii) neither structure is known (i.e., no phase angles are known)
  (iii) one structure is known, the other is not

In the following sections, we will consider each of these situations.

1. *Isomorphous Structures*

If two proteins are *very* similar, i.e., they differ by at most a few amino acid substitutions, then they may crystallize in the same space group with the same unit cell dimensions and with very similar X-ray diffraction patterns. Such structures are termed isomorphous, and it is immediately obvious that the respective three-dimensional structures must be very similar. If an electron density map of one of the two structures has been calculated to some resolution d, i.e., a set of phase angles is available, then the two structures may be compared to the same resolution by calculating a difference Fourier synthesis [Eq. (31)]. Insofar as the assumed phase angles are accurate, the difference electron density map will show, to a good approximation, any changes in the electron density between the two proteins. Obviously, the detail with which these changes will be defined will depend upon the resolution of the difference electron density map and also upon the extent to which the two structures are truly isomorphous, but in general this is the most sensitive way in which two structures can be compared. The errors in a difference map are usually substantially less than in the electron density map of the protein itself (Henderson and Moffat, 1971).

Using this method of comparison, Kraut *et al.* (1967) showed that δ-, π-, and γ-chymotrypsin, the successive autolysis products of chymotrypsinogen A, all have very similar structures. Furthermore, the locations of the dipeptides successively liberated during activation were readily apparent, even in low resolution difference maps calculated with preliminary phase angles.

Perutz and colleagues have also used the difference method to determine the structural modifications present in several mutant hemoglobins (Greer, 1971; Greer and Perutz, 1971; Perutz *et al.*, 1972).

Another difference function which may be useful in comparing isomorphous structures has amplitudes of the form $(2F_1 - F_2)$. Blow and co-workers (Matthews *et al.*, 1967) determined the structure of inhibited α-chymotrypsin and wanted to obtain the structure of the active enzyme. This was achieved (Steitz *et al.*, 1969; Alden *et al.*, 1971) by using coefficients

$$(2F_P - F_{PI}) \exp(-i\phi_{PI}) \tag{45}$$

where F_{PI} and ϕ_{PI} are the amplitudes and phases determined for the inhibited enzyme, and F_P is the structure amplitude for the active enzyme. Clearly, the coefficient (45) can be rewritten as

$$F_{PI} \exp(-\phi_{PI}) + 2(F_P - F_{PI}) \exp(-\phi_{PI}) \tag{46}$$

The first term gives the electron density of the inhibited enzyme, and the second term adds the difference in density between the inhibited and native enzymes, the factor of 2 putting both densities on approximately the same scale (see Section II,C,*4,b*). A more complicated function was used by Lipscomb *et al.* (1970) to allow for partial occupancy of an inhibitor and displacement of solvent molecules from carboxypeptidase A, but this technique requires knowledge of the calculated protein structure factors. If the structure factor phases are available, then it may be advantageous to use them, rather than the isomorphous phases, in calculation of conventional difference maps (cf. Lipscomb *et al.*, 1970; Alden *et al.*, 1971).

2. *Nonisomorphous Structures*

a. **Electron Density Maps Are Available for Both Structures.** At low resolution, electron density maps may be compared by using solid models cut out to enclose all regions of electron density above a convenient threshold. Myoglobin and hemoglobin were first compared in this manner (Perutz *et al.*, 1960). Obviously, the method has its limitations but it can be very useful for preliminary comparisons and for determining the approximate orientations of related molecules in different unit cells. Once the approximate orientations of the respective molecules are known, then more quantitative comparisons can be made.

Muirhead *et al.* (1967) compared the electron density of oxy- and reduced hemoglobin by optimizing the overlap between the electron densities in the respective subunits and were able to define quite precisely the changes in the quaternary structure of hemoglobin in going from the oxy- to the reduced form. Cohen *et al.* (1970) made similar comparisons of α- and γ-chymotrypsin. In both cases it was very clear that the structures being compared were very similar, but on the other hand, the comparisons suffered from the limitations of all low resolution electron density maps, i.e., they could not be used to quantitate possible small structural differences at the atomic level.

The "high resolution" comparisons which have been made are generally in terms of selected coordinates measured from atomic models, although for subtilisin a direct comparison has been made of two high resolution electron density maps (see Section III,D).

b. **Electron Density Maps Are Not Available for Either Structure—The Rotation Function.** In this situation, one has two crystals from which are measured structure amplitudes F_1 and F_2, but no phase

angles are known. A formally similar situation occurs when a crystal has two or more subunits or molecules that are not related by crystallographic symmetry elements, and one wishes to determine their relative orientation and their structural similarity.

In the absence of phases, one cannot compare electron density maps, but the two structures can be compared indirectly by determining the similarity of their Patterson functions [Eq. (25)].

Rossmann and Blow (1962), who pioneered the application of these techniques in protein crystallography, introduced the "rotation function"

$$R = \int_u P_2(\mathbf{x}_2) P_1(\mathbf{x}_1)\, d\mathbf{x}_1 \tag{47}$$

which measures the degree of coincidence within the volume u of the Patterson functions P_1 and P_2. The volume u encloses the origin of the Patterson functions and is chosen to include, as far as possible, all vector peaks arising from atoms within one molecule but to exclude peaks from atoms in neighboring molecules. Any point $\mathbf{x}_1$ in P_1 is related to any other point $\mathbf{x}_2$ in P_2 by a rotation defined by the matrix [C], where

$$\mathbf{x}_2 = [\mathbf{C}]\, \mathbf{x}_1 \tag{48}$$

Equation (47) can be used directly for "real space" comparison. Alternatively, Eq. (47) may be expressed in "reciprocal space" quantities as

$$R = \sum_{\mathbf{p}} F_{\mathbf{p}}^2 \sum_{\mathbf{h}} F_{\mathbf{h}}^2 G_{\mathbf{h}\mathbf{h}'} \tag{49}$$

where $F_{\mathbf{p}}$ and $F_{\mathbf{h}}$ are the observed structure amplitudes of the two crystals, and G is an interference function depending upon the orientation matrix [C] and upon u (Rossmann and Blow, 1962). The real space [Eq. (47)] and reciprocal space [Eq. (49)] alternatives are formally equivalent, although in application they allow rather different strategies to be used.

Rossmann and Blow (1962) used the rotation function to determine the relative orientation of the α and β chains of hemoglobin and to show the presence of a noncrystallographic symmetry axis relating the two molecules of α-chymotrypsin in the monoclinic crystal form (Blow *et al.*, 1964). In other applications to proteins, the rotation function has been used to study hemoglobin (Prothero and Rossmann, 1964), in-

sulin (Dodson *et al.*, (1966), myoglobin (Tollin, 1969; Lattman and Love, 1970; Lattman *et al.*, 1971), lysozyme (Joynson *et al.*, 1970), aldolase (Eagles *et al.*, 1969), and lactate dehydrogenase (Adams *et al.*, 1972).

More recent developments in the theory and application of the rotation function are included in papers by Sasada (1964), Tollin *et al.* (1966), Tollin and Rossmann (1966), Lattman and Love (1970), and Lattman (1972).

In a number of applications of the rotation function and related methods (see below), it has been found that relative orientations of structurally similar molecules can in fact be determined, but that the numerical value of the rotation function is not that much higher at its maximum than that of arbitrary relative orientations of the molecules being compared. In view of the "homogeneity" of most protein structures and peak overlap from symmetry-related molecules and from intermolecular Patterson vectors, this lack of contrast is not surprising. As a corollary, the rotation function cannot be used to quantitate the degree of structural similarity of two proteins. For example, in a recent comparison of lactate dehydrogenase and glyceraldehyde phosphate dehydrogenase (Rossmann *et al.*, 1972), it was not possible to show whether or not these two proteins were structurally similar.

c. **One Structure Is Known, The Other Is Not.** If one structure or a portion of a structure is known, then the X-ray scattering amplitudes for this entity, placed in an arbitrary (large) unit cell, can be calculated and used in a rotation function calculation [Eq. (47) or (49)] to compare the known with the unknown structure. The essential advantage of knowing one structure is that its Patterson function can be calculated without error due to the overlap from intermolecular vector peaks which always occurs in a real crystal. In addition to the rotation function, a number of related methods for comparing known with unknown structures have been proposed (Hoppe, 1957; Sparks, 1961; Huber, 1965, 1969; Tollin and Cochran, 1964; Hoppe and Paulus, 1967; Nordman and Nakatsu, 1963; Braun *et al.*, 1969; Hornstra, 1969; Schilling, 1969; Lattman and Love, 1970; Joynson *et al.*, 1970; Nordman, 1972).

Having determined the relative *orientation* of similar protein molecules in the same or in two different crystal forms, it still remains to find the *translation* of the molecules relative to a common origin. This problem has been considered by Nordman and Nakatsu (1963), Rossmann *et al.* (1964), Tollin (1966), Crowther and Blow (1967), and Hoppe and Paulus (1967).

F. Novel Methods for Structure Determination

1. Direct Methods

The electron density map of a protein is not an arbitrary function but has conditions imposed by the physical properties of the electron distribution which it represents. In particular, it is nonnegative and, if the individual atoms are resolved, the electron density function and its square are very nearly alike. A number of authors have shown that these properties imply relations between the structure factors which might be used to generate additional phases from a set of known phase angles (e.g., see the review by Karle, 1964).

Sayre (1952) showed that the signs of large reflections in a centrosymmetric structure are related by the equation

$$sF_{\mathbf{h}} = s \sum_{\mathbf{k}} F_{\mathbf{k}} F_{\mathbf{h}-\mathbf{k}} \tag{50}$$

where s represents "sign of." For a protein structure, which will necessarily be noncentrosymmetric, Sayre's equation may be generalized to obtain the "tangent formula" (Karle and Hauptman, 1956; Karle and Karle, 1966)

$$\tan \phi_{\mathbf{h}} \simeq \frac{\sum_{\mathbf{k}} |E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}| \sin(\phi_{\mathbf{k}} + \phi_{\mathbf{h}-\mathbf{k}})}{\sum_{\mathbf{k}} |E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}| \cos(\phi_{\mathbf{k}} + \phi_{\mathbf{h}-\mathbf{k}})} \tag{51}$$

where $\phi_{\mathbf{k}}$ is the phase of reflection $\mathbf{k}$. $E_{\mathbf{h}}$ is the normalized structure factor, derived from the observed structure $F_{\mathbf{h}}$ by the equation

$$E_{\mathbf{h}} = F_{\mathbf{h}} \left\{ \sum_{j=1}^{N} f_j^2 \right\}^{-1/2} \tag{52}$$

where f_j is the atomic scattering factor and the summation is over all atoms on the unit cell.

Most recent applications of direct methods to protein crystallography have dealt with the feasibility of using the tangent formula to derive additional phase information from an initial set of known phases or to improve the accuracy of the starting phases. A possible starting set of phase angles might be those provided by a single isomorphous replacement (Cochran, 1955; Coulter, 1965; Karle, 1966; Hendrickson, 1971) or those from a native anomalous scatterer. It is often true in crystallography that a reciprocal space method has a real space counterpart, and direct methods are no exception. Real space

analogues of the tangent formula have been described by Hoppe *et al.* (1963), Hoppe and Gassmann (1968), and Barrett and Zwick (1971).

Weinzierl *et al.* (1969) carried out a series of tests in which the tangent formula was used to refine phase angles for a model protein and for data for cytochrome *c*. The tests, which were restricted to about 4 Å resolution, were not conclusive, but reportedly gave an apparent improvement in the electron density maps tested and indicated that tangent formula refinement might work better at higher resolution. On the other hand, Reeke and Lipscomb (1969) tested the use of the tangent formula with the known carboxypeptidase A structure and came to the opposite conclusion, i.e., that direct methods were most effective at the *lowest* resolution and produced increasingly poorer results as the resolution was increased to 2.0 Å. Barrett and Zwick (1971) attempted to predict a set of myoglobin phases between 3 and 2 Å and reported some improvement in electron density maps including these terms, although the agreement between the calculated and expected phases was only marginally better than for random phases. Coulter (1971) reported that attempts to extend the myoglobin phases from 2.0 to 1.4 Å were not very successful.

In a recent carefully executed study, Hendrickson and Karle (1973) started with a set of essentially single isomorphous replacement phase angles for carp muscle binding protein (Kretsinger and Nockolds, 1973; Nockolds *et al.*, 1972) and attempted to refine these phase angles by successive application of the tangent formula. They reported that some features which were uninterpretable in the starting electron density map were only interpreted after refinement. This is a very encouraging result, although it is qualified by the finding that an electron-dense feature in the starting map, presumably an artifact, not only increased during the refinement process but increased by a greater relative amount than the physically meaningful peaks. Sections of the respective electron density maps before and after tangent refinement are illustrated in Fig. 13. It has been pointed out by several authors that, at least in part, the application of the tangent formula tends to "square" the electron density. This effect, which is equivalent to "sharpening" the original features, is apparent in Fig. 13 and might also explain the observed enhancement of the artifact peak. In summary, at the present time the use of direct methods for proteins remains of marginal value.

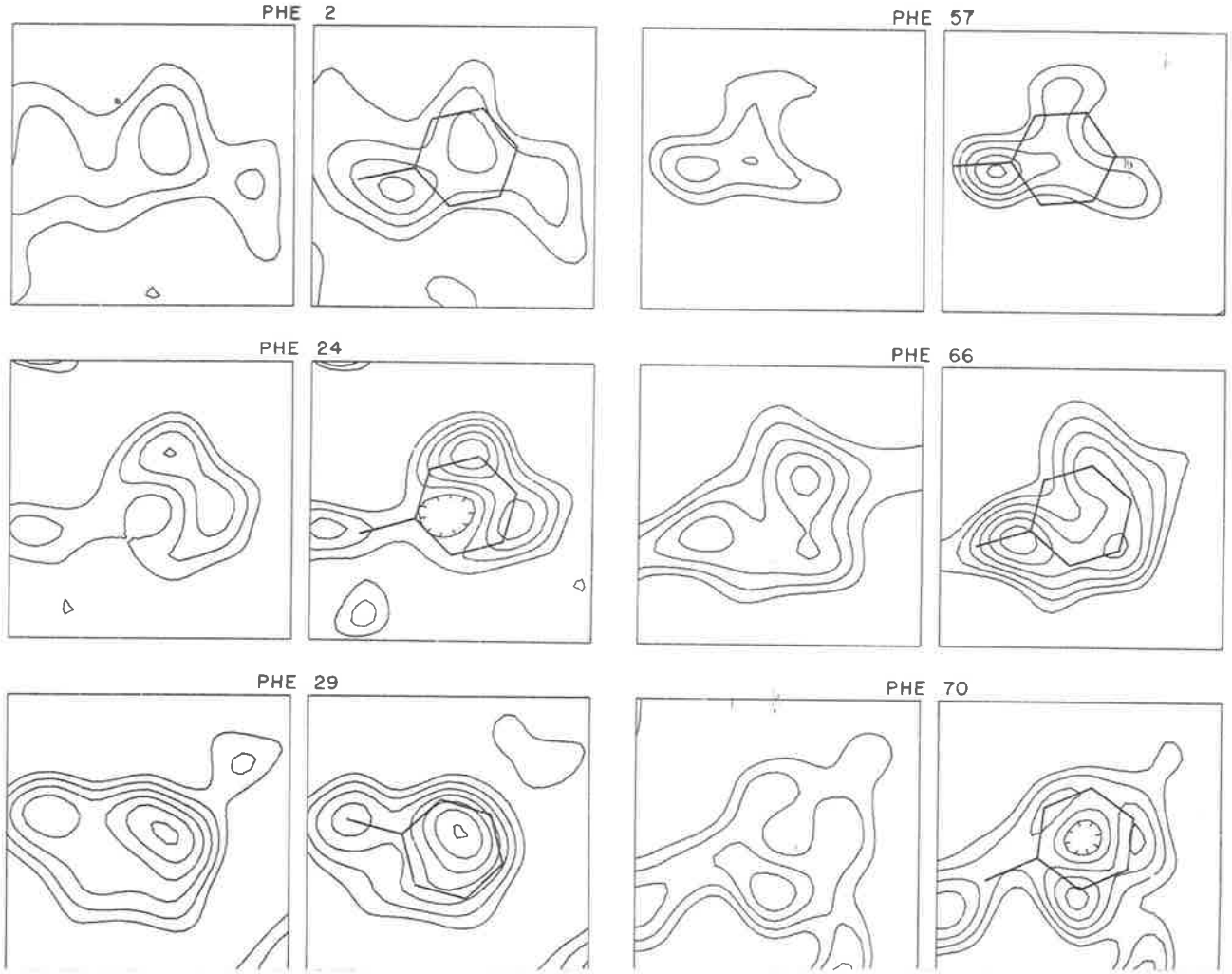
Steitz (1968) has suggested that Sayre's equation might be useful in locating heavy atoms bound to protein crystals and has tested the method with centrosymmetric data for derivatives of carboxypepti-

dase A. Steitz suggests that this method might be particularly well suited to studying multisite derivatives, as might be expected for proteins of large molecular weight. The method is unlikely to replace the calculation of difference Pattersons, but might be considered as a supplement in doubtful cases.

2. *Molecular Replacement*

A hope for the future is that it may be possible to use a known protein structure determined for one crystal form to determine the structure of similar molecules in other crystal forms, without the necessity of redetermining the phase angles by the conventional isomorphous replacement technique. The proposed procedure is first to determine the relative orientation and translation between the molecules in the two crystal forms, thus allowing the "known" molecule to be placed so that it essentially replaces the "unknown" molecule in its cell (see Section II,E,2). Structure factors and phase angles could then be calculated, followed by an inverse Fourier synthesis using the amplitudes observed for the unknown structure with phase angles calculated from the known molecule. Hopefully, the Fourier synthesis would show elements of the unknown structure which were not present in the known structure, or vice versa, and a series of Fourier refinements (see Section II,D,3) could then be utilized to eventually obtain an electron density map of the new molecule in its own unit cell. Comparisons of this sort have been commenced for the triclinic and tetragonal forms of lysozyme (Joynson *et al.*, 1970), and for sperm whale myoglobin and yellow fin tuna myoglobin (Lattman *et al.*, 1971). Recently, there has been substantial progress in this area. Schmid and Herriott (1976) have used the known structure of carboxypeptidase A to derive the structure of carboxypeptidase B and also Fehlhammer and Bode (1975) and Bode and Schwager (1975) have used the structure of trypsin as seen in the trypsin:trypsin-inhibitor complex to derive and refine the structure of crystalline-free trypsin.

Rossmann and Blow (1963) have proposed a rather different method of structure determination which may be applicable when one has a crystal form in which several molecules or subunits are crystallographically independent. In such cases, the structural redundancy leads to restrictions on the structure factors, and it is hoped that these may be used to enable direct phase determination (Rossmann and Blow, 1963, 1964; Main and Rossmann, 1966; Crowther, 1967, 1969). The practical usefulness of the method remains to be demonstrated.


PHE 2
PHE 57
PHE 24
PHE 66
PHE 29
PHE 70

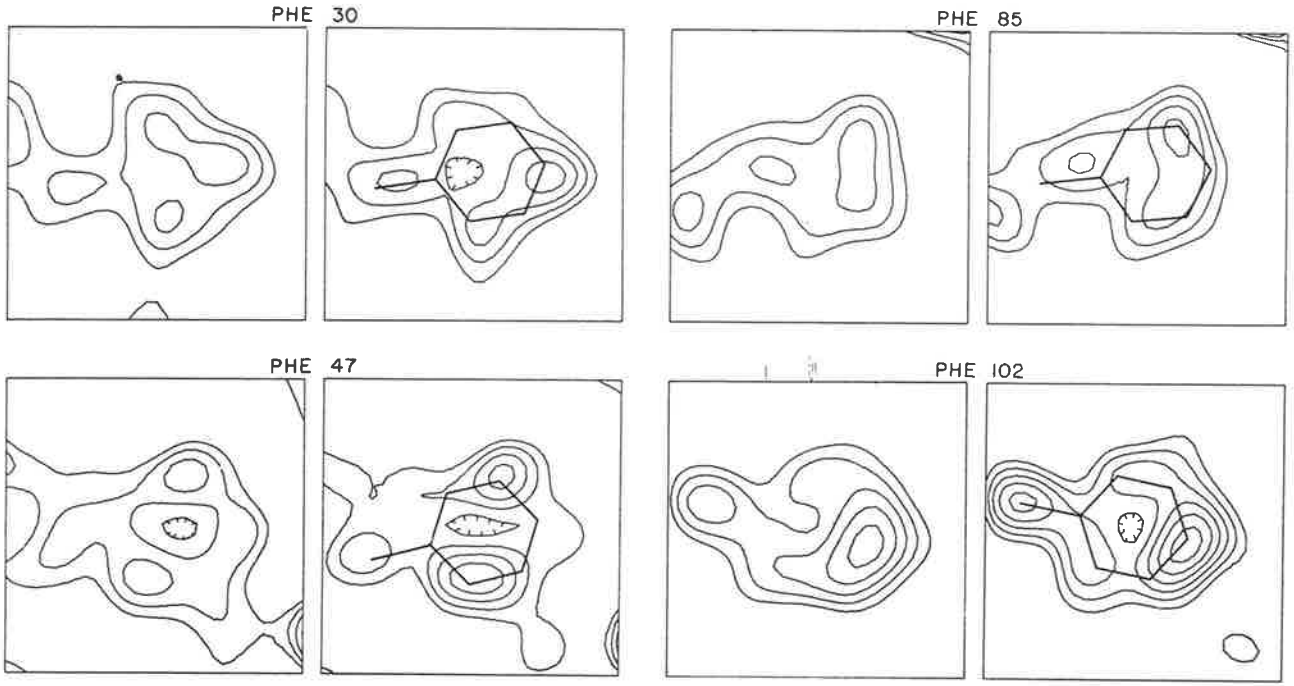


Fig. 13 Comparisons of sections through the electron density distribution of the phenylalanine side chains of carp myogen before (left frames) and after (right frames) tangent formula refinement of the phase angles. Taken from Hendrickson and Karle (1973).

G. Neutron Diffraction

As is well known, thermal neutrons have a de Broglie wavelength on the order of 1 Å and are diffracted from crystals in a manner entirely analogous to X-rays, except that the X-rays are scattered by the electrons in the crystal, whereas the neutrons interact with the nuclei. Virtually all of the theory of X-ray crystallography is applicable directly to neutron diffraction, although in practice there are a number of important differences between the two techniques. In particular, the strength of scattering of X-rays by a given atom is roughly proportional to the number of electrons associated with it, whereas neutron scattering factors (scattering lengths) can vary discontinuously with changes in atomic number or even among different isotopes. For example, the X-ray scattering amplitudes of hydrogen, deuterium, oxygen, and iron are 1, 1, 8, and 26 electrons, whereas for neutrons the scattering amplitudes of the same isotopes are −3.8, 6.5, 5.8, and 9.6 fermis. The negative value indicates that neutrons scattered from hydrogen undergo a phase shift of 360°, compared with 180° for the other isotopes listed.

An electron density map for a protein, calculated using X-ray structure amplitudes, has peak heights roughly proportional to scattering amplitude. Thus the iron would appear as a very dense feature, the oxygen would have moderate density, and hydrogen and deuterium would be indistinguishable, with density so low as to be virtually impossible to detect in current electron density maps. In contrast, a map calculated using scattering amplitudes and phase angles for neutrons would have numerically comparable peaks for each of the four atoms, but with the hydrogen density negative. Thus the unique features of the neutron diffraction technique are that it provides a method for locating hydrogen atoms and that it allows hydrogen to be clearly distinguished from deuterium. Other potential advantages of the neutron technique are that specimen damage due to irradiation is absent, nitrogen atoms may be distinguished from carbon and oxygen atoms, and large anomalous dispersion effects for isotopes such as ^{113}Ca and ^{149}Sm offer the possibility of determining the phase angles for the neutron diffraction data (Ramaseshan, 1966; Schoenborn and Nunes, 1972). The overriding disadvantage of the neutron diffraction method is that neutron sources are typically lower in intensity than conventional X-ray sources by a factor of about 10^5 (Nunes *et al.*, 1971).

Schoenborn (1969, 1971) pioneered the application of neutron scattering to crystalline proteins, with myoglobin as a test case. Using a large crystal (4 × 3 × 2 mm) it was possible to measure neutron dif-

fraction data to a resolution of 2.0 Å, although counting times of 5 min per reflection were necessary and near overlap of adjacent reflections made data collection technically difficult. Nunes *et al.* (1971) are attempting to develop a technique whereby not just one neutron wavelength is used, but rather a whole spectrum of energies. If successful, this technique would extend considerably the range of potential applications of the neutron technique.

Since the atomic scattering factors for neutrons are different from those for electrons, the neutron structure factor will differ from that for X-rays [Eq. (8)] both in magnitude and phase. Schoenborn (1971) obtained *approximate* neutron phase angles by calculation [Eq. (8)], using the myoglobin coordinates determined by X-ray analysis (Kendrew *et al.*, 1960; Watson, 1969). Since the purpose of the study was to determine whether or not hydrogen and deuterium atoms in proteins could be located by the neutron technique, these atoms were omitted from the phase calculation. By calculating a Fourier synthesis with *ob-*

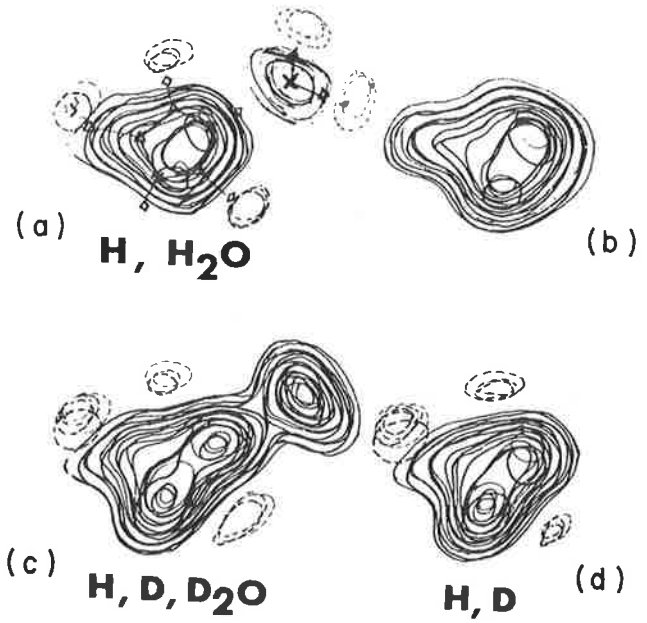


Fig. 14 Fourier representation of the neutron scattering density of four imidazole rings with different combinations of hydrogen, deuterium, and water. Hydrogen atoms appear as negative density (broken contours), and deuterium, oxygen, and nitrogen as positive density (solid contours). (a) H atoms are on all positions except $N^{\delta 1}$; H_2O is hydrogen-bonded to $N^{\epsilon 2}$. (b) No H or water indicated. (c) Same as in (1) except D is on $N^{\epsilon 2}$ and $N^{\delta 1}$. (d) Same as (3) with water omitted. Taken from Schoenborn (1971).

served neutron and structure amplitudes and *calculated* phases [cf. Eq. (1)], one should obtain a map of the approximate neutron scattering density, i.e., C, N, O, D would appear as positive peaks, but H negative. The myoglobin crystals used for data collection were soaked in a mother liquor containing about 75% deuterium and 25% hydrogen atoms to reduce the incoherent scattering. Thus some (hydrogen-bonding) hydrogens in the native myoglobin molecule were substituted by deuterium, whereas others (covalently bonded) were not (Fig. 14). In Fig. 14, negative features indicating hydrogen atoms are clearly visible. The water molecule has its oxygen positive and the two hydrogens negative, whereas the D_2O has all three peaks positive, although not resolved. Schoenborn (1971) suggests that it may be possible to improve the calculated neutron phase angles by a cyclic process analogous to conventional Fourier refinement (see Section II,D,3). Also, it is hoped that suitable myoglobin derivatives will allow direct phase determination by the anomalous scattering technique.

III. PROTEIN CRYSTALS

A. Solvent Content

The nature of a protein crystal may be illustrated by Escher's lithograph (Fig. 15), in which the birds can be considered as protein molecules and the fish as intervening solvent. The solvent region is in free and rapid equilibrium with the supernatant surrounding the crystal and is not to be compared with water of crystallization found in small molecule crystals. The contacts between the protein molecules (birds) are few in number and rather tenuous. As a consequence, the forces stabilizing the crystal are rather weak, which has the disadvantage that the crystals are much less well ordered than small molecule crystals and also are very fragile. On the other hand, since the intermolecular forces are weak, it is unlikely that they will substantially modify the three-dimensional structure of the protein as it crystallizes. Also, the presence of the solvent channels through the crystal may allow reactions to be carried out *in the intact crystal*. The symmetry aspects of Fig. 15 will be discussed in Section III,C,3.

In studying any new protein crystal, one wishes to determine the number of molecules per unit cell. This may be estimated from the unit cell volume and from crystal density measurements (see Section III,C,2), but for small or very fragile crystals such measurements may

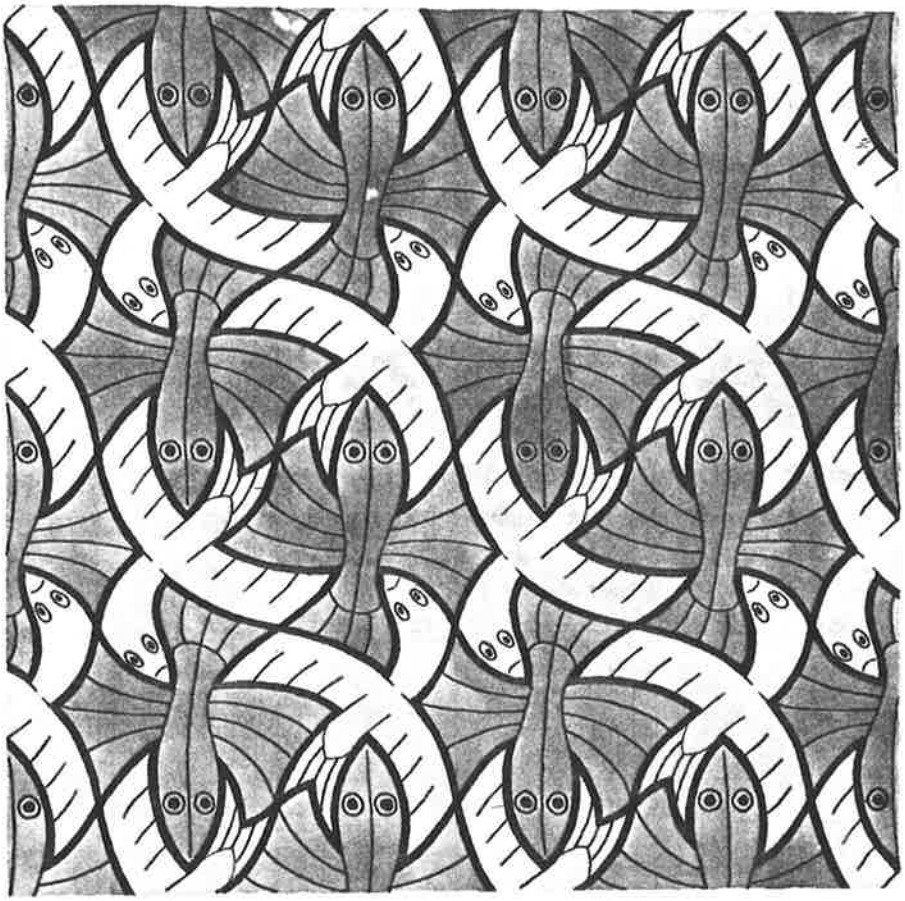


Fig. 15 Escher lithograph to illustrate the nature of a protein crystal. Taken from "Symmetry Aspects of M. C. Escher's Periodic Drawings," C. H. MacGillavry (1965). Original by M. C. Escher. (I am most grateful to the owners, The Escher Foundation, Haags Gemeentemuseum, The Hague, for permission to reproduce this figure.)

not be possible. It is therefore of interest to know the range of values that might be expected for the solvent content of a typical protein crystal, since this will place approximate upper and lower limits upon the number of protein molecules in the unit cell.

The solvent content of a protein crystal may be conveniently characterized by V_M, the ratio of the volume of the unit cell to the molecular weight of protein in the cell. It can readily be shown (Matthews, 1968) that V_p', the fraction of the crystal volume occupied by protein, is given by

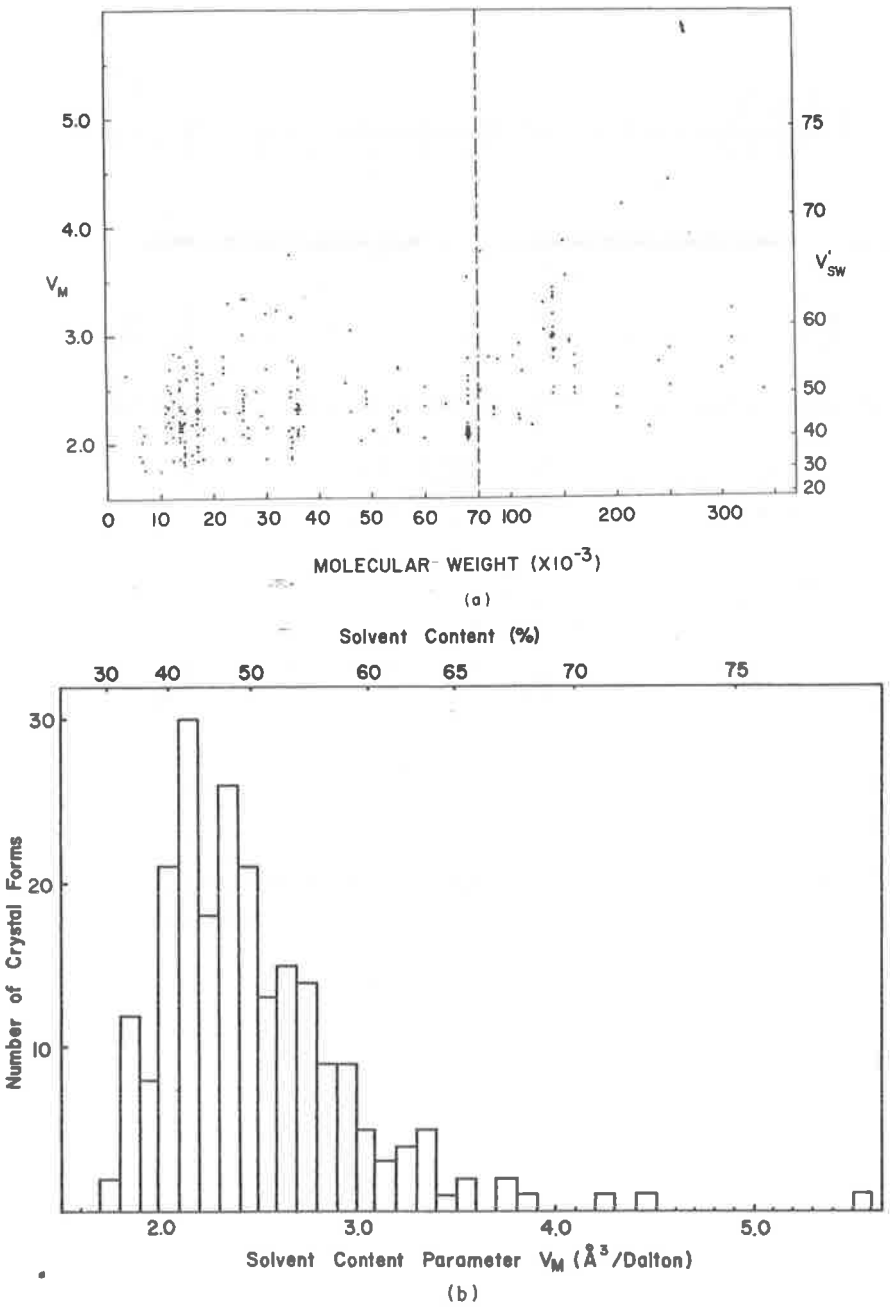


Fig. 16 (a) Solvent content of 226 protein crystals. V_M is the solvent content parameter (see text) and gives the number of cubic angstroms in the crystal per unit of protein molecular weight. V'_{sw} is the volume % of the crystal occupied by solvent. The abscissa is molecular weight per molecule rather than molecular weight per asymmetric unit, as was used in the survey of Matthews (1968). Note that the molecular weight scale changes at 70,000 daltons. (b) Histogram showing the distribution of V_M and of the solvent content of the 226 crystal forms plotted in Fig. 16a.

$$V'_p = 1.66\ \bar{v}_p/V_M \tag{53}$$

where $\bar{v}_p$ is the partial specific volume of the protein in the crystal, and V_M has dimensions Å^3/dalton. (We will use the prime to distinguish V_p, the *fractional* unit cell volume occupied by protein, from V_p, the *absolute* volume.) For most proteins v_p is approximately 0.74 ml/gm, so that unless there is reason to believe that the protein in question has an abnormally low or high partial specific volume, or that the partial specific volume has a different value in the crystal from that in dilute solution, the volume fraction occupied by protein is approximately given by

$$V'_p = 1.23/V_M \tag{54}$$

By difference, the fractional volume V'_{sw} occupied by solvent is

$$V'_{sw} = 1 - V'_p = 1 - 1.23/V_M \tag{55}$$

V'_{sw} includes both the "free solvent" and the "bound water" discussed in Section III,C,2. In the present context, the "solvent content" of a crystal is meant to include the total volume not occupied by protein.

In Fig. 16a, the solvent contents of 226 protein crystals reported prior to 1973 are summarized by plotting V_M as a function of protein molecular weight. The summary includes 116 crystal forms reported in an earlier survey (Matthews, 1968) and those listed in Tables I, II, and IV. Now that the sample is larger, particularly for proteins of higher molecular weight, it is much more obvious that V_M tends to increase with molecular weight, as was suggested previously (Matthews, 1968).

A histogram showing the frequency distribution of V_M for all the crystals surveyed is shown in Fig. 16b. The distribution is not symmetric, but as we shall show, this is largely due to the reciprocal relation between the protein content and V_M [Eq. (54)] and does not necessarily imply that the observed cutoff is sharper at low solvent content than at high solvent content. To obtain the frequency distribution of observed solvent contents we have used Eq. (55) to estimate V'_{sw} for each of the 226 crystal forms. This is obviously an approximation but will be quite adequate for the present purpose. In view of the apparent increase in solvent content with protein molecular weight, the molecular weights were grouped into four convenient ranges. The resultant distributions of solvent content within each range, and for all the crystals, are illustrated in Fig. 17 and summarized in Table III. It is apparent that the distribution of the solvent content is much more symmetrical than that of V_M. Also, it can be seen that within each molecular weight range the spread of the solvent content or of V_M is

TABLE II

Crystal Forms of Monomeric Proteins[a]

Protein	Source	Molecular weight[b]	Space group[c]	V_M (Å³/dalton)	Technique[d]	Reference[e]
Trypsin inhibitor	Bovine pancreas	6,500 (2)	$P2_1$	2.02		Huber *et al.* (1970)
Cobrotoxin	Cobra	6,950	$P4_12_12$	2.09		Wong *et al.* (1972)
Erabutoxin	Sea snake	7,000	$P2_12_12_1$	1.76		Low *et al.* (1971)
Ribonuclease	*B. subtilis*	10,700 (3)	$P3_1$	2.55		A. J. Wonacott and D. M. Blow, personal communication
Albumin	Carp	11,000 (4)	$P2_1$	2.29		Kretsinger *et al.* (1971a)
Cytochrome b_5	Rabbit	11,000 (2)	$P2_12_12_1$	2.36		Kretsinger *et al.* (1970)
Bence Jones fragment	Human	11,500 (2)	P2	2.35		Solomon *et al.* (1970)
Thioredoxin	*E. coli* β	11,700 (2)	C2	2.69		Holmgren and Söderberg (1970)
Cytochrome b_{562}	*E. coli*	12,000 (2)	$P4_122$	2.48		Czerwinski *et al.* (1972)
Cytochrome b_{562}		12,000 (?)	P1	–		Czerwinski *et al.* (1972)
Cytochrome b_{562}		12,000 (2)	P1	2.16		Czerwinski *et al.* (1972)
Cytochrome *c*	Rice	12,700	$P6_1$	2.40		Morita and Ida (1972)
Lysozyme	Human	14,500	$P2_12_12_1$	1.98	IR, 6 Å	Blake and Swan (1971)
Lysozyme	Human	14,500 (2)	$P2_12_12_1$	2.72		Osserman *et al.* (1969)
α-Lactalbumin	Goat's milk	14,500 (2)	I222	2.20		Aschaffenburg *et al.* (1972a)
α-Lactalbumin	Goat's milk	14,500 (2)	$P2_1$	2.29		Aschaffenburg *et al.* (1972b)
Lysozyme	Duck egg-white	14,600	$P2_1$	1.90		Berthou *et al.* (1972b)
α-Lactalbumin	Cow's milk	15,500 (?)	A2	–		Inman and Bryan (1966)
Flavodoxin	*Desulfovibrio vulgaris*	16,000	$P4_32_12_1$	2.9	IR, 2.5 Å	Watenpaugh *et al.* (1973b)
Myoglobin	Blue fin tuna	17,000	$P2_12_12_1$	2.31		Kretsinger (1968)

Myoglobin	Skipjack	17,000	$P2_1$ (?)	–		Kretsinger (1968)
Myoglobin	Albacore	17,000 (2)	$P2_1$	2.53		Kretsinger (1968)
Myoglobin	Yellow fin tuna	17,000	$P2_12_12_1$	2.47	RF	Lattman *et al.* (1971)
α-Lytic protease	*Sorangium* sp.	20,000	$P3_121$	2.56		James and Smillie (1969)
Protease	*Arthrobacter*	23,000 (4)	$P2_12_12_1$	2.02		Eklund *et al.* (1968)
Protease	*Arthrobacter*	23,000 (8)	$P2_1$	2.61		Eklund *et al.* (1968)
Bence Jones protein	"Rei"	23,000	$P6_1$	3.3	IR, PS, 4 Å	Epp *et al.* (1972)
Agglutinin	Wheat germ	23,000 (2)	C2	1.86		Wright *et al.* (1972)
Carbonic anhydrase β	Human	30,000	$P2_12_12_1$	1.86		Kannan *et al.* (1972)
Carbonic anhydrase	Bovine	30,000	$P6_122$	2.69		O. Lindqvist and U. Carlsson, private communication
Trypsin-trypsin inhibitor	Bovine	30,000	I222	3.2		Rühlmann *et al.* (1971)
Peroxidase *c*	Radish	32,000	P422(2)	3.23		Morita (1972)
Thermolysin	*B. thermoproteolyticus*	34,600	$P4_122$	3.74		Colman *et al.* (1972b)
Protease	*Penicillium janthinellum*	35,000	C2	2.15		Camerman *et al.* (1969)
Protease	*Endothia parasitica*	36,000	$P2_1$	2.36		Moews and Bunn (1970)
Protease	*Endothia parasitica*	36,000	$P2_1$	2.08		Moews and Bunn (1970)
Phosphoglycerate kinase	Yeast	45,000	C2	2.56	IR, 5 Å	Wendell *et al.* (1972)
Phosphoglycerate kinase	Horse	48,000	$P2_1$	2.03	IR, 6 Å	Blake *et al.* (1972)
Fab fragment "Hil"	Human	55,000 (2)	$P2_12_12_1$	2.13		Humphrey *et al.* (1969)

(*Continued*)

TABLE II (*Continued*)

Protein	Source	Molecular weight[b]	Space group[c]	V_M (Å³/dalton)	Technique[d]	Reference[e]
Fab fragment "Smo"	Human	55,000 (2)	P3,21	2.12		Humphrey *et al.* (1969)
Fab fragment "New"	Human	55,000	C2	2.29		Avey *et al.* (1968)

[a] This table does not include proteins listed in Table I or in the earlier survey of Matthews (1968). Also, eight crystal forms of lamprey hemoglobin (Hendrickson *et al.*, 1968) with six or more monomers per unit cell have been omitted. This compilation was made at the end of 1972.

[b] The number of monomers in the crystallographic asymmetric unit, if different from unity, is indicated in parentheses or, if doubtful, by a question mark.

[c] Enantiomorphic or alternative space groups are not included.

[d] If a three-dimensional electron density map has been calculated, the resolution is given. Other abbreviations are IR, isomorphous replacement; RF, rotation function; PS, Patterson search.

[e] In general, reference is given to the most recent publication.

[f] Fenna (1973) has described a triclinic crystal form of α-lactalbumin, pseudo A2, with cell dimensions almost identical to those reported by Inman and Bryan (1966). This is very probably the same crystal form. Inman and Bryan (1966) suggest that there are 16 α-lactalbumin molecules per cell, implying $V_M = 5.3$ Å³/dalton, which seems improbable. On the basis of crystal density measurements, Fenna (1973) favors 32 or 40 molecules per cell, i.e., $V_M = 2.7$ or 2.1 Å³/dalton.

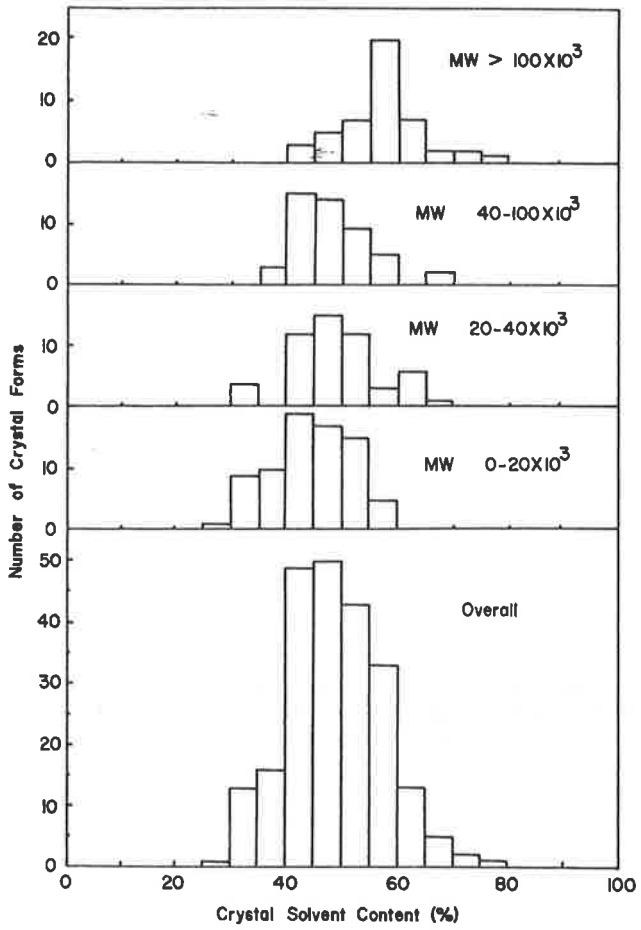


Fig. 17 The frequency distribution of the solvent contents of 226 different crystal forms of globular proteins, shown for different ranges of protein molecular weight, and for all forms.

less than for the data taken as a whole. Typically, the high and low values of V_M for each range (Table III) differ by a factor of about 2. This suggests that if one obtains a new crystal with a value of V_M (or solvent content) near the middle of the range, then in most cases a comparison with other protein crystals will probably be sufficient to determine the number of molecules in the unit cell, but if, on the other hand, values of V_M near the extremes of the ranges are possible, then the number of molecules per cell may be determined with less certainty. It is, of course, desirable to confirm the number of molecules by other methods whenever possible (see Section III,C,2).

TABLE III

Solvent Content of Protein Crystals

MW Range	No. in sample[a]	Minimum	Maximum	Average	Most common
Summary of values observed for V_M (Å³/dalton)					
0–20,000	76	1.75	2.90	2.24	2.1
20,000–40,000	56	1.86	3.74	2.45	2.3
40,000–100,000	47	2.03	3.76[b]	2.40	2.2
>100,000	47	2.14	5.60	2.96	2.9
Overall	226	1.75	5.60	2.47	2.2
Summary of values observed for fractional solvent content V'_{sw} (%)					
0–20,000	76	30	58	44.3	44
20,000–40,000	56	34	67	48.4	47
40,000–100,000	47	39	67[b]	48.1	46
>100,000	47	43	78	56.9	57
Overall	226	30	78	48.8	47

[a] The crystals included in this summary are taken from Tables I, II, and IV and from the earlier compilation of Matthews (1968). Crystal forms of ferritin, edestin, excelsin, and tobacco-seed globulin, included in the earlier survey, have been omitted.

[b] The structure of a mouse myeloma Fab fragment at 3.1 Å resolution has recently been reported by Segal *et al.* (1974). $V_M = 4.2$ Å³/dalton and the solvent content is about 71%, assuming a molecular weight of 55,000.

It is interesting to note that the most tightly packed crystal, *Chromatium* high-potential iron protein (Carter *et al.*, 1971), has an estimated solvent content of 30% (V_M = 1.75 Å³/dalton) which is still higher than the value of 26% expected for the closest packing of spheres. At the other end of the scale, yeast lysine transfer ribonucleic acid ligase (Rymo *et al.*, 1970; Lagerkvist *et al.*, 1972) has a V_M of 5.6 Å³/dalton, which corresponds to a solvent content of 78%. This very tenuous packing of the molecules has been confirmed by density measurements (Rymo *et al.*, 1970; Lagerkvist *et al.*, 1972; U. Lagerkvist, personal communication) which led to an X-ray molecular weight per asymmetric unit of 71,000 [Eqs. (61) and (62)], in good agreement with the values determined by centrifugation and by gel chromatography. Although the value of 5.6 Å³/dalton for V_M is well above that found for any other protein crystal (Fig. 16a), the corresponding solvent content is not that much higher than for catalase (72% solvent) (Rossmann and

Labaw, 1967; Longley, 1967) and for β-amylase (71% solvent) (Colman and Matthews, 1971). Like almost all protein crystals with high solvent content, the lysine tRNA ligase crystals do not give high resolution diffraction patterns and are unusually sensitive to X-ray exposure, consistent with the idea that crystalline perfection and stability are determined by the number and strength of the intermolecular interactions in the crystal rather than by the intrinsic stability of the protein molecule.

There is probably a correlation between the fact that the majority of protein crystals used for successful structure determinations (Table I) have low values of V_M and therefore tend to give better diffraction patterns, but it must be added that not all successful determinations fall into this category. Oxidized cytochrome *c* (Dickerson *et al.*, 1971) and lactate dehydrogenase (Adams *et al.*, 1970a), for example, both have a V_M near 3.0, corresponding to solvent contents of nearly 60%. Also, medium resolution studies have recently been completed for a mouse myeloma Fab fragment (Segal et al., 1974) and a bacteriochlorophyll protein (Fenna and Matthews, 1975), both of which have a V_M of 4.2 Å^3/dalton (71% solvent).

B. Crystallization

As Eisenberg (1970) has aptly put it, "Growing protein crystals is still more of an art than a science." In general terms, one must find a solvent in which the protein is soluble and reasonably stable, and then approach conditions of insolubility in such a way that the protein comes out of solution as a relatively small number of large crystals, rather than as tiny crystals, or as an amorphous precipitate. On occasion, however, crystals have been found to grow from the precipitate. Many variables can affect this process, including the nature of the precipitant, pH, temperature, presence of electrolytes and nonelectrolytes, protein concentration and purity, presence of nucleation sites, and chemical modification of the protein.

The great majority of crystals used for X-ray diffraction studies have been grown by "salting out" with solutions of ammonium sulfate. Use of alcohols and diols to reduce the protein solubility has also been successful in several cases (e.g., see King *et al.*, 1956, 1962; Drenth *et al.*, 1962; Cotton *et al.*, 1966). Some proteins have very low solubility in water, and use has been made of this fact to obtain crystals (e.g., see Perutz, 1946; Lipscomb *et al.*, 1966).

Experimentally, three main routes to crystallization have been employed. In each of these, crystals may grow overnight or may take weeks or longer.

1. Batch Crystallization

This is the traditional method. The protein solution, usually about 10 mg/ml, is divided among a number of vials, and the precipitating solution slowly added until precipitation, heralded by onset of a faint opalescence, is approach. The vial is then sealed and left along with other vials with slightly lower concentrations of precipitant. By this method it is possible to rapidly screen a variety of precipitants over a range of pH. Using narrow vials (capacity 1 ml), starting aliquots of 0.02 ml may be used, i.e., about 0.2 mg of protein per experiment.

It is often stated, but untrue, that batch crystallization requires tens of milligrams of protein per experiment. In fact, Dobler *et al.* (1972) have recently described a method of batch crystallization in which the effects of protein and ammonium sulfate concentrations were explored by observing 0.003-ml droplets sealed in 1-mm melting point capillary tubes. When conditions for optimal crystallization have been obtained, the crystallization process may be carried out with increased volumes in larger tubes or vials.

2. Vapor Diffusion

Another method, also requiring only small amounts of material, has evolved largely from attempts to crystallize tRNA, which was available only in milligram amounts. The macromolecule in solution is sealed as a droplet on a depression slide in a container along with a large reservoir. The concentration of the reservoir can be adjusted from time to time to promote diffusion of vapor into or out of the droplet (e.g., see Hample *et al.*, 1968). The crystal growth can be monitored under a microscope and, if unsatisfactory, the conditions may be manipulated to achieve dissolution of the protein, followed by subsequent recrystallization (P. B. Sigler, personal communication). In another variation of the vapor diffusion technique the droplet of protein solution is placed on a glass slide which is then inverted to form a "hanging drop."

3. Equilibrium Dialysis

Crystal growth by diffusion of salts through dialysis membranes has been used in the past but has been given impetus by the success of Zeppezauer *et al.* (1968) with microdiffusion cells requiring 0.01 ml of protein solution per experiment. Zeppezauer *et al.* (1968) and Weber and Goodkin (1970) have described two possible experimental arrangements and have discussed some of the factors influencing the growth of crystals and diffusion rates.

C. Preliminary Information

1. *Unit Cell and Space Group*

The unit cell and space group of a protein crystal can usually be determined in a straightforward way from Buerger precession photographs (see Section II,B). Typically, the smallest protein crystals usable for such experiments are no less than 0.05 mm in their smallest dimension. The size and shape of the unit cell and the distribution of the symmetry elements may give a clue as to the overall dimensions of the protein molecules, but such extrapolations must be made with the greatest caution, since the molecules may take up arbitrary orientations relative to the cell edges and almost invariably extend across the boundaries of the unit cell (or of the asymmetric unit).

2. *Molecular Weight*

Although a number of techniques exist for the determination of the molecular weights of proteins in solution, such determinations may be subject to error, particularly if the macromolecule under study tends to aggregate or dissociate. In contrast, with a protein crystal the same molecular species is present in each asymmetric unit and, as we shall describe, its molecular weight can be determined accurately from the basic crystallographic parameters and from crystal density measurement, usually made in a calibrated density gradient (Low and Richards, 1952). Colman and Matthews (1971) have described a convenient method of introducing the crystals.)

M_c, the molecular weight in daltons of the contents of one unit cell of a crystal, may be determined directly from the unit cell volume V and the crystal density D_c as

$$M_c = NVD_c \tag{56}$$

where N is Avogadro's number. If there were n molecules in the unit cell and nothing else, the molecular weight of one molecule would be

$$M_P = NVD_c/n \tag{57}$$

For a protein, the determination of molecular weight from the crystallographic data is complicated by the fact that the crystal contains not only the protein molecules but also the intervening solvent. Furthermore, the nonprotein part of the unit cell is not homogeneous, but, as shown by Perutz (1946), consists of two parts: a "free solvent" region accessible to external ions and in equilibrium with the external supernatant, and a "bound water" region, not accessible to salts,

which can be visualized as a hydration shell surrounding the protein.

The density D_c of the protein crystal can be written as the sum of three parts

$$D_c = D_p(V_p/V) + D_w(V_w/V) + D_s(V_s/V) \tag{58}$$

where D_p, D_w, and D_s are the densities of the protein, bound water, and free solvent; (V_p/V), (V_w/V), and (V_s/V) are the fractional volumes occupied by these three components; and V is the unit cell volume (e.g., see Colman and Matthews, 1971).

If the ratio of the bound water volume to the protein volume is denoted f, i.e.,

$$f = V_w/V_p \tag{59}$$

then it follows from Eq. (58) (Matthews, 1974) that M_p, the molecular weight of the protein, is given by

$$M_p = \frac{NVD_p(D_c - D_s)}{n\{D_p - D_s - f(D_s - D_w)\}} \tag{60}$$

Making the simplifying assumption that the density of the protein is equal to the reciprocal of its partial specific volume $\bar{v}_p$, Eq. (60) leads to

$$M_p = \frac{NV(D_c - D_s)}{n\{1 - \bar{v}_p D_s - w[(D_s - D_w)/D_w]\}} \tag{61}$$

where w $(=f\bar{v}_p D_w)$ is the *weight fraction of bound water, measured in* grams water per gram protein.

Occasionally, protein crystals are grown from water or from solutions of very low ionic strength for which the density is equal to that of water, in which case Eq. (61) reduces to

$$M_p = \frac{NV(D_c - D_w)}{n(1 - \bar{v}_p D_w)} \tag{62}$$

In this special case, the molecular weight of the protein can be obtained with quite high precision (±5% or less) from a knowledge of the unit cell volume, the number of molecules in the unit cell, the partial specific volume of the protein, and the density of the crystals. However, most protein crystals are grown from concentrated salt solutions and dissolve quite quickly in water, so that Eq. (62) is not applicable.

Following the early studies of Perutz (1946), Colman and Matthews (1971) pointed out that it is possible to determine the molecular weights of protein crystals, even in the presence of concentrated salt

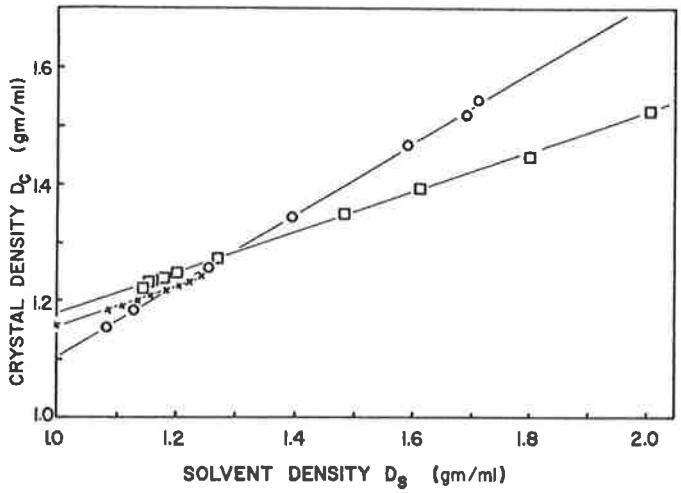


Fig. 18 Plots of crystal density as a function of supernatant density showing the linear relation between them. (□) γ-chymotrypsin; (○) β-amylase (from Colman and Matthews, 1971); (×) hemoglobin. Adapted from the data of Perutz (1946).

solutions, by making use of the fact that if the solvent density is varied, then from Eq. (58) the crystal density will also vary. A plot of D_c against D_s will yield a straight line of slope (V_s/V) and of intercept D_o, the density which the crystal would have if all the solvent could be replaced with water. Application of this method to γ-chymotrypsin and to β-amylase (Colman and Matthews, 1971) is illustrated in Fig. 18. Additional data for hemoglobin, adapted from the results of Perutz (1946), have been included. In each case, the observed points lie on straight lines which may be extrapolated to find the hypothetical crystal density D_o of the crystal in water. Substituting D_o for D_c in Eq. (62), the protein molecular weight may be obtained with an accuracy of about ±4% in a typical case. For hemoglobin, Perutz (1946) was able to obtain the same crystal form from water and from concentrated salt solutions, allowing direct verification of the density extrapolation.

The density plots (Fig. 18) also provide additional information about the unit cell components. The slope gives the volume fraction of free solvent in the unit cell, and the ratio of the volume of bound water to the volume of protein is given by

$$f = D_p w = (D_p - D_e)/(D_e - 1) \tag{63}$$

where D_e is the equilibrium density at which $D_c = D_s$ (Adair and Adair, 1936; Colman and Matthews, 1971).

Note that the crystal density measurements do not determine $\bar{v}_p$, the

partial specific volume of the protein, which is assumed in the equations given here to be equal to $1/D_p$.

Another approach to the determination of molecular weights of proteins crystallized from concentrated salt solutions is by the use of cross-linking reagents such as glutaraldehyde. Although cross-linking of protein crystals causes an increase in their density and therefore an increase in the apparent molecular weight of the protein, the judicious use of cross-linking reagents does provide a quick, easy, and moderately accurate method of estimating molecular weights (Cornick *et al.*, 1973; Matthews, 1974). Cross-linked protein crystals can often be placed into water and will maintain their integrity for at least several hours, as judged by the fact that they still give diffraction patterns similar to those of the native protein crystals (Fig. 1). Cornick *et al.* (1973) found that crystals of α-chymotrypsin soaked for 8–24 hr in a 0.25% solution of glutaraldehyde in amine-free phosphate were stable when transferred into water, allowing the unit cell dimensions and crystal density in the water environment to be determined. Using Eq. (62), the molecular weight of α-chymotrypsin was found to be within 3% of the value determined from the amino acid sequence. On the other hand, Matthews (1974) found that crystals of γ-chymotrypsin required overnight soaking in a 2% glutaraldehyde solution in 55% saturated ammonium sulfate to prevent dissolution upon transfer into water. The densities of such cross-linked γ-chymotrypsin crystals, plotted as a function of supernatant density, are shown in Fig. 19. As with the native crystals, the density relationship is linear, and when extrapolated to a solvent density of unity, is in good agreement with the observed density of the cross-linked crystals in water. For cross-linked γ-chymotrypsin crystals, the unit cell volume was observed to expand slowly as a function of immersion time in water, the fractional increase being 6% after one day. In contrast, the diffusion of salts into and out of the crystals proceeds rapidly, being essentially complete within 10–20 min (Wyckoff *et al.*, 1967a; Matthews, 1974).

The density plots in Fig. 19 show that the glutaraldehyde causes an apparent increase of 14% in the molecular weight of γ-chymotrypsin and of 25% in β-amylase. This is obviously much more than would be expected for simple cross-linking of lysine residues and suggests that a substantial amount of polymerization must occur or that more than one glutaraldehyde molecule reacts with each lysine ϵ-amino group (Quiocho and Richards, 1964; Richards and Knowles, 1968; Korn *et al.*, 1972).

In practice, the amount of glutaraldehyde used should be the minimum necessary to stabilize the crystals in water. Also, the crystal den-

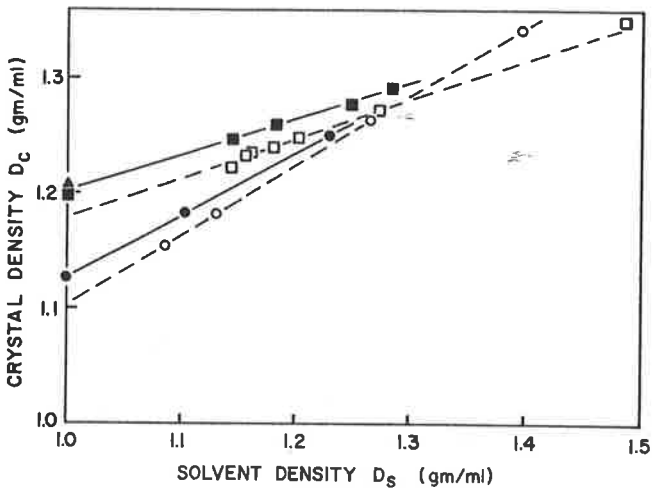


Fig. 19 Illustration of the effect of cross-linking with glutaraldehyde on the density of protein crystals. (□) γ-chymotrypsin; (○) β-amylase. The open symbols are for the normal crystals (data from Fig. 18), and the solid symbols are for the cross-linked crystals. The triangle shows the density of cross-linked γ-chymotrypsin crystals measured in water, uncorrected for the expansion of the crystal lattice. Taken from Matthews (1974).

sities and unit cell dimensions should be measured at approximately the same time after the crystals have been transferred into water. Although the cross-linking method must be used with caution if accurate molecular weights are desired, it has the advantage that fewer density measurements are required than for the extrapolation technique. The mechanical stability of the cross-linked crystals also facilitates the density measurements.

Often, crystal densities are measured at only one solvent density, typically that of the supernatant from which the crystals were grown. (Note that a crystal density quoted without the density of the supernatant is virtually meaningless.) Consideration of Eq. (61) shows that the molecular weight could be found from such measurements if w were known. It is clear from a number of studies, both crystallographic (e.g., Adair and Adair, 1936; Perutz, 1946; Colman and Matthews, 1971) and using various other techniques (e.g., see Fisher, 1965; Kuntz *et al.*, 1969) that for most proteins w will be in the range of 0.15–0.35 gm water/gm protein. Thus, in estimating molecular weight from a pair of crystal and solvent densities, it is clearly preferable to use Eq. (61), substituting a reasonable value such as 0.25 for w, rather than ignoring the bound water, as is sometimes done. Of course, w

may be obtained directly by increasing the supernatant density until the crystals remain just suspended [Eq. (63)]. Note that as the crystal density tends toward the solvent density then, from Eq. (63), both the numerator and denominator of Eq. (60) tend to zero, and the molecular weight becomes undefined. In practice, crystal densities measured only for dense supernatant solutions such as saturated ammonium sulfate will tend to give inaccurate molecular weights because of the increased experimental error in $(D_c - D_s)$ and the increased uncertainty introduced by the $w(D_s - D_w)$ term.

Another method sometimes used to estimate the solvent content of protein crystals, and thereby the protein molecular weight, is measurement of the loss of weight of crystals upon drying (North, 1959; Berthou *et al.*, 1972a; Matthews, 1974). The accuracy of the method is limited by the need to allow for any residual water left in the crystal, which may be 5%–10% of the protein weight (Haurowitz, 1950), and to allow for the salt which is left in the crystal. Although not recommended for accurate molecular weight determinations, this method is useful for finding the number of protein molecules per unit cell.

If the crystals are well enough formed for their volumes to be determined under a microscope, the crystal can be dissolved in a known volume of solvent, and the mass of protein in the crystal determined from the optical density and the extinction coefficient of the protein (Love, 1957; Heidner *et al.*, 1971). Hence the molecular weight of protein per unit cell can be found. This method has the advantage that the presence of salt or organic solvents in the crystal has no effect on the results. A related method in which the molecular weight is determined from the mass of single crystals or of microcrystals has been proposed by Matthews (1974).

3. *Molecular Symmetry*

The study of the structure and symmetry of oligomeric enzymes and proteins is an area of active investigation by a number of techniques. Recent reviews of the field include those by Koshland (1970), Frieden (1971), Matthews and Bernhard (1973), and Klotz *et al.* (1975). It is not the purpose of this chapter to review the role of symmetry in oligomeric enzymes, although it is an undisputed fact that the X-ray diffraction technique has provided the most direct and unequivocal evidence for the structure and symmetry of oligomeric proteins. We shall restrict the present discussion to some definitions that will be necessary for the description of specific structures.

The possible ways in which a finite number of *identical* subunits

may aggregate so that their environments are equivalent is very restricted. Each possibility is designated by a *point group symmetry* which, for proteins, consists of one or more rotation axes intersecting at a point. There are three classes of point symmetry: cyclic, dihedral, and cubic (e.g., see Weyl, 1952).

The possible types of *association* between protomers have been defined by Monod *et al.* (1965). *Isologous* association occurs about twofold axes, while other associations are termed *heterologous*. Matthews and Bernhard (1973) have categorized almost isologous interactions as *pseudo-isologous* and have defined interactions mediated through a symmetry axis (as, for example, through the zinc ion in insulin) as *axial* (see Section IV,E).

The X-ray diffraction method can sometimes directly reveal the point symmetry of an oligomeric protein and, by inference, the probable modes of association between subunits. The asymmetric unit of a protein crystal, defined in Section II,B as the smallest *unique* volume within the unit cell, is related to its neighboring asymmetric units by whatever symmetry elements are included in the space group symmetry. In Fig. 15, for example, one unit cell contains two birds and two fish and can be imagined as defined by lines joining the left wing tips of four birds all flying vertically. Each unit cell is related to the other unit cells by translation. The smallest *unique* area in the pattern, i.e., the asymmetric unit, contains a bird and a fish, the two birds (and the two fish) within the unit cell being related by twofold rotation axes. Now if Escher's lithograph were a crystal of a dimeric protein and it was determined that each unit cell contained only *one* molecule (i.e., one dimer), then it is clear that this molecule would have to consist of two subunits (birds), related to each other by a dyad axis, i.e., the dimer would have to possess a twofold axis of symmetry. On the other hand, if it were determined that the unit cell contained *two* molecules, then the *molecules* would be related by twofold axes, but nothing would be known about whether or not there was a twofold axis within the molecule (i.e., within a bird).

Thus, given the crystallographic space group and the number of molecules per unit cell, the symmetry of the molecule *may* be revealed directly. Whether or not a symmetrical oligomer will reveal its symmetry in this direct manner, i.e., whether or not the molecular symmetry elements are likely to coincide with the crystallographic symmetry elements, is not predictable. Kitaigorodskii (1961) has studied this problem in relation to small molecules and has concluded that a symmetrical molecule will crystallize with its symmetry axes coincident with the crystallographic symmetry elements, provided that this

mode of packing does not necessitate a decrease in the packing density.

In an early application of this method of symmetry determination, Bernal *et al.* (1938) showed the hemoglobin molecule to have a twofold symmetry axis. Also, Rossmann *et al.* (1967) demonstrated that the tetrameric enzyme lactate dehydrogenase from dogfish has exact 222 symmetry in the absence of coenzyme. Terry *et al.* (1968) used the same method to show that human immunoglobulins have diad axes of symmetry.

It sometimes happens that an oligomer with more than one symmetry axis will reveal only part of its full symmetry in a given crystal form. Of course, as with dimers, it is also possible that none of the potential symmetry axes needs be apparent. For example, Wiley and Lipscomb (1968) obtained a trigonal crystal form of aspartate transcarbamylase, in which the molecule exhibited a threefold axis of symmetry, and a tetragonal crystal form of the same enzyme, in which the molecule possessed a twofold axis. Taken together these two crystal forms suggested very convincingly that the enzyme with its six catalytic subunits and six regulatory subunits (Weber, 1968) had to have overall point symmetry 32. Subsequently, this result was confirmed by yet another crystal form in which the full 32 point symmetry was expressed, the asymmetric unit of the crystal consisting of one catalytic monomer and one regulatory monomer (Wiley *et al.*, 1971).

With higher symmetry space groups or with higher symmetry oligomers, the range of possible packing arrangements is larger, and it becomes more important to determine the protein content of the unit cell more precisely (see Section III,C,2). For example, Heidner *et al.* (1971) showed that a hexagonal crystal form of rabbit muscle aldolase had nine aldolase tetramers per cell rather than six as thought previously (Goryunov *et al.*, 1969; Eagles *et al.*, 1969). In this unusual packing arrangement, six of the aldolase molecules were on twofold axes, whereas the remaining three had their center at the point of intersection of three mutually perpendicular diads, i.e., they had 222 point symmetry. [As discussed by Matthews and Bernhard (1973), the fact that all the aldolase subunits may not be identical further complicates the interpretation of these results.]

Should the potential symmetry axes of an oligomeric protein not coincide with the crystallographic axes and therefore not be revealed directly, the determination of such symmetry elements becomes more complicated. Matthews and Bernhard (1973) have described the following crystallographic methods which may provide additional symmetry information.

a. **Pseudo-Symmetry in the Diffraction Pattern.** Sometimes a symmetry axis of an oligomer will be aligned in such a way that it causes obvious additional pseudo-symmetry in one or more of the diffraction photographs of the crystal. For example, Campbell *et al.* (1971) found that tetrameric crayfish triosephospate isomerase had an exact (crystallographic) diad axis and, in addition, the diffraction patterns had striking pseudo-symmetry which showed that the subunits were arranged in perfect or near perfect 222 symmetry. In cases like this there is no a priori reason to believe that the local or noncrystallographic symmetry axis is more likely to be inexact rather than exact.

b. **Pseudo Crystallographic Symmetry.** On occasion, a protein crystal has an apparent unit cell and space group at low resolution, but a different space group at higher resolution. Usually, the "low resolution" space group has additional symmetry elements not present in the high resolution case. This implies either that the oligomer has an approximate symmetry axis or that the oligomer has an exact symmetry axis which does not quite coincide with a crystallographic symmetry axis. For example, Gorjunov *et al.* (1972) found that the tetramer of human holo glyceraldehyde-phosphate dehydrogenase had an exact twofold symmetry axis and, at least at low resolution, 222 symmetry.

It seems that pseudo crystallographic symmetry is more likely due to inexact symmetry of the oligomer rather than to noncoincidence of the (exact) molecular symmetry axis with the crystallographic axis, although the evidence for inexact symmetry cannot be taken as conclusive (Matthews and Bernhard, 1973).

c. **The Rotation Function.** The rotation function (Rossmann and Blow, 1962), discussed in Section II,E,2, provides a general method of detecting noncrystallographic symmetry elements.

d. **Structure Determination.** Obviously, the most powerful crystallographic method of studying structural symmetry in an oligomeric enzyme is by the complete determination of its atomic structure. Specific studies are discussed in Section IV.

Table IV summarizes the symmetry information that has been obtained from X-ray studies of crystalline proteins. The outstanding feature of this compilation is the ubiquitous nature of twofold axes and, by implication, isologous associations in subunit–subunit interaction. There are three clear-cut cases in which isologous or pseudo-isologous interactions do not participate in the stabilization of an

TABLE IV

Crystal Forms and Symmetry of Oligomeric Proteins[a]

Protein	Source	Molecular weight[b]	No. of protomers[c]	Space group[d]
Proinsulin	Bovine	18,000	2	$P4_12_12_1$
C-Phycocyanin	*Mastigocladus-laminosus*	28,000 (2)	2	$P6_3$
Ribonuclease	Bovine, seminal	29,000	2	$P2_12_12_1$
Insulin	Cod	34,800	6_2	$P2_12_12_1$
Insulin, iodinated	Pig	34,800 ($\frac{1}{3}$)	6_2	R32
Insulin	Pig	34,800 ($\frac{1}{3}$)	6_2	R3
Insulin	Pig	34,800 (2)	6_2	$P2_1$
Insulin	Pig	34,800 (?)	6_2	I23 or I2,3
β-Lactoglobulin	Cow's milk	36,000 ($\frac{1}{2}$)	2	$B22_12$
Rhodanese	Bovine liver	37,000	2	C2
Bence Jones protein	Human	46,000	2	$P2_12_12$
Bence Jones protein	Human	46,000	2	$P3_121$
Triosephosphate isomerase	Chicken muscle	49,000	2	$P2_12_12_1$
Triosephosphate isomerase, DHAP added	Chicken muscle	49,000	2	$P2_12_12_1$
Triosephosphate isomerase, phosphoglycollate added	Chicken muscle	49,000	2	$2_12_12_1$
Prealbumin	Human plasma	54,000 ($\frac{1}{2}$)	4	$P22_12_1$?) (ps. 1222)
Fc fragment	Human myeloma	55,000	(2)	$P2_12_12_1$
Triosephosphate isomerase	Rabbit muscle	60,000 ($\frac{1}{2}$)	2	$P6_122$
Triosephosphate isomerase	Rabbit muscle	60,000	2	$P2_1$
Methionine-tRNA synthetase	*E. coli (modified)*	64,000	2	$P2_1$
Hemoglobin, met, oxy	Horse	66,500 ($\frac{1}{2}$)	2_2	C2
Hemoglobin, met, oxy	Human	66,500	2_2	$P4_12_1$
Hemoglobin, deoxy	Horse	66,500	2_2	$C222_1$
Hemoglobin, met	Fetal sheep	68,000 ($\frac{1}{2}$)	2_2	$B22_12$
Hemoglobin, met	Ox	68,000 ($\frac{1}{2}$)	2_2	$P4_132$

Symmetry determined crystallographically	Full symmetry of oligomer[e]	V_M (Å^3/dalton)	Technique[f]	Reference[g]
–	Uncertain	2.65		Fullerton *et al.* (1970)
–	Uncertain	2.48		Dobler *et al.* (1972)
–	2 or ps. 2	2.26		Capasso *et al.* (1972)
–	(ps. 32)	1.87		Baker and Dodson (1970)
3	(ps. 32)	1.88		Baker and Dodson (1970)
3	ps. 32	1.96	IR, 2.8 Å	Blundell *et al.* (1971a,b)
–	(ps. 32)	2.06		Baker and Dodson (1970)
2 or 3	(ps. 32)	3.17		Harding *et al.* (1966)
2	2	2.13		Aschaffenburg *et al.* (1965)
–	Uncertain	2.16		Drenth and Smit (1971)
–	Uncertain	2.29		Schiffer *et al.* (1970)
–	Uncertain	3.04	IR, 6 Å	Edmundson *et al.* (1972)
–	2 or ps. 2	2.48		Johnson and Wolfenden (1970)
			IR, 6 Å	Banner *et al.* (1971)
			IR, 2.5 Å	Banner *et al.* (1975)
–	2 or ps. 2	2.41		Banner *et al.* (1971)
–	2 or ps. 2	2.36		Banner *et al.* (1971)
2	222 or ps. 222	2.23		Blake *et al.* (1971)
–	2 or ps. 2	2.70	IR, 6 Å	Poljak *et al.* (1972)
2	2	2.52		Johnson and Waley (1967)
–	(2)	2.05		Johnson and Waley (1967)
–	Uncertain	2.36		Waller *et al.* (1971)
2	2	2.61	IR, 2.8 Å	Perutz *et al.* (1968a,b)
2	2	2.03		Bragg and Perutz (1952)
2	2	2.26	IR, 2.8 Å	Bolton and Perutz (1970)
2	2	2.49		Bragg and Perutz (1952)
2	2	3.53		North (1959)

(*Continued*)

TABLE IV (*Continued*)

Protein	Source	Molecular weight[b]	No. of protomers[c]	Space group[d]
Malate dehydrogenase	Pig heart	72,000	2	$P2_12_12$
Phosphogluconate aldolase	*Pseudomonas putida*	73,000 ($\frac{1}{3}$)	3	$P2_13$
Alkaline phosphatase	*E. coli*	80,000 ($\frac{1}{2}$)	2	$P3_121$
Alcohol dehydrogenase	Horse liver	84,000 ($\frac{1}{2}$)	2	$C222_1$
Alcohol dehydrogenase (NAD^+, pyrazole)	Horse liver	84,000	2	P1
Alcohol dehydrogenase (NADH, isobutyramide)	Horse liver	84,000	2	$P2_1$
Enolase	Yeast	88,000	2	P1
Hexokinase B	Yeast	102,000	2	$P2_122_1$
Concanavalin A	Jack bean	108,000 ($\frac{1}{4}$)	4	I222
Concanavalin A, demetallized	Jack bean	108,000 ($\frac{1}{2}$)	4	$P2_122_1$
Phosphoglucose isomerase	Pig	108,000 ($\frac{1}{2}$)	2	$P4_12_12$
Phosphoglycerate mutase	Yeast	110,700 ($\frac{1}{2}$)	4	C2
Leucine-tRNA synthetase	Yeast	120,000 ($\frac{1}{2}$)	2	I222 or $I2_12_12_1$
Lysine-tRNA synthetase	Yeast	138,000 ($\frac{1}{2}$)	2	$P3_121$
Lactate dehydrogenase	Dogfish	140,000 ($\frac{1}{4}$)	4	F422
Lactate dehydrogenase, NAD^+ diffused	Dogfish	140,000 ($\frac{1}{2}$)	4	$C4_22_12$
Lactate dehydrogenase, cocryst. with NAD^+, pyruvate	Dogfish	140,000 ($\frac{1}{4}$)	4	$P42_12$
Lactate dehydrogenase, cocryst. with NAD^+	Dogfish	140,000	4	I2
Lactate dehydrogenase, NAD^+, pyruvate	Pig	140,000 ($\frac{1}{2}$)	4	$P22_12_1$

Symmetry determined crystallographically	Full symmetry of oligomer[e]	V_M (Å^3/dalton)	Technique[f]	Reference[g]
–	ps. 2	2.48	IR, 5 Å	Tsernoglou *et al.* (1972)
			IR, 3 Å	Hill *et al.* (1972)
3	3	3.76		Vandlen *et al.* (1973)
2	2	2.79		Hanson *et al.* (1970)
2	2	2.26		Zeppezauer *et al.* (1967)
			IR, 6 Å	Brändén *et al.* (1969)
–	ps. 2 or 2	2.31		Zeppezauer *et al.* (1967)
–	ps. 2 or 2	2.33		Zeppezauer *et al.* (1967)
–	Uncertain	2.77		Campbell *et al.* (1971)
–	Screw	2.8	PF	Steitz (1971)
222	222	2.26	IR, 2.0 Å	Edelman *et al.* (1972)
			IR, 2.8 Å	Hardman and Ainsworth (1972)
2	ps. 222 or 222	2.22	RF	Jack *et al.* (1971)
2	2	2.91		Campbell *et al.* (1971)
2	222 or ps. 222	2.66	IR, 6 Å	Campbell *et al.* (1972)
2	2	2.16		Chirckjian *et al.* (1972)
2	2	5.60		Lagerkvist *et al.* (1972)
222	222	2.97	IR, 2.5 Å	Rossmann *et al.* (1971)
2	ps. 222 or 222	2.97		Adams *et al.* (1969b)
222	222	2.78		Leberman *et al.* (1969)
			IR, 3.0 Å	Rossmann *et al.* (1971)
–	(ps. 222 or 222)	2.86		Adams *et al.* (1972)
2	(222)	2.51		Adams *et al.* (1972)

(*Continued*)

TABLE IV (*Continued*)

Protein	Source	Molecular weight[b]	No. of protomers[c]	Space group[d]
Lactate dehydrogenase, cocryst. with NAD^+	Pig	140,000	4	C2
Lactate dehydrogenase, cocryst. with NAD^+, oxalate	Pig	140,000	4	C2
Glyceraldehyde-phosphate dehydrogenase	Crayfish	140,000 ($\frac{1}{2}$)	4	C2
Glyceraldehyde-phosphate dehydrogenase	Lobster	140,000	4	$P2_12_12_1$ (ps. $P4_12_12$)
Glyceraldehyde-phosphate dehydrogenase (PCMB)	Lobster	140,000 ($\frac{1}{2}$)	4	$C222_1$
Glyceraldehyde-phosphate dehydrogenase (holo)	Human	140,000 ($\frac{1}{2}$)	4	$C222_1$ (ps. $I4_122$)
Glyceraldehyde-phosphate dehydrogenase (holo)	Human	140,000 ($\frac{1}{2}$)	4	C2
Glyceraldehyde-phosphate dehydrogenase (apo)	Human	140,000	4	C2
Glyceraldehyde-phosphate dehydrogenase (apo)	*B. stearothermophilus*	140,000	4	$P2_1$ (ps. $B22_12$)
Glyceraldehyde-phosphate dehydrogenase (holo I)	*B. stearothermophilus*	140,000	4	$P2_12_12$ (ps. $B22_12$)
Glyceraldehyde-phosphate dehydrogenase (holo II)	*B. stearothermophilus*	140,000	4	$P2_1$ (ps. $B22_12$)
Ceroplasmin	Human	≃ 150,000 (2)	–	I4
Bacteriochlorophyll protein	*Chlorobium limicola*	127,000	3	$P6_3$
Immunoglobulin	Human myeloma	155,000 ($\frac{1}{2}$)	2_2	C2

Symmetry determined crystallographically	Full symmetry of oligomer[e]	V_M (Å³/dalton)	Technique[f]	Reference[g]
–	(ps. 222 or 222)	3.17		Adams *et al.* (1972)
–	(ps. 222 or 222)	2.45		Adams *et al.* (1972)
2	ps. 222 or 222	3.35		Campbell *et al.* (1971)
–	(ps. 222 or 222)	3.00	RF	Watson and Banaszak (1964) Rossmann *et al.* (1972)
2	(ps. 222 or 222)	2.99		Watson and Banaszak (1964)
2	ps. 222 or 222	3.34		Gorjunov *et al.* (1972)
2	Prob. ps. 222	3.43	IR, 6 Å	Watson *et al.* (1972)
–	222 or ps. 222	3.38		Gorjunov *et al.* (1972)
–	Uncertain	3.06		A. J. Wonacott and D. M. Blow, personal communication
–	Uncertain	2.99		A. J. Wonacott and D. M. Blow, personal communication
–	Uncertain	2.86		A. J. Wonacott and D. M. Blow, personal communication
–	Uncertain	3.86(?)		Magdoff-Fairchild *et al.* (1969)
–	3	4.2		Fenna and Matthews (1975)
2	2	2.92	IR, 6 Å	Sarma *et al.* (1971)

(*Continued*)

TABLE IV (*Continued*)

Protein	Source	Molecular weight[b]	No. of protomers[c]	Space group[d]
Immunoglobulin	Human myeloma	155,000 ($\frac{1}{2}$)	2_2	$C222_1$
Aldolase	Rabbit muscle	160,000	4	$P2_1$
Aldolase (PCMB)	Rabbit muscle	160,000	4	$P2_1$
Aldolase	Rabbit muscle	160,000 ($\frac{2}{4}$)	4	$P6_222$
Aldolase	Rabbit muscle	160,000 ($\frac{1}{4}$)	4	$P6_222$
Phosphorylase *b*	Rabbit muscle	200,000 (2)	2	$P2_1$
Phosphorylase *b*	Rabbit muscle	200,000 ($\frac{1}{2}$)	2	$P4_12_12$
β-Amylase	Sweet potato	206,000	4	$P4_122$
Pyruvate kinase	Human muscle	240,000 ($\frac{1}{4}$)	4	I222
Pyruvate kinase	Rabbit muscle	240,000	4	$P6_222$(?)
Pyruvate kinase	Rabbit muscle	240,000 ($\frac{1}{2}$)	4	$P6_222$
Catalase	Horse liver	250,000	4	$P22_12$
Catalase	Horse liver	250,000	4	$P2_12_12_1$
Catalase	Ox liver	250,000	4	$P3_121$
Hexon	Adenovirus type 5	300,000 ($\frac{1}{2}$)	6	I23 or $I2_13$
Aspartate transcarbamylase	*E. coli*	310,000 ($\frac{1}{6}$)	6_2	R32
Aspartate transcarbamylase (cocryst. with CTP)	*E. coli*	310,000 ($\frac{1}{3}$)	6_2	P321
Aspartate transcarbamylase	*E. coli*	310,000 ($\frac{1}{2}$)	6_2	$P4_222$ or $P4_22_12$
Hexon	Adenovirus type 2	340,000 ($\frac{1}{3}$)	6	$P2_13$
Glutamine synthetase	*E. coli*	592,000	12	Hexagonal
Glutamine synthetase	*E. coli*	592,000	12	Prob. $P4_2$
Dihydrolipoyl transsuccinylase	*E. coli*	–	24	F432

[a] This table includes most of the crystal forms of oligomeric proteins, except those in the survey of Matthews (1968) which did not yield any information about the symmetry of the oligomer. Only a selection of the many crystal forms of hemoglobin and β-lactoglobulin has been given. The compilation was made at the end of 1972.

[b] The number of molecular weight units in the crystallographic asymmetric unit, if different from unity, is indicated in parentheses. If this number is less than one, then information is obtained about the symmetry of the oligomer (see Section III,C,3) and is entered under "Symmetry determined crystallographically."

[c] "Protomer" is defined by Monod *et al.* (1965). A subscript indicates the number of polypeptide chains per protomer, if different from unity.

Symmetry determined crystallographically	Full symmetry of oligomer[e]	V_M (Å³/dalton)	Technique[f]	Reference[g]
2	2	2.94		Edmundson *et al.* (1970)
–	(222)	2.45		Eagles *et al.* (1969)
–	222 or ps. 222	2.5	PF, RF	Eagles *et al.* (1969)
222 and 2	222	2.7		Heidner *et al.* (1971)
222	222	2.8		Heidner *et al.* (1971)
–	(2)	2.43		Mathews (1967) Eagles *et al.* (1972)
2	2	2.32		Eagles *et al.* (1972)
–	⩾ ps. 2	4.2		Colman and Matthews (1971)
222	222	2.74		Campbell *et al.* (1971)
–	Uncertain	–		Hollenberg *et al.* (1971)
2	⩾2	2.14		McPherson and Rich (1972)
–	Uncertain	2.52		Glauser and Rossmann (1966)
–	Uncertain	2.86		Burgner and McGandy (1967)
–	Uncertain	4.42	–	Rossmann and Labaw (1967) Longley (1967) Vainshtein *et al.* (1969)
2	(Prob. 32)	2.68		MacIntyre *et al.* (1969)
32	32	2.79	IR, 5.5 Å	Wiley *et al.* (1971)
3	(32)	2.95		Wiley and Lipscomb (1968) Steitz *et al.* (1967)
2	(32)	2.76		Wiley and Lipscomb (1968)
3	(Prob. 32)	2.48		Franklin *et al.* (1971) Cornick *et al.* (1971)
–	–	–		Bywater *et al.* (1972) Bywater *et al.* (1969)
2	Exact or ps. 6 or 622	–		Eisenberg *et al.* (1971)
432	432	–		DeRosier *et al.* (1971)

[d] Enantiomorphic space groups are not included.

[e] The entry gives the best estimate of the point symmetry of the protein, based upon all the crystallographic data currently available. In a number of cases the symmetry of an oligomer is not determined for a given crystal but can be inferred from other crystal forms or from closely related species. If this has been done, the point symmetry is enclosed in parentheses.

[f] If a three-dimensional electron density map has been calculated, the resolution is given. Other abbreviations are: IR, isomorphous replacement; RF, rotation function; PF, Patterson function.

[g] In general, reference is given to the most recent publication.

oligomer, two of these being cyclic trimers. The first trimer is 2-keto-3-deoxy-6-phosphogluconate aldolase from *Pseudomonas putida* (Vandlen *et al.*, 1973); the second is the bacteriochlorophyll protein from *Chlorobium limicola*, previously thought to be tetrameric (Olson *et al.*, 1969) but now proved to be a trimer with threefold symmetry (Fenna *et al.*, 1974; Fenna and Matthews, 1975). The other example of heterologous association is yeast hexokinase B. In two different crystal forms, the monomers of this protein associate in distinct non-symmetrical modes (Steitz, 1971; Anderson *et al.*, 1974; Anderson and Steitz, 1975). Whether the modes of association seen in the crystals of hexokinase also occur in solution is uncertain.

It is interesting to note that although there is an abundance of crystallographic evidence for tetramers with 222 (i.e., D_2) symmetry, there is no evidence for a tetramer with cyclic symmetry.

D. Structure in the Crystal and in Solution

"Is the conformation of a protein the same in the crystal as it is in solution?" The answer, from a variety of lines of evidence, is an almost unqualified "yes."

Perhaps the most convincing arguments are of two types: first, the observation that (covalently) similar proteins crystallized under widely different conditions in different crystal forms have very similar tertiary structures, and second, the finding that many enzymes exhibit catalytic activity in the crystal.

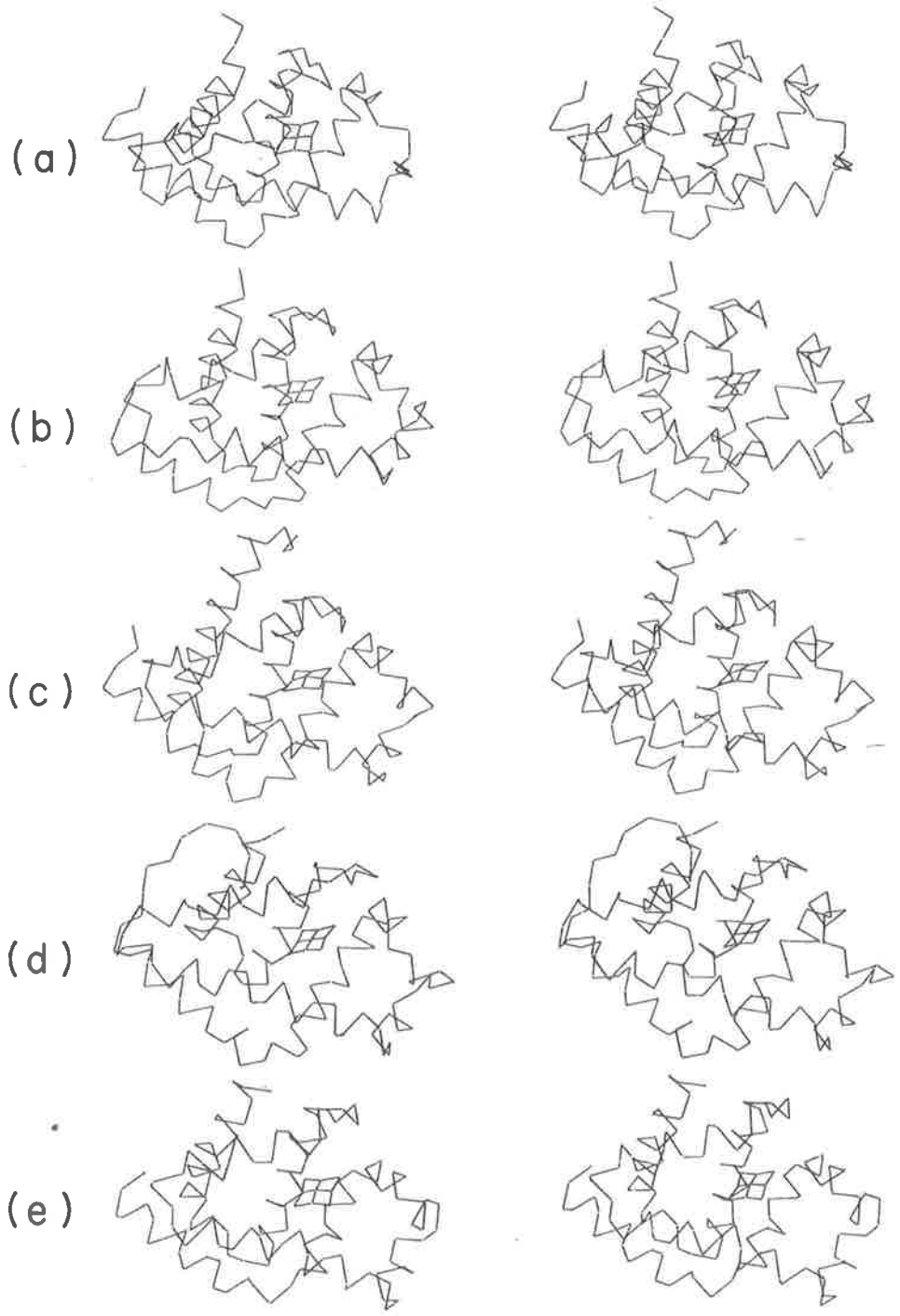
α-Chymotrypsin and γ-chymotrypsin have identical amino acid sequences but crystallize quite differently. The two structures, as determined crystallographically, were found to be not significantly different except for a few surface residues (Matthews *et al.*, 1968; Cohen *et al.*, 1970; Blow, 1971; D. R. Davies, personal communication). Both structures are similar to that determined for chymotrypsinogen A (Freer *et al.*, 1970). In spite of the removal of two dipeptides during activation, the mean deviation between respective α-carbon atoms of α-chymotrypsin and chymotrypsinogen A is 1.8 Å (Freer *et al.*, 1970; Kraut, 1971).

Another comparison can be made between substilisin BPN′, crystallized from 2.1 *M* ammonium sulfate, and subtilisin Novo, which has the same primary structure but was crystallized from 55% acetone–water (v/v) in a different space group (Wright *et al.*, 1969; Hol, 1971; Robertus *et al.*, 1971). The two electron density maps were interpreted independently, and the coordinates of the 1895 nonhydrogen atoms in the respective molecules, when compared, were

found to have a rms difference of 1.55 Å (Drenth *et al.*, 1971a). The backbone atoms were in closer agreement. Drenth *et al.* (1971a) concluded that there were no major structural differences between the two crystal forms, although some differences in the binding of cations were observed. Further examples of structural similarity are provided by ribonuclease S and ribonuclease A (Wyckoff *et al.*, 1967b; Kartha *et al.*, 1967); by the subunits of lactate and malate dehydrogenase (Rossmann *et al.*, 1971; Hill *et al.*, 1972); and by the globins from a variety of species, illustrated in Fig. 20. Other than some small adjustments of relatively flexible, exposed side chains, there are no structural differences that can be attributed to the different crystal environments. The evidence that these tertiary structures represent, to an excellent approximation, the tertiary structure in solution is therefore very compelling.

The second main line of evidence for equivalence of proteins in the crystal and in solution comes from studies of catalytic activity. Ribonuclease S, ribonuclease A, carboxypeptidase A, and γ- and α-chymotrypsin all retain catalytic activity after crystallization (Doscher and Richards, 1963; Bello and Nowoswiat, 1965; Quiocho and Richards, 1964, 1966; Sigler and Skinner, 1963; Sigler *et al.*, 1966). Also, a number of crystalline enzymes form complexes with substrates, inhibitors, and coenzymes. Rossi and Bernhard (1970) compared the conformation of indoleacryloyl-α-chymotrypsin in the crystal and in solution and concluded that, at least for acyl chymotrypsins, molecular inferences regarding the catalytic mechanism of enzyme action derived from crystallographic diffraction studies were validly extendable to aqueous solution. Sluyterman and De Graaf (1969) showed that, after correcting for diffusion, α-chymotrypsin in the crystalline form was not less active than the dissolved enzyme and that in two crystal forms papain was as active as in solution. However, for carboxypeptidase A and alcohol dehydrogenase, crystallization was found to result in a decrease in activity of at least several hundredfold (Quiocho and Richards, 1966; Theorell *et al.*, 1966). In a series of studies, Rupley and coworkers have concluded that crystallization induces no change in average protein conformation, although it may lead to reduced motility of the protein (e.g., see Rupley, 1968, 1969, and references therein).

An apparent difference between the conformation of carboxypeptidase A in solution and in the crystal seems to have been resolved. Johansen and Vallee (1971) found that carboxypeptidase A, specifically labeled at Tyr 248 with diazotized arsanilic acid, was red when in solution but yellow in the crystal form, indicating that the azophenol moiety complexes with the active site Zn^{2+} in solution but not in the


(a)
(b)
(c)
(d)
(e)

crystal. This is in apparent conflict with the movement of Tyr 248 observed in crystalline carboxypeptidase A (e.g., see Quiocho and Lipscomb, 1971). Subsequently, Quiocho *et al.* (1972) pointed out that the form of the enzyme used by Johansen and Vallee (CPA_γ) has seven residues less than that used for the structure determination (CPA_α) and crystallizes somewhat differently. Furthermore, modified CPA_α is red both in the crystal form and when dissolved, indicating that in both cases Tyr 248 is free to approach the zinc ion. Lipscomb (1973) has argued that differences in intermolecular contacts in the different crystal forms of carboxypeptidase could account for the differences observed, but Vallee and co-workers have disputed this conclusion (Johansen and Vallee, 1975).

E. Heavy Atom Derivatives

After obtaining suitable crystals of a protein, the major obstacle in the way of a successful structure determination is finding satisfactory heavy atom isomorphs. It is still true that the best derivatives are often found on a trial and error basis, although as the number of conclusive structure determinations increases, the repertoire of proved methods and useful compounds continues to increase. The problems associated with obtaining heavy atom derivatives have been extensively discussed by Blake (1968). Also, lists of compounds and approaches which have been used with some degree of success have been given by Phillips (1966), Holmes and Blow (1965), and Eisenberg (1970). In this section we shall briefly review the basic approaches.

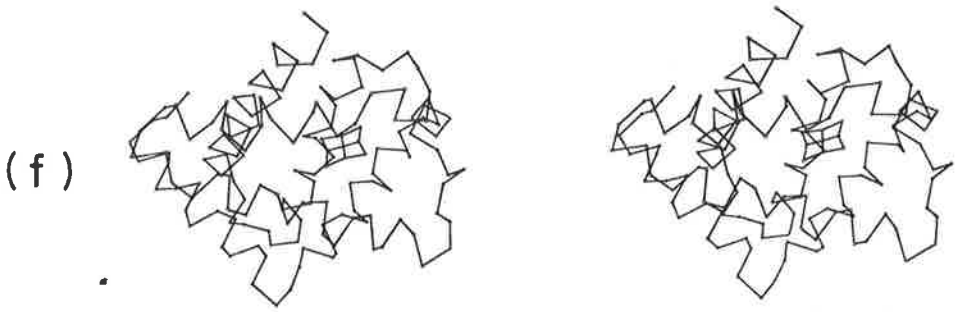


Fig. 20 Stereo drawings of the α-carbon backbone conformation for six globins. (a) α chain of horse methemoglobin; (b) β chain of horse methemoglobin; (c) sperm whale myoglobin; (d) lamprey hemoglobin; (e) *Chironomus* hemoglobin; and (f) *Glycera* hemoglobin. Taken from Love *et al.* (1971).

1. *Trial and Error*

The majority of successful isomorphous derivatives have been obtained by the brute force method of soaking pregrown crystals of the protein in many solutions containing different heavy atom ligands. After several days or weeks the crystals are photographed and examined for intensity differences which, if present, are further analyzed by difference Patterson and difference Fourier techniques (see Section II,C,*4*). Perhaps the most widely effective reagent is the $PtCl_4^{2-}$ ion, which has been found to bind to exposed methionine (Dickerson *et al.*, 1969) and histidine residues in a number of proteins. Eisenberg (1970) has summarized the sites of binding of $PtCl_4^{2-}$ and of a number of other useful reagents.

Sigler and Blow (1965) suggested that ligand binding to a protein may be enhanced by replacement of the ammonium sulfate, often used for crystallization, with an ammonia-free solution such as NaH_2PO_4 or $MgSO_4$, which would not tend to compete with the protein for ligand binding. It also follows that it may be easier to obtain heavy atom derivatives for proteins which have crystallized from solutions of low ionic strength. If a heavy atom reagent is insoluble, it may be helpful to add organic solvents. Also, variation of pH, within the range that the crystals maintain their integrity, may alter the specificity and avidity of heavy atom binding.

2. *Specific Chemical Modification*

With some proteins, it may be possible to obtain heavy atom isomorphs by exploiting chemically reactive groups. For example, Perutz and colleagues were the first to utilize the reactivity of cysteine to obtain mercury derivatives of hemoglobin (Cullis *et al.*, 1961a). This approach has also been successfully employed for other hemoglobin crystals and for papain (Drenth *et al.*, 1968), carbonic anhydrase (Tilander *et al.*, 1965), and lactate dehydrogenase (Adams *et al.*, 1969b). Use of sulfhydryl groups and possible sulfhydryl reagents have been discussed by Blake (1968) and Dunnill *et al.* (1966). Other reactions which have been shown to be at least partially useful include iodination of tyrosine residues and labeling of α-amino and ϵ-amino groups with the iodo derivative of Edman's reagent (Drenth *et al.*, 1968). Iodination of α-chymotrypsin was used to locate one of the tyrosine residues (Sigler, 1970), thereby helping in the interpretation of the electron density map (Matthews *et al.*, 1967). Iodination was also used with γ-chymotrypsin to obtain a heavy atom derivative and to facilitate comparison of α- and γ-chymotrypsin (Matthews *et al.*, 1968; Cohen *et al.*, 1970).

Another method of specific chemical labeling of enzymes is by heavy atom inhibitor analogues. For example, Tilander *et al.* (1965) used a number of heavy atom-containing sulfonamide inhibitors to label the active site of carbonic anhydrase. Also, Sigler *et al.* (1964) showed that the sulfonyl fluorides could be used in the crystal to specifically label the active sites of γ-chymotrypsin, an approach which also proved successful for α-chymotrypsin (Sigler *et al.*, 1966) and elastase (Shotton and Watson, 1970). Sigler *et al.* (1964) found that *p*-iodophenyl sulfonyl (pipsyl) γ-chymotrypsin was not very isomorphous with native γ-chymotrypsin, but that excellent isomorphism could be achieved by comparing pipsyl-γ-chymotrypsin with toluene sulfonyl (tosyl) γ-chymotrypsin. This was also true for the other serine proteases. Unfortunately, this approach yields phase angles for the inhibited enzyme rather than the native enzyme and tends to introduce uncertainties in the electron density map in the active site region. However, this may be a reasonable price to pay for a good derivative, especially if no other can be found. Also, a good approximation of the electron density of the uninhibited enzyme can be obtained by the use of Eq. (45).

In this section we have emphasized labeling reactions carried out in the pregrown crystal. If a desired modification requires conditions that would disrupt the crystal, it may be possible to carry out the reaction in solution and then subsequently crystallize the covalently labeled protein. This technique allows purification of the reacted species but requires more material and assumes that the modified protein can be crystallized in the same form as the unreacted material. The method has proved very successful in some cases and was used to obtain the first isomorphous derivative of a protein (Green *et al.*, 1954). In other cases, modification resulted in an entirely different crystal form or in no crystals at all.

3. *Substitution*

Some proteins, in their native form, bind metals which may be replaced, thus providing very satisfactory isomorphous derivatives. This approach has been of considerable use in studies of zinc-containing enzymes. For example, Tilander *et al.* (1965) removed the zinc ion from crystals of carbonic anhydrase and then, by dialysis of the zinc-free crystals against mercuric acetate, obtained a derivative with mercury atoms bound at the sulfhydryl site and substituted at about 0.7 Å from the zinc site. With carboxypeptidase A, it was possible to substitute mercury for the zinc ion simply by extended dialysis of the native crystals against mercuric chloride solutions (Hartsuck *et al.*, 1965).

With insulin, the zinc was first removed with EDTA and then replaced with lead by soaking the crystals in lead acetate (Adams *et al.*, 1969a). In the case of thermolysin, the zinc was first removed with a combination of EDTA and 1,10-phenanthroline, then mercury substitution was achieved by exposing the crystals to low concentrations of mercuric ion for only 15 min. Longer exposure to mercury (e.g., 45 min) produced a less satisfactory derivative, presumably due to additional mercury binding (Matthews *et al.*, 1972a, and unpublished results).

Calcium is another metal which binds to a number of proteins. For staphylococcal nuclease, an excellent derivative was obtained by replacement of the bound calcium ion with barium (Arnone *et al.*, 1971). Recently, it has been shown that the rare earths also provide potential calcium replacements for protein crystallography (Colman *et al.*, 1972b). In the case of thermolysin, lanthanide substitution at three of the four calcium-binding sites was achieved simply by soaking the crystals in solution containing the rare earth. The high degree of isomorphism coupled with significant anomalous scattering for selected rare earths makes lanthanide substitution very attractive (Matthews and Weaver, 1974).

F. Data Collection

1. Crystal Mounting

For X-ray photography, protein crystals are normally mounted in a thin-walled glass or quartz tube together with some supernatant to preserve the humidity. Glass capillaries are more fragile but give less background scatter (and are cheaper) than quartz. Techniques such as that described by King (1954) may be used to obtain a satisfactory mount, although most workers prefer to seal each end of the capillary with wax or Vaseline. Also, it is easier to insert the crystal into the capillary with a Pasteur pipette, rather than relying upon surface tension to draw the crystal up into the tube.

A general method of ensuring that the crystal does not move after mounting remains to be determined.

It has generally been regarded as important to have the crystal no larger than the diameter of the X-ray beam, but this may not be critical in photographic methods, particularly if the orientation of the crystal relative to the X-ray beam does not vary a great deal. Åkervall and Strandberg (1971) grew single crystals of satellite tobacco necrosis virus in a capillary tube until they were large enough to become wedged, then translated the crystal between exposures so as to bring

unexposed regions of the crystal into the beam, thus minimizing the effects of radiation damage. Also, in the determination of the structure of concanavalin A by photographic methods (Edelman *et al.*, 1972), exposure of different parts of a single large crystal was used to obtain several photographs per crystal (Becker *et al.*, 1975).

Wyckoff *et al.* (1967a) have described a rather different approach to crystal mounting, using a "flow cell" made of polyethylene tubing, in which the crystal is held in place by cotton linters or a bed of Sephadex. The crystal is bathed in its mother liquor which may be changed during X-ray observation by use of a series of tubes. The flow cell technique suffers from the disadvantage of increased absorption and scatter of the X-ray beam due to the polyethylene tube and additional mother liquor, although ways in which the experimental conditions can be optimized to obtain satisfactory counting statistics have been proposed by Wyckoff *et al.* (1967a). Using the flow cell it was possible to show directly that ammonium sulfate diffused into and out of the crystals of ribonucleases with a half-time of about 90 sec, whereas the inhibitor 5-iodouridine 2′(3′)-phosphate had a diffusion half-time of more than 11 hr. Upon changing the supernatant pH from 6.3 to 5.5, the cell dimensions were observed to change, apparently continuously, with a half-time of 2 hr.

2. *Radiation Damage*

It is well known that protein crystals deteriorate during exposure to X-rays and that some crystals are affected more than others. Also, different crystal forms of the same protein may show marked differences in radiation stability. The best procedure for dealing with radiation damage is still open to debate.

Blake and Phillips (1962) studied the consequences of X-irradiation of myoglobin crystals and found that the effective diffracting volume of the crystal decreased monotonically with time. They also showed that superimposed on this overall decay in diffraction intensity, some reflections either decreased or increased in intensity with crystal exposure. Thus, for myoglobin, it would not have been possible to correct for *all* the effects of irradiation by simply applying an overall correction factor, although such a correction would presumably be preferable to no correction at all. Traub and Hirshfeld (1960) also reported that with insulin the effects of radiation were not continuous with time.

On the other hand, Watenpaugh *et al.* (1973a) recently collected a complete set of data to a resolution of $d = 1.54$ Å (9350 reflections) from a single crystal of rubredoxin. Jensen and colleagues measured

three monitor reflections rather frequently and, in addition, a large number of standard reflections less often. The three monitor reflections were used to determine the overall intensity decay as a function of time (in this case, a maximum of 10%–15%), which was applied as a correction; the standard reflections were checked to see that intensity anomalies, as described by Blake and Phillips (1962), were within acceptable limits. Also, Jensen and colleagues have successfully applied even larger empirical radiation decay corrections, ranging up to 25%–33% for some of the heavy atom derivatives of ferredoxin and flavodoxin (Herriott *et al.*, 1970; Watenpaugh *et al.*, 1973a). Empirical corrections of this type were also employed by Mathews *et al.* (1972). Other workers have preferred to discard crystals once the intensity of a set of reference reflections has decreased by about 5%–10% (e.g., see Kraut *et al.*, 1962; Arnone *et al.*, 1971). Clearly, it is advantageous if the useful life of a crystal can be extended by radiation damage correction, but it seems that each protein should be tested to see to what extent this procedure is justified.

A number of workers have considered cooling the crystal as a possible method of reducing radiation damage and perhaps increasing crystalline order. The diffraction data used for the structure determinations of carboxypeptidase A and staphylococcal nuclease, for example, were collected at crystal temperatures between 0° and 4° (Lipscomb *et al.*, 1968; Arnone *et al.*, 1971). At lower temperatures, crystal disorder due to formation of ice crystals will normally occur, leading to gross deterioration of the diffraction pattern (Low *et al.*, 1966). Haas (1968) attempted to preserve crystal perfection by soaking crystals of lysozyme and of lactate dehydrogenase, cross-linked with glutaraldehyde, in glycerol or sucrose, then instantaneously transferring them into liquid nitrogen so that a glass was formed (Haas, 1968; Haas and Rossmann, 1970). Using this technique, a tenfold reduction in the rate of radiation damage was achieved, but the overall quality of the diffraction data measured at $-75°$ appears to be somewhat inferior to that collected at room temperature. The overall "temperature factor" was substantially unaltered. This method has not been widely adopted.

3. *Diffractometric Methods*

The fundamental difficulty in the measurement of diffraction data from protein crystals is that the intensity of diffraction is weak. Since the average reflection intensity is proportional to the reciprocal of the unit cell volume, the problem tends to become more serious with increasing protein molecular weight. Diffractometers, discussed in detail by Arndt and Willis (1966), were at one stage thought to provide

the answer to the problem of low intensity, since any desired accuracy could be achieved by counting a given reflection for a long enough period of time. However, for proteins, compromises must be made, partly because the crystal decays during X-ray exposure and also because hundreds of thousands of reflections have to be measured.

In addition to increased instrumental sophistication (Arndt and Willis, 1966), a number of improvements have been made in measurement strategy which have led to more accurate and more efficient data collection. A number of relatively recent developments are described in papers presented at the *International Meeting on Accurate Determination of X-Ray Intensities*, edited by Arndt and Mathieson (1969). In studies of many proteins, intensities have been measured by scanning through a reflection peak and subtracting the background intensity measured at the extremes of the scan. Wyckoff *et al.* (1967a) showed that a substantial improvement in efficiency could be achieved by measurement only at points close to the reflection maximum. A somewhat similar technique, in which the integrated intensity is estimated from a single measurement with the crystal and the detector both stationary, was used in the study of ribonuclease A (Furnas and Harker, 1955; Magdoff and Crick, 1955; Kartha *et al.*, 1967).

Methods in which integrated intensities are estimated from a set of measurements spanning the reflection maximum, using different types of ordinate analysis, have been described by Diamond (1969), Watson *et al.* (1970), and Watenpaugh *et al.* (1973b). An advantage of measuring only near the center of the peak is that the number of counts will be large and the counting statistics more favorable. Since the peak counts will be generally much greater than background counts, the latter can be measured as a function of scattering angle rather than individually for each reflection. The dangers of the method are that it assumes the peak shapes to be at least roughly constant and has little tolerance for crystal or instrumental misalignment.

Another potential method of increasing data collection efficiency is by simultaneous measurement of more than one reflection (Phillips, 1964; Arndt *et al.*, 1964). Since each of the reflections to be measured will have maximum intensity at slightly different crystal alignments, it is necessary with this technique to increase the scan width, causing some loss of accuracy. Also, the method is not readily compatible with the peak-maximum technique of Wyckoff *et al.* (1967a), although compromises have been proposed (Diamond, 1969). In principle, this method ought to substantially increase data collection efficiency, but it has not been widely adopted, perhaps due to the increased instrumental sophistication that is required. Also, Wyckoff has devised a

novel scheme whereby Friedel-related reflections may be measured simultaneously by having two X-ray sources and two symmetrically placed detectors on opposite sides of the crystal (Wyckoff *et al.*, 1967a; H. W. Wyckoff, personal communication).

It is usually necessary to correct diffractometer measurements for errors due to absorption, and convenient empirical methods have been given by Furnas (1955), North *et al.* (1968), Kopfmann and Huber (1968), and Huber and Kopfmann (1969).

4. Photographic Methods

Films have the great advantage over conventional X-ray counters that they are area detectors and can record many reflections simultaneously (Fig. 21). Although photographic methods were used to eluci-

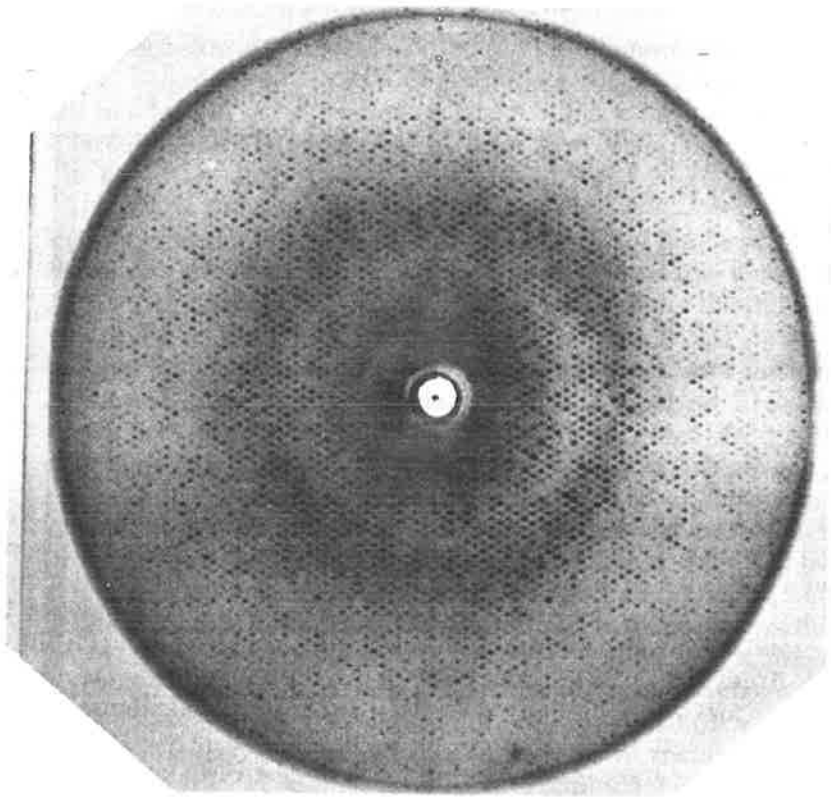


Fig. 21 Precession photograph of the (*hk*0) zone of crystalline thermolysin. The precession angle is 23°, corresponding to diffraction to Bragg spacings of about 2.0 Å. (Elliot rotating anode generator, 40 kV 40 mA, 200 μ focal spot, crystal 0.4 × 0.4 × 0.5 mm, exposure 40 hr.)

date the structures of myoglobin and α-chymotrypsin, diffractometers were employed for most of the early protein structure determinations (see Table I). Within the last few years there has been a revival of interest in the photographic technique due, in large part, to the development of a series of automated devices for film measurement (Drenth *et al.*, 1965; Abrahamsson, 1966; Arndt *et al.*, 1968; Xuong, 1969).

Conventional Buerger precession photographs (Buerger, 1964) of a protein crystal are not the most efficient way of recording data but do have a number of practical advantages for data collection, not the least of which are that the photographic record is permanent and that "bad" data are immediately obvious. Also, the Cartesian geometry allows very easy and efficient data measurement (Arndt *et al.*, 1968; Werner, 1970; Nockolds and Kretsinger, 1970; Matthews *et al.*, 1972c).

The film scanners in use at the present time are of two quite different types. The first class is typified by the computer-linked cathode ray instrument described by Arndt *et al.* (1968), in which the light spot of a high resolution cathode ray tube is focused on the film and the transmitted light measured to give the desired optical density. The position of the light spot is under computer control and can therefore be used to scan any portion of the film at will, so that the instrument is very flexible. In fact, the same instrument can be used to display both contoured electron density maps and drawings of molecular models. For film measurement, the cathode ray scanner has rather low positional accuracy, which was in practice partly compensated for by the computer, but it was still necessary to "refine" the position of the light spot onto each reflection before density measurement (Arndt *et al.*, 1968). Another limitation was introduced by a small amount of background light from a halo around the light source and from the face of the cathode ray tube, which caused optical densities greater than about 1.2 to become increasingly inaccurate.

The other type of film scanner most commonly used at the present time has the film wrapped around a rotating cylinder. An optical head mounted on a lead screw moves a light source inside and a photoelectric detector outside the film along a path parallel to its axis of rotation (Abrahamsson, 1966; Xuong, 1969). The film densities can be read sequentially and stored on a magnetic tape or disk for subsequent processing (Abrahamsson, 1966; Xuong, 1969; Werner, 1969; Nockolds and Kretsinger, 1970), but there are additional advantages if the film scanner is operated under computer control and intensities are measured "on line" (Matthews *et al.*, 1972c). The drum scanner has high positional accuracy so that once a set of "reference reflections" on a

film has been located, the remainder of the reflections can be located with enough accuracy to eliminate the time-consuming recentering process. Also, drum scanners can accurately measure optical densities up to 3.0, so that by making an appropriate parabolic correction to allow for the nonlinearity between X-ray intensity and photographic density, most reflections for a typical protein crystal can be measured from a pack of two films (Matthews *et al.*, 1972c).

Recently, other modes of photographic data collection for proteins have been developed in which the crystal is rocked through a small angle and all reflections generated are recorded on the film, not just those selected by a layer line screen, as in the conventional techniques. A number of rocking motions are feasible: for example, Milledge (1963), Abrahamsson *et al.* (1968), and Schrauber (1966) have used an oscillation geometry, whereas Xuong and co-workers have suggested that a precession motion is slightly more efficient (Xuong *et al.*, 1968; Xuong and Freer, 1971). The screenless precission method was used exclusively in the determination of the structure of *Chromatium* high-potential iron protein (Carter *et al.*, 1971) and used in conjunction with an automatic diffractometer for the structure determination of chymotrypsinogen (Freer *et al.*, 1970). The potential advantage of the oscillation, as opposed to the precession, technique is that it allows a series of contiguous exposures so that a useful measurement can be made even for a reflection which is partially recorded during two or more consecutive photographs. This method of data collection may be essential for measurement of data for proteins or viruses with large unit cells, which require small rocking angles. It has been tested with some success by Arndt *et al.* (1973) but has not yet been proved by a high resolution structure determination. Some factors that are important in optimizing the strategy of intensity measurement have been discussed by Arndt (1968).

5. *Rapid Data Acquisition*

There is great interest at the present time in the development of high resolution area detectors for X-rays which, if successful, would revolutionize intensity measurement for macromolecules. Using such a detector, one could in principle collect all the X-rays scattered from a single protein crystal as it slowly rotated in the X-ray beam, and by using a computer to continuously process the data, one could obtain in a very short period of time (minutes to hours) the entire diffraction pattern of the protein. Arndt (1969) and Minor *et al.* (1974) are exploring the use of a fluorescent screen, image intensifiers, and a television camera for high-speed data acquisition, and Xuong and co-

workers (Cork *et al.*, 1973) have developed a very promising system using a multiwire proportional chamber.

IV. PROTEIN STRUCTURES

In 1964, when Dickerson reviewed X-ray analysis and protein structure in the previous edition of "The Proteins," the structure of myoglobin had been determined to atomic resolution, and low resolution electron density maps had been calculated for hemoglobin, chymotrypsinogen, lysozyme, and ribonuclease. Now, more than 60 structures (including some duplication) have been independently determined to essentially atomic resolution (Table I) and, in addition, the crystallographic findings have been complemented and extended by many other biochemical and biophysical techniques. Protein crystallography has come of age.

An evening's perusal of the recent *Cold Spring Harbor Symposium on Quantitative Biology* (1971) on the structure and function of proteins will make clear the near impossibility of attempting to describe, in a single chapter, the enormous amount of structural data that has been obtained within the last several years. One can hope to do no more than briefly describe some of the highlights and refer the interested reader to the more recent literature for details.

A. Heme Proteins

1. Myoglobin

The general nature of the sperm whale metmyoglobin structure (Kendrew *et al.*, 1960; Kendrew, 1962), with its preponderance of α-helices and largely buried heme group, is perhaps best illustrated by Dickerson's (1964) well-known diagram (Fig. 22). Watson (1969) has discussed the detailed stereochemistry of the molecule, and Dickerson and Geis (1969) have given a profusely illustrated, nontechnical description of myoglobin and its relation to hemoglobin. Scouloudi (1969, and personal communications) has shown the structure of seal myoglobin to be very similar to that of sperm whale myoglobin.

Stryer *et al.* (1964) pioneered the use of the difference Fourier technique by showing that, in the azide complex of sperm whale metmyoglobin, the water molecule at the sixth coordination position of the heme iron is displaced but that the structure of the protein is unaffected. Deoxymyoglobin has the same conformation as metmyoglobin but lacks the coordinated water molecule (Nobbs *et al.*, 1966).

In an interesting study, Schoenborn *et al.* (1965) showed that xenon binds to an internal site in the myoglobin molecule and also to the α and β chains of hemoglobin, although the hemoglobin and myoglobin binding sites were quite different (Schoenborn, 1965). Xenon is an anesthetic agent in man, although it is not clear whether or not this is related to the observed xenon binding.

2. *Hemoglobin*

In contrast to myoglobin, a monomer, hemoglobin is a tetramer consisting of two α subunits and two β subunits. Myoglobin and hemoglobin both bind oxygen, but only the latter exhibits cooperativity, i.e., the binding of oxygen at one heme enhances the oxygen affinity of the

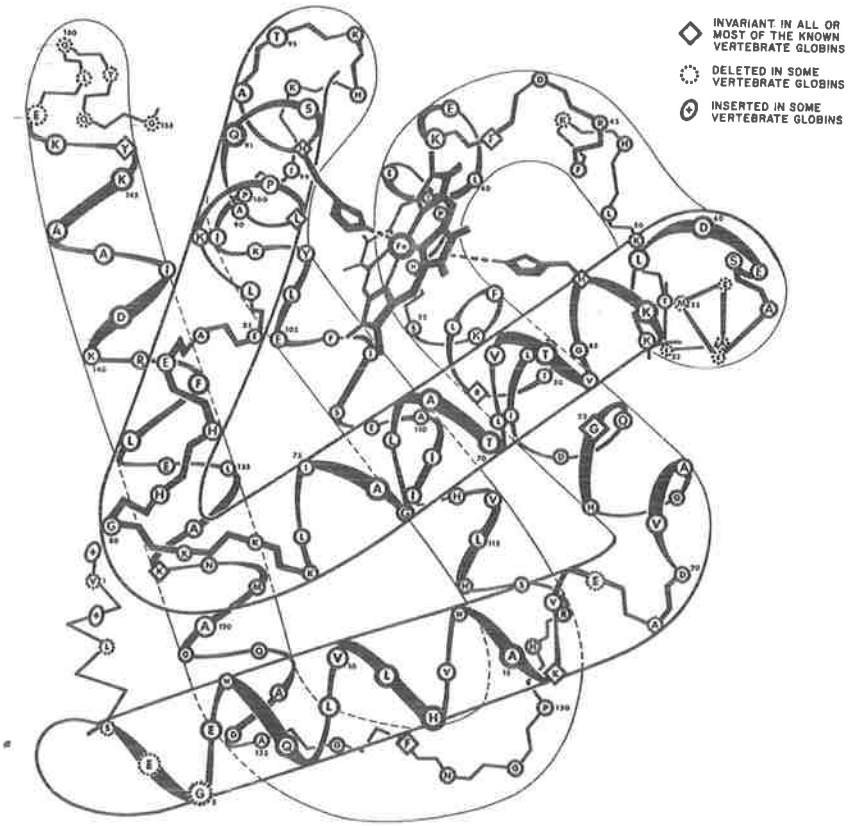


Fig. 22 The backbone of sperm whale myoglobin. After Dickerson (1964). Amino acid residues are identified by the single-letter code. Taken from Dayhoff (1972).

remaining unliganded subunits, suggesting interaction between the different hemes (e.g., see Pauling and Coryell, 1936). It was therefore of the greatest interest to find from the pioneering work of Kendrew and Perutz that the tertiary structures of the α and β subunits of hemoglobin were strikingly similar to that of myoglobin and that in hemoglobin the heme groups were far apart, thus ruling out direct interaction between them (Kendrew *et al.*, 1960; Perutz *et al.*, 1960). The realization that dependence among the ligand binding sites in hemoglobin must be communicated via intersubunit interactions promoted the development of the two alternate models for "allosteric" enzymes which have been the focus of a great deal of the thinking in current enzymology, i.e., the "symmetry driven" model of Monod *et al.* (1965) and the alternative model of Koshland (Koshland *et al.*, 1966; Koshland, 1970; see also Pauling, 1935; Pauling and Coryell, 1936).

The "myoglobin fold" is now known to be common not only to myoglobin and to the α and β chains of horse hemoglobin, but also to hemoglobin from man (Muirhead and Perutz, 1963), the sea lamprey (Hendrickson and Love, 1971), a marine annelid worm (Padlan and Love, 1968, 1974), a larval insect (Huber *et al.*, 1969, 1971a), and lupin root nodules (Vainshtein *et al.*, 1975), and to myoglobin from the harbor seal (Scouloudi, 1969) and the yellow fin tuna (Lattman *et al.*, 1971). Stereographic drawings comparing some of these species are shown in Fig. 20.

Horse oxyhemoglobin has a twofold symmetry axis which relates the α and β subunits in pairs (Bernal *et al.*, 1938). On deoxygenation, the molecule undergoes a substantial change in quaternary structure, accompanied by more subtle changes in tertiary structure within the subunits, but the twofold symmetry axis is preserved (Muirhead *et al.*, 1967). The principal interactions linking the subunits are between the α_1–β_1 subunits and the α_1–β_2 subunits and their symmetry-related counterparts. These contacts, which are extensive, have been described in detail for horse oxy- and deoxyhemoglobin (Perutz *et al.*, 1968b; Bolton and Perutz, 1970). On deoxygenation, the major change is at the α_1–β_2 interface where new sets of intersubunit contacts are formed. The symmetry-related β subunits are not in direct contact, but in the deoxy form move far enough apart for a molecule of 2,3-diphosphoglycerate (DPG) to be interposed, ionically linking the subunits (Arnone, 1972).

Other changes thought to be important include movements of the iron atoms relative to the heme plane. In deoxyhemoglobin, the iron is ferrous, high-spin, five-coordinated, and 0.75 Å from the plane of the four heme nitrogens. The "oxy" hemoglobin used for the X-ray analy-

sis is in fact methemoglobin, with six-coordinated, high-spin ferric iron liganded to water rather than oxygen, and its iron atom 0.3 Å from the heme plane. In oxyhemoglobin, the low-spin ferrous iron is presumed to be coplanar with the heme (Perutz, 1970b). Thus, on oxygenation, the heme-linked histidine would move about 0.75–0.95 Å toward the heme, possibly being the trigger responsible for the other changes in subunit conformation. On the basis of these and other observed structural differences between met- and deoxyhemoglobin, Perutz has proposed detailed stereochemical mechanisms for the heme–heme interaction, the Bohr effect (i.e., the influence of pH on oxygen binding affinity), and the effect of DPG (Perutz, 1969b, 1970b,c, 1972; Perutz and Ten Eyck, 1971). The essential features of the mechanism are that oxygen-free hemoglobin is constrained in its quaternary structure by salt bridges and, if present, by bound DPG. The presence of DPG tends to stabilize the oxygen-free form, thus lowering the oxygen affinity, which is physiologically advantageous (Benesch and Benesch, 1969). On binding of oxygen at either the α or β heme, conformational changes take place within the subunit, causing rupture of salt links constraining the quaternary structure and at the same time releasing Bohr protons. As successive oxygens are bound, additional ionic links are weakened, and at some point the symmetrical quaternary deoxy structure becomes unstable and clicks to the symmetrical oxy form, breaking the remainder of the constraining salt bridges, and expelling DPG.

The crucial information *not* provided by the crystallographic studies is the sequence of oxygen binding and the conformation of possible intermediates. Other techniques provide convincing evidence for intermediate tertiary conformations (Ogawa and McConnell, 1967; Huestis and Raftery, 1972a,b) but these do not exclude the possibility that the intersubunit contacts may be restricted to either the oxy or the deoxy forms.

3. *Cytochromes*

Cytochrome *c* is a relatively small heme protein, important as an electron carrier in the mitochondria of all aerobic organisms, and as such is widely distributed throughout nature. Dickerson and coworkers have determined the structure of horse cytochrome *c* in the oxidized (ferri) form, and of tuna cytochrome *c* in the reduced (ferro) form (Dickerson *et al.*, 1971; Takano *et al.*, 1971, 1973). (Tuna was chosen for the reduced study because it gave suitable crystals, whereas reduced horse cytochrome did not.) Mathews *et al.* (1971a) have determined the structure of calf liver cytochrome b_5, and Sa-

lemme *et al.* (1973), the structure of cytochrome c_2 of *Rhodospirillum rubrum*, both to 2.0 Å resolution.

In horse cytochrome *c*, the polypeptide chain of 104 amino acids wraps around the heme in such a way that the porphyrin ring is tightly enveloped in hydrophobic groups (Fig. 23). It appears that the principal influence on the folding of the protein is the heme itself, and the necessity of providing the observed tightly packed hydrophobic environment around the heme is thought to account for many of the unchanging features in the amino acid sequences of cytochromes from 30 different species (Dickerson *et al.*, 1971, and references therein).

In the electron density map of reduced cytochrome *c*, substantial conformational differences are observed (Takano *et al.*, 1973), although not as extensive as at first supposed (Takano *et al.*, 1971). In contrast, Kakudo *et al.* (1972) have determined the structure of bonito ferrocytochrome *c* at 2.3 Å resolution and report that, in the absence of a reducing agent, oxidation may be achieved without deterioration of the crystals. A difference Fourier synthesis showed that the oxidation–reduction reaction in the crystalline state did not bring about a large change in the main chain folding. Crystals of oxidized and reduced horse liver cytochrome b_5 are also isomorphous and exhibit no differences in main chain conformation or internal side chain orientation (Argos and Mathews, 1975).

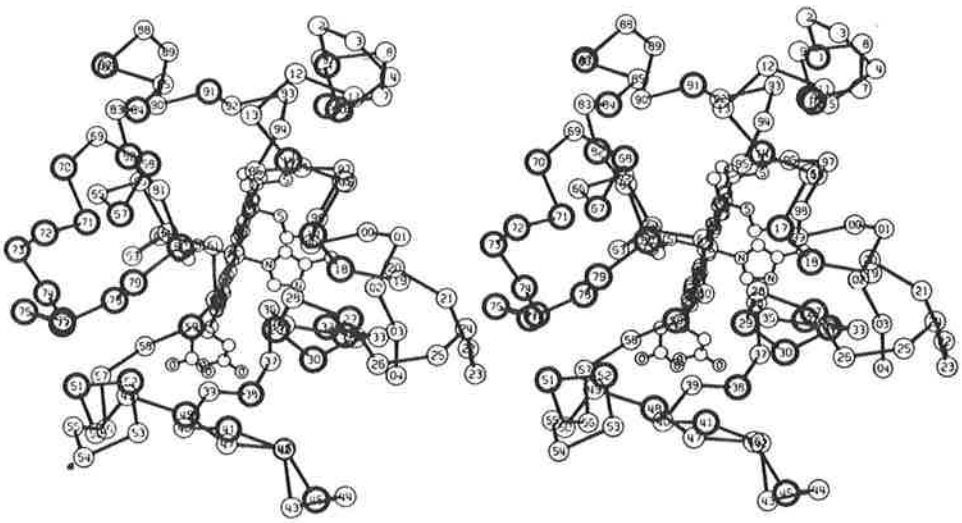


Fig. 23 Stereo drawing of horse cytochrome *c*. The 35 invariant residues among a sample of 29 cytochromes *c* are marked with heavy circles. Note the invariance at the left side of the heme crevice. Taken from Dickerson *et al.* (1971).

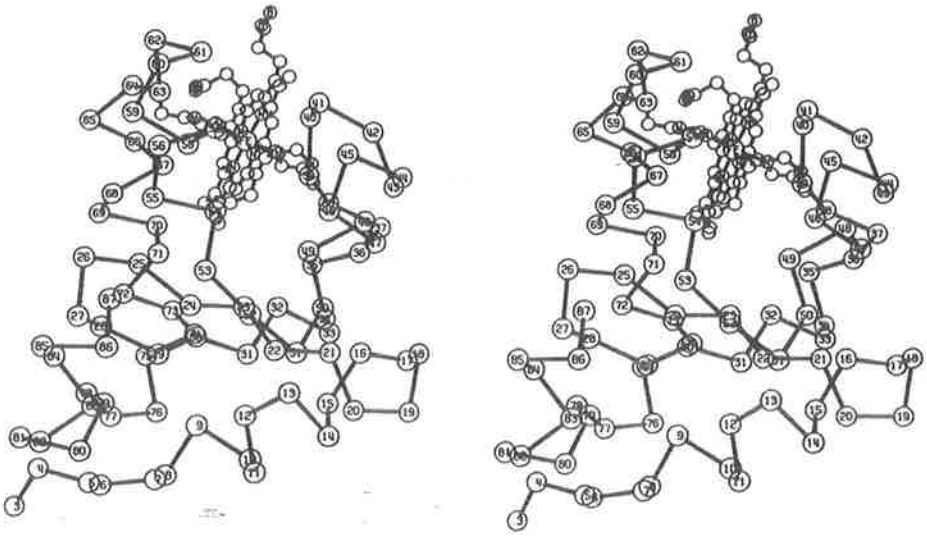


Fig. 24 Stereo drawing of the α-carbon chain of cytochrome b_5, including the heme group and the side chain atoms of His 39 and His 63. Taken from Mathews *et al.* (1971a).

The overall conformation of cytochrome b_5 is illustrated in Fig. 24 (Mathews *et al.*, 1971a). As in cytochrome *c* and myoglobin, the heme is buried, but the folding of the polypeptide chains for these three heme proteins is quite different. For example, myoglobin has 80% helix, cytochrome b_5 has about 50% helix and 25% β structure, and cytochrome *c* has very little regular secondary structure. The coordination of the iron in cytochrome b_5 is symmetrical, with imidazole or pyrole nitrogen atoms at each of the six octahedral ligand positions, whereas in myoglobin and cytochrome *c* the sixth coordination site is occupied respectively by oxygen and sulfur (of Met 80). In myoglobin and cytochrome b_5 the heme group is bound noncovalently with the propionic acid groups exposed to the solvent, whereas in cytochrome *c* the heme group is bound covalently by thioether bonds between vinyl groups and cysteine residues, with one propionic acid group hydrogen-bonded to the interior of the molecule (Mathews *et al.*, 1971a).

The overall three-dimensional structure of cytochrome c_2 from the procaryote *R. rubrum* is substantially similar to that of eukaryotic cytochrome *c* (Dickerson, 1971; Salemme *et al.*, 1973). Furthermore, Timkovich and Dickerson (1973) have observed the same folding to occur in the cytochrome from a nitrate-respiring bacterium.

B. Nonheme Electron Transport Proteins

1. Rubredoxin

The rubredoxins are nonheme iron proteins, isolated from anaerobic bacteria and assume to function in electron transport. It has been shown that rubredoxin can replace ferredoxin as an electron carrier in certain reactions (Lovenberg and Sobel, 1965), although unlike the ferredoxins the rubredoxins do not contain acid-labile sulfur atoms (Valentine, 1964).

The structure of rubredoxin from *Clostridium pasteurianum* (Lovenberg and Sobel, 1965) has a molecular weight of 6000 and contains a single iron atom and four cysteine residues per molecule. The three-dimensional structure was first determined at 3 Å resolution from a single isomorphous derivative, including anomalous scattering (see Section II,C,6). Subsequently, a second derivative was included in the phase determination, and the resolution was extended to 2.5 Å (Herriott *et al.*, 1970) and more recently to 1.5 Å (Watenpaugh *et al.*, 1971, 1973a). Although the amino acid sequence was not known, about forty of the fifty-four residues were correctly identified from a 2.0 Å resolution electron density map.

As was discussed in Section II,D,3, the successful refinement of the rubredoxin structure by Jensen and co-workers represents a substantial breakthrough in this aspect of macromolecular crystallography.

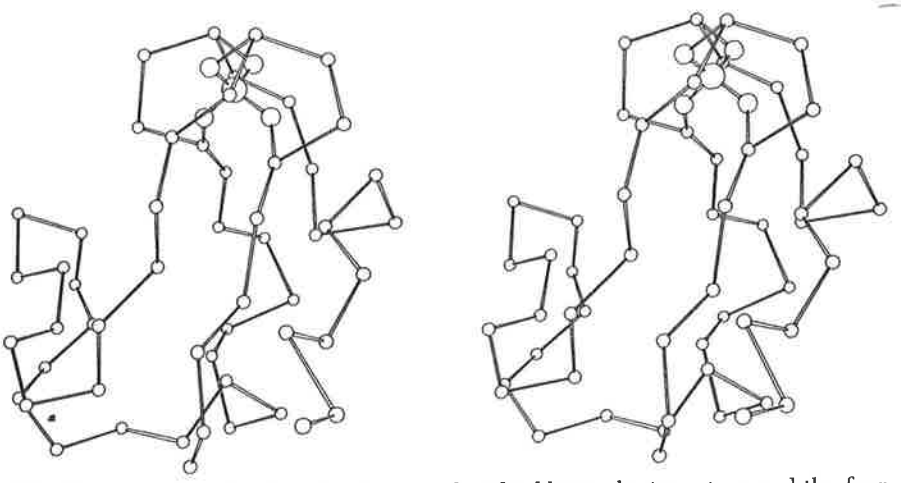


Fig. 25 Stereo drawing showing the α-carbon backbone, the iron atom, and the four iron–sulfur bonds of rubredoxin. Taken from Watenpaugh *et al.* (1973a).

The three-dimensional structure of rubredoxin is shown in Fig. 25. The four cysteines which participate in iron-binding occur at positions 6, 9, 39, and 42, only slightly different from the usual sequence of 6, 9, 38, and 41 observed in most rubredoxins (e.g., see Dayhoff, 1972), suggesting that their tertiary structures are very similar.

During the refinement of the rubredoxin coordinates (Watenpaugh *et al.*, 1973a) it appeared that one of the four chemically equivalent iron–sulfur bonds might be anomalously short, but in further refinement the bond lengths have become more nearly equal (L. H. Jensen, personal communication).

The refinement of the rubredoxin structure also permitted an analysis of peptide nonplanarity (Ramachandran, 1968; Winkler and Dunitz, 1971). It was concluded that, of the fifty-one peptides tested, four were nonplanar at the 0.99 confidence level, with observed distortions of about 10°–15° (Watenpaugh *et al.*, 1973a; L. H. Jensen, private communication).

2. *Ferredoxin*

The ferredoxins are nonheme iron proteins found in anaerobic and photosynthetic bacteria, in algae, and in higher plants. From a biochemical point of view, they have been much more extensively studied than the rubredoxins (e.g., see Dayhoff, 1972), yet for a number of years the X-ray work was totally frustrated by the lack of suitable isomorphous derivatives (Sieker and Jensen, 1965). Recently, however, Jensen and co-workers (Sieker *et al.*, 1972) overcame technical difficulties due to instability of the protein in the presence of oxygen and lability of the iron–sulfur complex and obtained a 2.5 Å resolution electron density map of the clostridial-type ferredoxin from *Peptococcus*, formerly called *aerogenes*. Improved electron density maps (Adman *et al.*, 1973) have confirmed the initial finding that this protein (MW 6000) contains two similar iron–sulfur clusters, about 12 Å apart, which can be fit rather well by a model of four iron and four sulfur atoms at alternate corners of a cube, with four more sulfur atoms projecting from the iron atoms. This is similar to the arrangement reported to occur in high potential iron protein (see below). The fact that there are two similar iron–sulfur clusters in ferredoxin and that in each the polypeptide chain is similarly folded (Sieker *et al.*, 1972; Adman *et al.*, 1973) is consistent with the evidence from sequence data that gene doubling occurred during the evolutionary history of the ferredoxins (e.g., see Dayhoff, 1972).

3. *High Potential Iron Protein*

High potential iron protein is distinguished from the other iron–sulfur proteins by having an oxidation–reduction potential of +0.35 V, in comparison to −0.06 V for rubredoxin and −0.41 V for clostridial ferredoxin; however, no enzymatic function for the protein has been demonstrated (e.g., see Carter *et al.*, 1971).

The protein from *Chromatium vinosum* has a molecular weight of 9652 (Dus *et al.*, 1971) and an inorganic prosthetic group composed of four sulfides and four iron atoms bound to the apoprotein by four cysteine sulfur atoms.

The early crystallographic studies by Kraut and colleagues (1968; Strahs and Kraut, 1968), which suggested that the four iron atoms were located in a compact cluster and were arranged tetrahedrally, have been confirmed by structure analysis and refinement at 2.0 Å resolution (Carter *et al.*, 1974a; Freer *et al.*, 1975).

The geometry of the iron–sulfur complex, as determined from the protein electron density map and from Bijvoet different functions (Strahs and Kraut, 1968), is shown in Fig. 26. The redox center is in approximately the center of the molecule in a highly hydrophobic environment and apparently is inaccessible to water. The iron–iron atom arrangement was close to tetrahedral and became geometrically more regular as it was refined. On oxidation and reduction, the protein crystals remain isomorphous and little if any conformational change was observed to take place, except a slight contraction of the iron–sulfur cluster on oxidation (Carter *et al.*, 1974b).

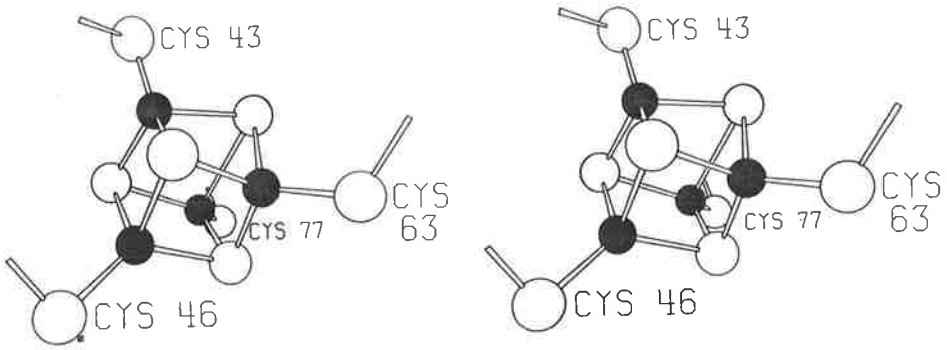


Fig. 26 Stereo representation of the iron–sulfur cluster in *Chromatium* high potential iron protein. Black spheres indicate iron atoms; white spheres indicate sulfur atoms. Taken from Carter *et al.* (1971).

4. *Flavodoxin*

Like the other proteins discussed in this section, the flavodoxins are electron-transport proteins; unlike the others, they bind one flavin mononucleotide prosthetic group per molecule rather than an iron–sulfur complex. Their molecular weights are about 16,000, and they have the ability to function interchangeably with ferredoxin in ferredoxin-dependent reactions.

Ludwig and co-workers (1969, 1971) have studied the protein from *Clostridium* MP and found that the oxidized and semiquinone forms crystallize isomorphously, although with significant intensity differences, indicating some structural change. Independent electron density maps of the semiquinone form at 3.25 Å resolution and of the fully oxidized enzyme at 1.9 Å resolution have been obtained (Ludwig *et al.*, 1971; Andersen *et al.*, 1972; Burnett *et al.*, 1974).

Using a single isomorphous derivative, Watenpaugh *et al.* (1972, 1973b) have obtained 2.5 Å and 2.0 Å resolution electron density maps of the oxidized form of flavodoxin from *Desulfovibrio vulgaris*.

The structures of the two flavodoxins appear to be very similar (Fig. 27a,b), each having a five-strand pleated sheet core with two long helices on either side of the sheet, although the orientation and interactions of the respective flavin rings differ strikingly. The oxidized and semiquinone forms of *Clostridium* MP flavodoxin have very similar conformations, despite large differences in X-ray intensities that are presumably due to bodily movements of the molecule within the unit cell.

C. Proteolytic Enzymes

The proteolytic enzymes rival the globins for the distinction of being the family of macromolecules most studied by X-ray diffraction (see Table I).

Structures of members of three families of proteolytic enzymes have been determined to essentially atomic resolution. First, the DFP-sensitive "serine" proteases, represented by α- and γ-chymotrypsin, their precursor chymotrypsinogen A, elastase, trypsin, and subtilisins BPN′ and Novo; second, the "sulfhydryl" proteases, represented by papain; and third, the "neutral" zinc-containing proteases, represented by carboxypeptidase A and thermolysin. The structures of one trypsin inhibitor and two inhibitor–enzyme complexes have also been determined.

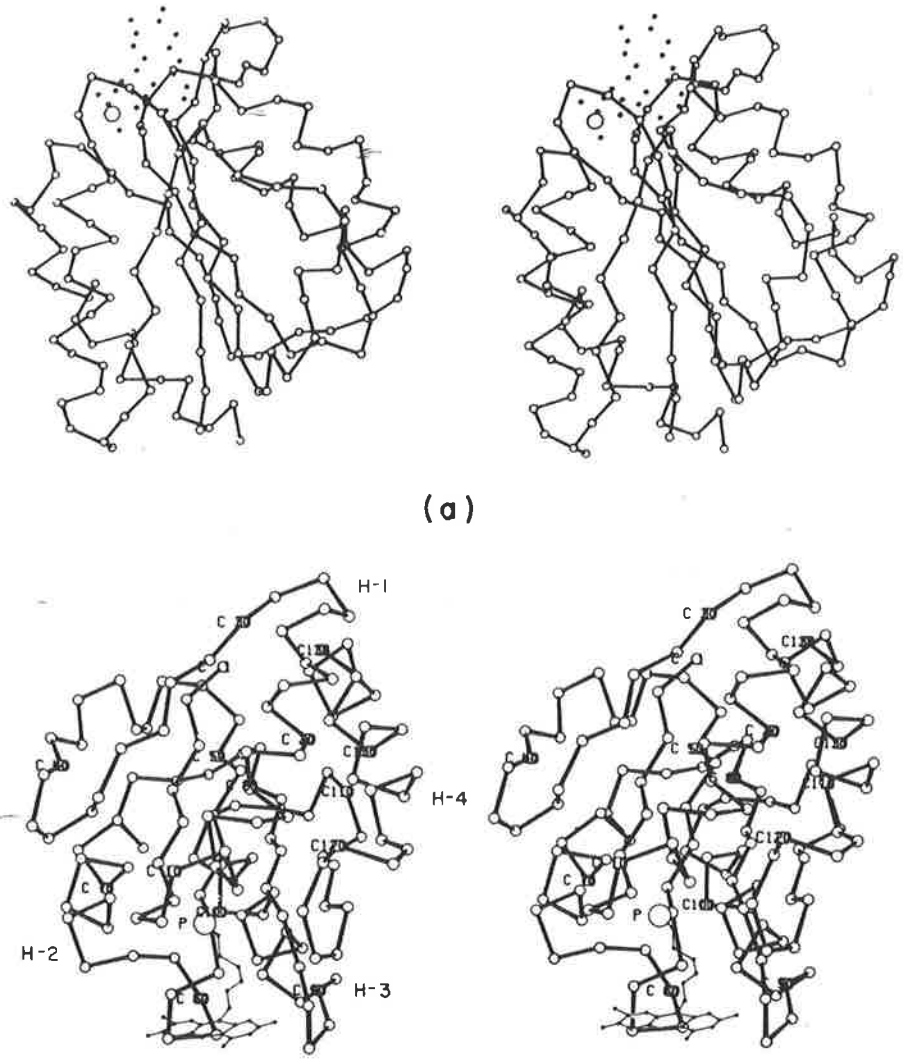


Fig. 27 (a) Stereo view of flavodoxin from *Desulfovibrio vulgaris*. The small solid circles indicate the flavin; the open circle indicates the phosphate. Note the sheet of β-structure through the center of the molecule. Taken from Watenpaugh *et al.* (1972). (b) Stereo representation of the polypeptide conformation proposed on the basis of a 3.25 Å resolution electron density map for flavodoxin from Clostridium MP. Note the similarity to (a). The orientation of the flavin is now thought to be as illustrated in (a) (M. L. Ludwig and L. H. Jensen, personal communication). Taken from Ludwig *et al.* (1971).

1. Serine Proteases

A number of proteolytic enzymes have the common property that they are totally inactivated by diisopropylfluorophosphate (DFP). Although in each case DFP reacts with a specific serine residue, the amino acid sequences of some of these proteases are quite different. First there are the vertebrate enzymes, including the chymotrypsins, trypsin, elastase, and thrombin, whose primary structures are obviously closely related (e.g., see Desnuelle *et al.*, 1970; Hartley, 1970; Dayhoff, 1972); second there are the bacterial serine proteases, such as subtilisin, whose amino acid sequence has no discernible similarity to those of the vertebrate enzymes (e.g., see Smith, 1970); and third there are additional serine proteases, such as the α-lytic protease from *Sorangium* (Smillie and Whitaker, 1967) and the trypsin-like protease from *Streptomyces griseus* (Jurašek *et al.*, 1969), whose amino acid sequences have some similarity to those of the mammalian enzymes, although the homology is extremely weak (Olson *et al.*, 1970; McLachlan and Shotton, 1971).

The mammalian proteases, in particular, have been the subject of extensive study for many years. Chymotrypsinogen A, the inactive precursor of α-chymotrypsin, can be converted by a series of limited proteolytic cleavages into any one of a family of related enzymes, including π-, δ-, α-, and γ-chymotrypsin, which are active, and the neochymotrypsinogens, which are not (e.g., see Blow, 1971).

The essential activation step is the tryptic cleavage of the Arg 15-Ile 16 bond. α-Chymotrypsin and γ-chymotrypsin, now known to be covalently identical, are operationally distinguished by the different ways in which they crystallize (Corey *et al.*, 1965).

The structure of tosyl-α-chymotrypsin, determined by Blow and collaborators from a 2.0 Å resolution electron density map and from the amino acid sequence (Hartley, 1964), is illustrated in Fig. 28 (Matthews *et al.*, 1967; Sigler *et al.*, 1968; Blow, 1971; Birktoft and Blow, 1972). In contrast to the globins, the polypeptide chain is largely nonhelical, most of the secondary structure consisting of irregular, extended antiparallel pleated sheets that form two large "cylinders," each consisting of six polypeptide strands (Blow, 1969). The cylinders are topologically similar, but the corresponding amino acid sequences have no apparent homology. The crystal structure confirmed many of the properties of the serine proteases predicted from earlier studies. For instance, His 57 and Ser 195 were found to be in close proximity, consistent with the view that they might participate catalytically in a concerted manner, and the free amino terminus at Ile 16, essential for activation of the zymogen, was observed to form an in-

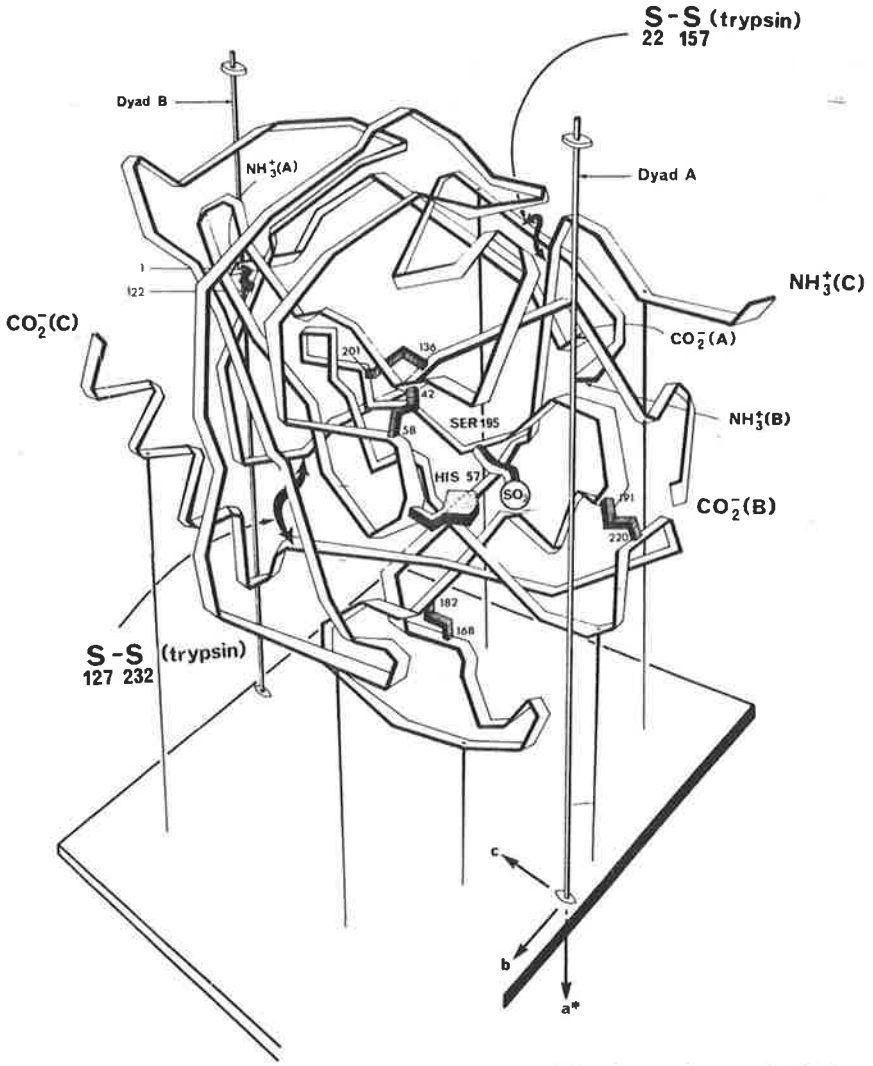


Fig. 28 Diagrammatic drawing showing the course of the three polypeptide chains of α-chymotrypsin. His 57 and a sulfonyl group bound at Ser 195 are also shown, along with the positions at which disulfide linkages (S—S) occur in trypsin, although not in α-chymotrypsin. Taken from Sigler *et al.* (1968).

ternal ionic interaction with Asp 194 (e.g., see Cunningham, 1965; Hess, 1971). Also, the three-dimensional structure of α-chymotrypsin confirmed the assumption that because the mammalian serine proteases had homologous amino acid sequences, their three-dimensional structures must also be very similar. The evidence was

particularly convincing for α-chymotrypsin and trypsin, in which α-carbon atoms linked by disulfide bridges in trypsin (but not in chymotrypsin) were found to be in positions that would allow formation of the disulfide links with minimal perturbation of the structure (Fig. 28).

In a later study, Blow *et al.* (1969) showed that residue 102 was in fact an aspartic acid and was close enough to His 57 to form either a hydrogen bond or an internal ion pair, depending upon the ionization state of the imidazole and the acid. Blow *et al.* (1969) proposed that this ion pair, together with Ser 195, formed a "charge relay system" by which the negatively charged carboxyl group of Asp 102 could transfer electrons via the imidazole of His 57 to the oxygen of Ser 195, thus making it a more powerful nucleophile.

As expected, subsequent structural studies of elastase (Shotton and Watson, 1970; Shotton *et al.*, 1971) and of trypsin (Stroud *et al.*, 1971, 1974) confirmed that their tertiary structures were in fact similar to that of α-chymotrypsin, although differing slightly at the respective substrate binding sites in ways that could account for the observed differences in specificity. The structure of γ-chymotrypsin has also been shown to be extremely similar to that of α-chymotrypsin (Matthews *et al.*, 1968; Wright *et al.*, 1968; Davies *et al.*, 1969; Cohen *et al.*, 1970; Segal *et al.*, 1971a; D. R. Davies, personal communication). Previous low resolution studies had shown that π- and δ-chymotrypsin, the other activation products of chymotrypsinogen, were isomorphous with γ-chymotrypsin and were structurally very similar (Kraut *et al.*, 1967).

Furthermore, Freer *et al.* (1970) determined the structure of chymotrypsinogen A at 2.5 Å resolution and found that its structure also was in general very similar to that of α-chymotrypsin, the mean displacement between corresponding α-carbon atoms, using unrefined coordinates, being 1.8 Å (Kraut, 1971). A few α-carbon atoms, particularly at points where the two dipeptides are absent in α-chymotrypsin, were observed to differ in position by over 10 Å. After tryptic cleavage, Ile 16 appeared to have moved about 11 Å from its position in the zymogen to that in the active enzyme where it interacts with the buried acid group of Asp 194. Contrary to the speculation of Sigler *et al.* (1968), Asp 194 was not found to be exposed to solvent in the zymogen but to be already buried, interacting with His 40. On activation, the movement of Asp 194 appears slight, but the neighboring residues, Gly 193 and Met 192, seem to undergo larger conformational adjustments so that in the zymogen the specificity cavity is lost or at least is incomplete and severely distorted. However, the "charge relay system" was found to be preformed in the zymogen, and the con-

formations of Asp 102, His 57, and Ser 195 in the inactive precursor were "almost indistinguishable" from those in the active enzyme (Kraut, 1971). Thus the question remains as to why the zymogen is not catalytically active, albeit with different specificity.

Recently, Tulinsky *et al.* (1973) obtained an independent electron density map at 2.8 Å resolution of the crystal form of α-chymotrypsin used by the Cambridge group. The overall structure of the enzyme determined by the two groups was found to be similar, but Tulinsky *et al.* (1973) found significant differences in conformation near the local twofold symmetry axis, which relates two crystallographically independent molecules (Blow *et al.*, 1964). In contrast, Blow and coworkers (Birktoft *et al.*, 1969; Birktoft and Blow, 1972) found that the electron density in their maps was poorly defined in these regions, suggesting a statistical disordering through the crystal. This difference in electron density between two studies of the same crystal form is rather surprising, but might be due to the slightly different crystallization conditions used by the respective groups (Matthews and Bernhard, 1973). Tulinsky *et al.* (1973) grew α-chymotrypsin crystals from approximately half-saturated ammonium sulfate solutions at a pH of about 4.0, whereas Blow and colleagues used similar conditions but included 2%–4% dioxane (v/v) which was found to suppress the twinning characteristic of the monoclinic crystal form (Sigler *et al.*, 1966). Dioxane is known to occupy part of the α-chymotrypsin substrate binding site and to cause conformational changes in the vicinity of the local twofold axis (Steitz *et al.*, 1969). It would be of interest to calculate the electron density difference, preferably both in the presence and absence of dioxane, between the Cambridge and Michigan maps.

The structures of two bacterial serine proteases have been independently determined by X-ray techniques; subtilisin BPN′ to a resolution of 2.5 Å in Kraut's laboratory (Wright *et al.*, 1969; Alden *et al.*, 1970, 1971; Kraut, 1971), and subtilisin Novo to 2.8 Å resolution by Hol in Drenth's laboratory (Hol, 1971; Drenth *et al.*, 1971a, 1972). The primary structures of these two enzymes are now known to be identical (Olaitan *et al.*, 1968; Robertus *et al.*, 1971) and, in spite of the fact that the conditions for crystallization used in the two studies were very different, the two independently determined tertiary structures were found to be extremely similar (see Section III,D).

The overall conformation of subtilisin (Fig. 29) is quite different from that of the mammalian serine proteases. Subtilisin contains eight helical segments and a twisted parallel-chain pleated sheet, in contrast to very little helix and exclusively antiparallel pleated sheets in α-chymotrypsin. The backbone chain folds into three separable

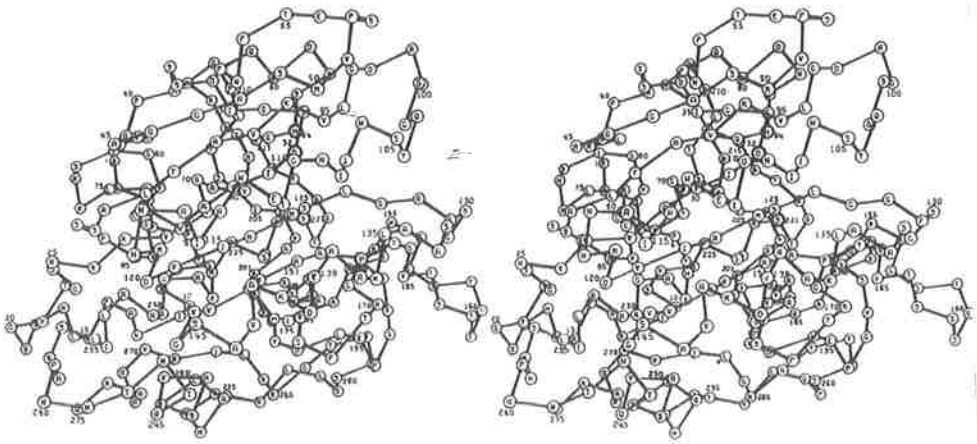


Fig. 29 Stereo drawing illustrating the backbone conformation of subtilisin BPN′. Taken from Dayhoff (1972).

regions, the active site being formed at one of the two places on the molecular surface where the three pieces come together (Wright *et al.*, 1969). In spite of these obvious differences in overall tertiary structure, the active site of subtilisin corresponded almost exactly to that of α-chymotrypsin and its homologues.

Subtilisin was observed to have counterparts to all elements of the "charge relay system" in α-chymotrypsin and, in addition, from studies of the binding of substrates and inhibitors to the two enzymes (Steitz *et al.*, 1969; Henderson, 1970; Kraut *et al.*, 1971; Wright, 1972) and to γ-chymotrypsin (Segal *et al.*, 1971a,b), it was found that the mode of substrate binding was entirely analogous. The similarity between the two classes of enzymes included not only the catalytic but also the binding apparatus: 27 atoms from the analogous active and binding site residues in α-chymotrypin and subtilisin were compared and found to have a rms difference in position of 0.81 Å. This suggested that the active sites were not only similar but were identical within the resolution limits of the electron density maps (Kraut *et al.*, 1971; Robertus *et al.*, 1972a). This truly remarkable degree of structural identity between the active site regions of α-chymotrypsin and subtilisin BPN′ is illustrated in Fig. 30.

In considering the premise that α-chymotrypsin and subtilisin are an example of convergent evolution, Robertus *et al.* (1972a) point out that not only is there no resemblance in their primary structures but also the linear sequences of the components of the respective active

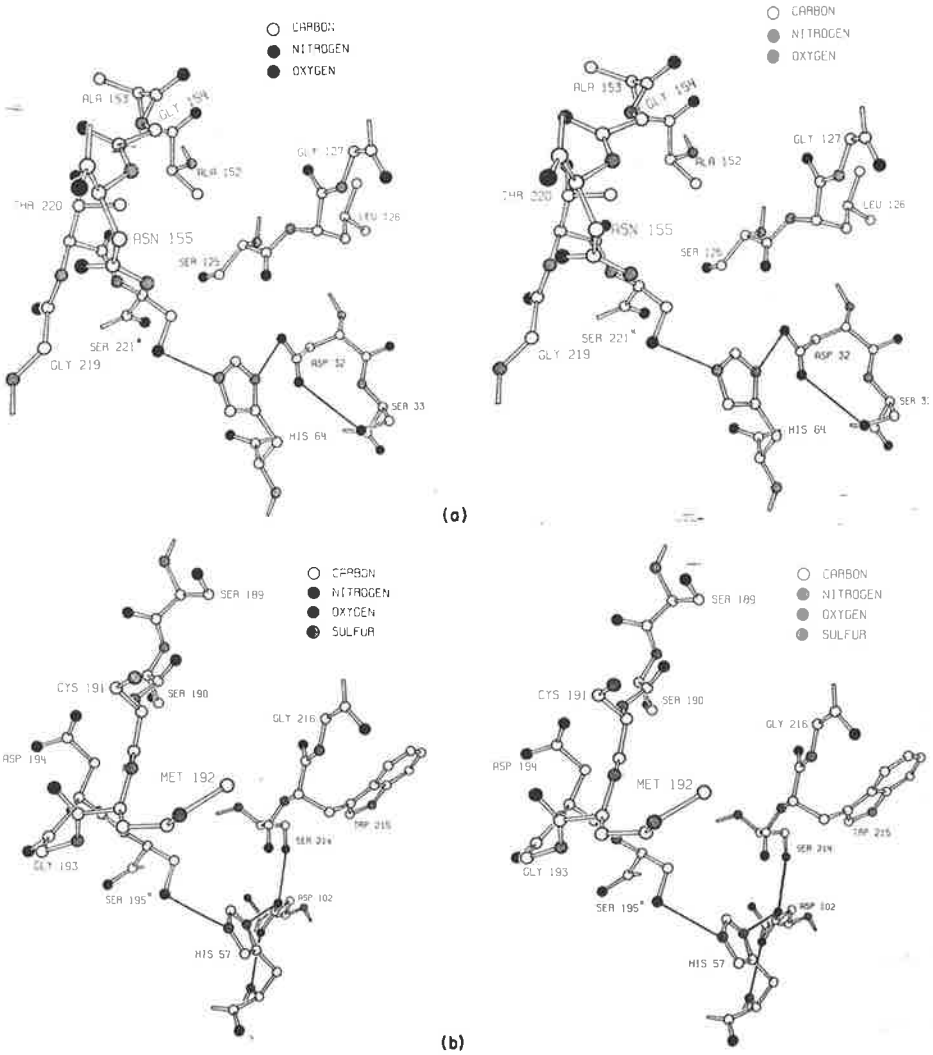


Fig. 30 Comparison of the active site regions of (a) subtilisin BPN′ and (b) α-chymotrypsin. Hydrogen bonds are depicted as thin solid lines. In subtilisin, the "charge relay system" includes Ser 221, His 64, and Asp 32; in α-chymotrypsin, the analogous residues are Ser 195, His 57, and Asp 102. Taken from Robertus *et al.* (1972a).

sites are quite different. It therefore seems highly unlikely that these two enzymes could have descended from a common ancestor in such a way that all the elements of the catalytic machinery were conserved but no vestige of structural or sequence homology should remain for

the rest of the enzyme. Robertus *et al.* (1972a) suggest that the presumed independent evolution of both the charge relay system *and* the same substrate binding apparatus may imply that these two entities must both function together for maximum enzyme efficiency, perhaps achieved by precise orientation of the substrate with respect to the catalytic groups (Storm and Koshland, 1970); see also Section V,B,2).

A number of studies of the binding of inhibitors and substrate analogues have led to a rather consistent picture of the probable mode of interaction of a real substrate with the serine proteases (Steitz *et al.*, 1969; Henderson, 1970; Kraut *et al.*, 1971; Segal *et al.*, 1971a,b; Shotton *et al.*, 1971; Henderson *et al.*, 1971; Robertus *et al.*, 1972a; Wright *et al.*, 1972; Wright, 1972; Krieger *et al.*, 1974). In crystals of α-chymotrypsin, the active site region is partly obstructed by the close proximity of another molecule, thus limiting studies to the binding of smaller inhibitors (Steitz *et al.*, 1969; Henderson, 1970; Henderson *et al.*, 1971). This was not the case with γ-chymotrypsin: Segal *et al.* (1971a,b) were able to make elegant use of N-acetylated mono-, di-, and tripeptide chloromethyl ketones, which reacted with His 57, to determine the probable mode of association of longer substrates. Although the inhibited γ-chymotrypsin crystals were not very isomorphous with the native crystals, they were isomorphous among themselves, thus making it possible to determine the location of the additional peptide in successively larger inhibitors. Parallel studies of subtilisin, inhibited by the chloromethyl ketones, proved equally successful and gave entirely consistent results (Robertus *et al.*, 1972a).

By combining the results obtained for α- and γ-chymotrypsin with those for subtilisin, Robertus *et al.* (1972b) have proposed a hypothetical Michaelis complex between subtilisin BPN′ and the extended peptide substrate, -Phe-Ala-Ala-Tyr↓Leu-Leu- (Fig. 31), where the arrow indicates the bond to be cleaved. It should be noted that the crystallographic evidence for the alignment of the four peptides on the amino side of the scissile bond is quite convincing, but the placement of the leaving portion, -Leu-Leu-, is somewhat speculative. One of the features of the substrate binding includes a region of antiparallel β-structure between the inhibitor and an extended segment of the main chain of the enzyme, consisting of Leu 126-Gly 127-Gly 128 in subtilisin, or Ser 214-Trp 215-Gly 216 in γ-chymotrypsin (Segal *et al.*, 1971a).

On the basis of the crystallographic findings and from model building, Robertus *et al.* (1972b) have presented a possible stereochemical mechanism for the serine proteases involving transition-state stabilization. The proposed mechanism is rather similar to that suggested earlier for α-chymotrypsin (Blow *et al.*, 1969), in which the transfer of

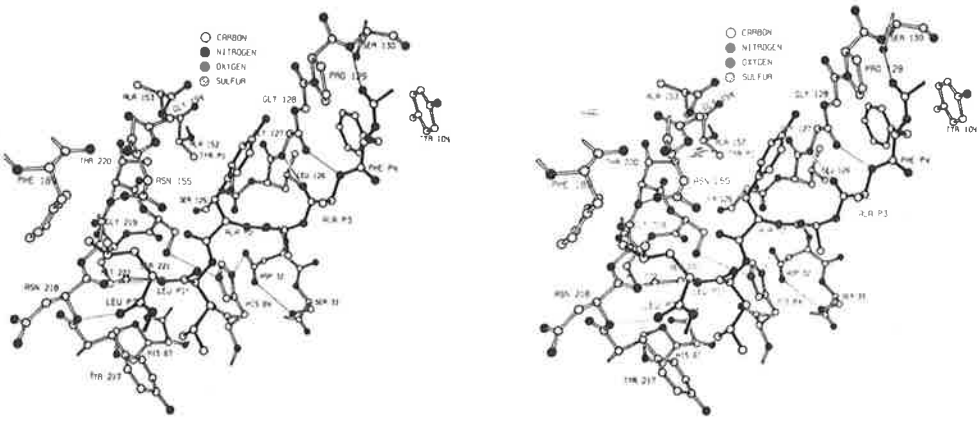


Fig. 31 Hypothetical Michaelis complex between the peptide substrate -Phe-Ala-Ala-Tyr↓Leu-Leu- and the binding channel region of subtilisin BPN'. Taken from Robertus *et al.* (1972b).

the proton from the γ-oxygen of Ser 195 to the amine nitrogen of the substrate is assumed to be facilitated by the charge relay system. Attack of the γ-oxygen on the carbonyl carbon of the substrate would then lead to an acyl enzyme, which would subsequently be deacylated in a reverse manner, with water in the position formerly occupied by the leaving group. The proposed mechanism of Robertus *et al.* (1972b) does, however, differ from that of the Cambridge group in some aspects of the stereochemistry (Blow *et al.*, 1969; Steitz *et al.*, 1969; Henderson, 1970; Henderson *et al.*, 1971; Blow, 1974).

Recently, the structures of two trypsin–trypsin inhibitor complexes have been determined and have helped confirm the probable mode of association of the enzyme with a real substrate. In both the complex of bovine trypsin with bovine pancreatic trypsin inhibitor (Rühlmann *et al.*, 1973; Huber *et al.*, 1974) and of porcine trypsin with soybean trypsin inhibitor (Blow *et al.*, 1974; Sweet *et al.*, 1974), a tetrahedral adduct is seen to form between the carbonyl carbon of the scissile bond of the inhibitor and the active site serine of the enzyme. The mode of association of the enzyme and inhibitor agrees well with models proposed on the basis of the known structures of the pancreatic trypsin inhibitor (Huber *et al.*, 1970, 1971b,c; Stroud *et al.*, 1971; Blow *et al.*, 1972; Deisenhofer and Steigemann, 1975).

2. *Sulfhydryl Proteases*

Papain is the most extensively studied member of the class of proteolytic enzymes requiring a free sulfhydryl group for activity. The three-dimensional structure of papain was determined from a 2.8 Å

resolution electron density map (Drenth *et al.*, 1968, 1971b). The interpretation of the map necessitated several changes in the tentative amino acid sequence (Light *et al.*, 1964). The crystals used for the X-ray analysis were grown from 60% methanol and were not active but had activity when put into 20% Na_2SO_4 (Sluyterman and De Graaf, 1969). This suggests that the structure determined crystallographically (Fig. 32) is very similar if not identical to that of the active enzyme.

As in a number of other proteins, the polypeptide chain folds not as a single unit but around two hydrophobic cores. There are five helical segments, a region of distorted pleated sheet including about thirty residues in parallel and antiparallel conformation, and also a buried ion pair between the side chains of Glu 50 and Arg 83 which, in contrast to the buried ion pairs in α-chymotrypsin, has no obvious catalytic function (Drenth *et al.*, 1968, 1971b,c). At the active center of papain, His 159 has its $N^{\gamma 1}$ 3.4 Å from the sulfur of Cys 25 and also forms a hydrogen-bond via $N^{\epsilon 2}$ to the side chain of Asn 175. It has been suggested that the combination SH...Im...Asn in papain might be an analogue of the arrangement OH...Im...Asp in the serine proteases, and that sulfur, being a stronger nucleophile than oxygen, would have its nucleophilicity increased enough by Im...Asn so that the presence of the acid would not be required (Drenth *et al.*, 1971b,c). Rate profiles for papain show pK_a values of about 4.5 and 8.5, but the ionizing groups to which these should be attributed are in dispute: alternative mechanisms have been proposed by Sluyterman and Wolthers (1969) and by Husain and Lowe (1968). Drenth *et al.* (1971b,c, 1975) and

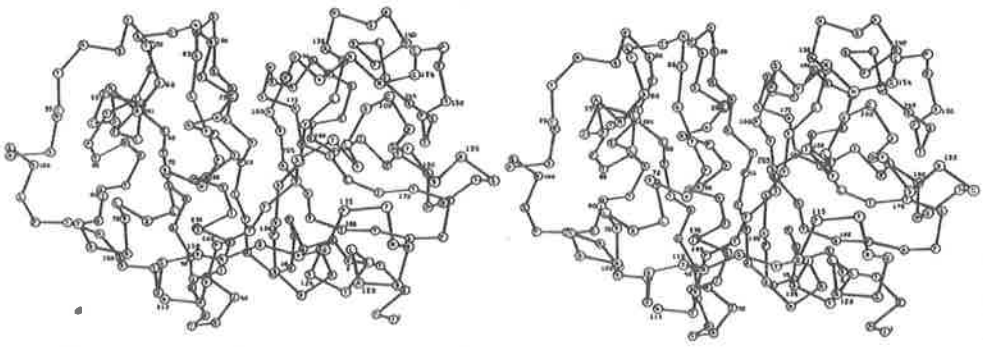


Fig. 32 The backbone conformation of papain. The catalytically important residues His 159 and Cys 25 are close together, although each is in one of the two different lobes of the molecule, which are clearly discernible in the figure. Taken from Dayhoff (1972).

Glazer and Smith (1971) have recently reviewed the structure and the chemistry of papain.

3. *Metalloproteases*

The third class studied crystallographically are proteolytic enzymes that are inhibited by chelating agents such as EDTA and require zinc for catalytic activity. The atomic structures of two members of this group, carboxypeptidase A and thermolysin, have been determined to essentially atomic resolution. Both enzymes have molecular weights of about 34,000 but they differ in that thermolysin is an endopeptidase, whereas carboxypeptidase A cleaves carboxyl-terminal peptide bonds. Both enzymes tend to cleave peptide bonds on the imino side



Fig. 33 The peptide chain of carboxypeptidase A showing the zinc (near center) with its three ligands from the protein, and the single disulfide bond at the right. Taken from Quiocho and Lipscomb (1971).

of bulky hydrophobic residues (excluding tyrosine and tryptophan in the case of thermolysin). This is in contrast to the serine proteases, in which the specificity is directed toward the residue that donates the carbonyl group of the bond to be cleaved.

Electron density maps of bovine carboxypeptidase A, culminating at 2.0 Å resolution, have been obtained by Lipscomb and co-workers (Reeke *et al.*, 1967; Lipscomb, *et al.*, 1968, 1969, 1970; Quiocho and Lipscomb, 1971; Hartsuck and Lipscomb, 1971). The amino acid sequence has been determined chemically by Neurath and co-workers (Bradshaw *et al.*, 1969b). The amino acid sequence of thermolysin has been determined in Neurath's laboratory (Titani *et al.*, 1972), and a 2.3 Å resolution electron density map has been obtained by Matthews *et al.* (1972a). Combination of the chemically determined sequence with the high resolution electron density map resulted in an atomic model of the structure of thermolysin (Matthews *et al.*, 1972b, 1974; Colman *et al.*, 1972a).

For both carboxypeptidase A and thermolysin, it was possible to identify many of the amino acids without reference to the chemically determined amino acid sequences, although, as discussed in Section II,D,2, such an interpretation of the electron density maps could not be used to identify catalytically important residues with certainty. The structure determinations of both carboxypeptidase A and thermolysin illustrate rather well the complementarity of the chemical and X-ray techniques.

The overall structure of carboxypeptidase, shown in Fig. 33, has as its most striking feature a twisted wall of β-structure, including eight strands of parallel and antiparallel pleated sheet extending through the molecule. There are nine helical segments distributed on either side of the central sheet (Lipscomb *et al.*, 1970). In contrast, the thermolysin molecule (Fig. 34) is folded into two distinct lobes, with β-structure predominating in the amino-terminal half of the polypeptide chain and helices predominating in the other half (Matthews *et al.*, 1972a; Colman *et al.*, 1972a). In thermolysin, the essential zinc atom is in a deep cleft formed at the junction of the two lobes. The cleft extends across the molecule, consistent with the prediction of Morihara and Oka (1968) that up to six residues may participate in the binding of thermolysin substrates. In contrast, in carboxypeptidase A the zinc atom lies in a shallow groove which terminates in a large hydrophobic pocket, as might be expected for an exopeptidase. The sheet of β-structure in carboxypeptidase A is twisted by 120° in going from the first to the last strand but is not markedly curved (Lipscomb *et al.*, 1968). In contrast, the extended β-structure in thermolysin is more

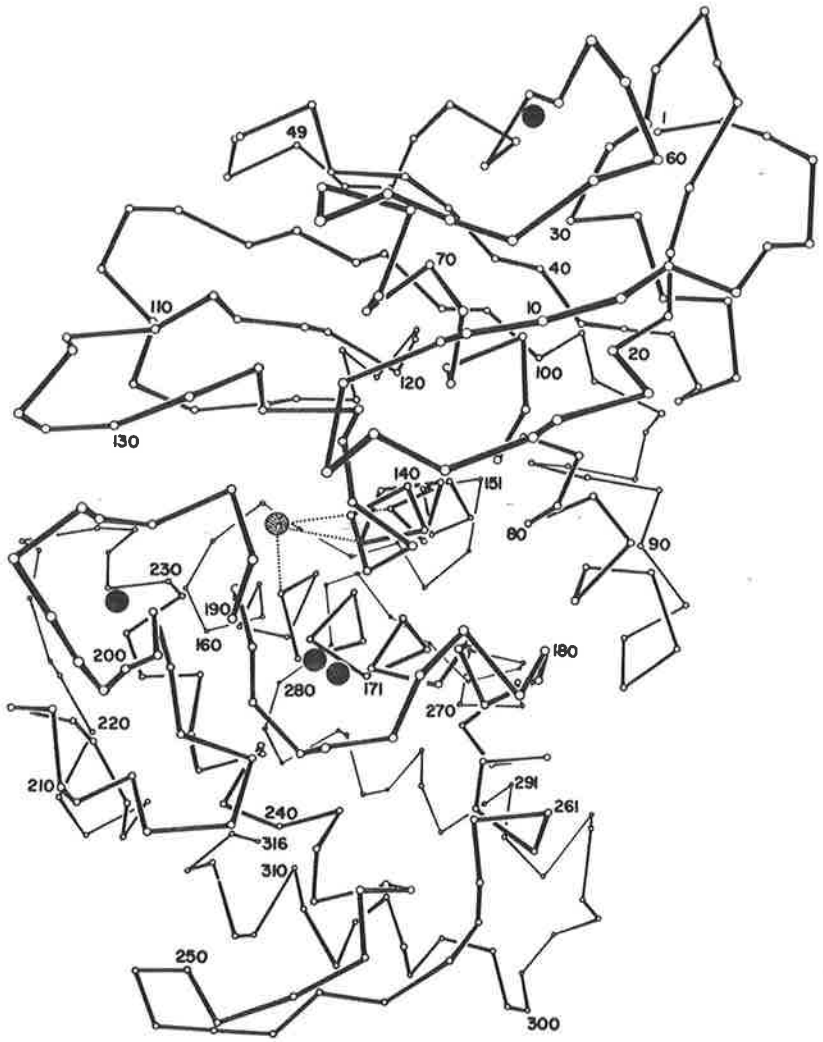


Fig. 34 The peptide chain of thermolysin. The single zinc at the active site is drawn stippled, and the four calcium ions drawn as black balls. The division of the molecule into two lobes is readily apparent.

convoluted and includes a series of five polypeptide chains which close via a calcium ion to form a twisted cylinder (Colman *et al.*, 1972a), reminiscent of those observed in the serine proteases (Blow, 1969; Shotton and Watson, 1970).

After careful study of the excellent electron density map of carboxy-

peptidase A, and taking into account the chemically defined primary structure, Lipscomb and colleagues have concluded that the peptide bond between Ser 197 and Tyr 198 is best interpreted as being in the cis rather than trans conformation (Hartsuck and Lipscomb, 1971). These two residues immediately follow His 196, one of the zinc ligands.

There also appear to be some unusual conformations in thermolysin, e.g., Pro 51, which probably has a cis conformation, and residues 226–229 (-Asp-Asn-Gly-Gly-), which form a turn of left-handed helix (Matthews *et al.*, 1974). Also, thermolysin includes an example of the "γ-turn" (Matthews, 1972) predicted by Némethy and Printz (1972), in which the direction of the polypeptide chain is reversed within the space of three α-carbons rather than four, as in the so-called β-turns (Venkatachalam, 1968; see also Section V,B,*1*).

The binding to carboxypeptidase A of the substrate glycyl-L-tyrosine has been studied in detail by difference Fourier methods at 2 Å resolution and is illustrated in Fig. 35a,b. On binding this dipeptide, several conformational changes are observed, the most striking of which is a rotation of about 120° of the side chain of Tyr 248 about its C^{α}—C^{β} bond, causing its phenolic hydroxyl to move by about 12 Å so that the OH group comes from the surface of the molecule to the vicinity of the peptide bond of the substrate. Additional structural changes include a movement of the guanidinium group of Arg 145 by about 2.0 Å, and a comparable movement in the carboxylate of Glu 270. The three groups that are observed to move all interact with the bound substrate. Arginine 145 forms a salt link with the carboxyl ion of the dipeptide, presumably explaining the specificity of carboxypeptidase A for cleavage of carboxyl-terminal peptide bonds. The oxygen of the phenolic hydroxyl of Tyr 248 is about 2.7 Å from the nitrogen of the scissile peptide bond and presumably forms a hydrogen bond. Finally, the carboxyl of Glu 270 binds through a water molecule to the α-amino group of the dipeptide (Lipscomb *et al.*, 1968, 1970; Hartsuck and Lipscomb, 1971). The observed interaction of Glu 270 is only possible for dipeptides, and it is assumed that with longer productive substrates the carboxyl of Glu 270 would be free to serve in a catalytic role. The carbonyl oxygen of the scissile peptide binds to the zinc, displacing a water molecule bound in the native enzyme. Cleavage of the peptide bond is thought to result from polarization of the carbonyl group by the zinc, followed by attack of the carbonyl carbon by Glu 270, either by a nucleophilic or general base mechanism, and donation of a proton to the peptide amide by Tyr 248.

In spite of the obvious dissimilarity between the overall conforma-

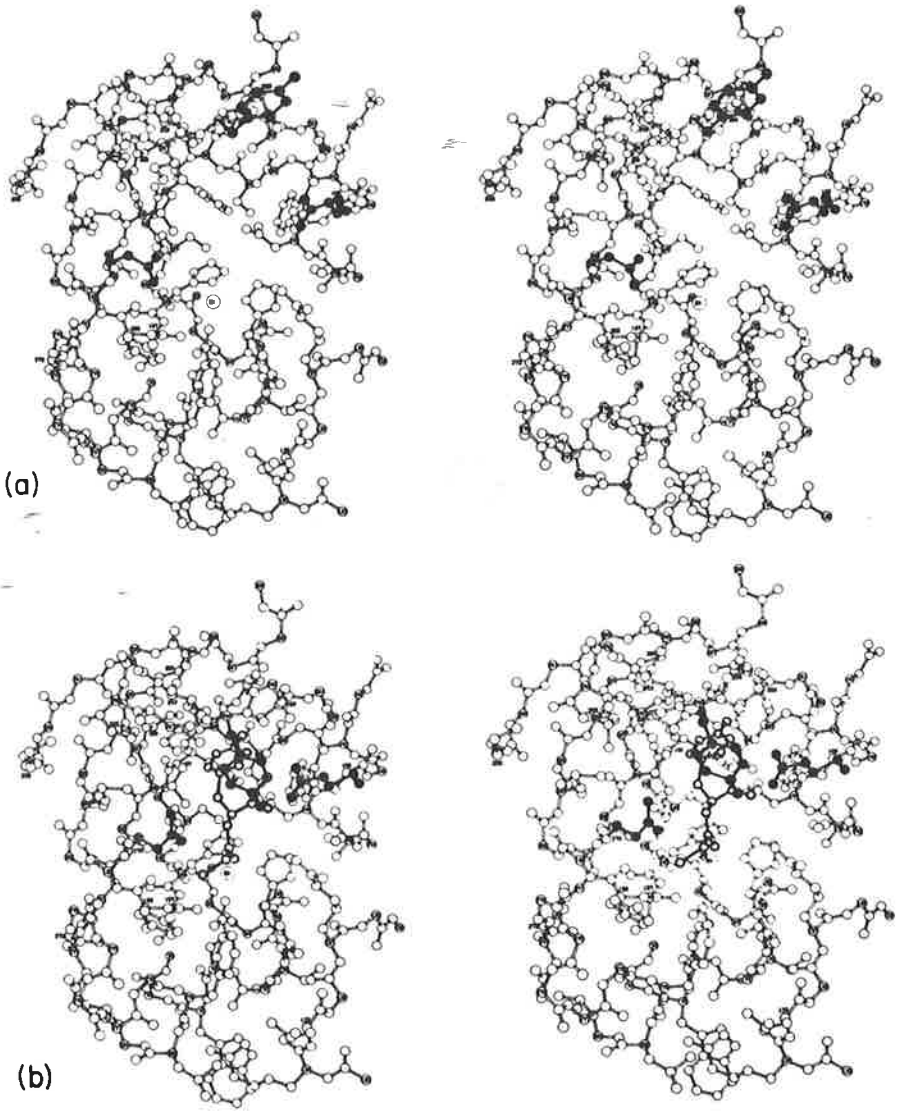


Fig. 35 Stereo drawings of carboxypeptidase A showing, in black, the residues which are thought to be catalytically important. In (a), the active site region is empty. On binding glycyl-L-tyrosine (b), a number of conformational changes take place, as described in the text. Taken from Quiocho and Lipscomb (1971).

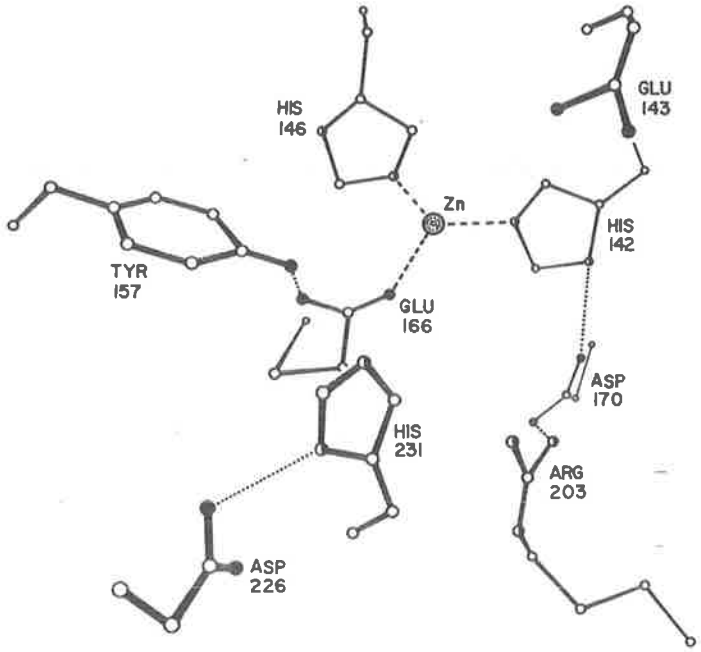


Fig. 36 Sketch illustrating the position of some of the residues in the active site of thermolysin. The direction of viewing is down the fourth ligand direction of the zinc. Taken from Colman *et al.* (1972a).

tions of carboxypeptidase A and thermolysin, there are some striking similarities in their respective active site regions. In both cases the zinc atom has, as ligands, two histidines and an aspartic acid in approximately tetrahedral coordination, with a water molecule presumed to be the fourth ligand. In addition, thermolysin has a glutamic acid (Glu 143) (Fig. 36) which appears to be in a position entirely analogous to that of Glu 270 in carboxypeptidase A. Further similarities between the two proteases include a system of salt links from one of the histidine zinc ligands through an aspartic acid to an arginine residue, i.e., His 69-Asp 142-Arg 145 in native carboxypeptidase A (Lipscomb *et al.*, 1968) and His 142-Asp 170-Arg 203 in thermolysin (Matthews *et al.*, 1972b; Colman *et al.*, 1972a). In view of the presumed role of Arg 145 in carboxypeptidase A in binding the carboxyl group of substrates, it was not anticipated that there would be a counterpart in thermolysin, an endopeptidase. In preliminary experiments (Matthews *et al.*, 1974) it appears that Arg 203 in thermolysin may interact with the carboxyl ion of competitive dipeptide inhibitors of the

enzyme. This finding need not be inconsistent with interpretations based upon the binding of glycyltyrosine to carboxypeptidase A, for it may be that Arg 203 in thermolysin is positioned to help bind extended substrates in a productive mode, whereas the presence of a free carboxyl, as in a dipeptide, could cause the substrate to bind nonproductively.

There is no tyrosine in thermolysin in a position analogous to Tyr 248 in carboxypeptidase A. In contrast, there is an additional histidine (His 231) in the active site region, which participates in an ionic interaction with Asp 226 and has $N^{\epsilon 2}$ of the imidazole 4 Å from the zinc (Fig. 36) (Matthews *et al.*, 1972b, 1974; Colman *et al.*, 1972a). It seems probable that histidine plays a role analogous to Tyr 248 in carboxypeptidase by donating a proton, either directly or through a water molecule, to the scissile peptide. The observed binding of silver ion to His 231, with concomitant loss of enzymatic activity, tends to confirm that it is catalytically important (Colman *et al.*, 1972a).

The structure of thermolysin is also of interest in the light of its unusual thermostability. When heated in aqueous solution for an hour at 80° the enzyme retains about half of its original activity, whereas at 65° practically no inactivation occurs (Endo, 1962; Matsubara, 1967). It has long been known that thermolysin requires calcium for maximum stability, but it has only recently been shown conclusively that calcium is not directly involved in catalytic activity (Feder *et al.*, 1971). Chemical studies indicated that three to four calcium ions were bound to each molecule of the enzyme (Latt *et al.*, 1969; Feder *et al.*, 1971). These results were confirmed and extended by the X-ray studies which demonstrated that four calcium ions were bound with high occupancy and that the closest of these was 13 Å from the active site zinc. Two of the calcium ions were close together, with a center-to-center distance of about 3.8 Å, and were surrounded by a cluster of water molecules, backbone carbonyl groups, and five acid groups (Matthews *et al.*, 1972b, 1974; Colman *et al.*, 1972a,b).

The unusual thermostability of thermolysin (and of thermostable proteins in general, relative to thermolabile ones) is thought to be due to a combination of rather subtle changes in ionic, hydrophobic, and hydrogen bond stabilization, and not to unusual folding or to other structural features absent in thermolabile structures (Matthews *et al.*, 1974). Perutz (1975) has recently reached essentially the same conclusion, emphasizing that the extra energy of stabilization can come in monomeric enzymes from a few extra salt bridges on the molecular surface, and in oligomeric enzymes from extra salt bridges, hydrogen bonds, or nonpolar bonds at the subunit interfaces.

D. Carbonic Anhydrase

Carbonic anhydrase is a zinc-containing enzyme, widely occurring in nature, which reversibly catalyses the hydration of carbon dioxide and certain aldehydes as well as the hydrolysis of esters. Two major forms of the enzyme, denoted carbonic anhydrase B and C, occur in human erythrocytes, and the structures of both have been determined to high resolution (Kannan *et al.*, 1971, 1972, 1975; Liljas *et al.*, 1972). As in carboxypeptidase A, the overall tertiary structures of both carbonic anhydrase B and C consist of a twisted sheet of β-structure composed of ten parallel and antiparallel chain segments, with the polypeptide chain on either side of the central "wall" folded to form a number of helices, all of which are partly exposed to the solvent (Kannan *et al.*, 1971, 1975; Liljas *et al.*, 1972) (Fig. 37). The polypeptide conformation of carbonic anhydrase is, however, quite different from that of either carboxypeptidase or thermolysin.

The active site of carbonic anhydrase is a funnel-shaped cavity, 15 Å deep, that is formed in part by the twisted pleated sheet, with the zinc ion at the base of the cavity (Fig. 37). There are three protein zinc ligands that, pending the complete sequence determination, have been tentatively identified as His 93, His 95, and His 117. Together with a water and hydroxyl ion, these three histidines provide the zinc ion with coordination within 20° of the tetrahedral angle (Kannan *et*

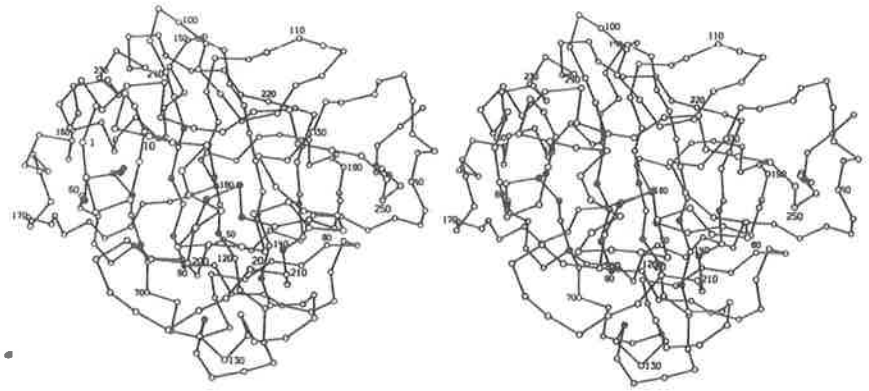


Fig. 37 The α-carbon backbone of carbonic anhydrase. The zinc (not shown) lies in the funnel-shaped cavity extending into the molecule from the left and has, as ligands, residues tentatively assigned as His 93, His 95, and His 117. Taken from Kannan *et al.* (1971).

al., 1971). A number of ordered solvent molecules, presumed to be water molecules, are observed in the active site. Bergstén *et al.* (1972) have studied the complexes of carbonic anhydrase with a number of anion and sulfonamide inhibitors and, in the light of their crystallographic findings, have discussed some of the mechanisms proposed for the action of the enzyme.

E. Insulin

In 1969, Hodgkin and collaborators (see Adams *et al.*, 1969a) reported the successful determination of the three-dimensional structure of pig insulin, culminating an X-ray analysis begun thirty-four years earlier (Crowfoot, 1935). Subsequently, the study has been extended to 1.9 Å resolution (Blundell *et al.*, 1971a,b, 1972). Also, independent studies of the same crystal form have been carried out to a resolution of 1.8 Å by the Peking Insulin Structure Group (1974) and to 3.1 Å resolution by Sakabe *et al.* (1972).

Under conditions used to obtain the rhombohedral crystal form used for the X-ray analysis, the insulin monomers, each consisting of two polypeptide chains (A and B) linked by two disulfide bridges, aggregate to form hexamers. Although it is not clear whether or not the formation of the hexamer is essential for biologic activity (Blundell *et al.*, 1971a), the mode of association of the monomers is of interest and may be a useful model for subunit interactions in oligomeric enzymes (Matthews and Bernhard, 1973). The hexamer consists of three dimers, related by an *exact* threefold symmetry axis. Within each dimer, the monomers are related by *approximate* twofold axes at right angles to the threefold axis, so that the hexamer has pseudo point symmetry 32 (Blundell *et al.*, 1971b). At both the monomer–monomer contacts and the dimer–dimer contacts, significant departures from exact symmetry were observed (cf. Tulinsky *et al.*, 1973). The hexamer binds two zinc ions, each situated on the threefold axis and coordinated by three histidines. In contrast to carboxypeptidase A, carbonic anhydrase, and thermolysin, in which tetrahedral coordination of zinc has been observed, the zinc coordination in insulin is thought to be octahedral. In another manifestation of the inexact twofold symmetry, the planes of the three histidine ligands of one zinc, relative to the threefold axis, are not exactly the same as the planes of the other three imidazoles coordinating the second zinc. Blundell *et al.* (1971b) report that the origin of the inexact symmetry is unclear, possibly being due to slightly different environments of the monomers in the crystal, or possibly arising in solution from nonequiv-

alence caused by sequential binding of the two zinc ions. Then again, the positioning of the monomer in exact twofold equivalence might result in impossibly close van der Waals approaches, and the observed nonequivalence might be the result of relieving these contacts, thus optimizing the energy of interaction.

F. Lysozyme

As is well known, the structure of hen egg-white lysozyme was determined to a resolution of 2.0 Å by Phillips and co-workers (Blake *et al.*, 1965, 1967a,b). In contrast to myoglobin, the oxygen storage protein, lysozyme had a less regular polypeptide conformation and exhibited a large cleft into which various inhibitors, including a trisaccharide, were observed to bind (Fig. 38). For the first time in a globular protein, anti-parallel pleated sheet and "hairpin bends" were observed. From model building considerations, it seemed likely that longer saccharides would be aligned in the cleft, with the bond to be

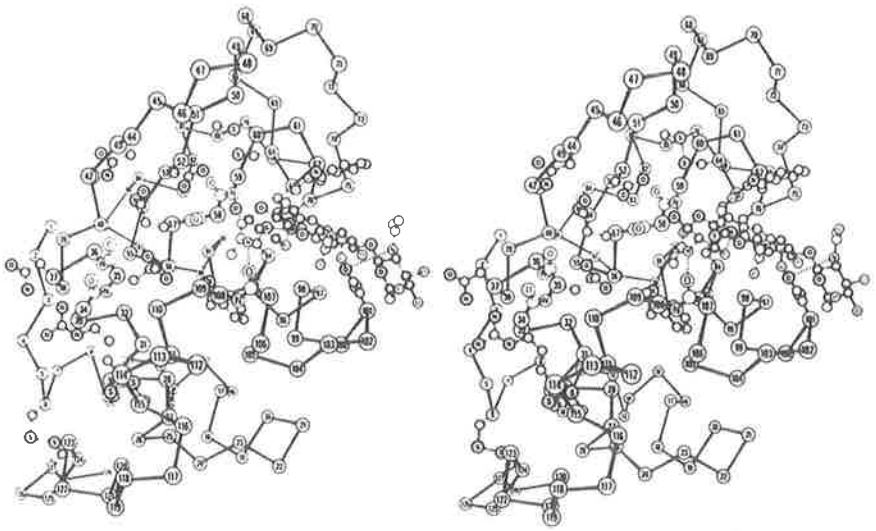


Fig. 38 Stereo drawing of the polypeptide chain of lysozyme, together with a trisaccharide inhibitor. The active site cleft runs from right to left, and longer saccharides are thought to extend across to the left-hand side of the molecule. The side chains of Glu 35, Asp 52, Trp 62, Trp 63, and Trp 108 are also shown. Reprinted by permission from R. E. Dickerson and I. Geis, Stereo Supplement to "The Structure and Action of Proteins." Harper & Row, New York, 1969. Copyright ©1969 by Richard E. Dickerson and I. Geis.

cleaved juxtaposed between Glu 35 and Asp 52, and that water would be excluded. It was postulated that on substrate binding, the p*K* values of the carboxylic groups would be perturbed by their respective environments in such a way that one group would act as a proton donor to the substrate and the other stabilize the resultant positive charge, thus facilitating cleavage of the glycosidic bond (Blake *et al.*, 1967b; North and Phillips, 1969). Subsequent chemical studies have tended to confirm that this is in fact the mechanism of catalysis.

The lysozyme structure has been illustrated by Xograph photographs of a space-filling model (Harte and Rupley, 1968), but this method of presentation cannot be recommended in preference to colored stereophotographs.

Recently, a 2.5 Å resolution electron density map of human lysozyme has been obtained (Blake and Swan, 1971; Banyard *et al.*, 1974) and, as anticipated from the sequence homology (59% of the residues are identical), the tertiary structures of the human and the hen egg-white enzymes are very similar, although the core of the human lysozyme molecule was altered to a much greater extent than expected simply from a consideration of the substitution of those residues present in the core of hen lysozyme.

A comparison has also been made between the structures of lysozyme and bovine α-lactalbumin, using models to match the known α-lactalbumin sequence (Brew *et al.*, 1967) to the three-dimensional hen egg-white lysozyme structure (Browne *et al.*, 1969). This comparison suggests that the structures of the two proteins are very similar. Elucidation of the differences between the two molecules must await the independent determination of the structure of α-lactalbumin, which has been hampered by technical difficulties (Aschaffenburg *et al.*, 1972a,b; Inman and Bryan, 1966).

The structure of the lysozyme from bacteriophage T4, which has similar specificity but different amino acid sequence to that of hen lysozyme, has been determined recently (Matthews and Remington, 1974). It is not clear whether or not the mechanisms of catalysis of the respective enzymes are related.

G. Nucleases

The structure of bovine pancreatic ribonuclease A was determined to a resolution of 2 Å by Kartha *et al.* (1967) and was followed shortly thereafter by the determination of the ribonuclease S structure at 3.5 Å (Wyckoff *et al.*, 1967b). Subsequently, the resolution of the latter study was increased to 2 Å (Wyckoff *et al.*, 1970). Also, Cotton and as-

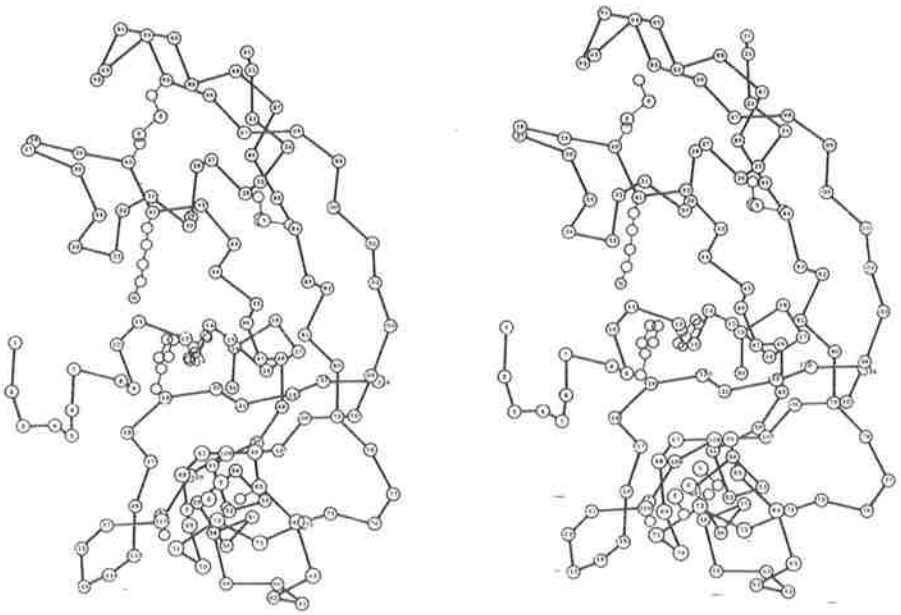


Fig. 39 Backbone conformation of ribonuclease S. The S-peptide (residues 1–20) is at the center (rear) and the active site, which includes His 12, His 119, and Lys 41, runs from the center left to right. Reprinted by permission from R. E. Dickerson and I. Geis, Stereo Supplement to "The Structure and Action of Proteins." Harper & Row, New York, 1969. Copyright ©1969 by Richard E. Dickerson and I. Geis.

sociates have obtained electron density maps of the extracellular nuclease from *Staphylococcus aureus* (Arnone *et al.*, 1971; Cotton *et al.*, 1971). In a study independent of the Buffalo group, Carlisle *et al.* (1972, 1974) have calculated a 2.5 Å resolution electron density map of ribonuclease A. Also, an analysis of a monoclinic crystal form of ribonuclease S is under way at Yale (Mitsui and Wyckoff, 1975). Comprehensive reviews of the extensive literature on both the pancreatic and the staphylococcal nucleases are available (Richards and Wyckoff, 1971; Cotton and Hazen, 1971; Anfinsen *et al.*, 1971).

The structure of ribonuclease S is illustrated in Fig. 39. According to Perutz (1969b), "The ribonuclease molecule looks somewhat like a bird, a cormorant or a vulture, with its head drawn in between its wings, carrying the active site, not in its head, but in its heart."

With the possible exception of residues 16–23, containing the cleaved bond in ribonuclease S, the structures of ribonuclease A and ribonuclease S are very similar.

Lee and Richards (1971) have developed a method of quantitating

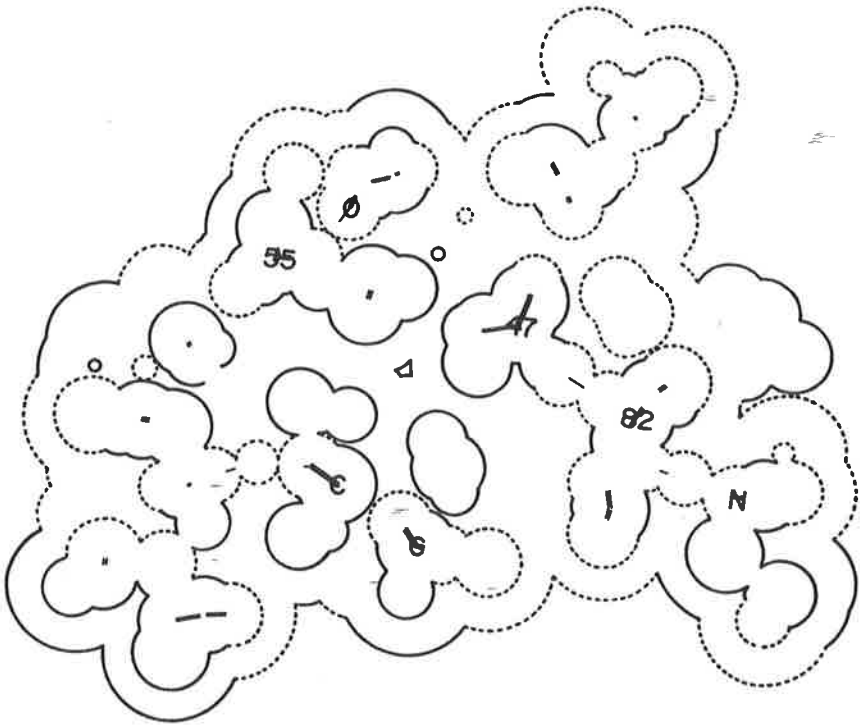


Fig. 40 Superposition of van der Waals contours and accessibility contours for a section of the ribonuclease S molecule. The outer contour defines the volume accessible to solvent. An internal cavity capable of accommodating a solvent molecule with a radius of 1.4 Å is indicated by an arrow. Taken from Lee and Richards (1971).

the "static accessibility" of each atom in a protein to water and have applied the technique to ribonuclease S (Fig. 40). The larger hydrophobic side chains and cysteine was found to be the most buried, whereas all other types of side chains had rather wide ranges of solvent accessibility (Fig. 41). It was noted that the residues in the substrate recognition site often were at the extremes of solvent accessibility for that class of side chain (e.g., Val 43 was the most exposed valine, and Lys 41 the most buried lysine). Plots of accessibility, twist, and bend of the ribonuclease S structure as a function of amino acid sequence number (Richards and Wyckoff, 1971) did not show any obvious pattern. Several proposals for the mechanism of ribonuclease activity have been critically reviewed in the light of the three-dimensional structure by Richards and Wyckoff (1971).

Cotton and co-workers have obtained 4 Å resolution electron den-

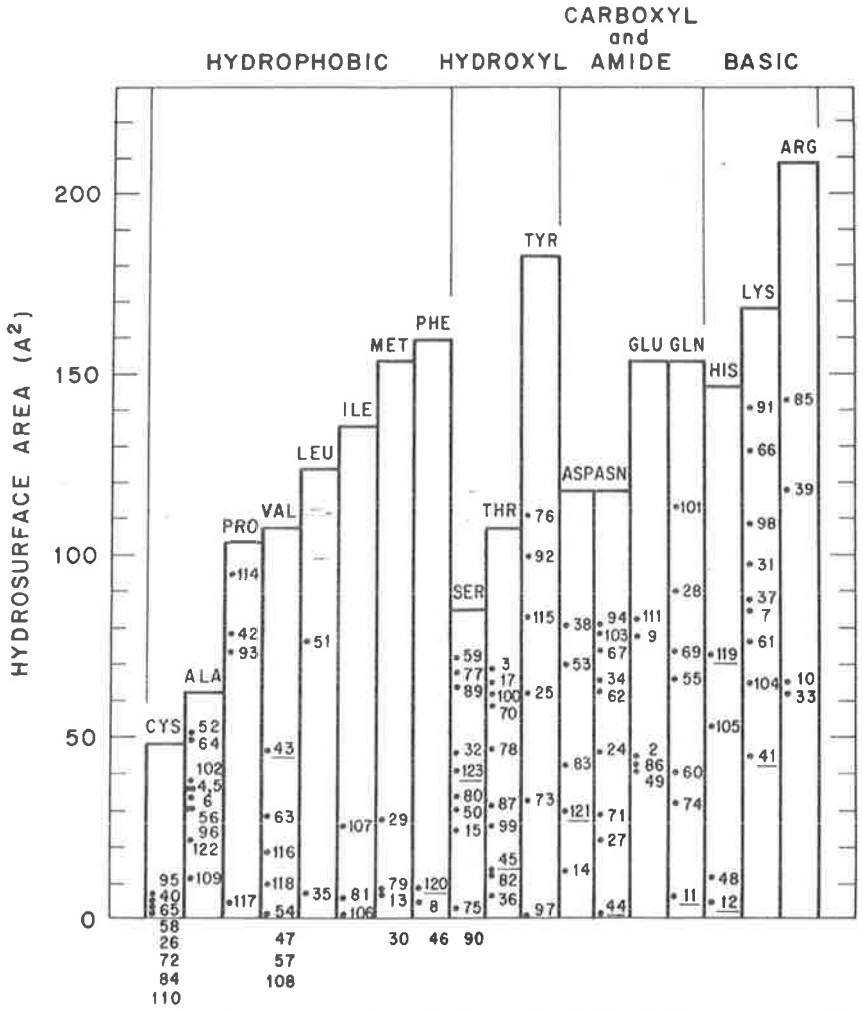


Fig. 41 Area accessible to solvent for each of the side chains of ribonuclease S. The upper bar for each amino acid is the maximum solvent accessibility of that residue in the hypothetical sequence Gly-X-Gly, assumed to be an extended conformation. Residues with no detectable exposure to solvent are listed below the axis. Some of the residues in the active site region are indicated by underlining. Taken from Richards and Wyckoff (1971).

sity maps for crystals of uninhibited staphylococcal nuclease and 2 Å resolution maps for the inhibited ternary complex of enzyme, thymidine 3′,5′-diphosphate, and Ca^{2+} (Arnone *et al.*, 1969, 1971; Cotton *et al.*, 1971). From a comparison of the two maps, the nuclease appears to undergo no gross conformational change on binding the

inhibitor. The calcium ion is complexed by the carboxylate side chains Asp 21, Asp 40, and Glu 43, with Asp 19 somewhat further away, forming a roughly square array. The metal ion is close to the 5′-phosphate but apparently too far for even a weak electrostatic bonding. Some correlations of the three-dimensional structure with studies in solution have been discussed by Cotton and Hazen (1971) and Anfinsen *et al.* (1971), and a tentative model for the sequential unfolding of the nuclease has been proposed by Jardetzky *et al.* (1971). There is no obvious relation between the three-dimensional structures of pancreatic ribonuclease (Fig. 39) and the staphylococcal enzyme (Fig. 42).

H. Dehydrogenases

1. Lactate Dehydrogenase

Lactate dehydrogenase (LDH) is an NAD^+-dependent tetrameric enzyme (MW 140,000) that catalyzes the interconversion of lactate and pyruvate. Rossmann and co-workers have chosen the enzyme from dogfish for detailed X-ray analysis and have obtained electron density maps at near-atomic resolution (Adams *et al.*, 1970a).

Crystals of the native enzyme from dogfish crystallize with one subunit per asymmetric unit, demonstrating that the tetramer must have point symmetry 222 (Rossmann *et al.*, 1967). However, when crystals of the apoenzyme are soaked in 1 m*M* solutions of NAD^+, or in 0.05 m*M* NADH, there is a change in space group, reversible on removal of the dinucleotide, so that the molecule has only a twofold axis. Whether the molecule has in fact lost its 222 symmetry or has rotated slightly about the remaining twofold axis cannot be determined from the space group symmetry alone (Adams *et al.*, 1969b). A probable reduction in symmetry in the presence of coenzyme has also been observed for the dimeric enzyme horse liver alcohol dehydrogenase (Zeppezauer *et al.*, 1967). The change in quaternary structure induced on binding coenzyme to LDH appears to be a bodily translation of each subunit by about 5 Å and presumably is not due to steric interference between the coenzyme molecules, which bind at least 19 Å apart (Adams *et al.*, 1969b, 1970b).

Although the binary complex of enzyme–coenzyme appears to have lower symmetry than the apoenzyme, the abortive ternary complex of lactate dehydrogenase, NAD, and pyruvate has full 222 point symmetry (Leberman *et al.*, 1969).

The approximate conformation of dogfish LDH, first determined

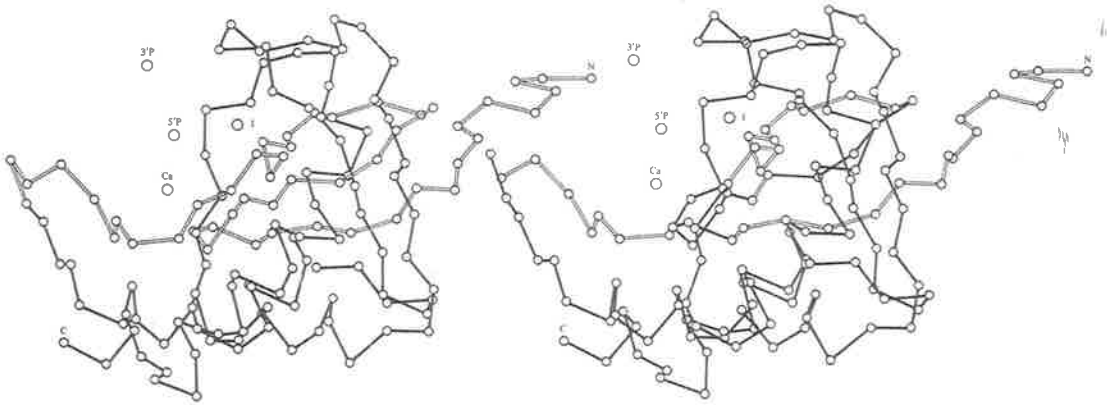


Fig. 42 Backbone conformation of staphylococcal nuclease. The positions of the two phosphates (3′P and 5′P) and the 5-methyl (I) of the inhibitor are indicated as well as that of Ca^{2+}. Taken from Cotton *et al.* (1971).

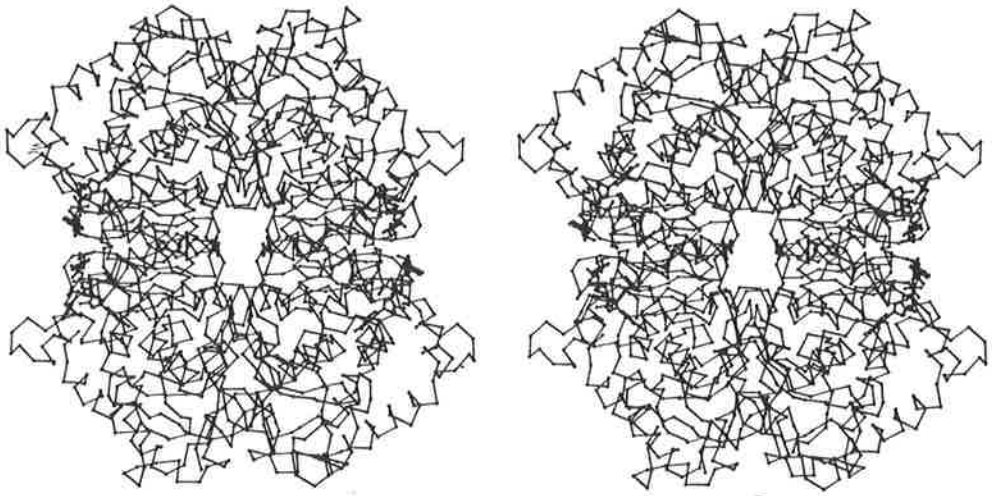


Fig. 43 Backbone conformation of the four subunits of lactate dehydrogenase. Taken from Adams *et al.* (1972).

from a 2.8 Å electron density map (Adams *et al.*, 1970a), was considerably improved from subsequent 2.5 and 2.0 Å resolution maps (Rossmann *et al.*, 1971; Adams *et al.*, 1973b).

The polypeptide chain within each subunit is folded largely into regular secondary structures (45% helix and 20%–25% β-structure) with shorter intervening conformations typical of a number of proteins. In general, the polypeptide chain is found to be folded compactly, but with the notable exception that the twenty-two amino-terminal residues form a long extended "arm" stretching far from the remainder of the subunit. In the tetramer (Fig. 43), this arm is observed to interact extensively with another subunit, presumably contributing to the stability of the oligomer.

The active center of LDH lies in a large cleft into which the coenzyme was observed to bind, apparently in an "open" conformation (Adams *et al.*, 1970b). Within the limitations of the approximate methods necessitated by the change in space group on binding coenzyme, it appears that the coenzyme binding is not accompanied by the substantial change in quaternary structure described above but by rather subtle changes in conformation within each subunit.

The crystal structure of the abortive ternary complex has been determined, at least approximately, by a 3.0 Å resolution map phased in part on a single isomorphous heavy atom derivative (Rossmann *et al.*, 1971). On comparing the maps of the apoenzyme and the abortive ter-

nary complex, it appears that in most respects the two structures are very similar, except for a loop containing residues 100–120 which moved up to 12 Å. In the vicinity of the active site, some atoms appear to move by 2–3 Å (Adams *et al.*, 1972). Not only are the tertiary conformations of the individual subunits very similar, but the quaternary arrangement of the subunits are almost identical, despite the fact that the crystal forms of the apoenzyme and the ternary complex are quite different.

Adams *et al.* (1973a) have discussed the constraints on possible mechanisms for lactate dehydrogenase imposed by the known three-dimensional structure.

2. *Malate Dehydrogenase*

The structure of soluble cytoplasmic malate dehydrogenase (s-MDH) from pig heart has been studied by Banaszak and co-workers and was first "solved" from an electron-density map at 3.0 Å resolution (Hill *et al.*, 1972; Webb *et al.*, 1973).

The enzyme is a dimer (MW 72,000) that catalyzes the NADH-NAD^+ dependent conversion of oxaloacetate and L-malate. The crystals chosen for detailed X-ray analysis were grown in the presence of a tenfold molar excess of NAD^+ and contained one dimer per asymmetric unit (Banaszak, 1966; Banaszak *et al.*, 1971). Therefore, the successful structure determination necessitated, in effect, the study of a 72,000 MW protein, but provided as a bonus the opportunity of determining whether or not the subunits were identical, particularly with respect to binding of NAD^+, which has been shown to have a stoichiometry of one NAD^+ per dimer (Harada and Wolfe, 1968).

The electron density map of s-MDH (Hill *et al.*, 1972) showed the conformations of the two independent subunits to be very similar. The estimated coordinates of 321 α-carbon atoms in each subunit were compared by rotation about the local twofold axis relating the monomers, and no significant differences were found. The average differences between the distances of pairs of related atoms from the diad axis was 1.3 Å.

The interesting finding obtained from the 3.0 Å resolution electron density map was that the backbone conformation of both of the subunits of malate dehydrogenase was strikingly similar to that of lactate dehydrogenase (see above), except that in s-MDH there was no counterpart to the twenty-two amino-terminal residues of LDH. The fact that the malate enzyme aggregates as a dimer, whereas the lactate enzyme aggregates as a tetramer, can presumably be attributed at least in part to this difference, for, as mentioned above, the extended amino-

terminal "arm" in LDH is thought to stabilize its association as a tetramer.

In spite of the convincing crystallographic and chemical evidence for equivalence or near equivalence of the two subunits of malate dehydrogenase, the crystals used for the structure determination were shown by chemical analysis to contain only one mole of NAD^+ per dimer (Glatthaar *et al.*, 1972). On soaking the pregrown crystals in solutions of 20 m*M* NAD^+ and 20 m*M* L-malate or oxaloacetate, 5.0 Å resolution difference maps showed a large continuous peak of density in one subunit that could be interpreted as an NAD^+ molecule in the open configuration (Tsernoglou *et al.*, 1972). At the twofold-related position in the other subunit, density was already present in the starting crystals, presumably corresponding to the single NAD^+ molecule bound to the crystalline dimer. The closest approach to the two NAD^+ binding sites was observed to be not less than about 20 Å, again ruling out the possibility, as in LDH, that direct steric interference might occur between the two bound coenzyme molecules.

In recent experiments it has been shown that soaking crystals of s-MDH in very high concentrations of NAD^+ results in density associated with only one of the two subunits. In all these soaking experiments, prolonged exposure of the crystals to NAD^+-free solutions will wash out the coenzyme from the "secondary" binding site but not from the "primary" site. It seems likely from 2.5 Å resolution native and difference electron density maps (L. J. Banaszak, personal communication; Webb *et al.*, 1973) that rather than one "apo" and one "holo" subunit, there may be a "higher occupancy" and a "lower occupancy" coenzyme-bound subunit. The most recent results also confirm the earlier conclusion that despite the obvious difference between the avidity of the two subunits for coenzyme, the conformational differences between them, if any, are subtle.

3. *Other Nucleotide-Binding Proteins*

During the last year or so there has been a veritable explosion in the number of nucleotide-binding proteins "solved" and it has become clear that a similar nucleotide-binding structural domain occurs in many of these enzymes. The implication of this finding for the evolution of these and other protein structures has generated considerable interest (e.g., see Rossmann, 1974; Rossmann *et al.*, 1974).

Other dehydrogenases which exhibit the "nucleotide binding fold," first seen in lactate and malate dehydrogenase, include liver alcohol dehydrogenase (Brändén *et al.*, 1973; Eklund *et al.*, 1974; Öhlsson *et al.*, 1974) and lobster D-glyceraldehyde-3-phosphate dehydrogenase

(Buehner *et al.*, 1973, 1974a,b). Similar, but not identical, domains form the presumed adenine binding sites of porcine muscle adenyl kinase (Schulz *et al.*, 1974a; Schulz and Schirmer, 1974), the flavin mononucleotide binding site of flavodoxin (Rossmann *et al.*, 1974), and possibly the ADP/ATP binding site of phosphoglycerate kinase (Blake and Evans, 1974; Bryant *et al.*, 1974). The widespread occurrence of this common structural element, which is based on a sheet of parallel β-structure with the binding site at the carboxy termini of the polypeptide strands, tends to suggest that proteins containing this domain have a common precursor (Rossmann *et al.*, 1974). On the other hand, it is surprising to find that a similar conformation also occurs in the vicinity of the substrate binding site of subtilisin. Furthermore, Steitz and co-workers (Fletterick *et al.*, 1975) have argued that the structure of the AMP binding site in yeast hexokinase is only superficially similar to that of lactate dehydrogenase and that it is unlikely that these nucleotide binding domains have evolved from a common ancestral gene. In addition, Campbell *et al.* (1974) have recently reported that yeast phosphoglycerate mutase, which has no nucleotide requirement, contains a domain which resembles that found in the dehydrogenases. At the present time it is not easy to rationalize all these results. While it is clear that a number of the structures listed above have a common evolutionary precursor, it is not clear that they all do (see also Section V,B).

I. Calcium-Binding Carp Myogen

Kretsinger and colleagues have determined and refined the structure of a calcium-binding protein prepared from carp muscle (Kretsinger *et al.*, 1971b; Nockolds *et al.*, 1972; Kretsinger and Nockolds, 1973; Moews and Kretsinger, 1975). The function of the protein is unknown, although it has structural homologies with myosin light chains and with troponin C (Weeds and McLachlan, 1974; Tufty and Kretsinger, 1975). Carp myogen is one of a series of closely related fish muscle proteins (MW about 12,000) with an unusual amino acid composition (10% phenylalanine, 20% alanine, one residue each of histidine, arginine, and cysteine, and no tryptophan, tyrosine, proline, or methionine). Two calcium ions are bound to the molecule. The structure was essentially determined by the single isomorphous replacement method, but the correct interpretation of the second calcium binding site was made after phase refinement by the tangent formula (Hendrickson and Karle, 1973; Kretsinger and Nockolds, 1973; see Section II,F.*1*).

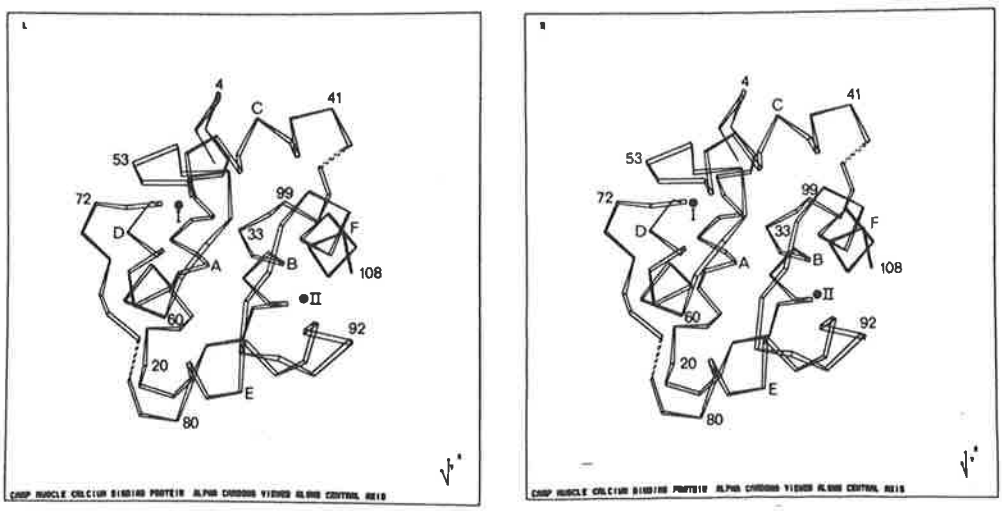


Fig. 44 Two bound calcium ions and part of the polypeptide backbone of carp myogen. Note the striking pseudo-symmetry about the local twofold axis which is normal to the figure. Taken from McLachlan (1972b).

The molecule is generally spherical, with a well-defined hydrophobic core one-seventh of its total volume. There are six helical regions designated A through F. Possibly the most interesting feature of the molecule is the structural similarity of the two calcium-binding sites. One calcium ion binds in a loop between the C and D helices, the other between the E and F helices. The structures of the C–D and E–F regions are not only strikingly similar but are related by an approximate twofold symmetry axis (Fig. 44). There is a further structural similarity with the A–B region of the molecule which does not bind calcium but has, as a possible analogue, an internal salt bridge. In spite of the obvious structural similarity between these three regions of the polypeptide chain, careful comparisons do not reveal any significant homology between the respective amino acid sequences. Kretsinger (1972) and McLachlan (1972b) propose that the observed near identity of tertiary structure results from gene duplication and subsequent divergent evolution of the amino acid sequence. The ability to bind calcium presumably has been retained in only one of the two postulated gene duplications (see Section V,B).

J. Concanavalin A

Concanavalin A (Con A) is a phytohemagglutanin that agglutinates a variety of cell types and binds mono-, oligo-, and polysaccharides. The

protein consists of subunits with a molecular weight of about 25,000, each containing one Mn^{2+} ion, one Ca^{2+} ion, and a sugar-binding site. Below pH 6, dimers are formed, while above pH 7 the subunits aggregate as tetramers.

Recently, the Rockefeller and Argonne laboratories have independently obtained 2.0 Å and 2.4 Å resolution electron density maps of Con A from the jack bean *Canavalia ensiformis* (Edelman *et al.*, 1972; Hardman and Ainsworth, 1972). The same crystal form, grown respectively at pH 6.8 and at about pH 6, was used in each case. In the crystals, the asymmetric unit contained one monomer, implying that the subunits were arranged with exact 222 point symmetry (Greer *et al.*, 1970; Hardman *et al.*, 1971; Reeke *et al.*, 1971). Since the crystals were grown below pH 7, it is not certain that the subunit aggregation seen in the crystal is the same as that exhibited by the tetramer in solution at higher pH values.

From the 2.0 Å electron density map and from the amino acid sequence, Edelman *et al.* (1972) were able to trace the course of the polypeptide chain and describe the subunit assembly. The results of Hardman and Ainsworth (1972), obtained from a 2.4 Å resolution map without reference to the complete amino acid sequence, are essentially in good agreement.

In the crystal, the four Con A subunits cluster about three mutally perpendicular twofold axes in a pseudo-tetrahedral arrangement. Each subunit has about 57% of the polypeptide chain segregated into three distinct sheets of β-structure with essentially no helices whatsoever (Fig. 45).

One of the sheets contains six polypeptide strands within one subunit and is related by twofold symmetry to a similar sheet in another subunit in such a way that in the tetramer a continuous sheet of twelve strands is formed. Presumably the intersubunit hydrogen bonds help stabilize the tetramer (or at least the dimer). It is interesting to note that two symmetry-related histidines occur close to the twofold axis relating the monomers; it may be that titration of these imidazole groups is intimately related to the association and dissociation of the tetramer (Hardman and Ainsworth, 1972; Edelman *et al.*, 1972; K. D. Hardman, private communication).

Jack *et al.* (1971) crystallized demetallized Con A at pH 5.0 and found that the unit cell dimensions were similar to those of native Con A, but the space group was $P2_122_1$ rather than I222, so that the asymmetric unit contained two subunits rather than one, as in the native enzyme. If apo-Con A was considered as a dimer under the conditions of crystallization, there was no symmetry requirement for the arrangement of the monomers in the dimer. However, if the molecule was

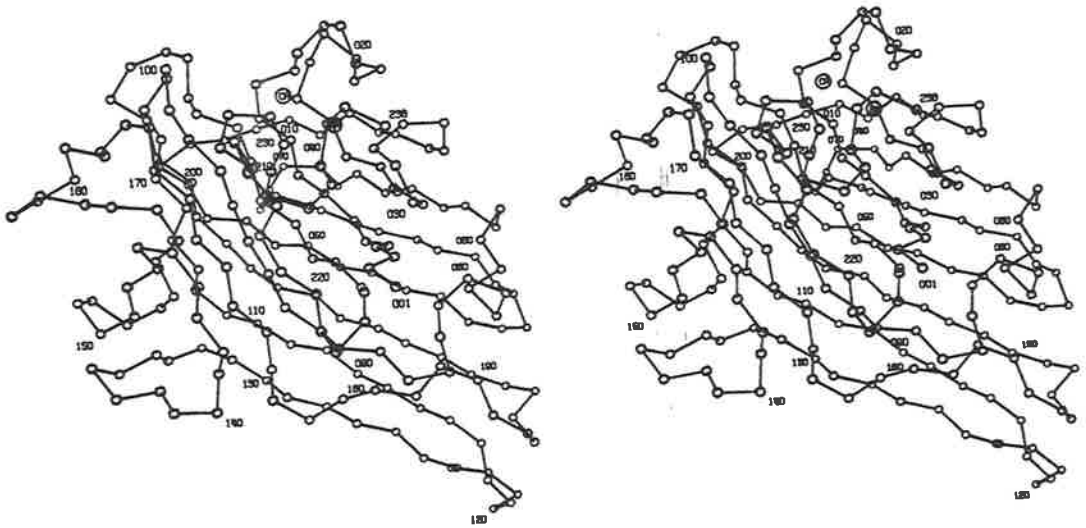


Fig. 45 Backbone conformation of one subunit of concanavalin A. Note the extensive β-structure. Taken from Edelman *et al.* (1972).

considered as a tetramer, it would need have only twofold symmetry rather than 222 point symmetry, as in the native enzyme. When the demetallized crystals were soaked in dilute solutions of transition metal ions and calcium ions, they yielded X-ray patterns nearly identical to those of the native crystals, suggesting that in the crystals, demetallized Con A aggregated in a manner very similar to native Con A. Jack *et al.* (1971) concluded from rotation and translation function calculations that in the demetallized crystals, the subunits had rotated about 7° and had translated about 6 Å relative to each other along the rotated axis. Hardman and Ainsworth (1972) report that in the absence of Mn^{2+} and Ca^{2+}, the polypeptide chain near the metal-binding site would be forced out into the solvent. They concluded that it is not surprising that the apoprotein cannot be crystallized in the same space group.

In the high resolution electron density map of the Rockefeller group (Becker *et al.*, 1975; Reeke *et al.*, 1975), the calcium and manganese ions are observed to be within 4.6 Å of each other. Both the Ca^{2+} and Mn^{2+} appear to be roughly octahedrally coordinated, Mn^{2+} apparently interacting with the side chains of Glu 8, Asp 10, Asp 19, His 24, and two water molecules, and Ca^{2+} with the carbonyl oxygen of Tyr 12, the side chains of Asn 14, Asp 10, and Asp 19, and two water molecules. Hardman and Ainsworth's (1972) interpretation of the metal-binding sites, made without the benefit of the complete amino acid sequence, is essentially similar, except that Ca^{2+} was reported to be pentacoordinated, and the identification of two of the calcium ligands was different. The approximate octahedral coordination and the sharing of acid groups are reminiscent of the double calcium-binding site in thermolysin (Matthews *et al.*, 1972a,b).

The binding site for β-(*o*-iodophenyl)-D-glucopyranoside, a saccharide that inhibits the agglutination reaction, was found by difference Fourier techniques (Becker *et al.*, 1971) to be in a relatively narrow but deep pocket at least 20 Å from the metal-binding region, although Hardman and Ainsworth (1973) showed that the binding at this site may well be due to the iodophenyl moiety. On the other hand, noncrystallographic experiments (Brewer *et al.*, 1973; Kalb and Levitzki, 1968) had suggested that the carbohydrate binding site was close to the metal-binding region.

This apparent conflict has recently been resolved by Becker *et al.* (1976) and Hardman and Ainsworth (1976) who have independently shown that the biologically relevant carbohydrate binding site is in fact about 13 Å from the metal ions, and that in the crystal form of concanavalin A used for the original structure determination, access to this site was prevented by a neighboring molecule in the crystal.

K. Other Proteins

The above survey includes most of the protein structures which are at the present time determined well enough to allow construction of a Kendrew–Watson skeletal model based upon an electron density map. Nevertheless, "new" structures are appearing so rapidly (Table I) that it is virtually impossible to provide up-to-date coverage within a single chapter. For example, no mention has been made of the rapid progress that has been made recently in studying immunoglobulin structure. Within the last year or so, the structures of no less than four immunoglobulin fragments have been reported to near atomic resolution (Poljak *et al.*, 1973, 1974; Schiffer *et al.*, 1973; Segal *et al.*, 1974; Epp *et al.*, 1974). Although the structure determination of an intact immunoglobulin has not yet been completed, the results which have been obtained have shed a great deal of light on the three-dimensional structure of antibodies and on the nature of the interaction between antibody and antigen (Davies *et al.*, 1975).

No attempt will be made here to discuss the numerous other crystallographic studies which have either been reported or are in progress. References to many of these can be obtained from Tables I, II, and IV, which were prepared in part with the hope that they might help minimize unnecessary duplication of effort.

V. CONCLUSIONS

A. Protein Conformation

Now that the structures of thirty or so distinctly different proteins are known, it is of interest to reconsider the generalizations made by Kendrew (1962) about the conformation of myoglobin.

> (i) The molecule is compact. There is no water inside the molecule, with the exception of a very small number (<5) of single water molecules presumably trapped at the time the molecule is folded up.
>
> (ii) Almost all the polar groups are on the surface. . . . The rare exceptions are polar side chains performing a special function within the molecule, e.g., the heme-linked histidine.
>
> (iii) The interior of the molecule is made up of nonpolar residues. . . . Not all the nonpolar residues are inside; some, especially the short ones (glycine, alanine), are at the surface.
>
> (iv) Bound water molecules are attached to all polar groups at the surface, including main chain CO and NH groups. Apart from these bound water molecules there is no obvious sign of order in the liquid regions.

Kendrew's first and fourth statements apply without modification to all subsequent protein structures. Most proteins with molecular weights greater than about 16,000 tend to form "lobes" which may represent convenient units into which the polypeptide can fold, with the side chains forming a hydrophobic core and the backbone wrapped around the outside (Matthews *et al.*, 1972a). Substructures of this general type have been observed for many proteins, including α-chymotrypsin and homologues, subtilisin, papain, lactate dehydrogenase, and thermolysin. It is interesting that carboxypeptidase, carbonic anhydrase, and concanavalin A, the three larger known proteins that do not have "lobes," have, in contrast, extensive β-structure that provides a very efficient method by which the polypeptide backbone can remain internal, yet have its hydrogen-bonding capacity satisfied. The occurrence of "lobes" notwithstanding, almost all the known protein structures can be described as compact, although in the case of phosphoglycerate kinase (Bryant *et al.*, 1974; Blake and Evans, 1974) two structural lobes are joined by a rather narrow neck.

The protein that departs most from spherical is the thermostable enzyme thermolysin, which has an axial ratio of about 2:1 (Matthews *et al.*, 1972a). It is interesting to note that in this case, heat stability of a protein is not achieved by a near-spherical shape and a well-developed hydrophobic interior, as might be naively anticipated.

Isolated water or solvent molecules have been observed trapped inside several proteins other than myoglobin. For example, lysozyme has three (D. C. Phillips, reported in Lee and Richards, 1971), carboxypeptidase A has ten (Hartsuck and Lipscomb, 1971), subtilisin BPN′ has seventeen located either in the neighborhood of the catalytic site or in the interior of the protein molecule (Alden *et al.*, 1971), and α-chymotrypsin has thirteen (Birktoft and Blow, 1972).

In several cases bound water or solvent molecules have been reported to occur at the protein surface, but in no case has an "iceberg" structure been observed to surround hydrophobic side chains exposed to the solvent.

Kendrew's statement that for myoglobin, polar groups were almost exclusively outside and nonpolar groups inside, as predicted by Kauzmann (1959), still contains a great deal of truth, but it must be modified somewhat in the light of more recent structure determinations. A detailed tabulation of the solvent exposure of each residue of myoglobin, lysozyme, and ribonuclease S has been given by Lee and Richards (1971). In the present context, this may be supplemented by the cruder classifications of "outside," "inside," and "surface."

In Table V we have summarized the observed side chain environ-

TABLE V

Side Chain Environment in α-Chymotrypsin, Carboxypeptidase A, and Thermolysin

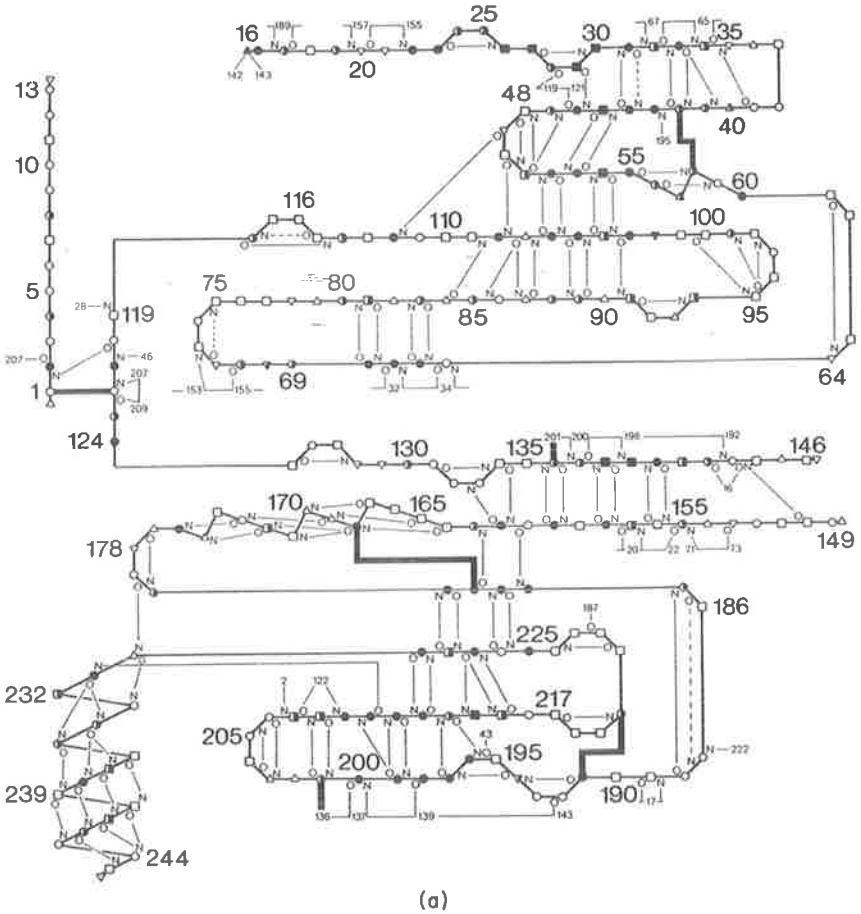
Side chain	Inside	Surface	Outside	Total
Gly	19	26	37	82
Ala	33	8	30	71
Val	29	19	13	61
Leu	26	14	18	58
Ile	28	13	8	49
Ser	17	16	52	85
Thr	16	24	33	73
Asp	5	16	26	47
Glu	3	19	26	48
Asn	3	19	26	48
Gln	2	16	17	35
Lys	0	11	29	40
Arg	0	9	15	24
His	1	9	8	18
Phe	17	11	4	32
Tyr	3	31	17	51
Trp	4	11	3	18
Pro	6	8	13	27
Met	4	1	2	7
Cys	3	5	4	12
Total	219	273	372	864

ments in α-chymotrypsin (Birktoft and Blow, 1972), carboxypeptidase A (Hartsuck and Lipscomb, 1971), and thermolysin (Colman *et al.*, 1972a). The summary is not intended to be exhaustive; it was made in order to obtain the largest possible sample from the readily available data.

It is clear that although the larger hydrophobic side chains such as valine, leucine, isoleucine, and phenylalanine tend to be buried, they are also often fully exposed to the solvent. Tyrosine and tryptophan, which are largely hydrophobic in character, are only occasionally buried and are more commonly found at least partly exposed to solvent. Glycine and alanine are about equally distributed between inside and outside. This survey confirms the finding of Lee and Richards (1971) that proline is the outstanding exception to the tendency of nonpolar residues to be buried.

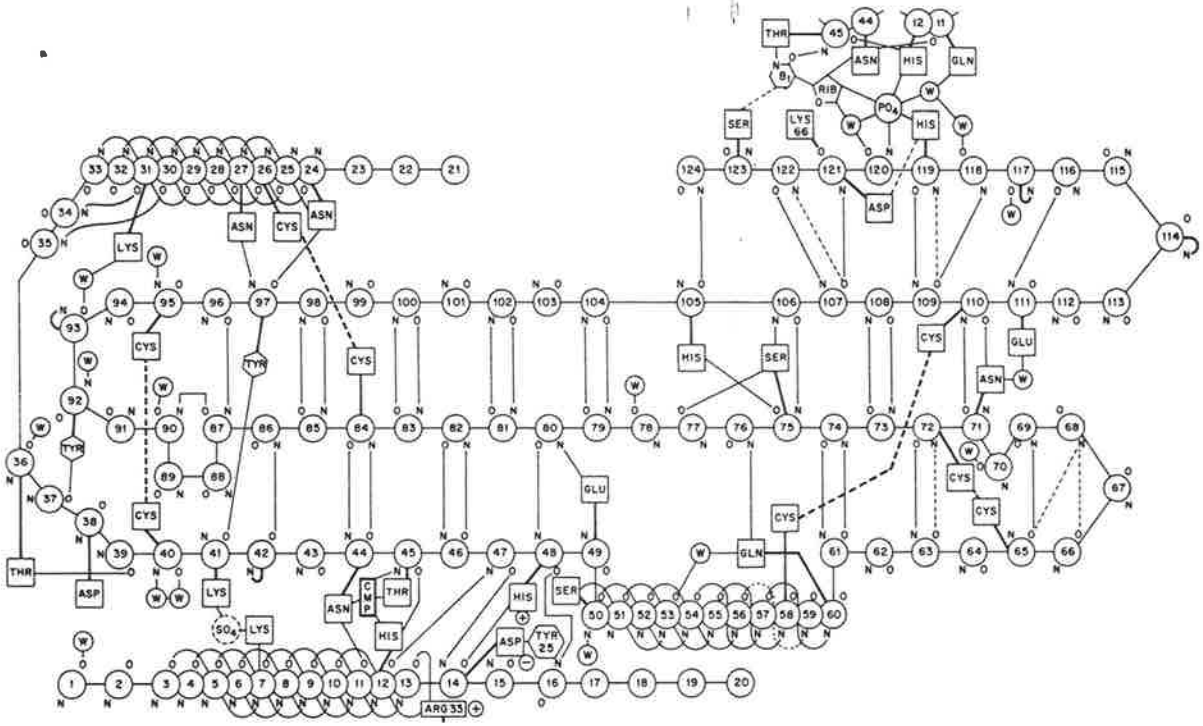
The tendency of polar residues to be outside is more pronounced than the tendency of hydrophobic residues to be inside, but there are frequent exceptions, especially in the case of serine and threonine,

where about one residue in five is buried. In fact for thermolysin, fourteen of the serine and threonine side chains are inaccessible to solvent, whereas only nine are fully exposed. It is rare to find asparagine or glutamine in an interior position and, with the possible exception of carp myogen (Kretsinger and Nockolds, 1973), a completely buried lysine or arginine has never been observed. On the other hand,



(a)

Fig. 46 (a) Schematic representation of main chain hydrogen-bonding for α-chymotrypsin. The thin lines representing hydrogen bonds are dashed when the bond is uncertain. Each hairpin loop is indicated by a kink in the peptide chain. The broad bars represent disulfide bridges. Symbols for side chains and chain termini are as follows: (▽), acidic; (△), basic, including histidine; (□), other polar side chains; (○), nonpolar. Also, the disposition of each side chain is indicated: filled, internal; half-filled, surface; open, external. Taken from Birktoft and Blow (1972).



(b)

Fig. 46 (b) Backbone hydrogen-bonding for ribonuclease S. Disulfide bridges are indicated in heavy dashed lines, and poor or uncertain hydrogen bonds in light dashed lines. W indicates a solvent molecule. Taken from Richards and Wyckoff (1971).

buried acid groups have now been observed in several proteins, although in each case they were neutralized by a nearby positively charged group. It can no longer be assumed that such buried charges are invariably associated with a special biologic function, such as activation in α-chymotrypsin, or the binding of metals. Carboxypeptidase A, for example, has three buried acid residues, Asp 104, Glu 108, and Glu 292, none of which is near the active site or has an apparent catalytic function (Hartsuck and Lipscomb, 1971).

Several methods have been used to summarize protein conformations. Hydrogen bonds can often be represented by two-dimensional diagrams such as those shown for α-chymotrypsin and ribonuclease S in Fig. 46a,b. With larger proteins it becomes increasingly difficult to include all the observed hydrogen bonds in a single figure of this type. The presence of the hydrogen-bonded cylinders (see Section IV,C,*1*) in α-chymotrypsin necessitates duplication in the figure of some portions of the polypeptide chain. Nevertheless, figures of this type provide a convenient method of summarizing a great deal of information. They are particularly useful in studying the topology and perfection of regions of β-structure and in looking for regions of possible structural similarity.

Another method of illustrating the overall conformation of a protein has been proposed by Phillips (1970) and is shown in Fig. 47 with data for lactate dehydrogenase (Adams *et al.*, 1972). In the figure, amino acids which are separated in the linear sequence, but have α-carbons less than 6 Å apart and therefore potentially interact, are indicated by off-diagonal points. Regions of antiparallel β-structure appear as clusters of points running at right angles to the diagonal; parallel β-structure is indicated by clusters of points parallel to the diagonal; and helices are indicated by a clustering of points close to the diagonal. It is immediately apparent that for lactate dehydrogenase most of the potential interactions are within the respective halves of the molecule, consistent with the observation that the polypeptide chain folds into two lobes.

The distance of all the α-carbon atoms from a chosen point in the structure may be conveniently presented by using a polar representation (Richards and Wyckoff, 1971). Examples for ribonuclease S and the lactate dehydrogenase subunit are illustrated in Fig. 48a,b. The diagram for ribonuclease S shows not only the distance to the α-carbon atoms from the phosphate position P, but also the distances to the tips of each side chain. Both structures are reasonably compact, except for residues 1–22 in lactate dehydrogenase, which extend to interact with another subunit of the tetramer (see Section IV,H,*1*).

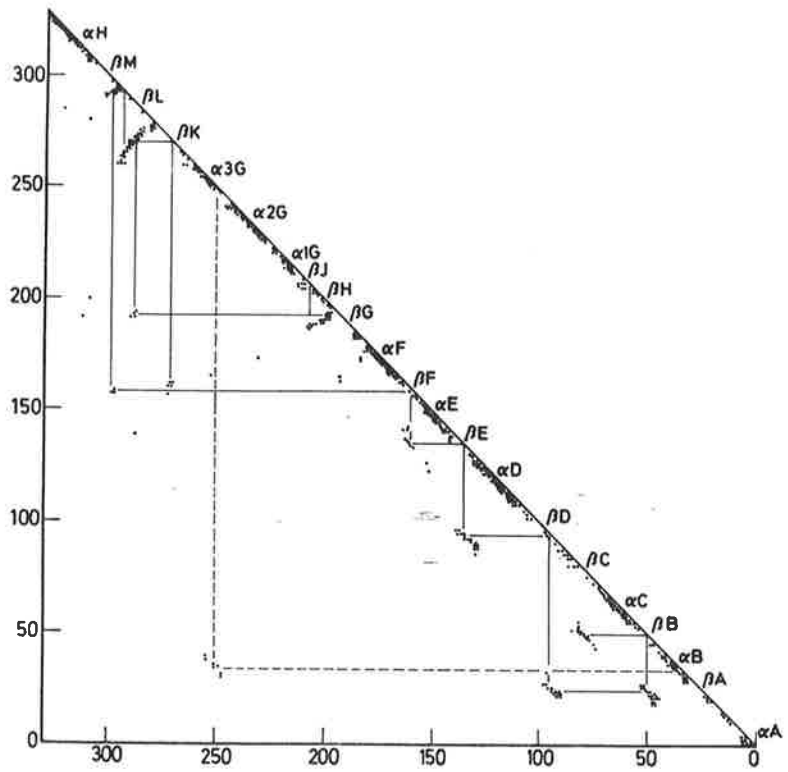
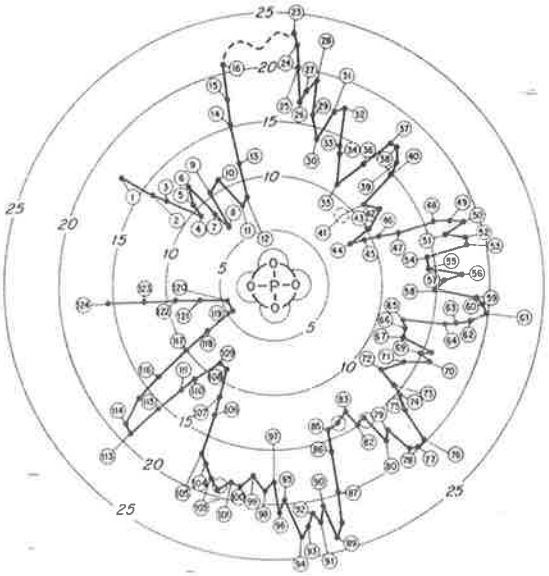


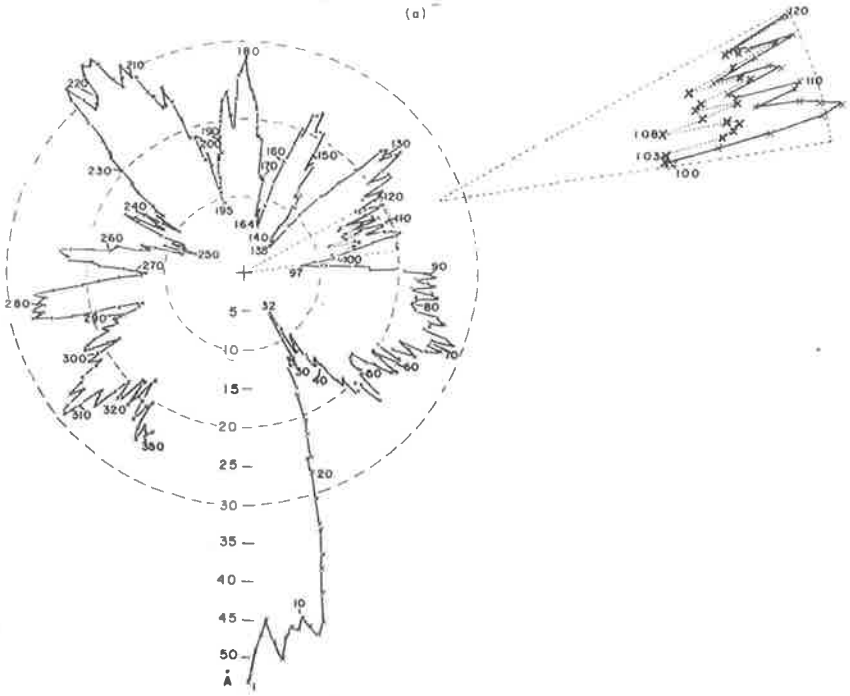
Fig. 47 Amino acids within one subunit of lactate dehydrogenase separated by less than 6.0 Å. Taken from Adams *et al.* (1972).

The conformations of the polypeptide backbone are usually given in terms of the dihedral angles (ϕ,ψ) introduced by Ramachandran *et al* (1963). Plots of (ϕ,ψ) are particularly useful for displaying doubtful or unusual conformations and for summarizing the conformations of a given amino acid; however, they contain no information about the neighboring residues in the polypeptide chain. Backbone conformations observed for seven proteins are summarized in Fig. 49. [The conventions given by the IUPAC-IUB Commission (1970) are used throughout this chapter.] As is expected, most of the observed (ϕ,ψ) values lie within the "allowed" regions, but there are a number of exceptions that even allow an uncertainty of 10°–20° in the conformational angles.

The structures of myoglobin and lysozyme showed that (ϕ,ψ) diagrams based simply upon an idealized hard-sphere peptide model


25
20
15
10
5
O
O—P—O
O

(a)


180
210
240
270
300
330
350
320
310
290
280
260
250
230
220
190
200
195
164
140
138
160
150
170
130
120
110
100
97
90
80
70
60
50
40
30
32
5
10
15
20
25
30
35
40
45
50
Å
10
20
120
110
106X
103X
100

(b)

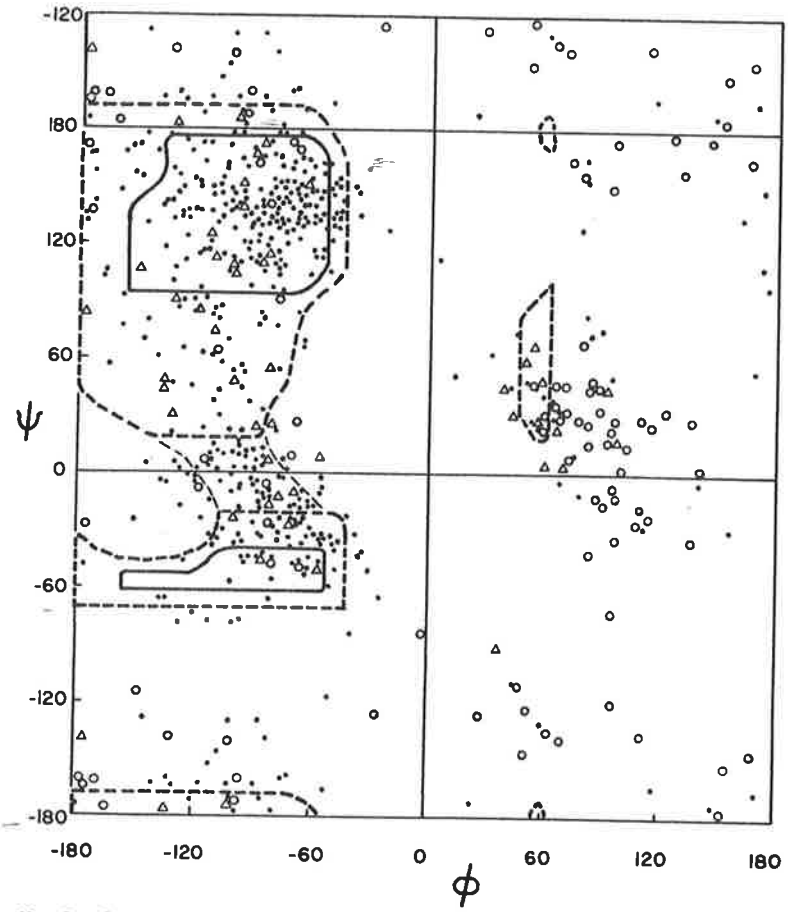


Fig. 49 Conformations observed for the amino acids in carboxypeptidase A, α-chymotrypsin, lysozyme, myoglobin, ribonuclease S, staphylococcal nuclease, and subtilisin. Glycine residues are designated by open circles, asparagine by triangles, and other residues by points. This plot does *not* include residues designated as being within helices or within extended β-structure. The help of Dr. C. G. Schellman in compiling this figure is gratefully acknowledged. References to the original work are given in Table I and in the text.

Fig. 48 (a) Diagram of the distance to each α-carbon atom and to the tip of each side chain of ribonuclease S from the phosphate position P. The angular coordinate increases with residue number but is not meant to have any other stereochemical significance. Taken from Richards and Wyckoff (1971). (b) Diagram showing the distances of successive α-carbon atoms in lactate dehydrogenase from a point near the substrate binding site. The insert shows an enlarged drawing of the loop region which, in the ternary complex (dotted), is observed to approach closer to the active center when coenzyme and substrate are bound. Taken from Adams *et al.* (1972).

were not adequate, since conformations near $\psi = 0°$ occurred quite commonly. Such conformations presumably indicate that the angle τ at the α-carbon atom is greater than the normal tetrahedral value of 109.5° and nearer the value of 112° often observed for amino acids and small polypeptides (Dunnill, 1965; Ramakrishnan and Ramachandran, 1965; Birktoft and Blow, 1972). In fact for the refined α-chymotrypsin coordinates, the mean value of $\tau(C^{\alpha})$ was 111.5° for most amino acids, 109.1° for the β-branched amino acids, and as high as 114.9° for proline. Also, it was necessary to expand the region near (−60, −60) for all residues, and near (60, 60) for glycine in order to allow for a possible hydrogen bond between the NH group of the second peptide and the carbonyl oxygen of the first (Ramachandran *et al.*, 1966; Leach *et al.*, 1966). It now appears that this hydrogen bond, once thought possible only for glycine, may be occasionally formed by other residues. For example, Phe 38 in lysozyme has been reported to have a conformation near (77°, −45°) (Blake *et al.*, 1967a), and Thr 26 in thermolysin a conformation near (86°, −57°) (Matthews, 1972).

Recently, a number of other methods of predicting the conformational energy of a pair of peptides have been proposed, some of which appear to depend upon the contribution of this hydrogen bond (Bystrov *et al.*, 1969; Popov *et al.*, 1968; Crippen and Scheraga, 1969; Pullman, 1971, and references therein).

The important conclusion seems to be that although diagrams of the conformational energy of dipeptides are useful in highlighting doubtful conformations, they cannot be used to totally exclude large areas of conformational space, as was at first thought. The protein structures determined to date have led to modifications and, presumably, improvements in the calculations of conformational energy which no doubt will be helpful in determining the role played by the different types of long-range interactions within globular proteins.

In order to define conformation as precisely as possible it is clearly desirable that the structures of a number of proteins be determined as accurately as possible. The ultimate accuracy attainable will be determined by the resolution limit of the diffraction pattern (see Section II,B,*1*). It has only recently been possible to refine protein structures so as to obtain a satisfactory agreement between their observed and calculated diffraction patterns and therefore approach the ultimate accuracy commensurate with the resolution and accuracy of the diffraction data. The striking success of Jensen and colleagues in refining the structure of rubredoxin at a resolution of 1.5 Å has already stimulated new attempts to refine other proteins (see Section II,D,*3*).

Although a few protein crystals diffract with measurable intensity beyond Bragg spacings of 1.5 Å, they are relatively rare. A number of

attempts to improve the quality of the diffraction patterns (e.g., by cooling or use of cross-linking agents) have been tested without notable success (see Section III,F,*2*). The transfer of crystals of lysozyme for aqueous to nonaqueous solvents reduced the volume of solvent in the unit cell but did not improve the diffraction pattern (Haas, 1969).

At the present time, rubredoxin is the most accurately known of any protein structure, with an estimated coordinate error less than 0.20 Å for most atoms. The considerable accomplishment of Jensen and coworkers notwithstanding, positional uncertainties of 0.1 Å will not satisfy those who are interested in the energetics of protein folding or precise enzyme and substrate alignment in enzyme catalysis. It is to be hoped that even higher accuracy can be attained for rubredoxin and a number of other proteins.

Within the accuracy of present structure determinations, it is clear that significantly nonplanar peptides occur in at least four places in rubredoxin (see Section IV,B,*1*) and probably also in lysozyme (Phillips, 1971).

Cis conformations have been reported for two proline residues in ribonuclease S (Wyckoff *et al.*, 1970), two in a Bence Jones protein (Huber and Steigemann, 1974), one in subtilisin (Wright *et al.*, 1969; Hol, 1971), and one is possible in thermolysin (see Section IV,C,3). More surprisingly, a nonproline cis peptide conformation is thought to occur in carboxypeptidase (Hartsuck and Lipscomb, 1971; see Section IV,C,3). If this postulated cis conformation does in fact occur, it would be unique among the structures of all globular proteins determined to date. However, it should be noted that most of the current electron density maps are poorer than that of carboxypeptidase A and might not be capable of differentiating a rare cis peptide.

The occurrence of regular secondary structure in proteins, including helices and extended β-structure (Pauling and Corey, 1951), is well established, although variable from protein to protein. Myoglobin, the first structure to be determined, has 75% helix and no β-structure, whereas Con A, one of the more recent, has 57% β-structure and no helix whatsoever (see Sections IV,A and J).

Inspection of Fig. 46a,b shows that for α-chymotrypsin and ribonuclease, the β-structure need not be idealized. This is the typical situation for most proteins, where regions of relatively perfect hydrogen bonding can be interrupted by more irregular non-hydrogen-bonded conformations. The most perfect and extensive regions of β-structure occur in carboxypeptidase, carbonic anhydrase, and Con A. In each case, the β-structure forms a twisted wall extending through the molecule and is obviously a major feature in the organization and stabilization of the protein. Both parallel and antiparallel β-structures are often

observed, the former more frequently. In both cases, the resultant β-sheet may be twisted, but this is not always the case.

Helices occur in most proteins, although they are rarely idealized, and individual residues may vary considerably from the idealized conformation of $(-57°, -48°)$ (Pauling and Corey, 1951). In one of the common distortions of the idealized α-helix, the carbonyl oxygens tilt outward from the helix axis to accept hydrogen bonds from neighboring side chains such as serine and threonine. Quite frequently a region of α-helix will "tighten" into a 3.0_{10} helical conformation (Bragg *et al.*, 1950; Donohue, 1953), particularly at its carboxyl terminus. Also, hydrogen bonds corresponding to the 3.6_{13} helix (Pauling and Corey, 1951) and the α_{II} helix (Neméthy *et al.*, 1967) have been reported (e.g., see Watson, 1969; Blake *et al.*, 1967a; Lipscomb *et al.*, 1970; Birktoft and Blow, 1972).

A single turn of left-handed helix is formed in thermolysin by the amino acid sequence -Asp-Asn-Gly-Gly- (see Section IV,C,*3*). It is interesting to note that asparagine has ϕ values near 60° much more frequently than any other amino acid except glycine. In fact it appears from the listings of conformational angles that excluding glycine, asparagine has $\phi \approx 60°$ half as frequently as all other residues combined.

In addition to helices and β-structure, "hairpin bends" occur quite often and can be considered as another category of secondary structure. Venkatachalam (1968) studied the conformations allowed for tripeptides in which a hydrogen bond could be formed between the first and third unit. He designated six general types of conformation, types I–III and types I′–III′, with approximate conformations given by the following pairs of values for $(\phi_2, \psi_2; \phi_3, \psi_3)$: type I $(-60°, -30°; -90°, 0°)$; type II $(-60°, 120°; 80°, 0°)$; type III $(-60°, -30°; -60°, -30°)$; type I′ $(60°, 30°; 90°, 0°)$; type II′ $(60°, -120°; -80°, 0°)$; type III′ $(60°, 30°; 60°, 30°)$. The different types are illustrated in Fig. 50.

Type I and II are related by a rotation of 180° of the central peptide unit. The conformations illustrated in Fig. 50a,b, in which it appears that the planes of the first and third peptides are parallel, are in fact sterically disallowed due to contacts between the carbonyl oxygen of the first peptide and the carbonyl carbon of the second. In the allowed type I and type II conformations, the first peptide lies mainly above the plane defined by C_2^α, C_3^α, and C_4^α and is tilted with its carbonyl bond out of the plane. The carbonyl oxygen remains approximately in the plane, so that the hydrogen bond N—H···O is approximately linear as it appears in Fig. 50c. Since the conformations at C_1^α in the type I and type III bends are identical and the differences at C_2^α are quite slight, type I and type III are almost indistinguishable. The con-

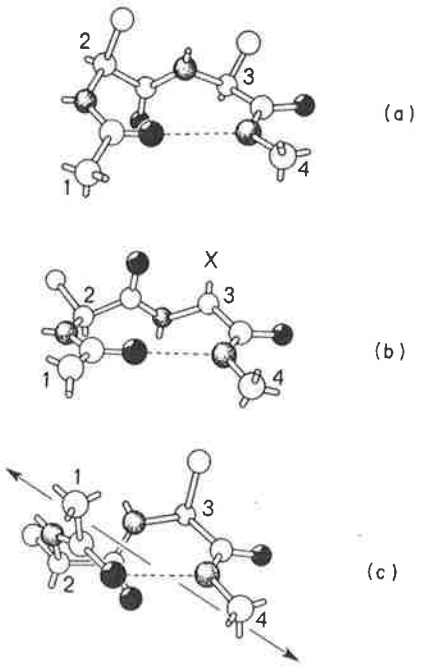


Fig. 50 Diagrammatic representation of turns possible for a tripeptide. Taken from Birktoft and Blow (1972) after Venkatachalam (1968). (a) Types I and II′ (type III is very similar). (b) Types II and I′ (type III′ is very similar). (c) Type III (types I and II′ are very similar). As discussed in the text, in type I and type II bends, the planes of the first and third peptide groups are not parallel. The type I and type III bends are in practice almost indistinguishable, as are type I′ and type III′.

formation of the type III bend corresponds to the 3_{10} helix. In the type I′ and type II′ conformations, the first peptide lies mainly below the plane of C_2^{α}, C_3^{α}, and C_4^{α} and is tilted in the opposite sense. As for types I and III, types I′ and III′ are practically indistinguishable.

To avoid steric interference and to preserve the hydrogen bond, only glycine is allowed at position 3 in type II turns, or at position 2 in type II′ turns, or at positions 2 and 3 in type I′ and III′ turns. The possible chain reversals predicted by Venkatachalam (1968) all include four α-carbons, yet a hairpin bend of only three α-carbons (Fig. 51) has been observed in thermolysin (Matthews, 1972; see Section IV,C,3).

Recently, Crawford *et al.* (1973) analyzed the coordinates of seven proteins and found that there were 46 examples of the different turns, including all six types, and that in addition there were 79 other conformations which had some but not all the properties of turns. Since each turn includes four residues, it follows that turns are a common structural component of globular proteins: Crawford *et al.* (1973) suggest

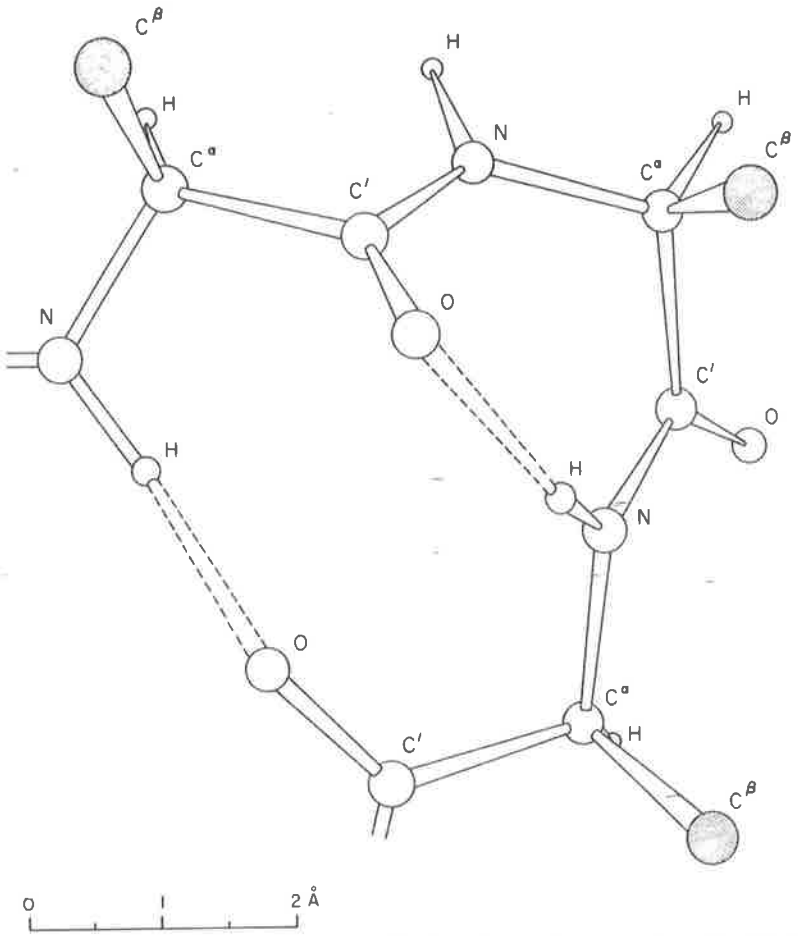


Fig. 51 The γ turn, a chain reversal possible for a dipeptide. Taken from Neméthy and Printz (1972).

that this may have important implications for the understanding of the tertiary conformation of proteins. That turns may be a directing influence in the folding of the polypeptide chain has been suggested by Lewis *et al.* (1971) and by Kuntz (1972).

Following earlier analyses by Cook (1967) and Ptitsyn (1969), Crawford *et al.* (1973) have further examined the amino acid composition of different secondary structures by analyzing data for eleven proteins. Five categories of secondary structure were tabulated: β-sheets, turns, the first three residues in a helix (called the N-helix), the last three residues in a helix (called the C-helix), and the middle residues of

TABLE VI

Distribution of Each Amino Acid between Turns, Helices, and β-Sheet[a]

Amino acid	Turn	N-Helix	M-Helix	C-Helix	β-Sheet
Asp	41.3	13.9	13.9	5.6	15.3
Glu	21.4	20.3	24.1	12.7	6.3
Lys	36.6	4.2	16.0	15.1	20.2
Arg	26.3	1.6	11.1	14.3	12.7
His	31.5	10.2	22.4	18.4	6.1
Asn	47.2	8.9	7.9	7.9	12.9
Gln	30.6	9.5	21.6	17.6	14.9
Cys	42.8	8.5	8.5	14.9	17.0
Thr	34.8	11.6	14.3	6.2	19.7
Ser	41.9	10.5	9.9	7.9	11.2
Tyr	44.8	7.2	6.0	3.6	24.1
Trp	40.7	17.1	25.7	2.9	14.3
Gly	45.8	6.4	7.1	6.4	9.6
Ala	20.0	8.5	28.1	9.1	17.1
Val	15.8	5.3	21.4	6.1	27.5
Leu	23.1	6.6	28.7	9.0	26.2
Ile	22.3	8.3	22.2	3.3	30.0
Phe	35.0	5.9	23.6	9.8	21.6
Pro	48.8	15.4	1.5	–	6.1
Met	22.2	4.0	28.0	16.0	24.0
Average	33.2	8.8	16.7	8.8	17.1

[a] The numbers are percentages and equal the number of residues of the amino acid in the given conformation divided by the total number of residues of that type. Frequencies for turns are based upon the published coordinates of carboxypeptidase A, α-chymotrypsin, ribonuclease S. lysozyme, myoglobin, subtilisin BPN′, and cytochrome b_5. Frequencies for helices and β-sheet include additional data for staphylococcal nuclease, trypsin inhibitor, and papain. Taken from Crawford *et al.* (1973).

helices (M-helix). The percentage distribution of each amino acid between the five conformations is summarized in Table VI. Unfortunately, Crawford *et al.* (1973) did not distinguish between turns in which the overall direction of the polypeptide chain is reversed and those in which it is not. The "Turn" category in Table VI includes some residues with type III (i.e., 3_{10}) conformations which might be more appropriately counted as helical. However, as is clear from the discussion above, the distinction between "hairpin bend" and "helix" need not depend simply upon the "type" of tripeptide conformation. For example, a dipeptide with type III (3_{10}) conformation can occur in a hairpin loop, and a dipeptide with type I conformation can occur in a helix. Also, there may be some duplication in Table VI between the

different categories, particularly for residues at the ends of helices (C. Schellman, private communication). With these reservations, it is clear that the turn is a common structural feature, since of all the residues compared, 33% are in turns, 34% in helices, and 17% in β-sheets. Over 75% of an average protein lies in one of these three conformations.

Glutamic acid, asparagine, glycine, and proline are of interest in that they have an apparent preference for one particular conformation. Glutamic acid is found mainly in helices (57%), while asparagine (47%) and glycine (46%) are found mainly in turns, with low occurrence in other conformations. Proline is found mainly in turns (49%) but also at the beginning of helices (15%) (Crawford *et al.*, 1973).

Although in the four cases quoted there is an apparent *preference* for one conformation or another, it is abundantly clear that most amino acids will be quite likely to fall into any one of the categories listed. It is therefore not surprising to find that all of the methods which have been proposed to predict secondary structure from amino acid sequence can claim some success but are, in every case, far from perfect. In a recent test of current predictive methods, the secondary structure of adenylate kinase was predicted without prior knowledge of the X-ray structure (Schulz *et al.*, 1974b). In general, the agreement between prediction and observation was moderately good, but in a subsequent test carried out for T4 phage lysozyme (Matthews, 1975), the agreement between prediction and observation was poor.

B. Protein Evolution

During the last few years the comparison of amino acid sequences has become a common tool for assessing possible structural and evolutionary relatedness (e.g., see Fitch, 1966; Cantor and Jukes, 1966; Margoliash *et al.*, 1968; Bradshaw *et al.*, 1969a; Neurath *et al.*, 1969; Haber and Koshland, 1970; Smith, 1970; McLachlan, 1972a; Dayhoff, 1972). It is now possible to augment such comparisons with the known tertiary structures of a number of proteins, and already a number of fascinating results have begun to emerge.

Consider, for example, the questions raised by the comparison of the known three-dimensional structures of α-chymotrypsin and subtilisin (see Section IV,C,*1*). These serine proteases have been shown to have quite different tertiary structures, yet in the vicinity of their active sites, they are strikingly similar. In each enzyme there are analogous residues thought to be important in both substrate binding and in catalysis which are spatially arranged in such a way that within the limitations of the current electron density maps, the active sites of the

respective enzymes appear to be identical. This would seem to be a clear case of convergent evolution of two different proteins that have independently evolved a catalytic apparatus which is optimal for the cleavage of internal peptide bonds.

The fact that not just a few residues but the whole of the active site region of the two enzymes are similarly arranged might be taken as evidence that evolution has led to the same molecular apparatus in two separate and independent instances and that this might be true in general of convergent evolution. However, from a comparison of the structure of thermolysin with that of carboxypeptidase A (see Section IV,C,*3*), it is clear that convergent evolution can lead to *similar* molecular machinery, but that it need not be *identical*.

There are also clear structural examples of *divergent* rather than convergent evolution; for example, the common nucleotide binding domain in lactate-, malate-, alcohol-, and glyceraldehyde-3-phosphate dehydrogenases (see Section IV,H,*3*). Also, the "calcium-binding loop" of a carp muscle protein appears to be present in myosin light chains and in troponin C (see Section IV,I).

The cases that are more difficult to categorize are those in which a topologically similar structural feature is present in different structures, but no evidence of sequence homology or functional relatedness exists. For example, the "nucleotide-binding domain" present in lactate dehydrogenase has an apparent counterpart in subtilisin and in phosphoglycerate mutase, neither of which has a nucleotide binding capability (see Section IV,H,*3*). Also, the calcium-binding protein from carp muscle has three "loop" regions that are structurally similar, but only two bind calcium. A region of T4 phage lysozyme also has a somewhat similar "loop" (Tufty and Kretsinger, 1975), although there is no evidence for metal binding (B. W. Matthews and S. J. Remington, unpublished observations). Another example of similar folding is the structure of superoxide dismutase (Richardson *et al.*, 1975) which is, in part, strikingly similar to that of an immunoglobulin domain. Topologically similar regions have also been observed in the two halves of the α-chymotrypsin molecule (Fig. 46), in which the polypeptide chain folds back and forth to form six-stranded antiparallel β-sheets which close to form two "cylinders" (see Section IV,C,*1*).

In most of the examples of structural relatedness quoted in the preceding paragraph, the corresponding amino acid sequences have no detectable homology. It is not known whether such related structures have derived from a common ancestral gene, or whether they represent examples of "super secondary structure" which have evolved independently.

It is always true that if two proteins have significant homology in their amino acid sequences, then no matter how far apart they may be from an evolutionary point of view, their tertiary structures will also be similar. For example, the structures of the hemoglobins from such diverse species as lupin, the sea lamprey, a marine worm, a larval insect, and man (see Fig. 20 and Section IV,A,2) have all been shown to be very similar. This suggests that if three-dimensional structure is preserved over extended periods of evolutionary time, then significant sequence homology tends to be retained. On the other hand, it is also clear that during evolution the amino acid sequence is much more variable and changes more rapidly than tertiary structure. The ability to make meaningful comparisons between proteins that are very distantly related must therefore come primarily from structural homology and only secondarily from sequence homology. When the amino acid sequences of two similar structural domains are aligned on the basis of structural similarity but no sequence homology is apparent, it is possible that the structural domains being compared have arisen through convergent or divergent evolution.

At present, enough structures have been determined to show that many proteins can belong to a specific structural family, the members of which have related tertiary structures. Whether the number of such families is likely to be small and whether all members within a family are evolutionarily related will only be decided through the determination of additional protein structures, chosen to bridge the gaps in our current understanding of protein evolution.

C. The Future

It seems to be a truism that every answer leaves a question. Obviously, a great deal more is known about protein structure and function now than ten years ago, yet there are many unanswered questions.

The structure of oxy- and deoxyhemoglobin are now both known in considerable detail and have led Perutz (1970b,c) to propose detailed molecular models for heme–heme interaction and the Bohr effect (see Section IV,A,2). Yet the crucial question concerning the structures of the partially oxygenated intermediates remains unresolved and is not likely to be answered directly by X-ray methods.

α-Chymotrypsin and its precursor chymotrypsinogen A have been shown to have very similar structures as well as all the structural elements thought to be important for peptide cleavage (see Section IV,C,), yet one is fully active and the other is catalytically impotent.

The three-dimensional structures of these two enzymes, as they are currently known, do not provide a satisfactory explanation for this difference, suggesting that the distinction between active and inert may be due to structural differences smaller than those detectable at the present limits of accuracy.

Perhaps even more fascinating are the evolutionary questions raised by the comparison of structurally similar proteins, as discussed in the preceding section.

In spite of the amount of structural information which has been accumulated, only modest progress has been made toward predicting the conformation of a protein from its amino acid sequence.

The proteins chosen for the early crystallographic studies were selected because they were readily available and could be crystallized in a form suitable for X-ray analysis. Now that the basic techniques have been worked out, the emphasis is changing, and crystallographers are looking to the more challenging problems of maximum biochemical interest.

The time-limiting steps are often those of crystal growth and of obtaining suitable heavy atom derivatives. Although a great deal of experience has been gained, many structural studies have been thwarted by one or the other of these hurdles. It is to be hoped that more rational approaches to each of these problem areas will be developed.

Given crystals, the amount of protein available need not be a major factor, as exemplified by the fact that Watenpaugh *et al.* (1972) recently grew crystals and obtained an interpretable electron density map of flavodoxin using only 50 mg of protein.

Although the development of new methods for rapid data acquisition (see Section III,F,5) can be expected, it seems likely that at least for the immediate future the X-ray diffraction technique will be restricted to the study of enzyme complexes which are stable for periods that exceed the half-life of typical enzyme–substrate complexes.

Nevertheless, it is encouraging to note that rather than being a technique apart from (or perhaps in competition with) the other biochemical and biophysical methods, X-ray crystallography is becoming accepted as just one of a possible arsenal of weapons which may be brought to bear on a particular problem. It is to be hoped that this trend, which will enhance the strengths and minimize the limitations of the respective techniques, will continue.

At the beginning of this chapter, we commented on the increasing number of new protein structures being determined each year. Inspection of Fig. 52 suggests that since 1965, the increase has, on the

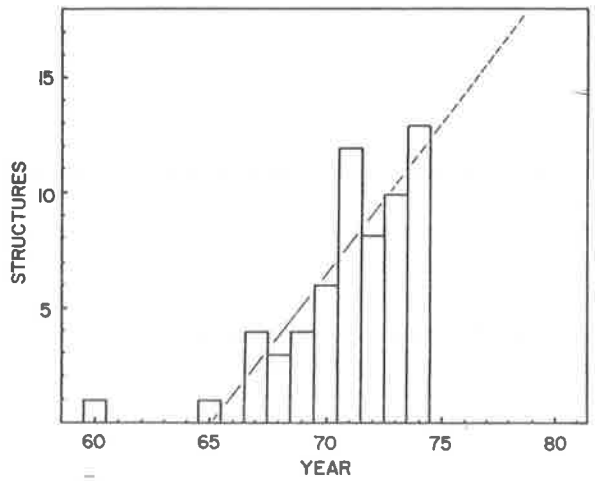


Fig. 52 Protein structures determined by X-ray crystallography (cf. Table I).

average, been roughly linear. There is a precedent for wrong predictions in protein crystallography (Bragg, 1965), but if Fig. 52 is any indication one can expect about thirteen new structures to be determined in 1975 and twenty in 1980.

ACKNOWLEDGMENTS

Many colleagues have contributed to this chapter in different ways, but I would like to express my particular thanks to those with whom I have had the privilege of close collaboration, particularly Drs. D. M. Blow, P. B. Sigler, D. R. Davies, and J. N. Jansonius.

I am also grateful to Drs. J. N. Jansonius and J. Herriott for their care in reading the original manuscript and for a number of helpful suggestions.

The preparation of this chapter was supported in part by grants from the National Science Foundation (BMS74-18407), the National Institutes of Health (GM20066, GM15423, and GM21967), and the award to the author of an Alfred P. Sloan Research Fellowship and a Public Health Service Career Development Award (GM70585) from the Institute of General Medical Sciences.

I am most grateful to a number of correspondents for sending material in advance of publication and for permission to reproduce figures.

REFERENCES

Abrahamsson, S. (1966). *J. Sci. Instrum.* **43**, 931.
Abrahamsson, S., Aleby, S., and Innes, J. (1968). *J. Sci. Instrum.* [2] **1**, 869.
Adair, G. S., and Adair, M. E. (1936). *Proc. R. Soc. London, Ser. B* **120**, 422.

Adams, M. J., Blundell, T. L., Dodson, E. J., Dodson, G. G., Vijayan, M., Baker, E. N., Harding, M. M., Hodgkin, D. C., Rimmer, B., and Sheat, S. (1969a). *Nature (London)* **244,** 491.

Adams, M. J., Haas, D. J., Jeffery, B. A., McPherson, A., Jr., Mermall, H. L., Rossmann, M. G., Schevitz, R. W., and Wonacott, A. J. (1969b). *J. Mol. Biol.* **41,** 159.

Adams, M. J., Ford, G. C., Koekoek, R., Lentz, P. J., McPherson, A., Rossmann, M. G., Smiley, I. E., Schevitz, R. W., and Wonacott, A. J. (1970a). *Nature (London)* **227,** 1098.

Adams, M. J., McPherson, A., Jr., Rossmann, M. G., Shevitz, R. W., and Wonacott, A. J. (1970b). *J. Mol. Biol.* **51,** 31.

Adams, M. J., Buehner, M., Chandrasekhar, K., Ford, G. C., Hackert, M. L., Liljas, A., Lentz, P., Jr., Rao, S. T., Rossmann, M. G., Smiley, I. E., and White, J. L. (1972). *In* "Protein-Protein Interactions" (J. Jaenicke and E. Helmreich, eds.), p. 139. Springer-Verlag, Berlin and New York.

Adams, M. J., Buehner, M., Chandrasekhar, K., Ford, G. C., Hackert, M. L., Liljas, A., Rossmann, M. G., Smiley, I. E., Allison, W. S., Everse, J., Kaplan, N. O., and Taylor, S. S. (1973a). *Proc. Natl. Acad. Sci. U.S.A.* **70,** 1968.

Adams, M. J., Ford, G. C., Liljas, A., and Rossmann, M. G. (1973b). *Biochem Biophys. Res. Commun.* **53,** 46.

Adman, E. T., Sieker, L. C., and Jensen, L. H. (1973). *J. Biol. Chem.* **248,** 3987.

Åkervall, K., and Strandberg, V. (1971). *J. Mol. Biol.* **62,** 625.

Alden, R. A., Wright, C. S., and Kraut, J. (1970). *Philos. Trans. R. Soc. London, Ser. B* **257,** 119.

Alden, R. A., Birktoft, J. J., Kraut, J., Robertus, J. D., and Wright, C. S. (1971). *Biochem. Biophys. Res. Commun.* **45,** 337.

Andersen, R. D., Apgar, P. A., Burnett, R. M., Darling, G. D., LeQuesne, M. E., Mayhew, S. G., and Ludwig, M. L. (1972). *Proc. Natl. Acad. Sci. U.S.A.* **69,** 3189.

Anderson, W. F., and Steitz, T. A. (1975). *J. Mol. Biol.* **92,** 279.

Anderson, W. F., Fletterick, R. J., and Steitz, T. A. (1974). *J. Mol. Biol.* **86,** 261.

Anfinsen, C. B., Cuatrecasas, P., and Taniuchi, H. (1971). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 4, p. 177. Academic Press, New York.

Argos, P., and Mathews, F. S. (1975). *J. Biol. Chem.* **250,** 747.

Arndt, U. W. (1968). *Acta Crystallogr., Sect. B* **24,** 1355.

Arndt, U. W. (1969). *Acta Crystallogr., Sect. A* **25,** 161.

Arndt, U. W., and Mathieson, A. McL. (1969). *Acta Crystallogr., Sect. A* **25,** 1.

Arndt, U. W., and Willis, B. T. M. (1966). "Single Crystal Diffractometry." Cambridge Univ. Press, London and New York.

Arndt, U. W., North, A. C. T., and Phillips, D. C. (1964). *J. Sci. Instrum.* **41,** 421.

Arndt, U. W., Crowther, R. A., and Mallett, J. F. W. (1968). *J. Sci. Instrum.* [2] **1,** 510.

Arndt, U. W., Champness, J. N., Phizackerley, R. P., and Wonacott, A. J. (1973). *J. Appl. Crystallogr.* **6,** 457.

Arnone, A. (1972). *Nature (London)* **237,** 146.

Arnone, A., Bier, C. J., Cotton, F. A. Hazen, E. E., Jr., Richardson, D. C., and Richardson, J. S. (1969). *Proc. Natl. Acad. Sci. U.S.A.* **64,** 420.

Arnone, A., Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. E., Jr., Richardson, D. C., Richardson, J. S., and in part Yonath, A. (1971). *J. Biol. Chem.* **246,** 2302.

Aschaffenburg, R., Green, D. W., and Simmons, R. M. (1965). *J. Mol. Biol.* **13,** 194.

Aschaffenburg, R., Fenna, R. E., Handford, B. O., and Phillips, D. C. (1972a). *J. Mol. Biol.* **67,** 525.

Aschaffenburg, R., Fenna, R. E., and Phillips, D. C. (1972b). *J. Mol. Biol.* **67,** 529.

Avey, H. P., Poljak, R. J., Rossi, G., and Nisonoff, A. (1968). *Nature (London)* **220,** 1248.

Baker, E. N., and Dodson, G. (1970). *J. Mol. Biol.* **54**, 605.
Banaszak, L. J. (1966). *J. Mol. Biol.* **22**, 389.
Banaszak, L. J., Tsernoglou, D., and Sade, M. (1971). *Probes Struct. Funct. Macromol. Membr., Proc. Colloq. Johnson Res. Found., 5th, 1969* Vol. 2, p. 71.
Banner, D. W., Bloomer, A. C., Petsko, G. A., and Phillips, D. C. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 151.
Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D., and Waley, S. G. (1975). *Nature (London)* **255**, 609.
Banyard, S. H., Blake, C. C. F., and Swan, I. D. A. (1974). *In* "Lysozyme" (E. F. Osserman, R. E. Canfield, and S. Beychok, eds.), p. 71. Pergamon, Oxford.
Barrett, A. N., and Zwick, M. (1971). *Acta Crystallogr., Sect. A* **27**, 6.
Barry, C. D., and North, A. C. T. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 577.
Becker, J. W., Reeke, G. N., Jr., Cunningham, B. A., and Edelman, G. M. (1976). *Nature (London)* **259**, 406.
Becker, J. W., Reeke, G. N., Jr., and Edelman, G. M. (1971). *J. Biol. Chem.* **246**, 6123.
Becker, J. W., Reeke, G. N., Jr., Wang, J. L., Cunningham, B. A., and Edelman, G. M. (1975). *J. Biol. Chem.* **250**, 1513.
Bello, J., and Nowoswiat, E. F. (1965). *Biochim. Biophys. Acta* **105**, 325.
Benesch, P., and Benesch, R. (1969). *Nature (London)* **221**, 618.
Bergstén, P.-C., Waara, I., Lövgren, S., Liljas, A., Kannan, K. K., and Bengtsson, J. (1972). *Oxygen Affinity Hemoglobin Red Cell Acid Base Status, Proc. Alfred Benzon Symp., 4th, 1971* p. 363.
Bernal, J. D. (1939). *Nature (London)* **143**, 633.
Bernal, J. D., and Crowfoot, D. (1934). *Nature (London)* **133**, 794.
Bernal, J. D., Fankuchen, I., and Perutz, M. F. (1938). *Nature (London)* **141**, 523.
Berthou, J., Cesbron, F., and Laurent, A. (1972a). *J. Mol. Biol.* **71**, 809.
Berthou, J., Laurent, A., and Jollès, P. (1972b). *J. Mol. Biol.* **71**, 815.
Bijvoet, J. M. (1954). *Nature (London)* **173**, 888.
Birktoft, J. J., and Blow, D. M. (1972). *J. Mol. Biol.* **68**, 187.
Birktoft, J. J., Matthews, B. W., and Blow, D. M. (1969). *Biochem. Biophys. Res. Commun.* **36**, 131.
Blake, C. C. F. (1968). *Adv. Protein Chem.* **23**, 59.
Blake, C. C. F., and Evans, P. R. (1974). *J. Mol. Biol.* **84**, 585.
Blake, C. C. F., and Phillips, D. C. (1962). *In* "Biological Effects of Ionizing Radiations at the Molecular Level," p. 183. JAEA, Vienna.
Blake, C. C. F., and Swan, I. D. A. (1971). *Nature (London), New Biol.* **232**, 12.
Blake, C. C. F., Fenn, R. H., Koenig, N. F., North, A. C. T., Phillips, D. C., and Poljak, R. J. (1963). *Acta Crystallogr.* **16**, A77.
Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C., and Sarma, V. R. (1965). *Nature (London)* **206**, 757.
Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C., and Sarma, V. R. (1967a). *Proc. R. Soc. London, Ser. B* **167**, 365.
Blake, C. C. F., Johnson, L. N., Mair, G. A., North, A. C. T., Phillips, D. C., and Sarma, V. R. (1967b). *Proc. R. Soc. London, Ser. B* **167**, 378.
Blake, C. C. F., Swan, I. D. A., Rerat, C., Berthou, J., Laurent, A., and Rerat, B. (1971). *J. Mol. Biol.* **61**, 217.
Blake, C. C. F., Evans, P. R., and Scopes, R. K. (1972). *Nature (London), New Biol.* **235**, 195.
Blake, C. C. F., Geisow, M. J., Swan, I. D. A., Rerat, C., and Rerat, B. (1974). *J. Mol. Biol.* **77**, 1.

Blow, D. M. (1958). *Proc. R. Soc. London, Ser. A* **247**, 302.
Blow, D. M. (1969). *Biochem. J.* **112**, 261.
Blow, D. M. (1971). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 3, p. 185. Academic Press, New York.
Blow, D. M. (1974). *Isr. J. Chem.* **12**, 483.
Blow, D. M., and Crick, F. H. C. (1959). *Acta Crystallogr.* **12**, 794.
Blow, D. M., and Matthews, B. W. (1973). *Acta Crystallogr., Sect. A* **29**, 56.
Blow, D. M., and Rossmann, M. G. (1961). *Acta Crystallogr.* **14**, 1195.
Blow, D. M., and Rossmann, M. G. (1962). *Acta Crystallogr.* **15**, 1060.
Blow, D. M., and Steitz, T. A. (1970). *Annu. Rev. Biochem.* **39**, 63.
Blow, D. M., Rossmann, M. G., and Jeffery, B. A. (1964). *J. Mol. Biol.* **8**, 65.
Blow, D. M., Birktoft, J. J., and Hartley, B. S. (1969). *Nature (London)* **221**, 337.
Blow, D. M., Wright, C. S., Kukla, D., Rühlmann, A., Steigemann, W., and Huber, R. (1972). *J. Mol. Biol.* **69**, 137.
Blow, D. M., Janin, J., and Sweet, R. M. (1974). *Nature (London)* **249**, 5452.
Blundell, T. L., Cutfield, J. F., Cutfield, S. M., Dodson, E. J., Dodson, G. G., Hodgkin, D. C., Mercola, D. A., and Vijayan, M. (1971a). *Nature (London)* **231**, 506.
Blundell, T. L., Cutfield, J. F., Cutfield, S. M., Dodson, E. J., Dodson, G. G., Hodgkin, D. C., Mercola, D. A., and Vijayan, M. (1971b). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 233.
Blundell, T. L., Cutfield, J. F., Cutfield, E. J., Dodson, E. J., Dodson, G. G., Hodgkin, D. C., Mercola, D. A., and Vijayan, M. (1972). *Acta Crystallogr., Sect. A* **28**, S34.
Bode, W., and Schwager, P. (1975). *J. Mol. Biol.* **98**, 693.
Bokhoven, C., Schoone, J. C., and Bijvoet, J. M. (1951). *Acta Crystallogr.* **4**, 275.
Bolton, W., and Perutz, M. F. (1970). *Nature (London)* **228**, 551.
Bradshaw, R. A., Neurath, H., and Walsh, K. A. (1969a). *Proc. Natl. Acad. Sci. U.S.A.* **63**, 406.
Bradshaw, R. A., Ericsson, L. H., Walsh, K. A., and Neurath, H. (1969b). *Proc. Natl. Acad. Sci. U.S.A.* **63**, 1389.
Bragg, L. (1965). *Rep. Prog. Phys.* **28**, 1.
Bragg, W. L. (1913). *Proc. Cambridge Philos. Soc.* **17**, 43.
Bragg, W. L., and Perutz, M. F. (1952). *Acta Crystallogr.* **5**, 323.
Bragg, W. L., Kendrew, J. C., and Perutz, M. F. (1950). *Proc. R. Soc. London, Ser. A* **203**, 321.
Brändén, C.-I., Holmes, K. C., and Kendrew, J. C. (1963). *Acta Crystallogr.* **16**, A175.
Brändén, C.-I., Zeppezauer, E., Boiwe, T., Söderberg, B.-O., and Nordström, B. (1969). *In* "Pyridine Nucleotide-Dependent Dehydrogenases" (H. Sund, ed.), p. 129. Springer-Verlag, Berlin and New York.
Brändén, C.-I., Eklund, H., Nordström, B., Boiwe, T., Söderlund, G., Zeppezauer, E., Ohlsson, I., and Åkeson, Å. (1973). *Proc. Natl. Acad. Sci. U.S.A.* **70**, 2439.
Braun, P. B., Hornstra, J., and Leenhouts, J. I. (1969). *Philips Res. Rep.* **24**, 85.
Brew, K., Vanaman, T. C., and Hill, R. L. (1967). *J. Biol. Chem.* **242**, 3747.
Brewer, C. F., Sternlicht, H., Marcus, D. M., and Grollmann, A. P. (1973). *Biochemistry* **12**, 4448.
Browne, W. J., North, A. C. T., Phillips, D. C., Brew, K., Vanaman, T. C., and Hill, R. L. (1969). *J. Mol. Biol.* **42**, 65.
Bryant, T. N., Watson, H. C., and Wendell, P. L. (1974). *Nature (London)* **247**, 14.
Buehner, M., Ford, G. C., Moras, D., Olsen, K. W., and Rossmann, M. G. (1973). *Proc. Natl. Acad. Sci. U.S.A.* **70**, 3052.
Buehner, M., Ford, G. C., Moras, D., Olsen, K. W., and Rossmann, M. G. (1974a). *J. Mol. Biol.* **82**, 563.

Buehner, M., Ford, G. C., Moras, D., Olsen, K. W., and Rossmann, M. G. (1974b). *J. Mol. Biol.* **90**, 25.
Buerger, M. J. (1964). "The Precession Method." Wiley, New York.
Burgner, J. W., and McGandy, E. L. (1967). *Abstr., Am. Crystallogr. Assoc.*, January, 1967. Abstr. C11.
Burnett, R. M., Darling, G. D., Kendall, D. S., LeQuesne, M. E., Mayhew, S. G., Smith, W. W., and Ludwig, M. L. (1974). *J. Biol. Chem.* **249**, 4383.
Bystrov, V. F., Portnova, S. L., Tsetlin, V. I., Ivanov, V. T., and Ovchinnikov, Y. A. (1969). *Tetrahedron* **25**, 493.
Bywater, R. P., Carlisle, C. H., and Jackson, R. B. (1969). *J. Mol. Biol.* **45**, 429.
Bywater, R. P., Carlisle, C. H., Jackson, R. B., Mackay, A. L., and Timmins, P. A. (1972). *Acta Crystallogr., Sect. A* **28**, S37.
Camerman, N., Hosmann, T., Jones, S., and Nyburg, S. C. (1969). *J. Mol. Biol.* **44**, 569.
Campbell, J. W., Duée, E., Hodgson, G., Mercer, W. D., Stammers, D. K., Wendell, P. L., Muirhead, H., and Watson, H. C. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 165.
Campbell, J. W., Hodgson, G. I., and Watson, H. C. (1972). *Nature (London), New Biol.* **240**, 137.
Campbell, J. W., Watson, H. C., and Hodgson, G. L (1974). *Nature (London)* **250**, 301.
Cantor, C. R., and Jukes, T. H. (1966). *Proc. Natl. Acad. Sci. U.S.A.* **56**, 177.
Capasso, S., Giordano, F., Mazzarella, L., and Ripamonti, A. (1972). *J. Mol. Biol.* **64**, 311.
Carlisle, C. H., Gorinsky, B. A., Bazumdar, S. K., Palmer, R. A., and Yeates, D. G. R. (1972). *Acta Crystallogr., Sect. A* **28**, S36.
Carlisle, C. H., Palmer, R. A., Bazumdar, S. K., Gorinsky, B. A., and Yeates, D. G. R. (1974). *J. Mol. Biol.* **85**, 1.
Carter, C. W., Jr., Freer, S. T., Xuong, Ng. H., Alden, R. A., and Kraut, J. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 381.
Carter, C. W., Jr., Kraut, J., Freer, S. T., Xuong, Ng. H., Alden, R. A., and Bartsch, R. G. (1974a). *J. Biol. Chem.* **249**, 4212.
Carter, C. W., Jr., Kraut, J., Freer, S. T., and Alden, R. A. (1974b). *J. Biol. Chem.* **249**, 6339.
Chirckjian, J. C., Wright, H. T., and Fresco, J. R. (1972). *Proc. Natl. Acad. Sci. U.S.A.* **69**, 1638.
Cochran, W. (1955). *Acta Crystallogr.* **8**, 473.
Cohen, G. H., Matthews, B. W., and Davies, D. R. (1970). *Acta Crystallogr., Sect. B* **26**, 1062.
Cold Spring Harbor Symposium on Quantitative Biology. (1971). Vol. 36. Cold Spring Harbor Lab., Cold Spring Harbor, New York.
Colman, P. M., and Matthews, B. W. (1971). *J. Mol. Biol.* **60**, 163.
Colman, P. M., Jansonius, J. N., and Matthews, B. W. (1972a). *J. Mol. Biol.* **70**, 701.
Colman, P. M., Weaver, L. H., and Matthews, B. W. (1972b). *Biochem. Biophys. Res. Commun.* **46**, 1999.
Cook, D. A. (1967). *J. Mol. Biol.* **29**, 167.
Corey, R. B., Battfay, O., Brueckner, D. A., and Mark, F. G. (1965). *Biochim. Biophys. Acta* **94**, 535.
Cork, C., Fehr, D., Hamlin, R., Vernon, W., Xuong, Ng. H., and Perez-Mendez, V. (1973). *J. Appl. Crystallogr.* **7**, 319.
Cork, J. M. (1927). *Philos. Mag.* [7] **4**, 688.
Cornick, G., Sigler, P. B., and Ginsberg, H. S. (1971). *J. Mol. Biol.* **57**, 397.
Cornick, G., Sigler, P. B., and Ginsberg, H. S. (1973). *J. Mol. Biol.* **73**, 533.

Cotton, F. A., and Hazen, E. E., Jr. (1971). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 4, p. 153. Academic Press, New York.
Cotton, F. A., Hazen, E. E., Jr., and Richardson, D. C. (1966). *J. Biol. Chem.* **241**, 4389.
Cotton, F. A., Bier, C. J., Day, V. W., Hazen, E. E., Jr., and Larsen, S. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 243.
Coulter, C. L. (1965). *J. Mol. Biol.* **12**, 292.
Coulter, C. L. (1971). *Acta Crystallogr., Sect. B* **27**, 1730.
Crawford, J. L., Lipscomb, W. H., and Schellman, C. G. (1973). *Proc. Natl. Acad. Sci. U.S.A.* **70**, 538.
Crick, F. H. C., and Kendrew, J. C. (1957). *Adv. Protein Chem.* **12**, 133.
Crick, F. H. C., and Magdoff, B. S. (1956). *Acta Crystallogr.* **9**, 901.
Crippen, G. M., and Scheraga, H. A. (1969). *Proc. Natl. Acad. Sci. U.S.A.* **64**, 42.
Crowfoot, D. M. (1935). *Nature (London)* **135**, 591.
Crowther, R. A. (1967). *Acta Crystallogr.* **22**, 758.
Crowther, R. A. (1969). *Acta Crystallogr., Sect. B* **25**, 2571.
Crowther, R. A., and Blow, D. M. (1967). *Acta Crystallogr.* **23**, 544.
Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G., and North, A. C. T. (1961a). *Proc. R. Soc. London, Ser. A* **265**, 15.
Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G., and North, A. C. T. (1961b). *Proc. R. Soc. London, Ser. A* **265**, 161.
Cunningham, L. (1965). *Compr. Biochem.* **16**, 85.
Czerwinski, E. W., Mathews, F. S., Hollenberg, P., Drickamer, K., and Hager, L. P. (1972). *J. Mol. Biol.* **71**, 819.
Davies, D. R., Cohen, G. H., Silverton, E. W. Braxton, H. P., and Matthews, B. W. (1969). *Acta Crystallogr., Sect. A* **25**, S182.
Davies, D. R., Padlan, E. A., and Segal, D. M. (1975). *Annu. Rev. Biochem.* **44**, 639.
Dayhoff, M. O. (1972). "Atlas of Protein Sequence and Structure, 1972." Natl. Biomed. Res. Found., Washington, D.C.
Deisenhofer, J., and Steigemann, W. (1975). *Acta Crystallogr., Sect. B* **31**, 238.
DeRosier, D. J., Oliver, R. M., and Lester, J. R. (1971). *Proc. Natl. Acad. Sci. U.S.A.* **68**, 1135.
Desnuelle, P., Neurath, H., and Ottesen, M., eds. (1970). "Structure-Function Relationships of Proteolytic Enzymes." Munksgaard, Copenhagen.
Diamond, R. (1966). *Acta Crystallogr.* **21**, 253.
Diamond, R. (1969). *Acta Crystallogr., Sect. A* **25**, 43.
Diamond, R. (1971). *Acta Crystallogr., Sect. A* **27**, 436.
Diamond, R. (1974). *J. Mol. Biol.* **82**, 371.
Dickerson, R. E. (1964). *In* "The Proteins" (H. Neurath, ed.), 2nd ed., Vol. 2, p. 603. Academic Press, New York.
Dickerson, R. E. (1971). *J. Mol. Biol.* **57**, 1.
Dickerson, R. E. (1972). *Annu. Rev. Biochem.* **41**, 815.
Dickerson, R. E., and Geis, I. (1969). "The Structure and Action of Proteins." Harper, New York.
Dickerson, R. E., Kendrew, J. C., and Strandberg, B. E. (1961). *Acta Crystallogr.* **14**, 1188.
Dickerson, R. E., Kopka, M. L., Varnum, J. C., and Weinzierl, J. E. (1967). *Acta Crystallogr.* **23**, 511.
Dickerson, R. E., Weinzierl, J. E., and Palmer, R. A. (1968). *Acta Crystallogr., Sect. B* **24**, 997.
Dickerson, R. E., Eisenberg, D., Varnum, J., and Kopka, M. L. (1969). *J. Mol. Biol.* **45**, 77.

Dickerson, R. E., Takano, T., Eisenberg, D., Kallai, O. B., Samson, L., Cooper, A., and Margoliash, E. (1971). *J. Biol. Chem.* **246**, 1511.
Dobler, M., Dover, S. D., Laves, K., Binder, A., and Zuber, H. (1972). *J. Mol. Biol.* **71**, 785.
Dodson, E., Harding, M. M., Hodgkin, D. C., and Rossmann, M. G. (1966). *J. Mol. Biol.* **16**, 227.
Donohue, J. (1953). *Proc. Natl. Acad. Sci. U.S.A.* **39**, 470.
Doscher, M., and Richards, F. M. (1963). *J. Biol. Chem.* **238**, 2399.
Drenth, J., and Smit, J. D. G. (1971). *Biochem. Biophys. Res. Commun.* **45**, 1320.
Drenth, J., Jansonius, J. N., Koekoek, R., Marrink, J., Munnim, J., and Wolthers, B. G. (1962). *J. Mol. Biol.* **5**, 398.
Drenth, J., Kloosterman, D., van der Woude, J., Croon, H. C., and van Zwet, L. C. M. (1965). *J. Sci. Instrum.* **42**, 222.
Drenth, J., Jansonius, J. N., and Wolthers, B. G. (1967). *J. Mol. Biol.* **24**, 449.
Drenth, J., Jansonius, J. N., Koekoek, R., Swen, H. M., and Wolthers, B. G. (1968). *Nature (London)* **218**, 929.
Drenth, J., Hol. W. G. J., Jansonius, J. N., and Koekoek, R. (1971a). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 107.
Drenth, J., Jansonius, J. N., Koekoek, R., and Wolthers, B. G. (1971b). *Adv. Protein Chem.* **25**, 79.
Drenth, J., Jansonius, J. N., Koekoek, R., and Wolthers, B. G. (1971c). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 3, p. 485. Academic Press, New York.
Drenth, J., Hol, W. G. J., Jansonius, J. N., and Koekoek, R. (1972). *Eur. J. Biochem.* **26**, 177.
Drenth, J., Swen, H. M., Hoogenstraaten, W., and Sluyterman, L. A. AE. (1975). *Proc. K. Ned. Akad. Wet. Ser. C* **78**, 104.
Dunnill, P. (1965). *Sci. Prog. (London)* **53**, 609.
Dunnill, P., Green, D. W., and Simmons, R. M. (1966). *J. Mol. Biol.* **22**, 135.
Dus, K., Tedro, S., Bartsch, R. G., and Kamen, M. D. (1971). *Biochem. Biophys. Res. Commun.* **43**, 1239.
Eagles, P. A. M., Johnson, L. N., Joynson, M. A., McMurray, C. H., and Gutfreund, H. (1969). *J. Mol. Biol.* **45**, 533.
Eagles, P. A. M., Iqbal, M., Johnson, L. N., Mosley, J., and Wilson, K. S. (1972). *J. Mol. Biol.* **71**, 803.
Edelman, G. M., Cunningham, B. A., Reeke, G. N., Jr., Becker, J. W., Waxdal, M. J., and Wang, J. L. (1972). *Proc. Natl. Acad. Sci. U.S.A.* **69**, 2580. Edmundson, A. B. (1965). *Nature (London)* **205**, 883.
Edmundson, A. B., Wood, M. K., Schiffer, M., Hardman, K. D., Ainsworth, C. F., and Ely, K. R. (1970). *J. Biol. Chem.* **245**, 2763.
Edmundson, A. B., Schiffer, M., Ely, K. T., and Wood, M. K. (1972). *Biochemistry* **11**, 1822.
Eisenberg, D. (1970). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 1, p. 1. Academic Press, New York.
Eisenberg, D., Heidner, E. G., Goodkin, P., Dastoor, M. N., Weber, B. H., Wedler, P., and Bell, J. D. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 291.
Eklund, H., Zeppezauer, E., and Brändén, C.-I. (1968). *J. Mol. Biol.* **34**, 193.
Eklund, H., Nordström, B., Zeppezauer, E., Söderlund, G., Ohlsson, I., Boiwe, T., and Brändén, C.-I. (1974). *FEBS Lett.* **44**, 200.
Endo, S. (1962). *J. Ferment. Technol.* **40**, 346.
Epp, O., Palm, W., Fehlhammer, H., Rühlmann, A., Steigemann, W., Schwager, P., and Huber, R. (1972). *J. Mol. Biol.* **69**, 315.

Epp, O., Colman, P., Fehlhammer, H., Bode, W., Schiffer, M., and Huber, R. (1974). *Eur. J. Biochem.* **45,** 513.
Feder, J., Garrett, L. R., and Wildi, B. S. (1971). *Biochemistry* **10,** 4552.
Fehlhammer, H., and Bode, W. (1975). *J. Mol. Biol.* **98,** 683.
Fenna, R. E. (1973). Ph.D. Thesis, University of Oxford, Oxford, England.
Fenna, R. E., and Matthews, B. W. (1975). *Nature (London)* **258,** 573.
Fenna, R. E., Matthews, B. W., Olson, J. M., and Shaw, E. K. (1974). *J. Mol. Biol.* **84,** 231.
Fisher, H. (1965). *Biochim. Biophys. Acta* **109,** 544.
Fitch, W. M. (1966). *J. Mol. Biol.* **16,** 1.
Fletterick, R. J., Bates, D. J., and Steitz, T. A. (1975). *Proc. Natl. Acad. Sci. U.S.A.* **72,** 78.
Franklin, R. M., Harrison, S. C., Pettersson, U., Philipson, L., Brändén, C.-I., and Werner, P.-E. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36,** 503.
Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T., and Xuong, Ng. H. (1970). *Biochemistry* **9,** 1997.
Freer, S. T., Alden, R. A., Carter, C. W., and Kraut, J. (1975). *J. Biol. Chem.* **250,** 46.
Frieden, C. (1971). *Annu. Rev. Biochem.* **40,** 653.
Friedrich, W., Knipping, P., and von Laue, M. (1912). *Sitzungsber. Math.-Phys. Kl. Bayer. Akad. Wiss. München* p. 303.
Fullerton, W. W., Potter, R., and Low, B. W. (1970). *Proc. Natl. Acad. Sci. U.S.A.* **66,** 1213.
Furnas, T. C. (1955). Reported in Magdoff and Crick (1955).
Furnas, T. C., and Harker, D. (1955). *Rev. Sci. Instrum.* **26,** 449.
Glatthaar, B. E., Banaszak, L. J., and Bradshaw, R. A. (1972). *Biochem. Biophys. Res. Commun.* **46,** 757.
Glauser, S., and Rossmann, M. G. (1966). *Acta Crystallogr.* **21,** 175.
Glazer, A. N., and Smith, E. L. (1971). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 3, p. 501. Academic Press, New York.
Gorjunov, A. I., Andreeva, N. S., Baranowski, T., and Wolny, M. (1972). *J. Mol. Biol.* **69,** 421.
Goryunov, A. I., Andreeva, N. S., and Spitsberg, V. L. (1969). *Biofizika* **14,** 1116.
Green, D. W., Ingram, V. M., and Perutz, M. F. (1954). *Proc. R. Soc. London, Ser. A.* **225,** 287.
Greer, J. (1971). *J. Mol. Biol.* **62,** 241.
Greer, J. (1974). *J. Mol. Biol.* **82,** 279.
Greer, J., and Perutz, M. F. (1971). *Nature (London), New Biol.* **230,** 261.
Greer, J., Kaufman, H. W., and Kalb, A. J. (1970). *J. Mol. Biol.* **48,** 365.
Haas, D. J. (1968). *Acta Crystallogr., Sect. B* **24,** 604.
Haas, D. J. (1969). *Acta Crystallogr., Sect. A* **25,** S183.
Haas, D. J. (1970). *Biopolymers* **9,** 1547.
Haas, D. J., and Lentz, P. J. (1969). *Biopolymers* **7,** 809.
Haas, D. J., and Rossmann, M. G. (1970). *Acta Crystallogr., Sect. B* **26,** 998.
Haber, J. E., and Koshland, D. (1970). *J. Mol. Biol.* **50,** 617.
Hample, A., Labanauskas, M., Connors, P. G., Kirkegard, L., RajBhendary, U. L., Sigler, P. B., and Bock, R. M. (1968). *Science* **162,** 1384.
Hanson, A. W., Applebury, M. L., Colman, J. E., and Wyckoff, H. (1970). *J. Biol. Chem.* **245,** 4975.
Harada, K., and Wolfe, R. G. (1968). *J. Biol. Chem.* **243,** 4123.
Harding, M. M., Hodgkin, D. C., Kennedy, A. F., O'Connor, A., and Weitzmann, P. D. J. (1966). *J. Mol. Biol.* **16,** 212.
Hardman, K. D., and Ainsworth, C. F. (1972). *Biochemistry,* **11,** 4910.
Hardman, K. D., and Ainsworth, C. F. (1973). *Biochemistry* **12,** 4442.

Hardman, K. D., and Ainsworth, C. F. (1976). *Biochemistry* **15,** 1120.
Hardman, K. D., Wood, M. K., Schiffer, M., Edmundson, A. B., and Ainsworth, C. F. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36,** 271.
Harker, D. (1956). *Acta Crystallogr.* **9,** 1.
Hart, R. G. (1961). *Acta Crystallogr.* **14,** 1194.
Harte, R. A., and Rupley, J. A. (1968). *J. Biol. Chem.* **243,** 1663.
Hartley, B. S. (1964). *Nature (London)* **201,** 1284.
Hartley, B. S. (1970). *Philos. Trans. R. Soc. London, Ser. B* **257,** 77.
Hartsuck, J. A., and Lipscomb, W. N. (1971). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 3, p. 1. Academic Press, New York.
Hartsuck, J. A., Ludwig, M. L., Muirhead, H., Steitz, T. A., and Lipscomb, W. N. (1965). *Proc. Natl. Acad. Sci. U.S.A.* **53,** 396.
Haurowitz, F. (1950). "Chemistry and Biology of Proteins," p. 91. Academic Press, New York.
Heidner, E. G., Weber, B. H., and Eisenberg, D. (1971). *Science* **171,** 677.
Henderson, R. (1970). *J. Mol. Biol.* **54,** 341.
Henderson, R., and Moffat, J. K. (1971). *Acta Crystallogr., Sect. B* **27,** 1414.
Henderson, R., Wright, C. S., Hess, G. P., and Blow, D. M. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36,** 63.
Hendrikson, W. A. (1971). *Acta Crystallogr., Sect. B* **27,** 1474.
Hendrikson, W. A., and Karle, J. (1973). *J. Biol. Chem.* **248,** 3327.
Hendrikson, W. A., and Lattman, E. E. (1970). *Acta Crystallogr., Sect. B* **26,** 136.
Hendrikson, W. A., and Love, W. E. (1971). *Nature (London), New Biol.* **232,** 197.
Hendrikson, W. A., Love, W. E., and Murray, G. C. (1968). *J. Mol. Biol.* **33,** 829.
Hermans, J., Jr., and McQueen, J. E., Jr. (1974). *Acta Crystallogr., Sect. A* **30,** 730.
Herriott, J. R., Sieker, L. C., and Jensen, L. H. (1969). *Acta Crystallogr., Sect. A* **25,** S186.
Herriott, J. R., Sieker, L. C., and Jensen, L. H. (1970). *J. Mol. Biol.* **50,** 391.
Herriott, J. R., Watenpaugh, K. D., Sieker, L. C., and Jensen, L. H. (1973). *J. Mol. Biol.* **80,** 423.
Hess, G. P. (1971). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 3, p. 213. Academic Press, New York.
Hill, E., Tsernoglou, D., Webb, L., and Banaszak, L. J. (1972). *J. Mol. Biol.* **72,** 577.
Hodgkin, D., and Riley, R. (1968). *In* "Structural Chemistry and Molecular Biology" (A. Rich and N. Davidson, eds.), p. 15. Freeman, San Francisco, California.
Hol, W. G. J. (1971). Ph.D. Thesis, University of Groningen, Groningen, The Netherlands.
Hollenberg, P. F., Flashner, M., and Coon, M. J. (1971). *J. Biol. Chem.* **246,** 946.
Holmes, K. C., and Blow, D. M. (1965). *Methods Biochem. Anal.* **13,** 113.
Holmgren, A., and Söderberg, B.-O. (1970). *J. Mol. Biol.* **54,** 387.
Hoppe, W. (1957). *Z. Elektrochem.* **61,** 1076.
Hoppe, W. (1959). *Acta Crystallogr.* **12,** 665.
Hoppe, W., and Gassmann, J. (1968). *Acta Crystallogr., Sect. B* **24,** 97.
Hoppe, W., and Paulus, E. F. (1967). *Acta Crystallogr.* **23,** 339.
Hoppe, W., Huber, R., and Gassmann, J. (1963). *Acta Crystallogr.* **16,** A4.
Hornstra, J. (1969). *In* "Crystallographic Computing" (F. R. Ahmed, ed.), p. 103. Munksgaard, Copenhagen.
Huber, R. (1965). *Acta Crystallogr.* **19,** 353.
Huber, R. (1969). *In* "Crystallographic Computing" (F. R. Ahmed, ed.), p. 96. Munksgaard, Copenhagen.

Huber, R., and Kopfmann, G. (1969). *Acta Crystallogr., Sect. A* **25**, 143.
Huber, R., and Steigemann, W. (1974). *FEBS Lett.* **48**, 235.
Huber, R., Formanek, H., and Epp, O. (1969). *Naturwissenschaften* **56**, 362.
Huber, R., Kukla, D., Rühlmann, A., Epp, O., and Formanek, H. (1970). *Naturwissenschaften* **57**, 389.
Huber, R., Epp, O., Steigemann, W., and Formanek, H. (1971a). *Eur. J. Biochem.* **19**, 42.
Huber, R., Kukla, D., Rühlmann, A., and Steigemann, W. (1971b). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 141.
Huber, R., Kukla, D., Rühlmann, A., and Steigemann, W. (1971c). *Proc. Int. Res. Conf. Proteinase Inhibitors, 1st, 1970* p. 56.
Huber, R., Kukla, D., Bode, W., Schwager, P., Bartels, K., Deisenhofer, J., and Steigemann, W. (1974). *J. Mol. Biol.* **89**, 73.
Huestis, W. H., and Raftery, M. A. (1972a). *Biochemistry*, **11**, 1648.
Huestis, W. H., and Raftery, M. A. (1972b). *Proc. Natl. Acad. Sci. U.S.A.* **69**, 1887.
Humphrey, R. L., Avey, H. P., Becka, L. N., Poljak, R. J., Rossi, G., Choi, T. K., and Nisonoff, A. (1969). *J. Mol. Biol.* **43**, 223.
Husain, S. S., and Lowe, G. (1968). *Biochem. J.* **108**, 855.
Inman, J. K., and Bryan, R. F. (1966). *J. Mol. Biol.* **15**, 683.
"International Tables for X-Ray Crystallography." (1952). Vol. 1. (N. F. M. Henry and K. Lonsdale, eds.). Kynoch Press, Birmingham, England.
IUPAC-IUB Commission. (1970). *J. Mol. Biol.* **52**, 1–17.
Jack, A., Weinzierl, J., Kalb, A. J. (1971). *J. Mol. Biol.* **58**, 389.
James, M. N. G., and Smillie, L. B. (1969). *Nature (London)* **224**, 695.
James, R. W. (1948). *Acta Crystallogr.* **1**, 132.
Jardetzky, O., Thielmann, H., Arata, Y., Markley, J. L., and Williams, M. N. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 257.
Johansen, J. T., and Vallee, B. L. (1971). *Proc. Natl. Acad. Sci. U.S.A.* **68**, 2532.
Johansen, J. T., and Vallee, B. L. (1975). *Biochemistry* **14**, 649.
Johnson, L. N., and Waley, S. G. (1967). *J. Mol. Biol.* **29**, 321.
Johnson, L. N., and Wolfenden, R. (1970). *J. Mol. Biol.* **47**, 93.
Joynson, M. A., North, A. C. T., Sarma, V. R., Dickerson, R. E., and Steinrauf, L. K. (1970). *J. Mol. Biol.* **50**, 137.
Jurašek, L., Fackre, D., and Smillie, L. B. (1969). *Biochem. Biophys. Res. Commun.* **37**, 99.
Kakudo, M., Ashida, T., Tanaka, N., Tsukihara, T., and Yamene, T. (1972). *Acta Crystallogr., Sect. A* **28**, S34.
Kalb, A. J., and Levitzki, A. (1968). *Biochem. J.* **109**, 669.
Kannan, K. K., Liljas, A., Waara, I., Bergstén, P.-C., Lövgren, S., Strandberg, B., Bengtsson, J., Carlbom, U., Fridborg, K., Jarup, L., and Petef, M. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 221.
Kannan, K. K., Fridborg, K., Bergstén, P.-C., Liljas, A., Lövgren, S., Petef, M., Strandberg, B., Waara, I., Adler, L., Falkbring, S. O., Gothe, P. O., and Nyman, P. O. (1972). *J. Mol. Biol.* **63**, 601.
Kannan, K. K., Notstrand, B., Fridborg, K., Lövgren, S., Ohlsson, A., and Petef, M. (1975). *Proc. Natl. Acad. Sci. U.S.A.* **72**, 51.
Karle, J. (1964). *Adv. Struct. Res. Diffr. Methods* **1**, 55.
Karle, J. (1966). *Acta Crystallogr.* **21**, 273.
Karle, J., and Hauptman, H. (1956). *Acta Crystallogr.* **9**, 635.
Karle, J., and Karle, I. L. (1966). *Acta Crystallogr.* **21**, 849.
Kartha, G. (1961). *Acta Crystallogr.* **14**, 680.

Kartha, G. (1965). *Acta Crystallogr.* **19**, 883.
Kartha, G. (1969). *In* "Crystallographic Computing" (F. R. Ahmed, ed.), p. 132. Munksgaard, Copenhagen.
Kartha, G., and Parthasarathy, R. (1965a). *Acta Crystallogr.* **18**, 745.
Kartha, G., and Parthasarathy, R. (1965b). *Acta Crystallogr.* **18**, 749.
Kartha, G., Bello, J., and Harker, D. (1967). *Nature (London)* **213**, 862.
Kauzmann, W. (1959). *Adv. Protein Chem.* **14**, 1.
Kendrew, J. C. (1962). *Brookhaven Symp. Biol.* **15**, 216.
Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). *Nature (London)* **181**, 662.
Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., and Davies, D. R. (1960). *Nature (London)* **185**, 422.
Kendrew, J. C., Watson, H. C., Strandberg, B. E., Dickerson, R. E., Phillips, D. C., and Shore, V. C. (1961). *Nature (London)* **190**, 666.
King, M. V. (1954). *Acta Crystallogr.* **7**, 601.
King, M. V., Magdoff, B. S., Adelman, M. B., and Harker, D. (1956). *Acta Crystallogr.* **9**, 460.
King, M. V., Bello, J., Pagnataro, E. H., and Harker, D. (1962). *Acta Crystallogr.* **15**, 144.
Kitaigorodskii, A. (1961). "Organic Chemical Crystallography." Consultants Bureau, New York.
Klotz, I. M., Darnall, D. W., and Langerman, N. R. (1975). *In* "The Proteins" (H. Neurath and R. L. Hill, eds.), 3rd ed., Vol. 1, p. 293. Academic Press, New York.
Kopfmann, G., and Huber, R. (1968). *Acta Crystallogr., Sect. A* **24**, 348.
Korn, A. H., Feairheller, S. H., and Filachione, E. M. (1972). *J. Mol. Biol.* **65**, 525.
Koshland, D. E., Jr. (1970). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 1, p. 342. Academic Press, New York.
Koshland, D. E., Jr., Némethy, G., and Filmer, D. (1966). *Biochemistry* **5**, 365.
Kraut, J. (1968). *J. Mol. Biol.* **35**, 511.
Kraut, J. (1971). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 3, p. 165. Academic Press, New York.
Kraut, J., Sieker, L. C., High, D. F., and Freer, S. T. (1962). *Proc. Natl. Acad. Sci. U.S.A.* **48**, 1417.
Kraut, J., Wright, H. T., Kellerman, M., and Freer, S. T. (1967). *Proc. Natl. Acad. Sci. U.S.A.* **58**, 304.
Kraut, J., Strahs, G., and Freer, S. T. (1968). *In* "Structural Chemistry and Molecular Biology" (A. Rich and N. Davidson, eds.), p. 55. Freeman, San Francisco, California.
Kraut, J., Robertus, J. D., Birktoft, J. J., Alden, R. A., Wilcox, P. E., and Powers, J. C. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 117.
Kretsinger, R. H. (1968). *J. Mol. Biol.* **38**, 141.
Kretsinger, R. H. (1972). *Nature (London), New Biol.* **240**, 85.
Kretsinger, R. H., and Nockolds, C. E. (1973). *J. Biol. Chem.* **248**, 3313.
Kretsinger, R. H., Hagihara, B., and Tsugita, A. (1970). *Biochim. Biophys. Acta* **200**, 421.
Kretsinger, R. H., Dangelat, D., and Bryan, R. F. (1971a). *J. Mol. Biol.* **59**, 213.
Kretsinger, R. H., Nockolds, C. E., Coffee, C. J., and Bradshaw, R. A. (1971b). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 217.
Krieger, M., Kay, L. M., and Stroud, R. M. (1974). *J. Mol. Biol.* **83**, 209.
Kuntz, I. D., Jr. (1972). *J. Am. Chem. Soc.* **94**, 4009.
Kuntz, I. D., Jr., Brassfield, T. S., Law, G. D., and Purcell, G. V. (1969). *Science* **163**, 1329.

Lagerkvist, U., Rymo, L., Lindquist, O., and Andersson, E. (1972). *J. Biol. Chem.* **247**, 3897.
Latt, S. A., Holmquist, B., and Vallee, B. L. (1969). *Biochem. Biophys. Res. Commun.* **37**, 333.
Lattman, E. E. (1972). *Acta Crystallogr., Sect. B* **28**, 1065.
Lattman, E. E., and Love, W. E. (1970). *Acta Crystallogr., Sect. B* **26**, 1854.
Lattman, E. E., Nockolds, C. E., Kretsinger, R. H., and Love, W. E. (1971). *J. Mol. Biol.* **60**, 271.
Leach, S. J., Némethy, G., and Scheraga, H. A. (1966). *Biopolymers* **4**, 369.
Leberman, R., Smiley, I. E., Haas, D. J., and Rossmann, M. G. (1969). *J. Mol. Biol.* **46**, 217.
Lee, B., and Richards, F. M. (1971). *J. Mol. Biol.* **55**, 379.
Levinthal, C. (1966). *Sci. Am.* **214**, 42.
Levitt, M. (1974). *J. Mol. Biol.* **82**, 393.
Levitt, M., and Lifson, S. (1969). *J. Mol. Biol.* **46**, 269.
Lewis, P. N., Momany, F. A., and Scheraga, H. A. (1971). *Proc. Natl. Acad. Sci. U.S.A.* **68**, 2293.
Light, A., Frater, R., Kimmel, J. R., and Smith, E. L. (1964). *Proc. Natl. Acad. Sci. U.S.A.* **52**, 1276.
Liljas, A., and Rossmann, M. G. (1974). *Annu. Rev. Biochem.* **43**, 475.
Liljas, A., Kannan, K. K., Bergstén, P.-C., Waara, I., Fridborg, K., Strandberg, B., Carlbom, U., Jarup, L., Lövgren, S., and Petef, M. (1972). *Nature (London), New Biol.* **235**, 131.
Lipscomb, W. N. (1973). *Proc. Natl. Acad. Sci. U.S.A.* **70**, 3797.
Lipscomb, W. N., Coppola, J. C., Hartsuck, J. A., Ludwig, M. L., Muirhead, H., Searl, J., and Steitz, T. A. (1966). *J. Mol. Biol.* **19**, 423.
Lipscomb, W. N., Hartsuck, J. A., Reeke, G. N., Jr., Quiocho, F. A., Bethge, P. H., Ludwig, M. L., Steitz, T. A., Muirhead, H., and Coppola, J. C. (1968). *Brookhaven Symp. Biol.* **21**, 24.
Lipscomb, W. N., Hartsuck, J. A., Quiocho, F. A., and Reeke, G. N., Jr. (1969). *Proc. Natl. Acad. Sci. U.S.A.* **64**, 28.
Lipscomb, W. N., Reeke, G. N., Hartsuck, J. A., Quiocho, F. A., and Bethge, P. H. (1970). *Philos. Trans. R. Soc. London, Ser. B* **257**, 177.
Lipson, H., and Cochran, W. (1966). "The Determination of Crystal Structures." Bell, London.
Longley, W. (1967). *J. Mol. Biol.* **30**, 323.
Love, W. E. (1957). *Biochim. Biophys. Acta* **23**, 465.
Love, W. E., Klock, P. A., Lattman, E. E., Padlan, E. A., Ward, K. B., Jr., and Hendrikson, W. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 349.
Lovenberg, W., and Sobel, B. E. (1965). *Proc. Natl. Acad. Sci. U.S.A.* **54**, 193.
Low, B. W., and Richards, F. M. (1952). *J. Am. Chem. Soc.* **74**, 1660.
Low, B. W., Chen, C. C. H., Berger, J. E., Singman, L., and Pletcher, J. F. (1966). *Proc. Natl. Acad. Sci. U.S.A.* **56**, 1746.
Low, B. W., Potter, R., Jackson, R. B., Tamiya, N., and Sato, S. (1971). *J. Biol. Chem.* **246**, 4366.
Ludwig, M. L., Hartsuck, J. A., Steitz, T. A., Muirhead, H., Coppola, J. C., Reeke, G. N., Jr., and Lipscomb, W. N. (1967). *Proc. Natl. Acad. Sci. U.S.A.* **57**, 511.
Ludwig, M. L., Andersen, R. D., Mayhew, S. G., and Massey, V. (1969). *J. Biol. Chem.* **244**, 6047.
Ludwig, M. L., Andersen, R. D., Apgar, P. A., Burnett, R. M., LeQuesne, M. E., and Mahew, S. G. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 369.

Lundberg, B. (1965). *Acta Crystallogr.* **18,** 576.
Luzzati, P. V. (1953). *Acta Crystallogr.* **6,** 142.
MacIntyre, W. M., Pereira, H. G., and Russell, W. C. (1969). *Nature (London)* **222,** 1165.
McLachlan, A. D. (1972a). *J. Mol. Biol.* **64,** 417.
McLachlan, A. D. (1972b). *Nature (London), New Biol.* **240,** 83.
McLachlan, A. D., and Shotton, D. M. (1971). *Nature (London), New Biol.* **229,** 202.
McPherson, A., Jr., and Rich, A. (1972). *J. Biol. Chem.* **247,** 1334.
Magdoff, B. S., and Crick, F. H. C. (1955). *Acta Crystallogr.* **8,** 461.
Magdoff-Fairchild, B., Lovell, F. M., and Low, B. W. (1969). *J. Biol. Chem.* **244,** 3497.
Main, P., and Rossmann, M. G. (1966). *Acta Crystallogr.* **21,** 67.
Margoliash, E., Fitch, W. M., and Dickerson, R. E. (1968). *Brookhaven Symp. Biol.* **21,** 259.
Mathews, F. S. (1967). *Fed. Proc., Fed. Am. Soc. Exp. Biol.* **26,** 831.
Mathews, F. S., Argos, P., and Levine, M. (1971a). *Cold Spring Harbor Symp. Quant. Biol.* **36,** 387.
Mathews, F. S., Levine, M., and Argos, P. (1971b). *Nature (London), New Biol.* **233,** 15.
Mathews, F. S., Levine, M., and Argos, P. (1972). *J. Mol. Biol.* **64,** 449.
Matsubara, H. (1967). *In* "Molecular Mechanisms of Temperature Adaption" (C. L. Prosser, ed.), Publ. No. 84, p. 283. Am. Assoc. Adv. Sci., Washington, D. C.
Matthews, B. W. (1966a). *Acta Crystallogr.* **20,** 82.
Matthews, B. W. (1966b). *Acta Crystallogr.* **20,** 230.
Matthews, B. W. (1968). *J. Mol. Biol.* **33,** 491.
Matthews, B. W. (1969). *In* "Crystallographic Computing" (F. R. Ahmed, ed.), p. 146. Munksgaard, Copenhagen.
Matthews, B. W. (1972). *Macromolecules* **5,** 818.
Matthews, B. W. (1974). *J. Mol. Biol.* **82,** 513.
Matthews, B. W. (1975). *Biochim. Biophys. Acta* **405,** 442.
Matthews, B. W. (1976). *Annu. Rev. Phys. Chem.* (in press).
Matthews, B. W., and Bernhard, S. A. (1973). *Annu. Rev. Biophys. Bioeng.* **2,** 257.
Matthews, B. W., and Remington, S. J. (1974). *Proc. Natl. Acad. Sci. U.S.A.* **71,** 4178.
Matthews, B. W., and Weaver, L. H. (1974). *Biochemistry* **13,** 1719.
Matthews, B. W., Sigler, P. B., Henderson, R., and Blow, D. M. (1967). *Nature (London)* **214,** 652.
Matthews, B. W., Cohen, G. H., Silverton, E. W., Braxton, H., and Davies, D. R. (1968). *J. Mol. Biol.* **36,** 179.
Matthews, B. W., Jansonius, J. N., Colman, P. M., Schoenborn, B. P., and Dupourque, D. (1972a). *Nature (London), New Biol.* **238,** 37.
Matthews, B. W., Colman, P. M., Jansonius, J. N., Titani, K., Walsh, K. A., and Neurath, H. (1972b). *Nature (London), New Biol.* **238,** 41.
Matthews, B. W., Klopfenstein, C. E., and Colman, P. M. (1972c). *J. Phys. E* **5,** 353.
Matthews, B. W., Weaver, L. H., and Kester, W. (1974). *J. Biol. Chem.* **249,** 8030.
Milledge, H. J. (1963). *Acta Crystallogr.* **16,** 72.
Minor, T. C., Milch, J. R., and Reynolds, G. T. (1974). *J. Appl. Crystallogr.* **7,** 323.
Mitsui, Y., and Wyckoff, H. W. (1975). *J. Mol. Biol.* **94,** 17.
Moews, P. C., and Bunn, C. W. (1970). *J. Mol. Biol.* **54,** 395.
Moews, P. C., and Bunn, C. W. (1971). *Acta Crystallogr., Sect. B* **27,** 1780.
Moews, P. C., and Kretsinger, R. H. (1975). *J. Mol. Biol.* **91,** 201.
Monod, J., Wyman, J., and Changeux, J.-P. (1965). *J. Mol. Biol.* **12,** 88.
Morihara, K., and Oka, T. (1968). *Biochem. Biophys. Res. Commun.* **30,** 625.
Morita, Y. (1972). *Acta Crystallogr., Sect. A* **28,** S52.
Morita, Y., and Ida, S. (1972). *J. Mol. Biol.* **71,** 807.

Muirhead, H., and Greer, J. (1970). *Nature (London)* **228**, 516.
Muirhead, H., and Perutz, M. F. (1963). *Nature (London)* **199**, 633.
Muirhead, H., Cox, J. M., Mazzarella, L., and Perutz, M. F. (1967). *J. Mol. Biol.* **28**, 117.
Némethy, G., and Printz, M. P. (1972). *Macromolecules* **5**, 755.
Némethy, G., Phillips, D. C., Leach, S. J., and Scheraga, H. A. (1967). *Nature (London)* **214**, 363.
Neurath, H., Bradshaw, R. A., and Arnon, R. (1969). "International Symposium on Structure-Function Relationships of Proteolytic Enzymes." Munksgaard, Copenhagen.
Nobbs, C. L., Watson, H. C., and Kendrew, J. C. (1966). *Nature (London)* **209**, 339.
Nockolds, C. E., and Kretsinger, R. H. (1970). *J. Phys. E* **3**, 842.
Nockolds, C. E., Kretsinger, R. H., Coffee, C. J., and Bradshaw, R. A. (1972). *Proc. Natl. Acad. Sci. U.S.A.* **69**, 581.
Nordman, C. E. (1972). *Acta Crystallogr., Sect. A* **28**, 134.
Nordman, C. E., and Nakatsu, K. (1963). *J. Am. Chem. Soc.* **85**, 353.
North, A. C. T. (1959). *Acta Crystallogr.* **12**, 512.
North, A. C. T. (1965). *Acta Crystallogr.* **18**, 212.
North, A. C. T., and Phillips, D. C. (1969). *Prog. Biophys. Mol. Biol.* **19**, 1.
North, A. C. T., Phillips, D. C., and Mathews, F. S. (1968). *Acta Crystallogr., Sect. A* **24**, 351.
Nunes, A. C., Nathans, R., and Schoenborn, B. P. (1971). *Acta Crystallogr., Sect. A* **27**, 284.
Ogawa, S., and McConnell, H. (1967). *Proc. Natl. Acad. Sci. U.S.A.* **58**, 19.
Ohlsson, I., Nordström, B., and Brändén, C.-I. (1974). *J. Mol. Biol.* **89**, 339.
Okaya, Y., and Pepinsky, R. (1960). *In* "Computing Methods and the Phase Problem in X-Ray Crystal Analysis" (R. Pepinsky, J. M. Robertson, and J. C. Speakman, eds.), p. 266. Pergamon, Oxford.
Olaitan, S. A., DeLange, R. J., and Smith, E. L. (1968). *J. Biol. Chem.* **243**, 5296.
Olson, J. M., Koenig, D. F., and Ledbetter, M. C. (1969). *Arch. Biochem. Biophys.* **129**, 42.
Olson, M. O. J., Nagabhushan, N., Dzwiniel, M., Smillie, L. B., and Whitaker, D. R. (1970). *Nature (London)* **228**, 438.
Osserman, E. F., Cole, S. J., Swan, I. D. A., and Blake, C. C. F. (1969). *J. Mol. Biol.* **46**, 211.
Padlan, E. A., and Love, W. E. (1968). *Nature (London)* **220**, 376.
Padlan, E. A., and Love, W. E. (1974). *J. Biol. Chem.* **249**, 4067.
Patterson, A. L. (1934). *Phys. Rev.* **46**, 372.
Pauling, L. (1935). *Proc. Natl. Acad. Sci. U.S.A.* **21**, 186.
Pauling, L., and Corey, R. B. (1951). *Proc. Natl. Acad. Sci. U.S.A.* **37**, 729.
Pauling, L., and Coryell, C. D. (1936). *Proc. Natl. Acad. Sci. U.S.A.* **22**, 210.
Peking Insulin Structure Group. (1974). *Sci. Sin.* **17**, 752.
Perutz, M. F. (1946). *Trans. Faraday Soc.* **42B**, 187.
Perutz, M. F. (1969a). *Eur. J. Biochem.* **8**, 455.
Perutz, M. F. (1969b). *Proc. R. Soc. London, Ser. B* **173**, 113.
Perutz, M. F. (1970a). *Acta Crystallogr., Sect. A* **26**, 184.
Perutz, M. F. (1970b). *Nature (London)* **228**, 726.
Perutz, M. F. (1970c). *Nature (London)* **228**, 734.
Perutz, M. F. (1972). *Nature (London)* **237**, 495.
Perutz, M. F. (1975). *Nature (London)* **255**, 256.
Perutz, M. F., and Ten Eyck, L. F. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 295.

Perutz, M. F., Rossmann, M. G., Cullis, A., Muirhead, H., Will, G., and North, A. C. T. (1960). *Nature (London)* **185,** 416.
Perutz, M. F., Muirhead, H., Cox, J. M., Goaman, L. C. G., Mathews, F. S., McGandy, E. L., and Webb, L. E. (1968a). *Nature (London)* **219,** 29.
Perutz, M. F., Muirhead, H., Cox, J. M., and Goaman, L. C. G. (1968b). *Nature (London)* **219,** 131.
Perutz, M. F., Pulsinelli, P. D., and Ranney, H. M. (1972). *Nature (London), New Biol.* **237,** 259.
Phillips, D. C. (1964). *J. Sci. Instrum.* **41,** 123.
Phillips, D. C. (1966). *Adv. Struc. Res. Diffr. Methods* **2,** 75.
Phillips, D. C. (1970). *Biochem. Soc. Symp.* **30,** 11.
Phillips, D. C. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36,** 589.
Poljak, R. J., Amzel, L. M., Avey, H. P., Becka, L. N., and Nisonoff, A. (1972). *Nature (London),* **235,** 137.
Poljak, R. J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizerlackerly, R. P., and Saul, F. (1973). *Proc. Natl. Acad. Sci. U.S.A.* **70,** 3305.
Poljak, R. J., Amzel, L. M., Chen, B. L., Phizerlackerly, R. P., and Saul, F. (1974). *Proc. Natl. Acad. Sci. U.S.A.* **71,** 3440.
Popov, E. P., Lipkind, G. M., Azkhipova, S. F., and Dashevskii, U. S. (1968). *Mol. Biol.* **2,** 622.
Prothero, J. W., and Rossmann, M. G. (1964). *Acta Crystallogr.* **17,** 768.
Ptitsyn, O. B. (1969). *J. Mol. Biol.* **42,** 501.
Pullman, B. (1971). *Int. J. Quantum Chem.* **4,** 319.
Quiocho, F. A., and Lipscomb, W. N. (1971). *Adv. Protein Chem.* **25,** 1.
Quiocho, F. A., and Richards, F. M. (1964). *Proc. Natl. Acad. Sci. U.S.A.* **52,** 833.
Quiocho, F. A., and Richards, F. M. (1966). *Biochemistry* **5,** 4062.
Quiocho, F. A., McMurray, C. H., and Lipscomb, W. N. (1972). *Proc. Natl. Acad. Sci. U.S.A.* **69,** 2850.
Raiz, V. Sh., and Andreeva, N. S. (1970). *Sov. Phys.—Crystallogr. (Engl. Transl.)* **15,** 206.
Ramachandran, G. N. (1968). *Biopolymers* **6,** 1494.
Ramachandran, G. N., and Raman, S. (1956). *Curr. Sci.* **25,** 348.
Ramachandran, G. N., Ramakrishnan, C., and Sasiekharan, V. (1963). *J. Mol. Biol.* **7,** 95.
Ramachandran, G. N., Venkatachalam, C. M., and Krimm, S. (1966). *Biophys. J.* **6,** 849.
Ramakrishnan, C., and Ramachandran, G. N. (1965). *Biophys. J.* **5,** 909.
Ramaseshan, S. (1966). *Curr. Sci.* **35,** 87.
Reeke, G. N., Jr., and Lipscomb, W. N. (1969). *Acta Crystallogr., Sect. B* **25,** 2614.
Reeke, G. N., Jr., Hartsuck, J. A., Ludwig, M. L., Quiocho, F. A., Steitz, T. A., and Lipscomb, W. N. (1967). *Proc. Natl. Acad. Sci. U.S.A.* **58,** 2220.
Reeke, G. N., Jr., Becker, J. W., and Quiocho, F. A. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36,** 277.
Reeke, G. N., Jr., Becker, J. W., and Edelman, G. M. (1975). *J. Biol. Chem.* **250,** 1525.
Richards, F. M. (1968). *J. Mol. Biol.* **37,** 225.
Richards, F. M., and Knowles, J. R. (1968). *J. Mol. Biol.* **37,** 231.
Richards, F. M., and Wyckoff, H. W. (1971). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 4, p. 647. Academic Press, New York.
Richardson, J. S., Thomas, K. A., Rubin, B. H., and Richardson, D. C. (1975). *Proc. Natl. Acad. Sci. U.S.A.* **72,** 1349.
Robertus, J. D., Alden, R. A., and Kraut, J. (1971). *Biochem. Biophys. Res. Commun.* **42,** 334.

Robertus, J. D., Alden, R. A., Birktoft, J. J., Kraut, J., Powers, J. C., and Wilcox, P. E. (1972a). *Biochemistry* **11**, 2439.
Robertus, J. D., Kraut, J., Alden, R. A., and Birktoft, J. J. (1972b). *Biochemistry* **11**, 4293.
Rossi, G. L., and Bernhard, S. A. (1970). *J. Mol. Biol.* **49**, 85.
Rossmann, M. G. (1960). *Acta Crystallogr.* **13**, 221.
Rossmann, M. G. (1961). *Acta Crystallogr.* **14**, 383.
Rossmann, M. G. (1974). *New Sci.* **61**, 266.
Rossmann, M. G., and Blow, D. M. (1961). *Acta Crystallogr.* **14**, 641.
Rossmann, M. G., and Blow, D. M. (1962). *Acta Crystallogr.* **15**, 24.
Rossmann, M. G., and Blow, D. M. (1963). *Acta Crystallogr.* **16**, 39.
Rossmann, M. G., and Blow, D. M. (1964). *Acta Crystallogr.* **17**, 1474.
Rossmann, M. G., and Labaw, L. W. (1967). *J. Mol. Biol.* **29**, 315.
Rossmann, M. G., Blow, D. M., Harding, M. M., and Coller, E. (1964). *Acta Crystallogr.* **17**, 338.
Rossmann, M. G., Jeffery, B. A., Main, P., and Warren, S. (1967). *Proc. Natl. Acad. Sci. U.S.A.* **57**, 615.
Rossmann, M. G., Adams, M. J., Buehner, M., Ford, G. C., Hackert, M. L., Lentz, P. J., Jr., McPherson, A., Jr., Schevitz, R. W., and Smiley, I. E. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 179.
Rossmann, M. G., Ford, G. C., Watson, H. C., and Banaszak, L. J. (1972). *J. Mol. Biol.* **64**, 237.
Rossmann, M. G., Moras, D., and Olsen, K. W. (1974). *Nature (London)* **250**, 194.
Rubin, B., and Richardson, J. S. (1972). *Biopolymers* **11**, 2381.
Rühlmann, A., Schramm, H. J., Kukla, D., and Huber, R. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 148.
Rühlmann, A., Kukla, D., Schwager, P., Bartels, K., and Huber, R. (1973). *J. Mol. Biol.* **77**, 417.
Rupley, J. A. (1968). *J. Mol. Biol.* **35**, 455.
Rupley, J. A. (1969). *In* "Structure and Stability of Biological Macromolecules" (S. N. Timasheff and G. D. Fasman, eds.), p. 291. Dekker, New York.
Rupley, J. A., and Bruice, T. C. (1968). *J. Mol. Biol.* **37**, 521.
Rymo, L., Lagerkvist, U., and Wonacott, A. J. (1970). *J. Biol. Chem.* **245**, 4308.
Sakabe, N., Sakabe, K., and Katayama, C. (1972). *Acta Crystallogr., Sect. A* **28**, S34.
Salemme, F. R., Freer, S. T., Xuong, Ng. H., Alden, R. A., and Kraut, J. (1973). *J. Biol. Chem.* **248**, 3910.
Sarma, V. R., Silverton, E. W., Davies, D. R., and Terry, W. D. (1971). *J. Biol. Chem.* **246**, 3753.
Sasada, Y. (1964). *Acta Crystallogr.* **17**, 611.
Sayre, D. (1952). *Acta Crystallogr.* **5**, 843.
Sayre, D. (1972). *Acta Crystallogr., Sect. A* **28**, 210.
Sayre, D. (1974). *Acta Crystallogr., Sect. A* **30**, 180.
Schiffer, M., Hartman, K. D., Wood, M. K., Edmundson, A. B., Hook, M. E., and Ely, K. R. (1970). *J. Biol. Chem.* **245**, 728.
Schiffer, M., Girling, R. L., Ely, K. R., and Edmundson, A. B. (1973). *Biochemistry* **12**, 4620.
Schilling, J. W. (1969). *In* "Crystallographic Computing" (F. R. Ahmed, ed.), p. 115. Munksgaard, Copenhagen.
Schmid, M. F., and Herriott, J. R. (1976). *J. Mol. Biol.* **103**, 175.
Schoenborn, B. P. (1965). *Nature (London)* **208**, 760.
Schoenborn, B. P. (1969). *Nature (London)* **224**, 143.

Schoenborn, B. P. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36,** 569.
Schoenborn, B. P., and Nunes, A. C. (1972). *Annu. Rev. Biophys. Bioeng.* **1,** 529.
Schoenborn, B. P., Watson, H. C., and Kendrew, J. C. (1965). *Nature (London)* **207,** 28.
Schrauber, H. (1966). *Acta Crystallogr.* **21,** A222.
Schulz, G. E., and Schirmer, R. H. (1974). *Nature (London)* **250,** 142.
Schulz, G. E., Elzinga, M., and Schirmer, R. H. (1974a). *Nature (London)* **250,** 120.
Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, O. B., Kabat, E. A., Wu, T. T., Levitt, M., Robson, B., and Nagano, K. (1974b). *Nature (London)* **250,** 140.
Scouloudi, H. (1969). *J. Mol. Biol.* **40,** 353.
Segal, D. M., Cohen, G. C., Davies, D. R., Powers, J. C., and Wilcox, P. E. (1971a). *Cold Spring Harbor Symp. Quant. Biol.* **36,** 85.
Segal, D. M., Powers, J. C., Cohen, G. C., Davies, D. R., and Wilcox, P. E. (1971b). *Biochemistry* **10,** 3728.
Segal, D. M., Padlan, E. A., Cohen, S. H., Rudikoff, S., Potter, M., and Davies, D. R. (1974). *Proc. Natl. Acad. Sci. U.S.A.* **71,** 4298.
Shotton, D. M., and Watson, H. C. (1970). *Nature (London)* **225,** 811.
Shotton, D. M., White, N. J., and Watson, H. C. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36,** 91.
Sieker, L. C., and Jensen, L. H. (1965). *Biochem. Biophys. Res. Commun.* **20,** 33.
Sieker, L. C., Adman, E., and Jensen, L. H. (1972). *Nature (London)* **235,** 40.
Sigler, P. B. (1970). *Biochemistry* **9,** 3609.
Sigler, P. B., and Blow, D. M. (1965). *J. Mol. Biol.* **14,** 640.
Sigler, P. B., and Skinner, H. C. W. (1963). *Biochem. Biophys. Res. Commun.* **13,** 236.
Sigler, P. B., Skinner, H. C. W., Coulter, C. L., Kallos, J., Braxton, H., and Davies, D. R. (1964). *Proc. Natl. Acad. Sci. U.S.A.* **51,** 1146.
Sigler, P. B., Jeffery, B. A., Matthews, B. W., and Blow, D. M. (1966). *J. Mol. Biol.* **15,** 175.
Sigler, P. B., Blow, D. M., Matthews, B. W., and Henderson, R. (1968). *J. Mol. Biol.* **35,** 143.
Sim, G. A. (1960). *Acta Crystallogr.* **13,** 511.
Singh, A. K., and Ramaseshan, S. (1966). *Acta Crystallogr.* **21,** 279.
Sluyterman, L. A. AE., and De Graaf, M. J. M. (1969). *Biochim. Biophys. Acta* **171,** 277.
Sluyterman, L. A. AE., and Wolthers, B. G. (1969). *Proc. K. Ned. Akad. Wet., Ser. B* **72,** 14.
Smillie, L. B., and Whitaker, D. R. (1967). *J. Am. Chem. Soc.* **89,** 3350.
Smit, J. D. G., Ploegman, J. H., Pierrot, M., Kalk, K. H., Jansonius, J. N., and Drenth, J. (1974). *Isr. J. Chem.* **12,** 287.
Smith, E. L. (1970). *In* "The Enzymes" (P. D. Boyer, ed.), 3rd ed., Vol. 1, p. 267. Academic Press, New York.
Solomon, A., McLaughlin, C. L., Wei, C. H., and Einstein, J. R. (1970). *J. Biol. Chem.* **245,** 5289.
Sparks, R. A. (1961). *Abstr., Am. Crystallogr. Assoc.* p. 37.
Srinivasan, R. (1968). *Z. Kristallogr. Kristallgeom., Kristallphys., Kristallchem.* **126,** 175.
Steinrauf, L. K. (1963). *Acta Crystallogr.* **16,** 317.
Steitz, T. A. (1968). *Acta Crystallogr., Sect. B* **24,** 504.
Steitz, T. A. (1971). *J. Mol. Biol.* **61,** 695.
Steitz, T. A., Wiley, D. C., and Lipscomb, W. N. (1967). *Proc. Natl. Acad. Sci. U.S.A.* **58,** 1859.

Steitz, T. A., Henderson, R., and Blow, D. M. (1969). *J. Mol. Biol.* **46**, 337.
Storm, D. R., and Koshland, D. E., Jr. (1970). *Proc. Natl. Acad. Sci. U.S.A.* **66**, 445.
Stout, G. H., and Jensen, L. H. (1968). "X-Ray Structure Determination." Macmillan, New York.
Strahs, G., and Kraut, J. (1968). *J. Mol. Biol.* **35**, 503.
Stroud, R. M., Kay, L. M., and Dickerson, R. E. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 125.
Stroud, R. M., Kay, L. M., and Dickerson, R. E. (1974). *J. Mol. Biol.* **83**, 185.
Stryer, L., Kendrew, J. C., and Watson, H. C. (1964). *J. Mol. Biol.* **8**, 96.
Sweet, R. M., Wright, H. T., Janin, J., Chothia, C. H., and Blow, D. M. (1974). *Biochemistry* **13**, 4212.
Takano, T., Swanson, R., Kallai, O. B., and Dickerson, R. E. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 397.
Takano, T., Kallai, O. B., Swanson, R., and Dickerson, R. E. (1973). *J. Biol. Chem.* **248**, 5234.
Terry, W. D., Matthews, B. W., and Davies, D. R. (1968). *Nature (London)* **220**, 239.
Theorell, H., Chance, B., and Yonetani, T. (1966). *J. Mol. Biol.* **17**, 513.
Tilander, B., Strandberg, B., and Fridborg, K. (1965). *J. Mol. Biol.* **12**, 740.
Timkovich, R., and Dickerson, R. E. (1973). *J. Mol. Biol.* **79**, 39.
Titani, K., Hermodson, M. A., Ericsson, L. H., Walsh, K. A., and Neurath, H. (1972). *Nature (London), New Biol.* **238**, 35.
Tollin, P. (1966). *Acta Crystallogr.* **21**, 613.
Tollin, P. (1969). *J. Mol. Biol.* **45**, 481.
Tollin, P., and Cochran, W. (1964). *Acta Crystallogr.* **17**, 1322.
Tollin, P., and Rossmann, M. G. (1966). *Acta Crystallogr.* **21**, 872.
Tollin, P., Main, P., and Rossmann, M. G. (1966). *Acta Crystallogr.* **20**, 404.
Traub, W., and Hirshfeld, F. L. (1960). *Acta Crystallogr.* **13**, 753.
Tsernoglou, D., Hill, E., and Banaszak, L. J. (1972). *J. Mol. Biol.* **69**, 75.
Tufty, R. M., and Kretsinger, R. H. (1975). *Science* **187**, 167.
Tulinsky, A., Vandlen, R. L., Morimoto, C. N., Mani, N. V., and Wright, L. H. (1973). *Biochemistry* **12**, 4185.
Vainshtein, B. K., Barynin, V. V., and Gurskaya, G. V. (1969). *Acta Crystallogr., Sect. A* **25**, 5181.
Vainshtein, B. K., Harutyunyan, E. H., Kuranova, I. P., Borisov, V. V., Sosfenov, N. I., Pavlovsky, A. G., Grebenko, A. I., and Konareva, N. V. (1975). *Nature (London)* **254**, 163.
Valentine, R. C. (1964). *Bacteriol. Rev.* **28**, 497.
Vandlen, R. L., Ersfeld, D. L., Tulinsky, A., and Wood, W. A. (1973). *J. Biol. Chem.* **248**, 2251.
Venkatachalam, C. M. (1968). *Biopolymers* **6**, 1425.
Waller, J.-P., Risler, J.-L., Monteilhet, C., and Zelwer, C. (1971). *FEBS Lett.* **16**, 186.
Watenpaugh, K. D., Sieker, L. C., Herriott, J. R., and Jensen, L. H. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 359.
Watenpaugh, K. D., Sieker, L. C., Jensen, L. H., LeGall, J., and Dubourdieu, M. (1972). *Proc. Natl. Acad. Sci. U.S.A.* **69**, 3185.
Watenpaugh, K. D., Sieker, L. C., Herriott, J. R., and Jensen, L. H. (1973a). *Acta Crystallogr., Sect. B* **29**, 943.
Watenpaugh, K. D., Sieker, L. C., and Jensen, L. H. (1973b). *Proc. Natl. Acad. Sci. U.S.A.* **70**, 3857.
Watson, H. C. (1969). *Prog. Stereochem.* **4**, 299.
Watson, H. C., and Banaszak, L. J. (1964). *Nature (London)* **204**, 918.

Watson, H. C., Kendrew, J. C., Coulter, C. L., Brändén, C.-I., Phillips, D. C., and Blake, C. C. F. (1963). *Acta Crystallogr.* **16**, A81.
Watson, H. C., Shotton, D. M., Cox, J. M., and Muirhead, H. (1970). *Nature (London)* **225**, 806.
Watson, H. C., Duée, E., and Mercer, W. D. (1972). *Nature (London), New Biol.* **240**, 130.
Webb, L. E., Hill, E. J., and Banaszak, L. J. (1973). *Biochemistry* **12**, 5101.
Weber, B. H., and Goodkin, P. E. (1970). *Arch. Biochem. Biophys.* **141**, 489.
Weber, K. (1968). *Nature (London)* **218**, 1116.
Weeds, A. G., and McLachlan, A. D. (1974). *Nature (London)* **252**, 646.
Weinzierl, J. E., Eisenberg, D., and Dickerson, R. E. (1969). *Acta Crystallogr., Sect. B* **25**, 380.
Wendell, P. L., Bryant, T. N., and Watson, H. C. (1972). *Nature (London), New Biol.* **240**, 134.
Werner, P.-E. (1969). *Ark. Kemi* **31**, 305.
Werner, P.-E. (1970). *Acta Crystallogr., Sect. A* **26**, 489.
Weyl, H. (1952). "Symmetry." Princeton Univ. Press, Princeton, New Jersey.
Wiley, D. C., and Lipscomb, W. N. (1968). *Nature (London)* **218**, 1119.
Wiley, D. C., Evans, D. R., Warren, S. G., McMurray, G. H., Edwards, B. F. P., Franks, W. A., and Lipscomb, W. N. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 285.
Wilson, A. J. C. (1950). *Acta Crystallogr.* **3**, 397.
Winkler, F. K., and Dunitz, J. D. (1971). *J. Mol. Biol.* **59**, 169.
Wong, C.-H., Chang, T. W., Lee, T. J., and Yang, C. C. (1972). *J. Biol. Chem.* **247**, 608.
Woolfson, M. M. (1956). *Acta Crystallogr.* **9**, 804.
Wright, C. S. (1972). *J. Mol. Biol.* **67**, 151.
Wright, C. S., Alden, R. A., and Kraut, J. (1969). *Nature (London)* **221**, 235.
Wright, C. S., Sato, T., Nagata, Y., McMillan, J., Langridge, R., and Burger, M. M. (1972). *Acta Crystallogr., Sect. A* **28**, S35.
Wright, H. T., Kraut, J., and Wilcox, P. E. (1968). *J. Mol. Biol.* **37**, 363.
Wyckoff, H. W., Doscher, M., Tsernoglou, D., Inagami, T., Johnson, L. N., Hardman, K. D., Allewell, N. M., Kelly, D. M., and Richards, F. M. (1967a). *J. Mol. Biol.* **27**, 563.
Wyckoff, H. W., Hardman, K. D., Allewell, N. M., Inagami, T., Johnson, L. N., and Richards, F. M. (1967b). *J. Biol. Chem.* **242**, 3984.
Wyckoff, H. W., Tsernoglou, D., Hanson, A. W., Knox, J. R., Lee, B., and Richards, F. M. (1970). *J. Biol. Chem.* **245**, 305.
Xuong, Ng. H. (1969). *J. Phys. E* **2**, 485.
Xuong, Ng. H., and Freer, S. T. (1971). *Acta Crystallogr., Sect. B* **27**, 2380.
Xuong, Ng. H., Kraut, J., Seely, O., Freer, S. T., and Wright, C. S. (1968). *Acta Crystallogr., Sect. B* **24**, 289.
Yankeelov, J. A., Jr., and Coggins, J. R. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 585.
Zeppezauer, M., Söderberg, B. O., and Brändén, C.-I. (1967). *Acta Chem. Scand.* **21**, 1099.
Zeppezauer, M., Eklund, H., and Zeppezauer, E. S. (1968). *Arch. Biochem. Biophys.* **126**, 564.