Purification and Analysis of the Trichohyalin Gene:

An Examination of the Role of Trichohyalin in the Inner Root Sheath

Michael James Fietz B.Sc. (Hons.)

A thesis submitted for the degree of Doctor of Philosophy in the University of Adelaide

Department of Biochemistry

January, 1991

for Janet

Table of Contents

Thesis Summary

This thesis reports on a study of the follicle protein trichohyalin. Trichohyalin is produced in the developing cells of the follicle inner root sheath and medulla and is stored in non-membrane bound granules. It is the substrate for two post-translational modifications and is finally incorporated into the hardened structures of both tissues. Within the inner root sheath the hardened structure involves a closely-packed array of filaments, which are of intermediate filament size, aligned parallel to the direction of hair growth. It is presently uncertain whether trichohyalin forms the filaments of the hardened inner root sheath or the matrix into which the filaments are embedded. The major aim of the work reported in this thesis was to gain an increased understanding of the role of trichohyalin by determining the complete trichohyalin amino acid sequence.

The initial aspects described in this thesis were involved in the production of an improved purification procedure for trichohyalin. The incorporated changes gave an increased recovery of extracted trichohyalin which allowed sufficient trichohyalin to be purified for proteolytic analysis.

Attempts were made to obtain amino acid sequences from peptides produced by the cleavage of trichohyalin. These attempts were unsuccessful with respect to sheep trichohyalin, but proteolytically produced peptides from guinea pig trichohyalin were purified and sequenced. Homology between a number of the sequences obtained suggested that trichohyalin contains a repeat structure.

A cDNA clone to sheep trichohyalin was separately isolated and provided for analysis. The characterisation and sequencing of the clone is described. Analysis of the deduced protein sequence, which corresponds to the carboxy-terminal 30% of the complete protein, has shown that the majority of the partial protein can take up an α-helical conformation and that 95% of the sequence is based on a 23 amino acid peptide repeat.

Genomic Southern analysis, using the cDNA clone as a probe, showed that the trichohyalin gene exists as a single copy gene within the sheep genome.

Genomic library clones containing what was believed to be the sheep trichohyalin gene were purified and characterised. The trichohyalin gene was then sequenced and analysed. Although the 5' end of the gene could not be experimentally located, it is predicted that trichohyalin is 1407 amino acids long and that the gene consists of three exons. Of outstanding note, over 80% of the residues constituting the deduced protein sequence are either glutamic acid, arginine, glutamine or leucine. The carboxy-terminal 60% of the molecule consists of tandem repeats based on a 24 amino acid consensus sequence which is very similar to that seen in the cDNA clone. Within the remaining amino-terminal region most of the sequence is unique, although there are a number of copies of a 40 amino acid repeat. As with the cDNA deduced protein sequence, almost all of the trichohyalin protein is predicted to form α-helix.

Unfortunately the determination of the complete amino acid sequence of trichohyalin has not enabled its function within the inner root sheath to be deduced. Although it does not contain any regions homologous with the conserved core sequence of the intermediate filament proteins, it may be able to form a long α-helical rod which could produce a filament with similar dimensions to the intermediate filaments. Alternatively, the α-helical rod of trichohyalin may be involved in the formation of the matrix of the inner root sheath cells.

Additionally, the amino-terminus of trichohyalin has been found to contain two calcium-binding EF hand motifs. This strongly suggests that in addition to its structural role, trichohyalin is involved in the signalling required for the differentiation of the inner root sheath and medulla.

Finally, the thesis describes <u>in situ</u> hybridization experiments in which the extent of trichohyalin expression within keratinised tissues is examined. Trichohyalin mRNA has been detected in the epithelia of tongue, hoof and rumen as well as in the expected regions of the hair follicle, *i.e.* the differentiating medulla and inner root sheath cells.

DECLARATION

This thesis contains no material which has been accepted for the award of any other degree of diploma in any University. To the best of my knowledge, this thesis contains no material that has been previously published or written by another person, except where due reference is made in the text.

Signed:

Michael James Fietz

NAME: Michael Fietz **COURSE:** Doctor of Philosophy

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

SIGNED: **DATE:** 11th November 1991

Acknowledgements

I wish to thank Professor George E. Rogers for the opportunity to undertake the research described in this thesis in the Department of Biochemistry, University of Adelaide. I would also like to sincerely thank him for his supervision of my research project as well as for his advice, enthusiasm and encouragement during the course of this work and preparation of this thesis.

I am especially grateful to Drs. Barry Powell, who critically read this thesis, and Richard Presland for their advice and comments throughout the course of this work and to Dr. Richard D'Andrea for critically reading a portion of this thesis.

I would like to thank the past and present members of Keratin Korna whose individual and collective efforts have ranged from stimulating to witty to distracting. In particular I would like to thank Lel Whitbread and Rebecca Keough for the regular coffee breaks and for their enlivening conversation, and Simon Bawden for his attempts at keeping me sane during many of the long and frustrating hours.

I must also thank the following people:

Guo Xiao Hui for her assistance in the sequencing of the trichohyalin gene;

Michael Calder for his help with the collection of wool follicles;

Drs. A.V. Sivaprasad and Ian Dodd for their assistance with much of the computing analysis;

Toni Nesci for her advice and assistance with the in situ hybridisation procedure;

Dr. Grant Booker for advice regarding the protein chemistry;

and last but not least, Brandt Clifford for his tireless efforts in the production of the prints for this thesis.

Finally, I am eternally grateful to Janet, for her love, patience and neverending care throughout the course of work for this thesis and for her late night efforts in the production of this volume.

I was supported throughout this study by a Commonwealth Postgraduate Research Award.

Abbreviations

bp:	base pair
BCIG:	5-bromo-4-chloro-3-indolyl-3-D-β-galactoside
BCIP:	5-bromo-4-chloro-3-indolyl phosphate
BSA:	bovine serum albumin
C:	carboxy
cDNA:	DNA complementary to mRNA
Ci:	curie
cRNA:	RNA complementary to DNA (synthesised in vitro)
DNA:	deoxyribonucleic acid
dNTP:	deoxyribonucleoside triphosphates
EDTA:	ethylenediaminetetraacetic acid
FPLC:	fast performance liquid chromatography
HPLC:	high performance liquid chromatography
IF:	intermediate filament
IFAP:	intermediate filament associated protein
IgG:	immunoglobulin
IPTG:	isopropylthiogalactoside
IRS:	inner root sheath
kb:	kilobase pair
kdal:	kilodalton
M_r:	molecular weight
mRNA:	messenger RNA
N:	amino
NBT:	nitroblue tetrazolium
NTP:	nucleoside triphosphates
ORF:	open reading frame
poly(A):	polyadenylic acid

RNA:	ribonucleic acid
SDS:	sodium dodecyl sulphate
TFA:	trifluoroacetic acid
TLCK:	1-chloro-3-tosylamido-7-amino-2-heptanone
Tris:	2-amino-2(hydroxymethyl)-1,3-propandiol
Tween-20:	polyoxyethylenesorbitan monolaurate
UV:	ultraviolet
UWGCG:	University of Wisconsin Genetics Computer Group

Chapter 1

Introduction

1.1 General Introduction

The hair follicle is a specialised derivative of the mammalian epidermis. Due to the economic importance of fibres such as wool, much research has been performed on the proteins of wool and hair. In contrast, little is known of the obligatory follicle layer, the inner root sheath (IRS), which surrounds the developing fibre, or of the medulla which forms the central core of some hairs. Recently a major structural protein of the IRS, termed trichohyalin, has been purified from sheep wool follicles (Rothnagel and Rogers, 1986) and has been shown to be immunologically related to proteins within the hair medulla. The work presented in this thesis attempts to determine the functional role of trichohyalin by obtaining the complete protein sequence. This will be achieved by purifying and sequencing the trichohyalin gene. Additionally, the experimental work will also examine the localisation of trichohyalin expression within the wool follicle and other keratinised tissues.

This chapter will give the background required for the evaluation of the experimental work presented in this thesis. It will present information on the development and structure of the hair follicle, particularly in relation to the IRS and medulla, and on the structure and function of the intermediate filament (IF) proteins and the intermediate filament associated proteins (IFAPs). In addition it will outline the work detailed in the thesis.

1.2 The Hair Follicle

1.2.1 Morphogenesis

During embryogenesis the hair follicle is formed directly from the developing epidermis. Dermal cells aggregate immediately below the epidermis and interact with the adjacent epidermal region leading to the formation of an epidermal appendage. The epidermal cells then proliferate, forming a plug of cells which pushes down into the dermis. The original aggregation of dermal cells are pushed downward by the developing hair peg which eventually invaginates and surrounds the dermal cells, which then form what is termed the dermal papilla (Hashimoto, 1970).

The follicle matrix cells lining the dermal papilla proliferate forming a cone of cells which move up through the middle of the hair peg toward the skin surface. This cone of cells will eventually develop into the differentiated hair (Robins and Breathnach, 1969).

Once the mature follicle has been formed (Fig. 1.1) it enters into the hair growth cycle (Fig. 1.2). The follicle is initially in the active phase, anagen, during which the hair fibre is being produced. It then enters a short transitional phase, catagen, in which the synthesis of the hair ceases and the club which is formed on the end of the fibre, together with the dermal papilla, moves up through the follicle. This is succeeded by a resting phase, telogen, in which the club end of the fully formed hair remains anchored in the follicle. Eventually the follicle will re-enter the anagen phase and a fresh hair fibre will be produced.

Although the division of the follicle matrix cells provides the cells required for hair development during anagen, the follicular stem cells have recently been shown to be located in the follicle bulge (Cotsarelis *et al.*, 1990; see Fig. 1.2). These cells are essential for the re-entry into the anagen phase after telogen and for the subsequent production of the new hair. During telogen the stem cells are activated by a factor believed to be produced by the dermal papilla cells. The stem cells then divide, pushing the dermal papilla down the follicle, leading to the initiation of the new anagen stage. The stem cells remaining in the bulge are then believed to return to their normal slow-cycling state in preparation for the next hair cycle.

The dermal papilla is vital to the hair follicle; in addition to its role in the hair cycle, it is also believed to regulate the size, length and rate of growth of the hair fibre. This was initially suggested by Cohen in 1961 and was conclusively demonstrated by Oliver (1970) when he showed that adult dermal papillae, when transplanted to foreign epidermal tissues, were able to stimulate follicle formation and lead to the production of hairs which were similar to those from which the papillae were removed. Even though the dermal papilla is of great importance in hair development, very little is known of the factors produced by the dermal papilla or how these effect the division and differentiation of the surrounding follicle bulb cells.

Figure 1.1 Schematic longitudinal section through the proximal portion of a non-medullated follicle. The differentiation and keratinisation of each of the cell layers is shown.

(Adapted from Auber, 1950.)

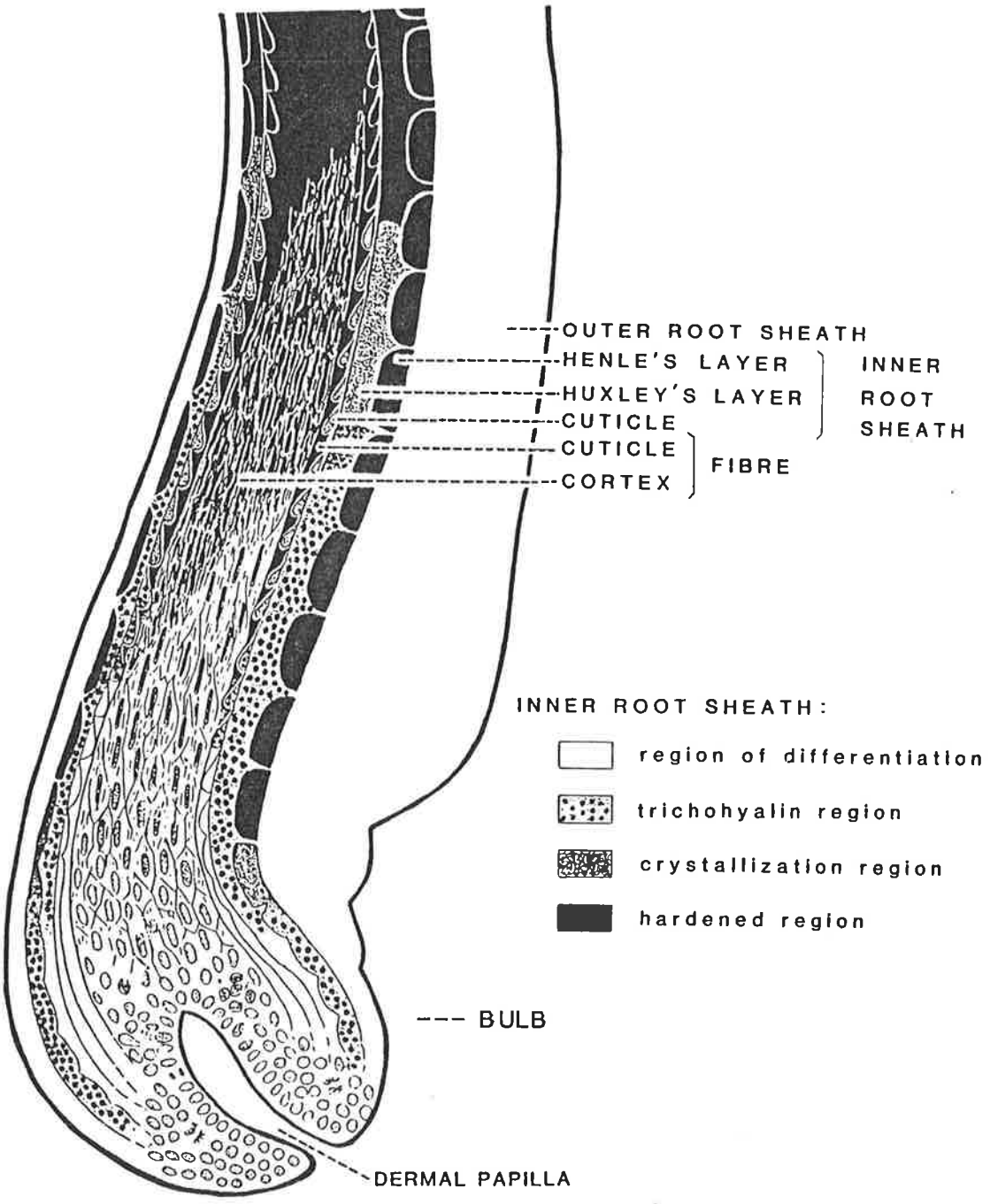
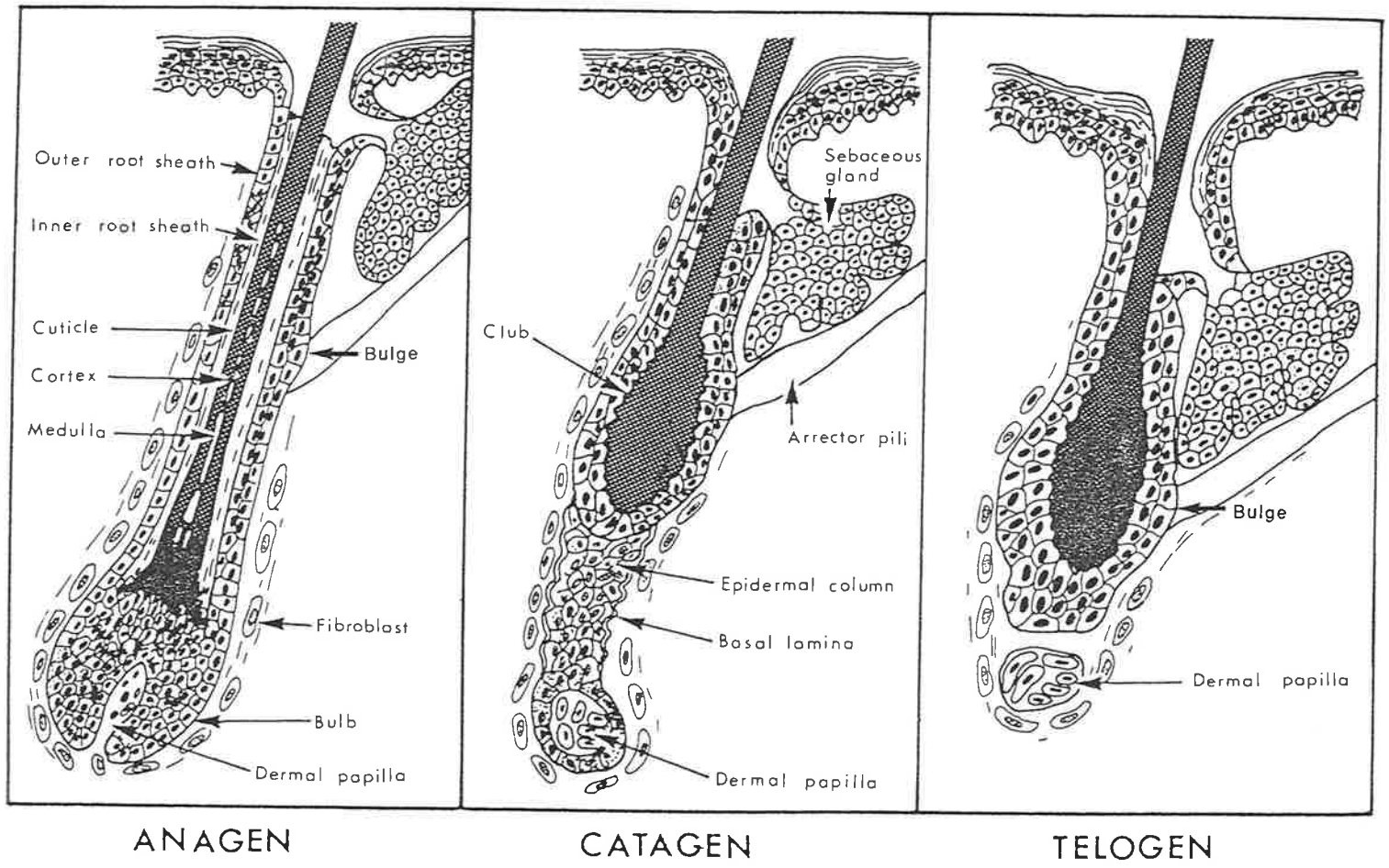

OUTER ROOT SHEATH
HENLE'S LAYER
HUXLEY'S LAYER
CUTICLE
CUTICLE
CORTEX
INNER ROOT SHEATH
FIBRE
INNER ROOT SHEATH:
region of differentiation
trichohyalin region
crystallization region
hardened region
BULB
DERMAL PAPILLA

Figure 1.2 Diagrammatic representation of the hair cycle. For description, see text. (Reproduced from Ebling, 1988.)



ANAGEN CATAGEN TELOGEN

1.2.2 Hair Follicle Structure

a. Follicle Cell Development

The hair follicle is one of the few proliferating tissues of the adult mammal. The dividing cells are positioned within the follicle bulb (Fig. 1.1) and the position of these cells in relation to the dermal papilla determines into which differentiated follicle layer the daughter cells will develop: the three layers of the IRS, namely the Henle layer, the Huxley layer and the cuticle; the cuticle of the fibre, the fibre cortex or, if present in the fibre, the medulla (Fig. 1.1). As the cells are pushed up the follicle shaft they begin to differentiate and start producing the proteins required to form the particular hardened tissue. The cells then undergo a phase of major protein synthesis, the synthesised protein contents begin to aggregate and eventually fill the cell. The proteins are cross-linked, cytoplasmic organelles are degraded and water is removed producing the mature hardened tissue.

b. Follicle Morphology

(i) Cortex

The cortex is the main constituent of the hair fibre developing from cells laterally surrounding the dermal papilla. Hardened cortical cells are filled with 8-10 nm microfibrils (intermediate filaments) aligned parallel with the direction of hair growth. These microfibrils are embedded within a proteinaceous matrix, the proteins of which belong to the class of intermediate filament associated proteins (IFAPs). Within the hardened cells the microfibrillar and matrix proteins are cross-linked by disulphide bonds.

On the basis of its structure and composition the hair cortex is a typical example of a keratinised tissue. Nevertheless the use of the term keratin proteins with respect to the cortex has remained ambiguous. Do the keratin proteins consist simply of the microfibrillar proteins or do they also include the matrix proteins? In accord with the definition of Fraser *et al.* (1972) and in agreement with previous work presented by this group, the term "keratin proteins" will be used to refer to both the microfibrillar and matrix proteins of the cortex.

(ii) Fibre Cuticle

Immediately surrounding the cortex is the fibre cuticle which forms a protective outer coating for the fibre. The outer two-thirds of each cuticle cell consists of a

specialised hardened region which is termed the exocuticle, whilst the inner third is termed the endocuticle.

(iii) Medulla

The medulla, when present in hair fibres, forms the central core of the fibre and contains hardened cells which are interspersed with large air spaces. The presence of medullary air spaces is believed to improve thermal insulation and also decreases the mass of thick fibres (Fraser *et al.*,1980). The medulla develops from cells positioned immediately over the dome of the dermal papilla and the developing cells are characterised by the presence of electron-dense non-membrane bound granules, termed trichohyalin granules (Fig. 1.3). As the cells move up the follicle the number and size of trichohyalin granules increases and eventually the granules merge forming the amorphous contents of the hardened medulla cells. The mature medullary proteins contain two post-translational modifications. The proteins are cross-linked by ε-(γ-glutamyl)lysine bonds which are formed between glutamine and lysine residues by the follicle enzyme transglutaminase (Fig. 1.4a). The second modification involves the conversion of a proportion of the arginine residues to citrulline residues by the enzyme peptidylarginine deiminase (Fig. 1.4b).

(iv) Inner Root Sheath

During its development the hair fibre is encompassed by the inner root sheath (IRS) which is composed of three layers, the Henle layer, the Huxley layer and the IRS cuticle. The IRS is thought to play a supporting role within the follicle, passively guiding and directing the growth of the hair fibre (Straile, 1965).

Developing IRS cells, which are derived from dividing cells positioned on the periphery of the follicle bulb, contain trichohyalin granules which have very similar histochemical and immunological characteristics to the granules found within the medulla (Fig. 1.3). As the cells move up the follicle the size and number of granules increases and 8-10 nm diameter filaments become associated with the granules (Rogers, 1964a). Eventually, the granules disappear, the IRS cells become completely filled with intermediate-like filaments aligned parallel with the direction of hair growth (Rogers, 1964a) and the filaments harden forming the insoluble contents of the mature IRS cell (Birbeck and Mercer, 1957).

Although both the mature IRS and cortical cells are filled with intermediate filaments, the two are readily distinguishable in that the IRS proteins are different in

Figure 1.3 Electron micrograph of a medullated follicle, showing the trichohyalin granules in the medulla (Medullary Granules) and in the Huxley layer and cuticle of the inner root sheath (Trichohyalin). Henle's layer has already been converted from the granular to the filamentous form. 5,000 x

(Obtained from J.A. Rothnagel.)

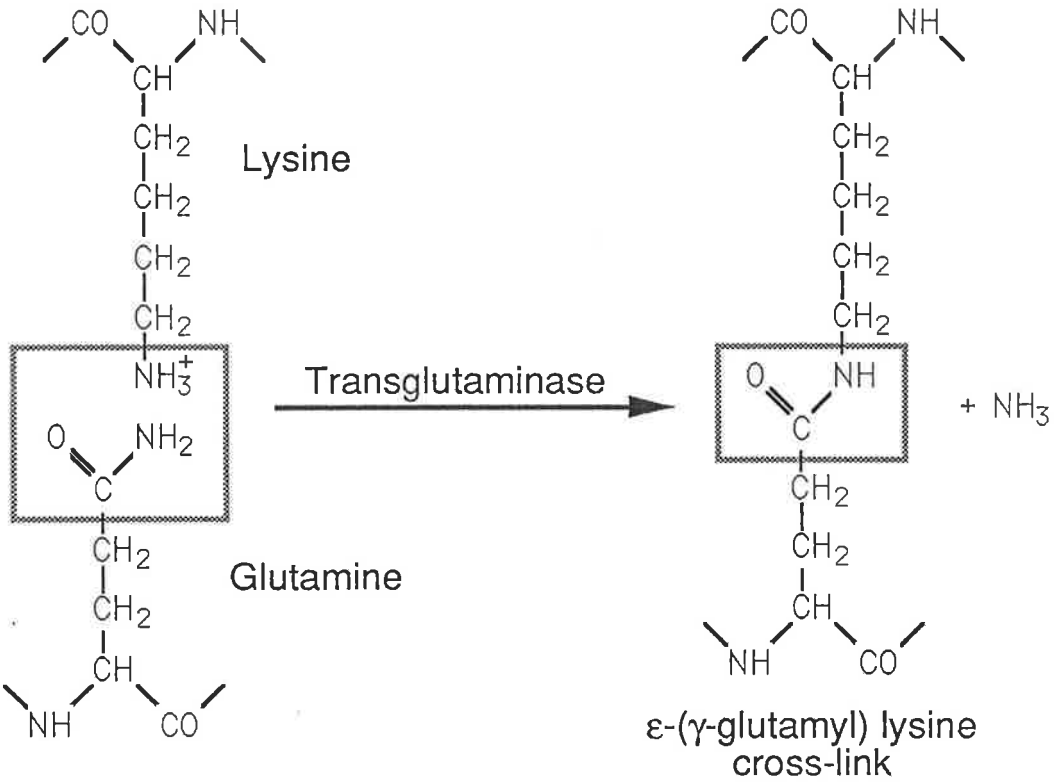

Medulla
Medullary Granules
Filaments
Cortex
Cuticle
Cuticle
Huxley's layer
Trichohyalin
Henle's layer
Glycogen
Outer root sheath

Figure 1.4 Post-translational modifications present within the proteins of the hardened medulla and IRS.

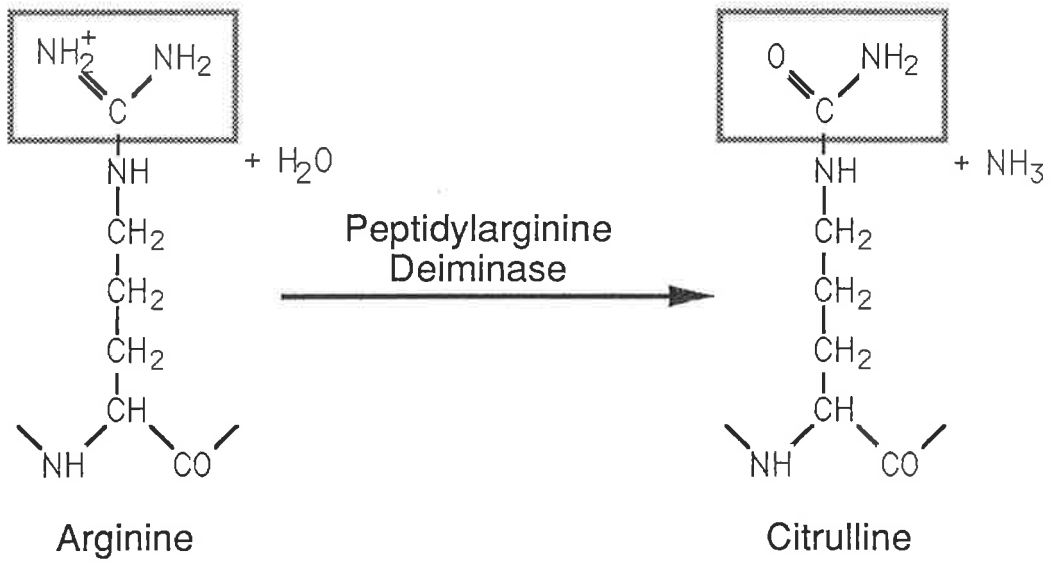
(a) Cross-linking of glutamine and lysine residues by follicle transglutaminase.

(b) Conversion of arginine to citrulline residues by follicle peptidylarginine deiminase.

a.


Lysine
Glutamine
Transglutaminase
+ NH3
ε-(γ-glutamyl) lysine
cross-link

b.


+ H2O
Peptidylarginine
Deiminase
+ NH3
Arginine
Citrulline

amino acid composition. The IRS proteins contain peptidylcitrulline and are cross-linked by ε-(γ-glutamyl)lysine iso-peptide bonds which are the two post-translational modifications also occurring in the medulla.

1.2.3 Cortical Proteins

The keratin proteins of the hardened cortical cells are cross-linked by disulphide bonds. The reduction of these bonds in a thiol/urea solution is required for the extraction of the cortical keratin proteins. Subsequent blockage of the sulphydryl groups by carboxymethylation with iodoacetate (Crewther, 1976) has allowed analysis of the hair keratins.

On the basis of amino acid composition and molecular weight, the keratin proteins can be sub-divided into three groups: the IF proteins (low-sulphur proteins) and two groups of IFAPs, the high-sulphur proteins, and the high-glycine/tyrosine proteins. The three classes can be readily visualised after separation by two-dimensional polyacrylamide gel electrophoresis (Fig. 1.5).

a. Cortical IF Proteins

The IF or low-sulphur proteins form the cortical microfibrils (Jones, 1975, 1976) and are characterised by low levels of cysteine (1-3 moles%) and molecular weights ranging from 38,000 to 58,000 (see Table 1.1). On the basis of amino acid sequence the low-sulphur proteins can be sub-divided into two sub-families, type I (originally termed component 8) and type II (components 5 and 7). The type II proteins have a higher molecular weight and are more basic in nature than the type I proteins. The low-sulphur proteins belong to the superfamily of IF proteins and contain the α-helical core region and non-helical terminal domains which are typical for IF proteins (see Section 1.5.2). The type I and type II low-sulphur proteins belong to the type I and type II IF subclasses, respectively (see Section 1.5.1).

b. The High-Sulphur Proteins

The high-sulphur proteins range in molecular weight from 10,000 to 30,000 and are characterised by the presence of up to 25 moles% cysteine (Gillespie and Frenkel, 1974a). On the basis of chromatographic and electrophoretic separation the wool high-sulphur proteins have been divided into four groups, namely, SCMKB1, SCMKB2, SCMKBIIIA and SCMKBIIIB (Crewther, 1976). No ordered structure has been detected within the high-sulphur proteins and this is probably due to the high levels of

Figure 1.5 Two-dimensional polyacrylamide gel electrophoresis of total wool protein.

Total wool proteins were extracted, S-carboxymethylated using ^{14}C-iodoacetic acid and separated on a two-dimensional gel. The electrophoresis was performed at pH 8.9 in the first dimension and in the presence of SDS in the other. A fluorograph of the gel is shown. LS, low-sulphur proteins; HS, high-sulphur proteins; UHS, ultra-high-sulphur proteins; HGT, high-glycine/tyrosine proteins.

(Reproduced from Rogers *et al.*, 1989.)

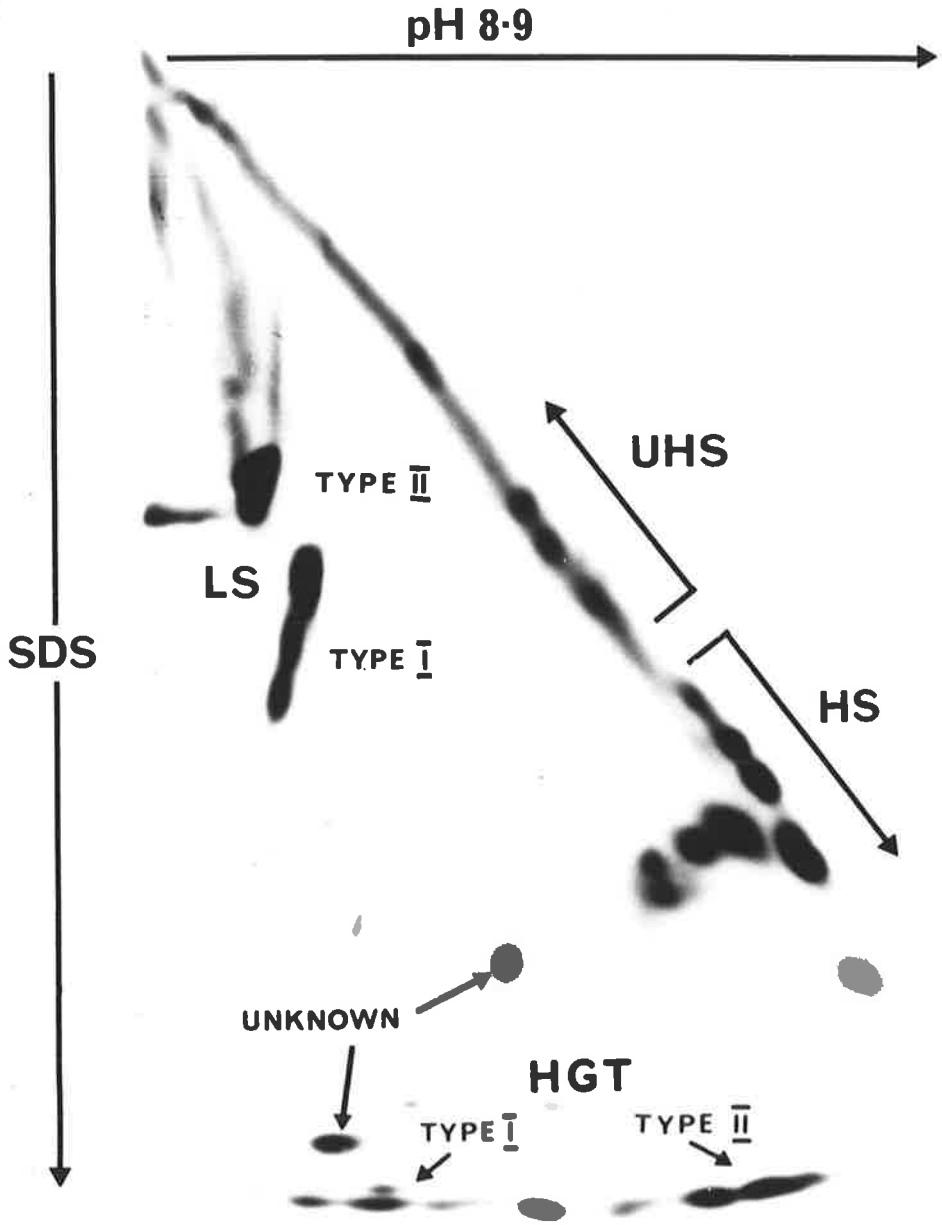

pH 8·9
SDS
TYPE II
LS
TYPE I
UHS
HS
UNKNOWN
HGT
TYPE I
TYPE II

Table 1.1 Characteristics of the major wool keratin proteins.

The protein data within this table are from Crewther (1976), Swart *et al.* (1976), Gillespie (1983) and MacKinnon *et al.* (1990).

1. The prefix SCMK indicates that the proteins are S-carboxymethylated derivatives of the native protein. The number in brackets represents an estimate for the number of proteins within each family.
2. The proteins within this group have not been characterised.
3. Some protein-chemical data suggests that the type I and II high-glycine/tyrosine proteins should be considered as subclasses, containing several families of proteins. However, the numbers of different polypeptide chains may be a conformational anomaly.

(Adapted from Powell and Rogers, 1986.)

Protein class and location in fibre	Characteristic amino acid content	Protein families within the class	Family notation[1]		Size (kd)
Intermediate Filament (microfibrils)	Cysteine (1-3 moles%)	2	type I type II	(4) (4)	38-43,000 56-58,000
High-Sulphur (matrix)	Cysteine (20-25 moles%)	4	SCMKB1 SCMKB2 SCMKBIIIA SCMKBIIIB	(?)[2] (7) (11) (4)	23-26,000 19,000 16,000 11,000
Ultra-High-Sulphur (matrix and cuticle)	Cysteine (>30 moles%)	≥2	-	(≥2)	>16,000
High-Gly/Tyr	Glycine (20-40 moles%) Tyrosine (12-21 moles%)	2	type I type II	(10)[3] (5)[3]	6-9,000 6-9,000

serine, threonine, cysteine and proline, all of which are known to inhibit helix formation (Gillespie, 1983). Nevertheless the presence of pentapeptide and decapeptide repeats within almost all known high-sulphur sequences suggests that the cysteine residues within the repeats may allow the formation of an ordered cross-linked structure with the low-sulphur proteins of the microfibrils (Powell and Rogers, 1986).

A sub-family of the high-sulphur proteins, the ultra-high-sulphur proteins, is also present in the cortex and the cuticle. These proteins have been shown to contain up to 35 moles% cysteine (Table 1.1). Although the overall structure is similar to the high-sulphur proteins, secondary structure analyses have predicted that short regions of the ultra-high-sulphur proteins may be able to form ß-sheets (MacKinnon, 1989).

c. The High-Glycine/Tyrosine Proteins

The high-glycine/tyrosine proteins are the smallest keratins within the hair and range in molecular weight from 6,000 to 10,000. They are divided into two sub-families (type I and II) on the basis of amino acid content and solubility and are characterised by high levels of glycine and tyrosine, which can range up to 25 moles% and 35 moles% respectively (Gillespie and Frenkel, 1974a,b; Table 1.1). Although the high-glycine/tyrosine proteins are present within the matrix, their significance is still uncertain.

d. Cortical Keratin Genes

(i) IF Genes

The genes encoding one type I (Wilson *et al.*, 1988) and two type II (B.C. Powell, personal communication) wool IF proteins have been sequenced. These genes are characterised by the presence of 8 introns and an overall length of approximately 7-8kb (Wilson *et al.*, 1988; Powell and Rogers, 1990a), a structure typical for epithelial IF genes. IF gene structure will be discussed in more detail in Section 1.5.3.

(ii) Cortical IFAP Genes

To date seven genes encoding cortical IFAPs have been sequenced; four high-sulphur genes (Powell *et al.*, 1983; Frenkel *et al.*, 1989) and three high glycine/tyrosine genes (Kuczek and Rogers, 1985, 1987; A. Fratini, personal communication). In addition the sequences are also known for two genes encoding cuticle ultra-high-sulphur proteins (MacKinnon *et al.*, 1990) and a gene encoding a third ultra-high-sulphur protein whose location has not yet been determined (McNab *et al.*, 1989). The cortical IFAP genes are characterised by a lack of introns and the presence of an 18 bp conserved

sequence, termed the matrix box, immediately prior to the initiating ATG codon (Fig. 1.6). The role of this sequence is presently uncertain.

1.2.4 Proteins of the Mature Inner Root Sheath and Medulla

Study of the proteins within the mature fibre cortex has been greatly facilitated by the ability to release the intact proteins by cleavage of the the disulphide cross-linkages. Initial investigations of the medulla and IRS showed that their hardened contents were insoluble in the urea/thiol solutions used to dissolve the cortical proteins (see Mercer, 1961; Fraser *et al.*, 1972; Elöd and Zahn, 1944; Hardy, 1952; Rogers, 1958a, 1964b) and this lack of solubility was due to the presence of ε-(γ-glutamyl)lysine cross-links (Harding and Rogers, 1971,1972a; Fig. 1.4a). To date there is no known agent, chemical or enzymatic, which specifically cleaves the ε-(γ-glutamyl)lysine bond and thus intact IRS and medulla proteins cannot be released. Therefore studies have had to be performed on the complete hardened tissue or on proteolytic fragments derived from the cross-linked proteins.

Amino acid analysis has indicated that glutamic acid/glutamine, arginine, leucine and lysine are the most abundant amino acids in both the medulla and IRS whereas in the cortex they are glutamic acid/ glutamine, serine, cysteine and glycine (Table 1.2). Additionally, the IRS and medulla were also shown to contain the amino acid citrulline which had not previously been found as a constituent in proteins (Rogers, 1958a; Rogers and Simmonds, 1958; Rogers, 1963). Tryptic digestion of medulla and IRS tissues released citrulline-rich polypeptides showing that the citrulline within the tissues was protein bound (Rogers, 1962, 1964b). The presence of citrulline in normal peptide linkages was conclusively shown by the release of citrulline using the wide-specificity protease subtilisin, and by the isolation and sequencing of small citrulline containing peptides (Steinert *et al.*, 1969; Table 1.3). The similarity of the amino acid compositions of the IRS and medulla, together with the presence of the ε-(γ-glutamyl)lysine cross-links, suggests that although the two tissues produce different hardened forms, they have a very similar protein chemistry which is quite distinct from that of the fibre cortex.

Studies have also been performed on the individual filaments of the mature IRS (Steinert *et al.*, 1971). These filaments were isolated by limited proteolysis of the IRS tissue, contain extensive α-helix, as shown by circular dichroism and X-ray diffraction analysis, and were found to have a diameter of approximately 7-8 nm, typical for

Figure 1.6 The "Matrix Box" of the cortical IFAP genes.

Comparison of the 18 bp sequence ("matrix box") immediately upstream of the initiating methionine codon in six sheep high-sulphur and two high-glycine/tyrosine IFAP genes. The sequence of the B2A gene is used as the model sequence against which the others are compared. Deletions (-) and insertions (superscript) were introduced to maximise homology. Common nucleotides are indicated by a star (*).

Data from Powell *et al.*, 1983; Kuczek and Rogers, 1985, 1987; Frenkel *et al.*, 1989; Powell *et al.*, in preparation.

(Adapted from Kuczek and Rogers, 1987.)

	-10 -1
B2A:	AACCCAACTCCTGACACC**ATG**
B2C:	*G**T***C*********ATG
B2D:	*G****************ATG
BIIIA:	TCGG*CC*A*ACAC****ATG
BIIIB"2":	******G**********ATG
BIIIB"4":	***A**G****CA*****ATG
HGT-F:	CG******-**A******ATG
HGT-C2:	TT***-*******C***AATG

Table 1.2 Amino acid composition of the hardened proteins of the cortex, medulla and inner root sheath of the guinea pig hair follicle together with those of the guinea pig TR-PPT fraction and wool trichohyalin.

1. Rogers (1983).
2. TR-PPT is a urea-soluble protein fraction, present in the guinea pig IRS and medulla, which is incorporated into the respective hardened tissues. From Rogers *et al.* (1977).
3. Rothnagel and Rogers (1986).

SCM, S-carboxymethylated.

Amino Acid	SCM-Hair Keratin[1]	Medulla[1]	Inner Root Sheath[1]	TR-PPT Fraction[2]	Wool Trichohyalin[3]
	mole percent	*mole percent*	*mole percent*	*mole percent*	*mole percent*
SCM-cys	16.0	0.0	0.0	0.8	0.0
Asp/Asn	5.2	4.5	9.3	5.2	6.1
Thr	5.9	1.6	4.8	1.7	3.0
Ser	9.7	2.3	6.9	2.7	5.5
Glu/Gln	12.7	41.0	20.8	36.7	17.0
Pro	5.2	0.0	3.1	1.9	4.0
Gly	8.4	2.9	7.9	3.1	6.4
Ala	5.0	2.1	6.2	2.4	5.3
1/2-Cys	0.0	trace	1.0	0.0	0.7
Val	5.4	1.6	4.4	2.0	4.1
Met	0.0	0.2	4.4	0.4	1.0
Ile	3.1	1.0	3.7	1.3	2.9
Leu	6.4	6.7	8.9	10.4	8.5
Tyr	3.0	0.8	2.5	1.4	2.1
Phe	3.4	1.6	2.8	3.1	2.7
Lys	2.4	6.1	4.3	4.7	11.6
His	1.2	0.9	1.6	0.9	2.7
Arg	7.1	3.5	3.1	21.2	16.5
Citrulline	0.0	23.6	4.4	0.0	0.0

Table 1.3 N-terminal amino acid sequences of peptides containing citrulline.

Cit, citrulline.

Glx, glutamic acid or glutamine.

Data from Steinert *et al.* (1969).

PEPTIDE	SEQUENCE
IM-TP_4/6a	Asp-Cit-Phe-Cit-
IM-TP4/7d	Cit-Cit-Val-Cit-Cit-
IM-TP_6/7b	Leu-Leu-Glu-Cit-Cit-
IM-TP_7/I	Phe-Cit-Glx-Glx-
IIM-TP_5/3b	Leu-Cit-Gln-
IIM-TP_{10}/2a	Asp-Cit-Cit-Phe-
1,9-Citrulline Bradykinin	Cit-Pro-Pro-

intermediate filaments (Steinert, 1978). The α-helical regions contain little citrulline or ε-(γ-glutamyl)lysine cross-links, suggesting that the IRS filamentous proteins are only cross-linked in their non-helical domains.

1.3 Proteins of the Developing IRS and Medulla

Further analysis of the IRS and medulla structural proteins requires the purification and examination of the precursor proteins prior to the formation of the cross-links. In addition, analysis of the enzymes which perform the post-translational modifications may also aid in the understanding of the structure and function of both the inner root sheath and medulla.

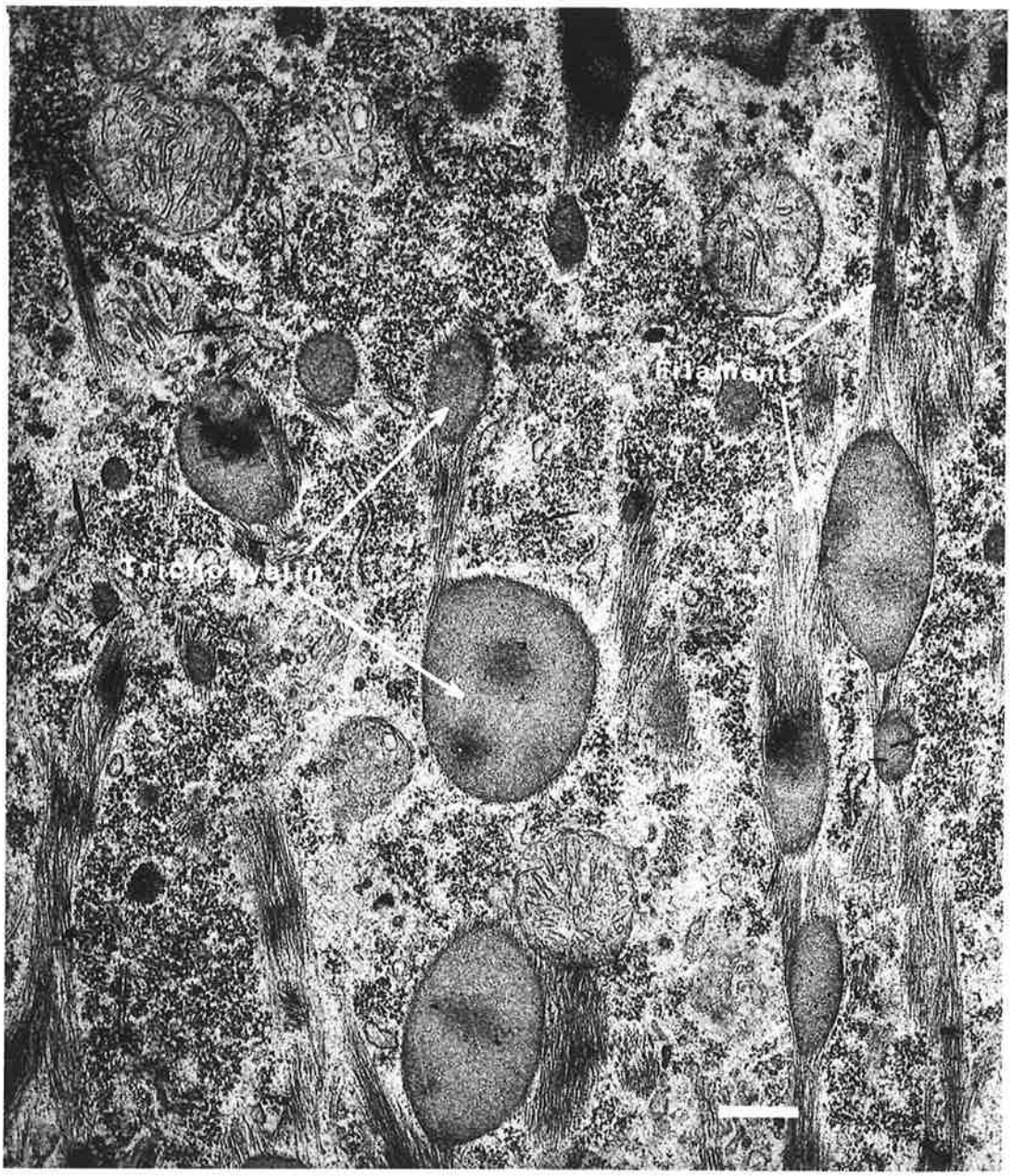
1.3.1 Trichohyalin

The major distinctive histological feature of the developing IRS and medulla cells is the presence of proteinaceous trichohyalin granules, first described by Vörner in 1903. The granules of the IRS and medulla have been shown to have similar histochemical properties (Auber, 1950). The role of the granular proteins has long been questioned. Electron microscopic studies have suggested that the trichohyalin granules transform into the filaments during IRS development (Birbeck and Mercer, 1957; Rogers, 1958b, 1964a,b; Fig. 1.7) and into the amorphous mass of the hardened medulla (Rogers and Harding, 1976a). Alternatively, others have suggested that the granular proteins provide the matrix into which the IRS filaments are embedded (Parakkal and Matoltsy, 1964; Parakkal, 1969; Steinert, 1981). A third possibility is that the granules play some other function and that the proteins are degraded and thus are not present in the mature cells.

Histochemical studies on the levels of arginine and citrulline within the follicle have shown that there are high levels of arginine within the trichohyalin granules of the developing IRS and medulla but little in the hardened tissues, whereas all of the citrulline is present within the hardened cells of the IRS and medulla (Rogers, 1963). This suggests that the granular arginine-containing proteins are the precursors for the citrulline-containing proteins of the mature tissues. In support of this, injected tritium-labelled arginine was initially detected in the IRS granules and later in the hardened IRS cells (Rogers and Harding, 1976a). These studies therefore suggested that a soluble arginine-rich precursor should be present within the granules.

Figure 1.7 Electron micrograph of an IRS cell at an early stage of differentiation. Intermediate filaments are seen in close association with the trichohyalin granules. Scale bar = 0.5 μm.

(Reproduced from Rothnagel, 1985.)


filament
trichohyalin

In an attempt to discover such an arginine-rich protein, urea-soluble follicle proteins were S-carboxymethylated and separated by ion-exchange chromatography (Rogers *et al.*, 1977). One of the resolved fractions, labelled TR, had an amino acid composition which closely resembled that of the hardened IRS and medulla (Table 1.2). Importantly, the TR fraction did not contain citrulline or ε-(γ-glutamyl)lysine cross-links, but did have high levels of arginine and glutamic acid/glutamine, i.e. the substrate amino acids for the post-translational changes. Antibodies raised against the TR fraction bound to the granules of both the medulla and IRS. Furthermore the proteins of the isolated TR fraction were shown by in vitro analysis to be able to act as substrates for both the follicle transglutaminase and peptidylarginine deiminase (Rogers *et al.*, 1977). Thus the TR fraction was deduced to be a granular precursor for the hardened IRS and medulla tissues.

Electrophoretic analysis originally suggested that the TR fraction contained a family of related proteins (Rogers, 1983). Subsequent work showed that when the extraction was performed at 4°C, without subsequent S-carboxymethylation, a single protein, having a molecular weight of 190k in sheep, was isolated (Rothnagel and Rogers, 1986; Fig. 1.8). Amino acid analysis showed that this protein has a very similar composition to the TR fraction (Table 1.2) and the protein has thus been termed trichohyalin.

To examine the role of trichohyalin within the hardened IRS, immunoelectron microscopy was performed using a polyclonal antibody raised against sheep IRS trichohyalin (Rothnagel and Rogers, 1986). The antibody bound to the granules of the IRS and medulla but not to the filaments streaming from the IRS granules or the hardened IRS or medulla cells. However, antibody binding was detected in a number of IRS cells in which the granules had just disappeared and the cell was filled with a newly organised array of filaments (Rothnagel and Rogers, 1986; Fig. 1.9). These results suggest that trichohyalin does not form the filaments but is involved in producing the matrix into which the IRS filaments are embedded.

It should be noted that the trichohyalin granules of the IRS and medulla have both been shown to contain immunologically related proteins (Rogers *et al.*, 1977; Rothnagel and Rogers, 1986) which are of a similar molecular weight (Rothnagel and Rogers, 1986). Although it has not been conclusively shown that the two proteins are identical, I will throughout this thesis jointly term them trichohyalin.

Figure 1.8 Electrophoretic separation of purified sheep trichohyalin.

Purified sheep trichohyalin was analysed by SDS polyacrylamide gel electrophoresis. Note that the purified protein runs as a doublet of approximately 190 and 185 kdal. The position of molecular weight markers is shown at the side of the gel.

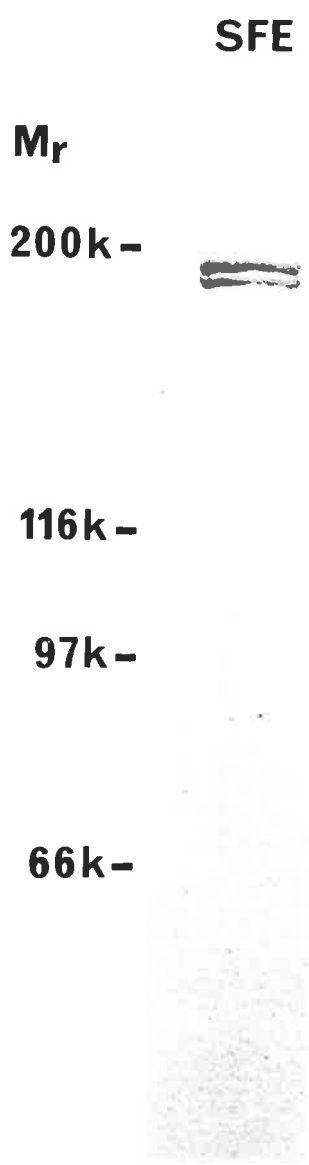
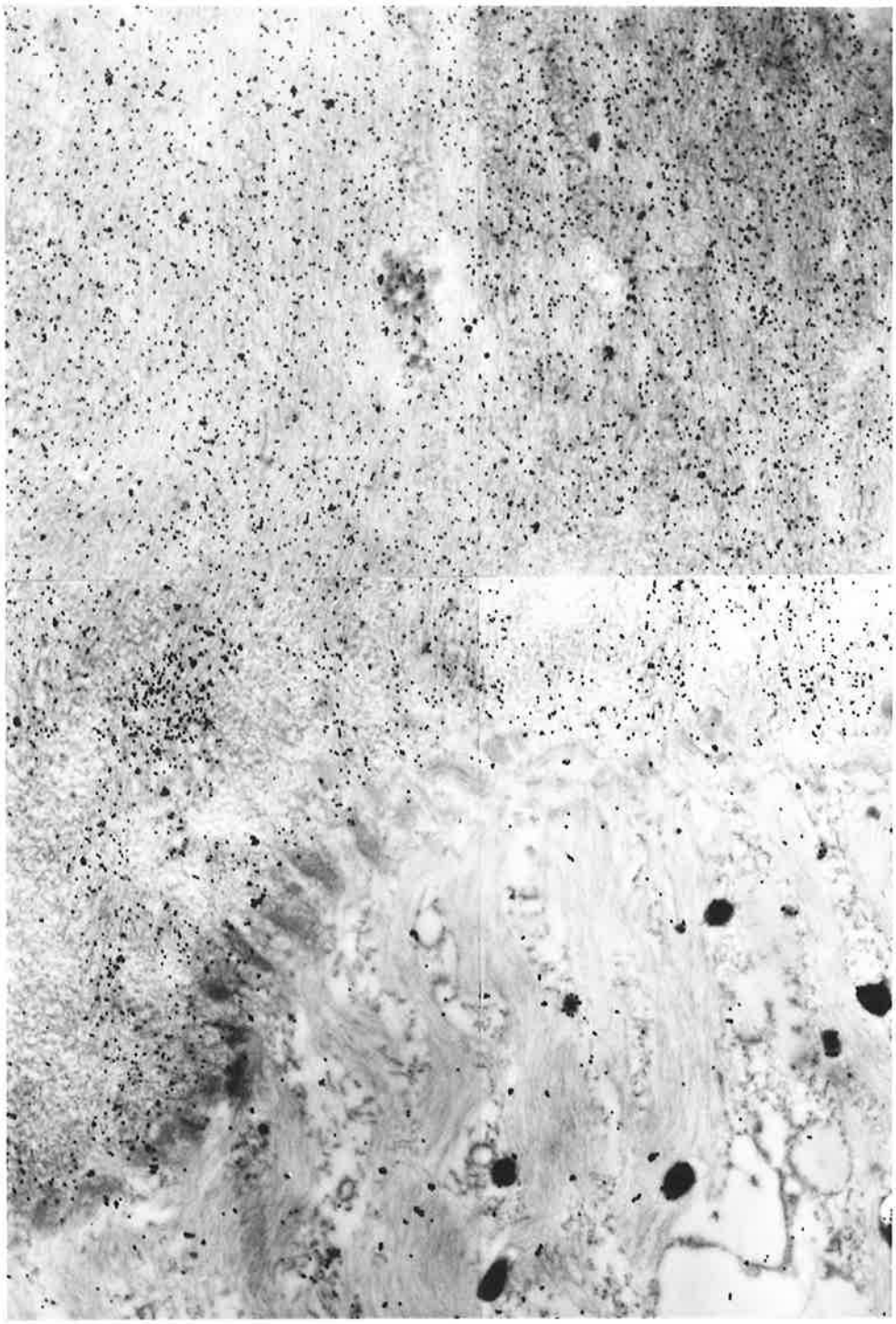

SFE
Mr
200k-
116k-
97k-
66k-

Figure 1.9 Immunoelectron microscopy of a rat vibrissa follicle reacted with anti-trichohyalin and detected with Protein A-gold. The montage shows a cell containing granules and filaments (bottom) and a cell in which the fully filamentous array has just been formed (top). Note the detection of bound antibody in the top cell. 22,000 x

(Reproduced from Rothnagel and Rogers, 1986.)



1.3.2 Peptidylarginine Deiminase

Rogers and Simmonds (1958), in showing the presence of citrulline in hair follicle IRS proteins, were the first to reliably demonstrate the presence of citrulline in any protein. Subsequent work conclusively showed that citrulline was incorporated in normal peptide linkages in proteins of the hardened IRS and medulla (see Section 1.2.4). As citrulline cannot be introduced into proteins during translation, its presence must arise from a post-translational enzymic modification of a normally incorporated amino acid. Such an activity, which is believed to remove the imino group from the guanidino side-chain of peptidylarginine residues producing peptidylcitrulline, was first detected in guinea pig hair follicles (Rogers and Harding, 1976b). The enzyme has subsequently been termed peptidylarginine deiminase (Fujisaki and Sugawara, 1981). Enzymic activity was then also detected in bovine and rat epidermis (Kubilus *et al.*, 1980; Fujisaki and Sugawara, 1981), rabbit skeletal muscle (Sugawara *et al.*, 1982), and bovine brain (Kubilus and Baden, 1983). Recently peptidylarginine deiminase activity has been detected in almost all mammalian tissues (Watanabe *et al.*, 1988; Takahara *et al.*, 1989) suggesting that the enzyme has a common functional role in most mammalian systems.

Peptidylarginine deiminase has been partially purified from epidermis (Kubilus *et al.*, 1980; Fujisaki and Sugawara, 1981) and from hair follicles (Rogers and Rothnagel, 1983), and has been purified to a single protein species from muscle (Takahara *et al.*, 1983) and brain (Kubilus and Baden, 1983). Immunological studies, using antibodies raised against purified skeletal muscle peptidylarginine deiminase, have shown that in all tissues, except the hair follicle and epidermis, the enzyme is of the skeletal muscle type (Watanabe *et al.*, 1988). Differences in substrate specificity between the epidermal and follicle enzymes (Watanabe *et al.*, 1988) suggest that there are three peptidylarginine deiminase types within mammalian tissues, i.e. follicle, epidermal and muscle types.

Calcium is an absolute requirement for all peptidylarginine deiminases, and with a crude extract of wool follicle enzyme 12 mM calcium was required to obtain maximal activity (Rothnagel, 1985). Optimal wool follicle peptidylarginine deiminase activity also requires 2.5 mM dithiothreitol, a pH of 7.5 and an incubation temperature of 37°C (Rothnagel, 1985). As stated earlier (Section 1.3.1), the follicle enzyme acts on the granular protein trichohyalin, but the functional role of this deimination is unknown.

1.3.3 Follicle Transglutaminase

The second post-translational modification within the follicle IRS and medulla is the formation of the ε-(γ-glutamyl)lysine cross-links. These cross-links are also found in several other tissues and are formed by a class of enzymes termed the transglutaminases. The transglutaminases, which have a calcium requirement for activity, catalyse the formation of a covalent bond between the γ-carbon of peptidylglutamine residues and a primary amino group (reviewed by Folk and Finlayson, 1977; Folk, 1980). Within the follicle the amino group is usually provided by the primary amino group of peptidyllysine residues (see Fig 1.4a).

The follicle ε-(γ-glutamyl)lysine iso-peptide bonds were first detected in guinea pig IRS and medulla by Harding and Rogers (1971). Subsequently, follicle transglutaminase activity was detected (Harding and Rogers, 1972b; Chung and Folk, 1972) and the enzyme was found to have a molecular weight of 54,000 which consisted of two identical 27 kdal subunits (Chung and Folk, 1972; Peterson and Buxman, 1981). Immunological studies have localised the follicle transglutaminase to the medulla and IRS cells (Peterson and Buxman, 1981). Transglutaminase activity has also been detected within the vertebrate epidermis (Goldsmith *et al.*, 1974) but the epidermal enzyme is immunochemically unrelated to the follicle transglutaminase (Ogawa and Goldsmith, 1977; Peterson and Buxman, 1981).

Interestingly, the reaction conditions for follicle transglutaminase are very similar to those required for follicle peptidylarginine deiminase, although it is possible that wool follicle transglutaminase is not calcium dependent (Harding and Rogers, 1972b). As the two enzymes are located in the follicle tissues it is possible therefore that the two may act in a concerted fashion to produce ε-(γ-glutamyl)lysine cross-linked proteins containing peptidylcitrulline residues.

1.3.4 IF Proteins

Although IF proteins have not been isolated from the IRS and medulla, numerous immunological studies using monoclonal antibodies have detected filamentous antigenic determinants in both tissues. In a number of cases (Ito *et al.*, 1986a,b; Lane *et al.*, 1985) the same determinants were detected in both the IRS and medulla. In contrast to this, Heid *et al.* (1988) have shown that the medulla and IRS bind different antibodies, with

the medulla binding an antibody to K19 and the IRS binding antibodies to K6 and K16. Thus the identity of the medulla and IRS IF proteins remains uncertain.

1.4 Epidermal Structural Proteins which are Functionally Related to Trichohyalin

1.4.1 Filaggrin

The fully differentiated mammalian epidermal cells (keratinocytes) are filled with an organised array of intermediate filaments. These filaments are aggregated by the epidermal IFAP filaggrin (Dale *et al.*, 1978; Steinert *et al.*, 1981a; Lonsdale-Eccles *et al.*, 1982; Harding and Scott, 1983; Lynley and Dale, 1983). Filaggrin is initially synthesised as a short-lived, very high molecular weight (>300 kdal; see Dale *et al.*, 1989), phosphorylated precursor termed profilaggrin (Harding and Scott, 1983). During epidermal cell development profilaggrin is dephosphorylated and rapidly cleaved to produce the smaller, functional filaggrin protein. The small filaggrin unit is then able to aggregate the epidermal IFs and has been suggested to promote disulphide bond formation amongst the IFs (Steinert, 1983).

Filaggrin has been shown to be a substrate for epidermal peptidylarginine deiminase (Harding and Scott, 1983; see Section 1.3.2). As the production of citrulline produces a less basic filaggrin unit, it has been proposed that citrulline formation lowers the affinity of filaggrin for the IF proteins, thus increasing the availability of filaggrin to proteolytic enzymes (Harding and Scott, 1983). The amino acids produced by filaggrin degradation may help to provide the high osmolarity of the outer epidermal layer, which is required for water retention and flexibility (Scott and Harding, 1981; Scott *et al.*, 1982; Horii *et al.*, 1983).

Partial cDNA clones to mouse and rat profilaggrin have been isolated and sequenced (Rothnagel *et al.*, 1987; Haydock and Dale, 1990). Each of these clones encodes a tandem amino acid repeat structure with each encoded repeat corresponding to the respective filaggrin units. Southern analysis of genomic DNA has suggested that the mouse filaggrin coding region does not contain any introns (Rothnagel *et al.*, 1987). Examination of the 5' end of the human profilaggrin gene has shown that an intron is present within the 5' non-coding region (R. Presland, personal communication).

1.4.2 Involucrin

During the terminal differentiation of the epidermal keratinocyte a chemically-resistant protein envelope is formed beneath the plasma membrane. Analysis of the cell envelope has shown that the constituent proteins are extensively cross-linked by ε-(γ-glutamyl)lysine iso-peptide bonds (Sun and Green, 1976; Rice and Green, 1977; Sugawara, 1977, 1979a) which are formed by an epidermal transglutaminase attached to the cell membrane (Thacher and Rice, 1985; Simon and Green, 1985).

Involucrin, a cytoplasmic substrate protein for epidermal transglutaminase, is incorporated into the cross-linked envelope (Rice and Green, 1979). Amino acid analysis determined that 46% of the residues within involucrin are either glutamic acid or glutamine (Rice and Green, 1979), indicating the likely presence of a high number of glutamine residues which could act as substrate amino acids for transglutaminase. Sedimentation and gel filtration studies have indicated that involucrin appears to have a rod-like structure (Rice and Green, 1979).

The human involucrin gene has since been sequenced and found to encode a 68kdal protein which contains 39 tandem repeats of a 10 amino acid sequence (Eckert and Green, 1986). Three of the amino acids within the repeat are glutamines, which could act as substrates for transglutaminase, and two of the remaining residues are glutamic acid. Interestingly, the involucrin gene, like the filaggrin gene (see Section 1.4.1) and the hair cortical IFAP genes (Section 1.2.3d(ii)), does not have an intron within the coding region, although it does have a 1.2 kb intron in the 5' untranslated region (Eckert and Green, 1986). The lemur involucrin gene has subsequently been sequenced and, although the non-repetitive sequences are similar, it contains a differing repetitive region. The lemur repetitive region has 19 tandem repeats of a predominantly 13 amino acid sequence (Tseng and Green, 1988). From comparison of the non-repetitive regions of the human and lemur involucrin genes it would appear that during involucrin gene evolution the repetitive regions of the human and lemur genes have arisen independently from separate sections of the ancestral involucrin gene (Tseng and Green, 1988).

1.5 Intermediate Filaments

1.5.1 IF Classification

The cytoskeleton of the eukaryotic cell contains three components; microtubules (22-25 nm in diameter), microfilaments (5-7 nm) and intermediate filaments (8-10 nm). Although intermediate filaments are present in most vertebrate cells their role is, in many cases, still uncertain. IF proteins range in size from 40-200 kdal and on the basis of their amino acid sequences have been divided into five distinct types (reviewed in Steinert and Roop, 1988; Table 1.4): type I consists of acidic keratins (epithelial tissue); type II consists of neutral-basic keratins (epithelia); type III are vimentin (mesenchymal cells), desmin (muscle) and glial fibrillary acidic protein (astroglia); type IV are neurofilament proteins (neural tissue); and type V consists of the nuclear lamins. In addition, a possible type VI IF protein, nestin, has recently been reported in the developing nervous system (Lendahl et al, 1990).

1.5.2 IF Protein Structure

Partial or complete amino acid sequence information is available for many IF proteins (see Parry and Fraser, 1985; Conway and Parry, 1988). Analysis of these sequences has shown that each IF protein is composed of a central α-helix-rich rod domain which consists of 310-314 (types I-IV) or 352-356 (type V and nestin) residues. The central domain is flanked by amino (N)-terminal and carboxy (C)-terminal domains which are of variable length and character (Fig. 1.10). Variation in the length of the terminal domains largely accounts for the different sizes of the IF proteins. In the organised IF arrays of epithelia and hair the terminal domains are cross-linked to form the hardened structure. Additionally, the properties of the end domains are thought to play a major role in defining the range of functions of the IF family (Steinert *et al.*, 1980, 1983, 1985b).

X-ray diffraction of IFs has produced reflection patterns typical for α-helices arranged in a coiled-coil structure (Fraser *et al.*, 1976; Steinert *et al.*, 1976; Renner *et al.*, 1981) which are of the form first described by Crick (1953). The central core of the IF proteins contains a characteristic heptad repeat (*a-b-c-d-e-f-g*), with *a* and *d* being hydrophobic residues, an arrangement typical for proteins forming two-chained coiled-

Table 1.4 Distribution of IF and IF-like proteins.

1. Proteins were detected using a "universal" IF antibody (Pruss *et al.*, 1981) which is believed to bind to a highly conserved region of the rod domain.

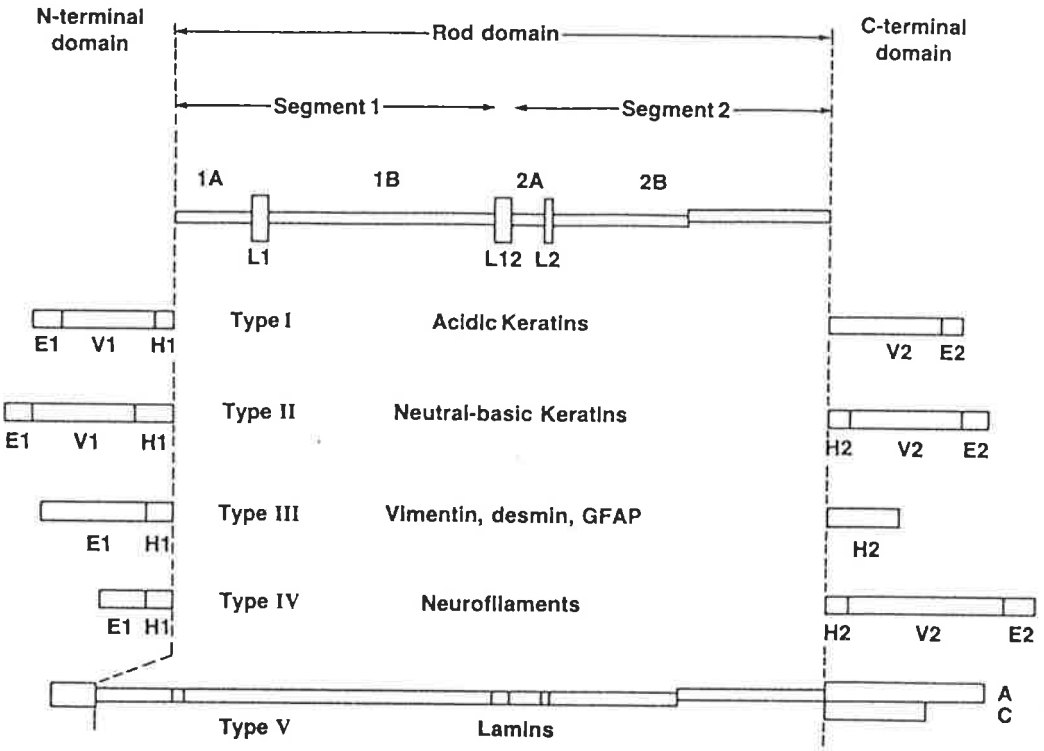
Data from Steinert and Roop (1988).

Origin	Common name	Sequence type	Estimated number of chains	Size (kd)
All epithelia	acidic keratins	I	~15	40-60
All epithelia	neutral-basic keratins	II	~15	50-70
Various "mesenchymal cells"	vimentin	III	1	53
Myogenic cells	desmin	III	1	52
Glial cells and astrocytes	glial fibrillary acidic protein	III	1	51
Most neurones	neurofilaments	IV	≥4 vertebrate ≥2 invertebrate	57-150 60-200
Nuclear lamina of all eukaryotes	lamins	V	≥4 vertebrate ≥1 invertebrate	60-70 60-70
Plant cells	[1]		-	
Yeast and other simple eukaryotes	[1]		-	
Flagellae	tektins	-	3	47-55

Figure 1.10 Organisation of the IF chain subdomains.

The proteins within each of the IF types consists of a central core domain flanked by differing end domains. The central domain of the type V proteins differs from those of types I-IV in that section 1B is longer and the linker regions are α-helical. The end domains can be divided into subdomains based on homologous (H), variable (V) or end (E) sequences.

(Reproduced from Steinert and Roop, 1988.)


N-terminal domain
Rod domain
C-terminal domain
Segment 1
Segment 2
1A
1B
2A
2B
L1
L12
L2
Type I
Acidic Keratins
E1
V1
H1
V2
E2
Type II
Neutral-basic Keratins
E1
V1
H1
H2
V2
E2
Type III
Vimentin, desmin, GFAP
E1
H1
H2
Type IV
Neurofilaments
E1
H1
H2
V2
E2
Type V
Lamins
A
C

coil structures (Parry and Fraser, 1985). The heptad repeat produces an apolar stripe inclined along the surface of each α-helix such that when two α-helices interact they can be stabilised by a regular "knobs-into-holes" packing of the hydrophobic side chains (Crick, 1953).

Four distinct regions of the rod domain contain the heptad repeat and have been designated as 1A, 1B, 2A and 2B (Fig. 1.10). These are interspersed by non-α-helical linkers termed L1 (joins 1A to 1B), L2 (2A to 2B) and L12 (1B to 2A). Together, the helical and linker regions produce a rod domain with an overall length of about 45 nm (Steinert *et al.*, 1983; Crewther *et al.*, 1983) which is close to the 47 nm axial repeat deduced from X-ray diffraction analysis (Fraser and MacRae, 1973).

Intermediate filaments appear to have the ability to assemble themselves without the aid of additional proteins. Keratin IFs are obligate heteropolymers, requiring both a type I and a type II chain for IF assembly both in vivo and in vitro (Steinert *et al.*, 1976, 1982; Hatzfeld and Franke, 1985). In contrast, type III and some type IV proteins have been shown to be able to assemble themselves into homopolymers (Cabral *et al.*, 1981; Geisler and Weber, 1981; Steinert *et al.*, 1981b, Liem and Hutchinson, 1982; Quinlan and Franke, 1982; Zackroff *et al.*, 1982).

The basic IF structural unit is believed to be a tetramer, consisting of a pair of two-stranded coiled-coil molecules (Geisler and Weber, 1982; Gruen and Woods, 1983; Woods and Inglis, 1984). From analysis of the heptad repeat lengths and possible ionic interactions it has been predicted that each coiled-coil molecule contains two proteins that are arranged in a parallel fashion and in register (Parry *et al.*, 1977; Steinert *et al.*, 1984, 1985a; Parry and Fraser, 1985; Parry *et al.*, 1985; see Fig. 1.11b). This has been shown for keratin IF (Woods and Inglis, 1984; Parry *et al.*, 1985) and the nuclear lamins (Aebi *et al.*, 1986). The arrangement of the coiled-coil pairs within the tetramer is less clearly understood. Although numerous models have been suggested, present data favours those in which the two coiled-coils are arranged in either an in register or half-staggered antiparallel fashion (for review, see Steinert and Roop, 1988; Fig 1.11c). Additionally, the organisation of the tetramers into an 8-10 nm diameter IF is also unclear, though analysis of scanning transmission electron microscope images has suggested that the major form of type I-III IFs contains 32 protein chains, i.e. 8 tetramers, in cross-section (Steven *et al.*, 1982; Eichner *et al.*, 1985; Aebi *et al.*, 1985).

Figure 1.11 Organisation of IF oligomers.

(a) The universal IF protein chain.

(b) Two-chain coiled-coil molecule formed by the alignment of two in-register parallel chains.

(c) Favoured models for the four-chain IF complex formed by either in-register (1) or half-staggered (2) anti-parallel complexes.

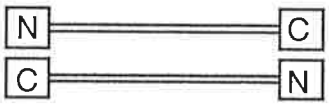
(Adapted from Steinert and Roop, 1988.)

a.

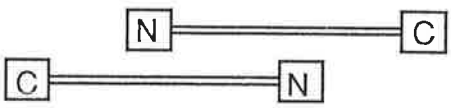
b.

c.

1



2



1.5.3 IF Gene Structure

In order to understand the evolution and expression of the IF proteins, much research has been directed towards isolating and analysing IF genes (see Steinert and Roop, 1988). Numerous type I-IV IF genes have been analysed and the positions of introns within representatives of all four types are shown in Figure 1.12. The type I-III IF genes all have a very similar organisation with the number and position of introns being relatively conserved both within and between types. On the other hand, the type IV genes are organised quite differently with fewer introns placed at different positions to any of the introns in the type I-III genes (Fig. 1.12).

It should be noted that although many genes have been sequenced, very little is known of the gene sequences involved in the control of gene expression.

1.5.4 Intermediate Filament Associated Proteins

Although IFs are autonomously assembled from their constituent proteins, the aggregation of IFs or the interaction of IFs with other regions of the cell requires auxiliary proteins, termed intermediate filament associated proteins (IFAPs). As shown in Table 1.5, the IFAPs can be sub-divided into four classes on the basis of their size and function (see Steinert and Roop, 1988).

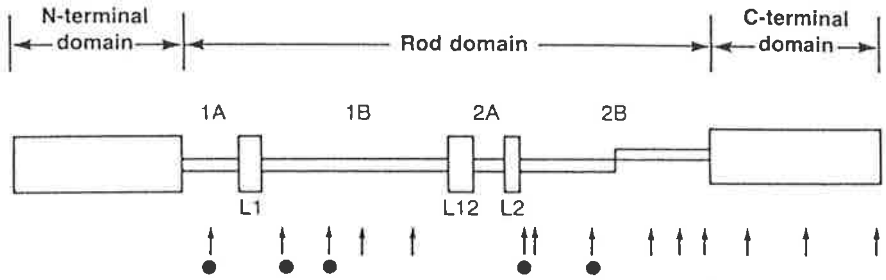
Class 1 consists of relatively low molecular weight proteins (6,000 - 45,000) involved in forming tightly bound IF aggregates. Examples include the high-sulphur and high-glycine/tyrosine matrix proteins of the hair fibre cortex (see Section 1.2.3) and filaggrin in the keratinised epidermis (see Section 1.4.1)

The second class of IFAPs contains high molecular weight proteins which cross-link IF into loose aggregates. This group includes paranemin (Price and Lazarides, 1983) and synemin (Granger and Lazarides, 1980) which both organise certain vimentin and desmin networks, microtubule-associated proteins (MAPs) which help aggregate neurofilaments (Williams and Aamodt, 1985), a 300 kdal protein present in many cultured fibroblasts (Lieska *et al.*, 1985; Yang *et al.*, 1985), and plectin which is believed to be identical to the 300 kdal protein (Herrmann and Wiche, 1987) and has been detected in numerous tissues and cell types (Wiche *et al.*, 1983). It has also been suggested that two of the neurofilament proteins, NF-M and NF-H, may also act as IFAPs (Liem and Hutchinson, 1982; Hirokawa *et al.*, 1984), although the presence of a characteristic IF

Figure 1.12 Location of the introns within the types I-IV IF genes.

The positions of introns within a number of the types I-IV IF genes (indicated by arrows) are shown relative to a model of an IF chain. The numbers refer to the residue position within the particular α-helical region or the C-terminal domain. Arrows with dots indicate introns which splice at the codon for the first residue in a heptad.

(Reproduced from Steinert and Roop, 1988.)



Mouse 59 kD (I)		14	42		94	4	46		119			113
Bovine 54 kD (I)		14	42		94	4	46		119			84
Human K14 52 kD (I)		14	42		94	4	46		120	15		
Human K17 46 kD (I)		14	42		94	4	46		120	15		
Bovine K19 44 kD (I)		14	42		94	4	46					
Human K1 67 kD (II)	17		42	62	94	7	46		119	12		
Human K3 63 kD (II)	17		42	62	94	4	46		119	12		
Human K6b 60 kD (II)	17		42	62	94	4	46		119	11		
Hamster Vimentin 53 kD (III)			42	62	94	4	46		120	14	42	
Hamster Desmin 52 kD (III)			42	62	94	4	46		120	14	41	
Mouse GFAP 50 kD (III)			42	62	94	4	46		120	14	42	
Mouse NF-L 68 kD (IV)								79111				95
Human NF-L 68 kD (IV)								79111				95
Human NF-M 105 kD (IV)								79111				

Table 1.5 Intermediate filament associated proteins.

[1,2,3] These are probably the same proteins

Data from Steinert and Roop (1988) and Ciment *et al.* (1986).

Protein name	Distribution	Size (kd)
1. Lateral aggregation		
high-sulphur family	hair and related tissues	10-30
high-glycine/tyrosine family	hair and related tissues	6-10
filaggrin family	orthokeratinising epithelia	16-45
2. Cross-linking		
paranemin	avian erythrocytes and muscle tissue	280
synemin	avian erythrocytes and muscle tissue	230
MAPs 1,2	neuronal tissue	300
300 kd[1]	fibroblasts	300
plectin[1]	fibroblasts and various other tissues	300
NF-M, NF-H (?)	neuronal tissues	105,305
3. Capping		
ankyrin	ubiquitous	220-240
desmoplakin	desmosomes, hemidesmosomes	220-240
lamin B	ubiquitous	65
spectrin (?)	ubiquitous	220-240
4. Others		
epinemin[2]	vimentin networks in many cultured cells	45
p50[2]	vimentin networks	50
p68[3]	vimentin networks	68
p95	vimentin networks	95
ß-internexin[3]	heat-shock proteins (?)	70
NAPA-73	avian neurones	73

rod domain indicates that they should be involved in the formation of filaments. Additionally, synemin (Granger and Lazarides, 1980) and the 300 kdal protein (Lieska *et al.*, 1985) may also contain an α-helical core, suggesting that this region may be integrated into IFs allowing the large end-domains to form spacers in the generation of the IF network (Steven *et al.*, 1985; Steinert and Roop, 1988). Trichohyalin, whether it forms both the IFs and the matrix of the follicle IRS or the matrix alone, will also belong to this class of IFAPs.

Class 3 IFAPS are involved in the anchoring of the ends of IFs. This class includes ankyrin (Georgatos and Marchesi, 1985; Georgatos *et al.*, 1985), the desmoplakins (Kartenbeck *et al.*, 1984; Jones and Goldman, 1985; Jones *et al.*, 1986), the lamin B chain of the nuclear lamina complex (Georgatos *et al.*, 1987; Georgatos and Blobel, 1987), and possibly spectrin (Granger and Lazarides, 1984; Mangeat and Burridge, 1984).

The final class involves proteins which do not appear to belong to any of the earlier three classes. Such proteins are epinemin (Lawson, 1983), ß-internexin (Napolitano *et al.*, 1984), NAPA-73 (Ciment *et al.*, 1986) and three proteins, p50 (Wang *et al.*, 1983), p68 (Wang *et al.*, 1980, 1981) and p95 (Lin and Feramisco, 1981), shown to be involved in vimentin networks.

1.6 Aims of the Project

The role of trichohyalin within the development of the hair follicle IRS has long evaded understanding. During the majority of IRS cell development trichohyalin is stored within large non-membrane bound granules. At a given stage within the cell maturation the granules disappear and are replaced by an organised filamentous arrangement which contains trichohyalin. What then is the function of trichohyalin within the filamentous array of the hardened IRS cell? Does it produce the IRS filaments or, does it act as an IFAP or, is it both an IF and an IFAP (see Fig. 1.13)? The overall aim of this thesis was to determine which of these roles is performed by trichohyalin based on its protein sequence and gene structure, and subsequently also to understand its role in the development of the hair follicle medulla.

The detailed aims were:

1. To obtain partial amino acid sequence from the trichohyalin protein to aid in the identification of cDNA clones.

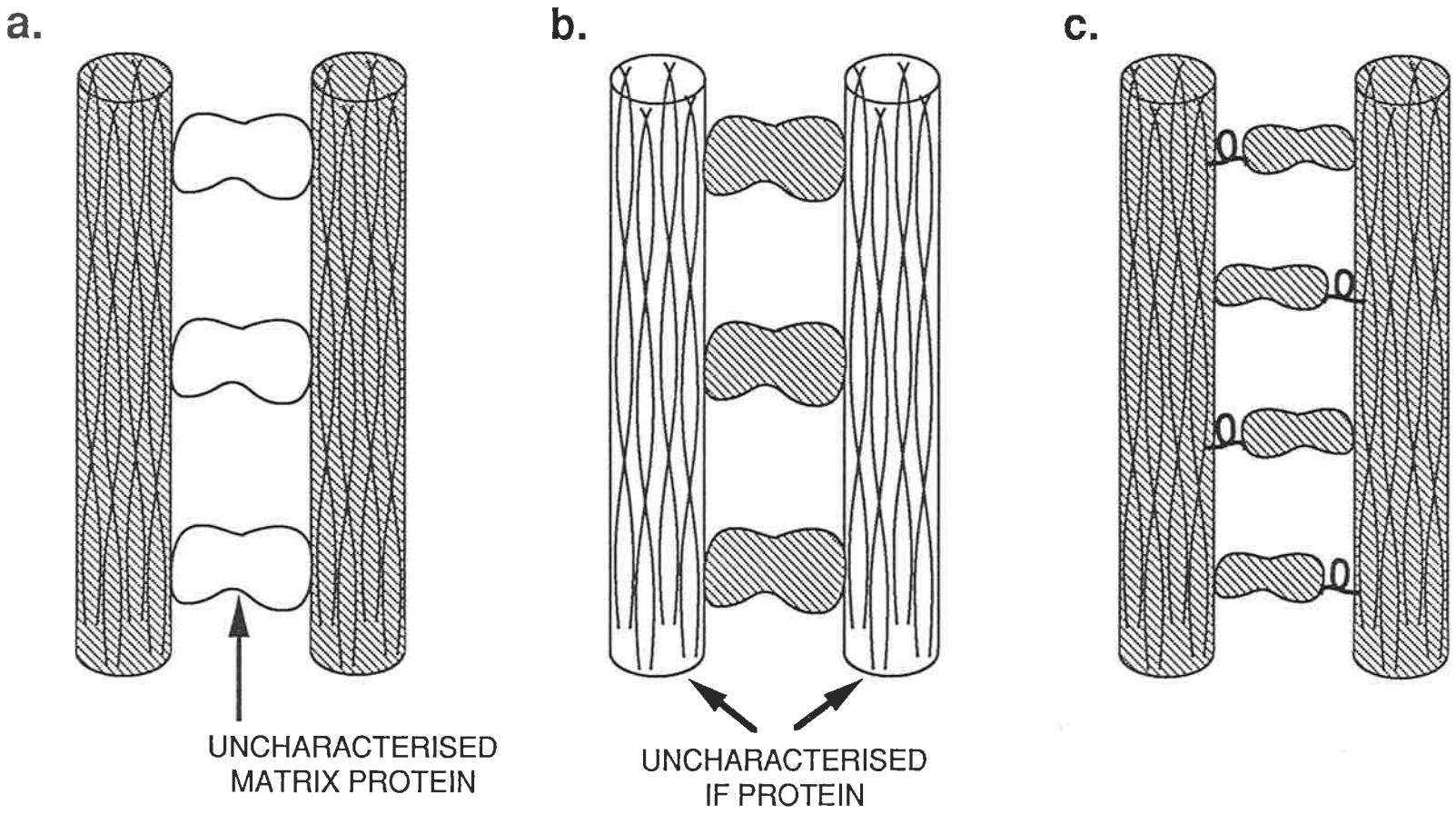
Figure 1.13 Possible roles of trichohyalin in the hardened IRS.

(a) Trichohyalin acts as a filamentous protein.

(b) Trichohyalin acts as an IFAP.

(c) Trichohyalin acts as both a filamentous and a matrix protein.

In each case the structure formed by trichohyalin is shaded.


a.
b.
c.
UNCHARACTERISED
MATRIX PROTEIN
UNCHARACTERISED
IF PROTEIN

2. To sequence a sheep follicle trichohyalin cDNA clone and subsequently to purify, sequence and analyse the sheep trichohyalin gene.

3. To perform in situ hybridisation analysis on various keratinised sheep tissues in order to determine whether trichohyalin is specifically expressed within the hair follicle.

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Tissue

Ovine protein and RNA samples used in the experiments described in this thesis were isolated from follicles obtained from domestic sheep Ovis aries of the breeds Merino-Dorset Horn x Border Leicester or Merino x Romney Marsh. Protein samples were also obtained from the follicles of guinea pigs. Experiments performed on animals were approved by the Animal Ethics Committee of the University of Adelaide.

2.1.2 Bacterial Strains

The following E. coli K12 strains were used in the experiments described in this thesis:

BB4: supF58 supE44 hsdR514 galK2 galT22 trpR55 metB1 tonA deltalacU169 F' [proAB$^+$ lacI lacZdeltaM15 Tn10(tet^r)] (Bullock *et al.* 1987)

ED8799: hsdS metB7 supE (glnV)44 supF (tyrT)58 lacZdeltaM15 r_k- m_k- (gift from Dr. S. Clarke, Biotechnology Australia)

JM101: supE thi delta(lac-proAB) F' [traD36 proAB$^+$ lacI lacZdeltaM15] (Messing, 1979)

LE392: supE44 supF58 hsdR514 galK2 galT22 metB1 trpR55 lacY1 (Borck *et al.*, 1976; Murray *et al.*, 1977)

2.1.3 Bacteriophage Strains

M13mp18 and M13mp19 (Norrander *et al.*, 1983) were used for subcloning and sequencing of restriction and Bal 31 deletion fragments.

2.1.4 Phagemid Strains

pGEM-3Zf(+) and pGEM-7Zf(+) were obtained from Promega Corporation and used for the subcloning of genomic fragments and for performing Exonuclease III deletions.

2.1.5 Plasmid Strains

pUC19: a derivative of pBR322 (Yanish-Peron *et al.*, 1985).

pGEM-1 and pGEM-2: derivatives of pSP65 and pSP64 respectively (Melton *et al.*, 1984) and were obtained from Promega Corporation.

2.1.6 Enzymes

Restriction endonucleases were purchased from either Toyobo Chemical Co. or Pharmacia.

E. coli DNA polymerase I (Klenow fragment), T4 DNA ligase and SP6 RNA polymerase were purchased from Bresatec Ltd.

Lysozyme, Ribonuclease T1 (RNase T1) and Ribonuclease A (RNase A) were purchased from Sigma Chemical Co..

T7 RNA polymerase, Exonuclease III and Taq Polymerase were purchased from Promega Corporation.

Endoproteinase Lysine-C was purchased from Boehringer Mannheim.

Bal 31 exonuclease was purchased from New England Biolabs.

2.1.7 Radiochemicals

[α-32P]dATP (specific activity, 3,000 Ci/mmole), [α-32P]dCTP (specific activity, 3,000 Ci/mmole) and [α-35S]UTP (specific activity, 1,000-1,500 Ci/mmole) were purchased from Bresatec Ltd.

2.1.8 Molecular Biology Kits

Dideoxy sequencing kits (DSK-A), oligo-labelling kits (OLK-A and OLK-C) and nick translation kits (NTK-B) were purchased from Bresatec Ltd.

Erase-a-Base and TaqTrack kits were purchased from Promega Corporation.

2.1.9 General Chemicals

The following chemicals were purchased from the Sigma Chemical Co.: acrylamide and bis-acrylamide, ampicillin, EDTA (ethylenediaminetetraacetic acid), goat anti-rabbit IgG/alkaline phosphatase conjugate, guanidine hydrochloride (grade 1), IPTG (isopropylthiogalactoside), 2-mercaptoethanol, salmon sperm DNA, SDS (sodium dodecyl sulphate), tetracyclin and Tween-20 (polyoxyethylenesorbitan monolaurate).

HPLC grade acetonitrile was purchased from Waters Associates Pty. Ltd..

Agarose (type 1) was obtained from BRL.

Polyethylene glycol (6,000) and formamide were purchased from BDH Laboratories.

Dextran sulphate and ficoll were purchased from Pharmacia.

Ultra-pure urea was obtained from Merck.

Both liquid and powder Vertex denture material was purchased from Dentimex Zeist, Holland.

BCIG (5-bromo-4-chloro-3-indolyl-3-D-β-galactoside) was purchased from Boehringer Mannheim.

BCIP (5-bromo-4-chloro-3-indolyl phosphate) and NBT (nitroblue tetrazolium) were purchased from Bresatec Ltd.

General chemicals not listed above were obtained from one of the following suppliers: Ajax Chemicals Pty. Ltd., BDH Chemicals Pty. Ltd., May and Baker Pty. Ltd., Merck, Pharmacia or Sigma Chemical Co. Chemicals were of the highest purity available.

2.1.10 Media and Buffers

a. Growth Media

The following media were made as described in Miller (1972).

LB Medium - used for the growth of E. coli ED8799 containing plasmids and phagemids.

Minimal Media - used for the growth of E. coli JM101.

Tryptone Broth - used for the growth of E. coli LE392.

2x YT - used for the growth of E. coli ED8799 and JM101 during the preparation of single-stranded phage.

SOC Medium was prepared as described by Sambrook *et al.* (1989).

Agar plates and agar/agarose overlay were made using LB medium, minimal medium or tryptone broth as described by Maniatis *et al.* (1982). Antibiotics were added where appropriate as described by Maniatis *et al.* (1982).

b. Buffers

2x SDS Loading Buffer - 125 mM Tris-HCl pH 6.8, 4% SDS, 20% glycerol, 2 mM EDTA, 2% 2-mercaptoethanol

SSC - 150 mM NaCl, 15 mM Na citrate, 1 mM EDTA pH 7.0

SSPE - 150 mM NaCl, 20 mM sodium dihydrogen phosphate, 1mM EDTA pH 7.4

TAE - 40 mM Tris-acetate pH 8.2, 20 mM Na acetate, 1 mM EDTA

TBE - 130 mM Tris, 50 mM Boric acid, 2.5 mM EDTA pH 8.3

TE - 10 mM Tris-HCl pH 8.0, 0.1 mM EDTA

2.1.11 Miscellaneous

The butyl reverse phase HPLC cartridge was purchased from Brownlee Labs.

Sepharose CL-4B medium and the Superose 12 column were obtained from Pharmacia.

Dialysis tubing was obtained from BDH Chemicals Ltd.

Freund's Complete Adjuvant was obtained from CSL.

YM10 and CF25 ultrafiltration membranes were purchased from Amicon Corporation.

Nitrocellulose and Nytran membranes were obtained from Schleicher and Schuell.

Zetaprobe was purchased from Bio-Rad.

X-ray film was obtained from Konica Corporation or Fuji Photo Film Corporation.

Sheep genomic DNA was a gift from Dr. G. Cam.

The Zoo blot, containing sheep, mouse and human genomic DNA cut with EcoR I, together with human, mouse and sheep skin sections and sheep hoof and tongue sections, were gifts from Dr. B. Powell.

Sheep rumen and oesophageal sections were gifts from Ms. L. Whitbread.

Ektachrome film and D19 developer (used for in situ hybridisations) were purchased from Kodak Ltd.

L4 emulsion (used for in situ hybridisations), Hypam Rapid fixer, and Pan F film were purchased from Ilford Ltd.

2.2 Methods

2.2.1 Collection of Follicle Tissue

Intact wool follicles were obtained using an adaptation of the method of Rothnagel and Rogers (1986). The sheep from which the follicles were to be removed was injected

intravenously with the general anaesthetic Nembutal. Once the sheep was unconscious, its flank was sheared using Oster clippers fitted with a size 40 comb. A mixture containing equal amounts of Vertex powder and liquid was applied at numerous sites on the exposed side of the sheep, with each application covering an area of approximately 8cm x 5cm. Each of these applications was immediately overlaid with a strip of glassfibre tape. Once the mix had set, the strips were removed and placed in liquid nitrogen. The strips, which contained the follicular material, were used immediately or stored at -80°C for future use.

2.2.2 Protein Methods

a. Extraction of Follicle Proteins

Frozen follicular material was scraped into 7 M guanidine-HCl, 50 mM Tris-HCl pH 7.5, 1 mM EDTA (10 ml buffer/g tissue) and stirred for 15 minutes at 0°C (Rothnagel and Rogers, 1986). The extract was centrifuged at 38,000 g for 10 minutes to remove insoluble follicular material and the supernatant was filtered through 0.2 µm Sartorius disposable filters to remove any contaminating follicles. The extract was then stored at -20°C.

b. Concentration of Protein Samples

Protein-containing samples, to which 2-mercaptoethanol had been added to 1%, were concentrated by either ultrafiltration under nitrogen pressure through an Amicon YM10 membrane or by centrifugal ultrafiltration through an Amicon CF25 filtration cone. If the sample contained guanidine-HCl and was to be used subsequently for SDS polyacrylamide gel electrophoresis, the concentrated sample was diluted three-fold in 8 M urea, 50 mM Tris-HCl pH 7.5, 1 mM EDTA, 1% 2-mercaptoethanol and was then re-concentrated. This step was repeated twice in order to reduce the concentration of guanidine to a sufficiently low level for subsequent polyacrylamide gel electrophoresis.

c. Gel Filtration Chromatography

Follicular protein extracts were chromatographed on a CL-4B gel filtration column (Pharmacia) by the method of Rothnagel and Rogers (1986). The column was equilibrated with 7 M guanidine-HCl, 50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM EDTA. 2 ml of concentrated follicular extract were chromatographed at 8 ml/h and 8 ml fractions were collected.

d. Dialysis

Dialysis tubing was boiled in 1% (w/v) sodium hydrogen carbonate, 1% (w/v) EDTA for 2-5 minutes to remove impurities and then washed in water before use (Thompson and O'Donnell, 1965). Samples were dialysed exhaustively against water at 4°C.

e. SDS Polyacrylamide Gel Electrophoresis

Polyacrylamide slab gels were prepared essentially by the method of Laemmli (1970). Samples were diluted with an equal volume of 2x SDS loading buffer and heated to 80°C for 5 minutes prior to loading.

f. Staining of Protein Gels

Proteins which had been fractionated on a polyacrylamide gel were detected with either Coomassie Brilliant Blue (Swank and Munkres, 1971) or by using a positive-image silver stain (Merril *et al.*, 1984).

g. Estimation of Protein Concentration

The concentration of individual proteins within a sample was determined by analysis of stained polyacrylamide gels. The thickness and intensity of staining of the desired band was compared with those of known amounts of chicken liver pyruvate carboxylase which had also been separated by polyacrylamide gel electrophoresis and stained with Coomassie Brilliant Blue.

h. Cleavage with Endoproteinase Lysine-C

Samples to be used for proteolysis were prepared in 2 M urea, 50 mM Tris-HCl pH 8.8, 1 mM EDTA, 1% 2-mercaptoethanol. Enzyme was added to 1% of the substrate mass and the reaction placed at 37°C. After an hour a further 1% of enzyme was added and the reaction left at 37°C overnight. The reaction was stopped by the addition of TLCK to 2 mM.

i. Cleavage with Cyanogen Bromide

Cleavage reactions were initially performed by the method of Walker and Mayes (1983) using a 100-fold excess of cyanogen bromide over the number of expected methionine residues. Later attempts used up to a 500-fold excess of cyanogen bromide.

j. Polyclonal Antibodies

Rabbit polyclonal antibodies were raised against purified sheep trichohyalin according to established procedures (Brown, 1967). Non-denaturing Ouchterlony

immunodiffusion was performed in 50 mM sodium phosphate pH 7.2, 0.02% sodium azide, 1% agarose as described by Radding and Shreffler (1966).

k. Western Transfer

After SDS polyacrylamide gel electrophoresis of samples, protein bands were transferred to nitrocellulose by the method of Svoboda *et al.* (1985) and reacted with antibodies. The anti-trichohyalin antisera was diluted 1 in 1000 before use. Bound antibody was detected by incubation with a goat anti-rabbit IgG/alkaline phosphatase conjugate followed by colorimetric staining (Forster *et al.*, 1985).

l. Gel Filtration FPLC

Peptides produced by cleavage of guinea pig trichohyalin with endoproteinase lysine-C were separated by gel filtration FPLC which was performed on a Pharmacia Superose 12 column (300mm x 10mm internal diameter). The column was equilibrated with 2 M urea, 50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM EDTA, chromatographed at a flow rate of 0.2 ml/min, and 0.5 ml fractions were collected.

m. Reverse Phase HPLC

Trichohyalin proteolytic peptides were separated by reverse phase HPLC which was conducted on a Brownlee Labs Aquapore Butyl cartridge (30mm x 2.1 mm internal diameter) equilibrated with 0.1% trifluoroacetic acid and chromatographed at 0.2 ml/min. Various gradients of acetonitrile were used in the different separations.

n. Protein Sequencing

The amino-terminal sequences of peptides were determined by automated Edman degradation with an Applied Biosystems gas phase sequencer (Hunkapiller *et al.*, 1983).

o. Amino Acid Analysis

Amino acid analyses of sheep and guinea pig trichohyalin were kindly performed by Dr. R.C. Marshall under standard conditions.

2.2.3 DNA Methods

a. General Methods

Ethanol precipitation was conducted by the method of Zeugin and Hartley (1985). Phenol extraction was conducted as described by Maniatis *et al.* (1982).

b. Screening a Cosmid Library

The sheep genomic cosmid library, which was a gift from Dr. G. Cam, was screened at a high density (approx. 30,000 colonies per 150 mm filter) for recombinant clones using the method of Hanahan and Meselson (1980).

c. Screening a Lambda Library

Sheep genomic λ libraries were screened at a high density (approx. 40,000 plaques per 150 mm filter) by the plaque hybridisation method described by Benton and Davies (1977). Positive plaques were re-screened until all of the plaques on a plate hybridised to the probe.

d. DNA Preparation

Recombinant plasmid DNA was prepared by the modified procedure of Birnboim and Doly (1979) described by Maniatis *et al.* (1982) using cesium chloride density gradient equilibrium centrifugation (Radloff *et al.*, 1967) for final purification of the plasmid DNA.

Small-scale preparation of plasmid DNA was conducted essentially by the method of Ish-Horowicz and Burke (1981).

Recombinant phage DNA was prepared by the liquid culture method of Kao *et al.* (1982).

e. Cleavage by Restriction Endonucleases

DNA was cleaved with restriction endonucleases using the conditions described by the manufacturer for each of the used enzymes. Usually 2-5 units of enzyme were used per μg of DNA and reactions were performed for approximately 2 hours.

f. Agarose Gel Electrophoresis

DNA was separated by agarose gel electrophoresis, which was performed in a horizontal apparatus, essentially by the method described by Maniatis *et al.* (1982). Analytical electrophoresis was performed on gels of 75mm x 55 mm whilst larger gels were used for DNA preparation and Southern analysis.

All gels were run in TAE buffer and samples were loaded in 2.5% Ficoll (type 400), 0.1% lauryl sarkosyl, 0.025% bromophenol blue, 0.025% xylene cyanol. Electrophoresis was performed at 60-100 mA. After electrophoresis DNA was detected by staining with 0.1% ethidium bromide for 5 minutes and viewing under short wave UV

light. Gels were photographed using a Polaroid camera loaded with either type 665 (ASA 80) or 667 (ASA 3000) Polaroid film.

g. Isolation of DNA from Agarose Gels

Restriction fragments ranging in size from 500 bp to 8 kb were extracted from agarose using a "Gene-clean" kit obtained from BIO 101 according to the manufacturers instructions. This method is based on the procedure described by Vogelstein and Gillespie (1979).

Smaller fragments, and those prepared by deletion with Bal 31 exonuclease, were eluted from the agarose in 0.5 M ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA, 0.1% SDS at 37°C for 16 hours. The DNA was then collected by precipitation with ethanol (Section 2.2.3a).

h. DNA Subcloning

(i) Vector Preparation

Plasmid, phagemid and M13 cloning vectors were prepared by linearisation with the appropriate restriction endonuclease(s) followed by removal of the 5' terminal phosphate groups as described by Maniatis *et al.* (1982). The vectors were then subjected to agarose gel electrophoresis, the linear vector band excised and the DNA prepared using the "Gene-clean" kit (Section 2.2.3g).

(ii) Ligation

Approximately 40 ng of vector DNA was combined with the desired DNA fragment in a 1:1 molar ratio, as estimated by the intensity of staining with ethidium bromide. The ligations were performed in 50 mM Tris-HCl pH 7.4, 10 mM magnesium chloride, 10 mM dithiothreitol, 1 mM ATP, 100 µg/ml BSA. Approximately 1 unit of T4 DNA ligase was added and ligation was allowed to proceed at either 4°C overnight or at room temperature for 4 hours.

(iii) Plasmid Transformation

E. coli strain ED8799 cells were made competent and transformed by a modification of the calcium chloride method of Sambrook *et al.* (1989) in which the first cell resuspension performed during the formation of competent cells was carried out in 0.1 M magnesium chloride rather than 0.1 M calcium chloride.

One half of the ligation mix was normally used for each transformation.

(iv) Phagemid Transformation

E. coli strain BB4 cells were made competent and transformed by a modification of the method used during plasmid transformation (Section 2.2.3h(iii)) in which the cells to be made competent were resuspended once rather than twice and this resuspension was performed in 0.1 M calcium chloride, 20 mM magnesium chloride.

(v) M13 Transfection

E. coli strain BB4 cells were made competent by the same procedure used for phagemid transformation (Section 2.2.3h(iv)). The transfection was carried out by a modification of the method of Sambrook *et al.* (1989) in which 0.9 ml SOC medium was added to the transfected cells immediately before they were plated out in melted LB agar containing IPTG and BCIG.

i. Preparation of Labelled DNA

(i) Oligo-Labelling

DNA restriction fragments were oligo-labelled using a kit obtained from Bresatec Ltd. according to the method provided by the manufacturers which is based on the procedure of Feinberg and Vogelstein (1983).

The labelled DNA was purified using the "Gene-clean" kit (Section 2.2.3g) or by Sepharose G-50 chromatography (Maniatis *et al.*, 1982).

(ii) Nick Translation

The nick translation of DNA was performed by the method of Rigby *et al.* (1977) with a kit purchased from Bresatec Ltd. used according to the manufacturers instructions.

The labelled DNA was purified by Sepharose G-50 chromatography (Maniatis *et al.*, 1982).

j. Deletions with Nuclease Bal 31

Cloned DNA was deleted with nuclease Bal 31 according to a modification of the procedure of Sambrook *et al.* (1989). In the modified procedure the deleted DNA was not end-filled after the Bal 31 deletions and relied upon the presence of blunt ends within a proportion of the deleted DNA fragments.

k. Deletions with Exonuclease III

Insert DNA which had been cloned into the desired phagemid, was deleted with Exonuclease III using the Erase-a-Base kit of Promega Corporation according to the manufacturers protocol.

l. Preparation of Single-stranded M13 DNA

Single-stranded M13 template DNA for dideoxy sequencing was prepared essentially as described by Winter and Fields (1980).

m. Preparation of Single-stranded Phagemid DNA

Single-stranded phagemid DNA was produced essentially as described by Sambrook *et al.* (1989) and the template DNA was then prepared in an identical fashion to the M13 template DNA (Section 2.2.3l).

n. DNA Sequencing

Single-stranded M13 and phagemid template DNA was sequenced using the dideoxy chain termination method (Sanger *et al.*, 1980; Messing *et al.*, 1981). The majority of the reactions were performed with the Klenow fragment of E. coli DNA Polymerase I using a kit obtained from Bresatec Ltd. essentially according to the manufacturers instructions.

A number of sequencing reactions were conducted with Taq polymerase using a TaqTrack kit purchased from Promega Corporation. These reactions, which used 7-deaza dGTP rather than dGTP, were performed according to the manufacturers instructions.

The products produced by both kits were denatured and electrophoresed on a 0.25 mm thick 6% denaturing polyacrylamide gel. The electrophoresis was performed at 1,000-1,500 V using TBE as the running buffer. After electrophoresis the gel was fixed in 12% acetic acid, washed in 20% ethanol and dried at 110°C. The gel was then autoradiographed at room temperature in the absence of an intensification screen.

o. Southern Transfer

Plasmid and genomic DNA, which had been cleaved by a restriction endonuclease(s), were transferred to Bio-Rad Zetaprobe membrane by the modified method of Southern (1977) reported by Reed and Mann (1985). The DNA was transferred from the gel to Zetaprobe using 0.4 M sodium hydroxide.

The filters were hybridised according to the instructions for Gene Screen filters.

The stringency of the individual washing conditions is reported for each of the hybridised filters.

The filters were autoradiographed at -70°C in the presence of an intensification screen.

Filters which were to be reused were washed in boiling 0.1x SSC, 0.1% SDS and shaken for fifteen minutes. This wash was then repeated and the filter autoradiographed prior to reuse.

2.2.4 RNA Methods

a. RNA Preparation

RNA was isolated from sheep wool follicles using the acid guanidine thiocyanate-phenol-chloroform extraction procedure of Chomczynski and Sacchi (1987).

b. Northern Transfer

(i) Glyoxal Gels

Glyoxolated RNA was fractionated through 1% agarose as described by Thomas (1983) and transferred to Zetaprobe by a modification of the method of Southern (1977). RNA was transferred to Zetaprobe using 5 mM sodium hydroxide as the transfer buffer. After transfer the filter was washed twice in 0.2x SSC, 0.1% SDS prior to hybridisation. Hybridisation and washing conditions were the same as those described for Southern transfer (Section 2.2.3o).

(ii) Formaldehyde Gels

Sheep follicle RNA was fractionated in a vertical agarose/formaldehyde gel according to the protocol of Hansen *et al.* (1989). RNA was transferred to Schleicher and Schuell Nytran membrane by capillary transfer using 10x SSC as the transfer buffer. The Nytran membrane was hybridised and washed using the conditions of Hansen *et al.* (1989).

c. In vitro Transcription

Inserts in either pGEM-1, pGEM-2 or pGEM-7Zf(+) were transcribed with either SP6 or T7 RNA polymerase according to the method of Krieg and Melton (1987) using a kit purchased from Bresatec Ltd. The RNA was labelled to high specific activity by the incorporation of [α-^{35}S]UTP.

Transcripts were phenol extracted and ethanol precipitated prior to further use.

d. Tissue in situ Hybridisation

The in situ hybridisation procedure was based on the method of Cox *et al.* (1984) with the modifications of Powell and Rogers (1990b). Tissue sections were stained using the SACPIC procedure of Auber (1950).

2.2.5 Computer Programmes

a. DNA and Protein Sequence Analysis

DNA sequence data was compiled and analysed on a VAX 11-785 computer using the programs ANALYSEQ (Staden, 1984), DIAGON (Staden, 1982) and a number of programs from the Sequence Analysis Software Package of the University of Wisconsin Genetics Computer Group (UWGCG) (Devereux *et al.*, 1984).

b. Database Searches

Searches of the Genbank, National Biomedical Research Foundation and Swiss Protein databases were performed using the UWGCG programmes FastA and TFastA (Devereux *et al.*, 1984) as well as the programmes MATCH, MATCH FAST and MATCH TRANSLATE (Wilbur and Lipman, 1983; Lipman and Pearson, 1985).

c. Secondary Structure Analysis

Secondary structure analysis was performed using the programme PREDICT which is a suite containing 10 secondary structure programmes assembled at the Dept. of Biophysics, University of Leeds. The component programmes are:-

Burgess (Burgess *et al.*, 1974).

Dufton (Dufton and Hider, 1977).

Fasman (Chou and Fasman, 1974).

Garnier (Garnier *et al.*, 1978).

Kabat (Kabat, 1973).

Lim (Lim, 1974a,b).

McLachlan.

Nagano (Nagano, 1973).

Joint (Eliopoulos *et al.*, 1982).

Seplot.

The output of the first 8 programmes was processed by JOINT producing a joint secondary structure prediction. This output was used by SEPLOT to produce a histogram of the predicted structure.

2.2.6 Containment Facilities

All work involving recombinant DNA was carried out under C1 containment conditions required for work involving viable organisms or C0 conditions required for

work not involving viable organisms, as defined and approved by the Australian Academy of Science Committee on Recombinant DNA and by the University Council of the University of Adelaide.

Chapter 3

Trichohyalin Peptide Purification

3.1 Introduction

In 1986, Rothnagel and Rogers reported the isolation of a 190 kdal sheep wool follicle protein which they termed trichohyalin. Trichohyalin corresponded to a previously described protein fraction shown to be located in the granules of the developing IRS and medulla which was incorporated into the hardened structures of both tissues (Rogers *et al.*, 1977; Section 1.3.1). In order to further examine its function and properties, Rothnagel and Rogers purified trichohyalin using a procedure which involved the extraction of follicular proteins with a guanidine-HCl buffer. After separation of the extracted proteins by gel filtration chromatography, up to 70% of the eluted trichohyalin was found to be at least 90% pure.

The research reported in this chapter was aimed at obtaining a partial amino acid sequence for trichohyalin. The determination of some sequence is essential in the purification of cDNA clones for trichohyalin, either to derive a complimentary oligodeoxynucleotide probe for the selection of cDNA clones or in the confirmation of the identity of separately purified cDNA clones. To obtain the amino acid sequence it was necessary to cleave the trichohyalin molecule such that some of the resultant peptides could be purified and sequenced. This approach required large amounts of pure trichohyalin and this chapter first describes improvements to the trichohyalin purification procedure of Rothnagel and Rogers (1986). It then reports on the proteolytic cleavage of the purified trichohyalin, purification of proteolytic peptides and discusses the resultant peptide sequences.

3.2 Results

3.2.1 Improved Purification of Sheep Trichohyalin

The initial attempts at the purification of sheep trichohyalin were performed using the procedure of Rothnagel and Rogers (1986). This involved the extraction of follicular proteins in a guanidine-HCl buffer and the concentration of the extract in an Amicon stirred ultrafiltration cell. The concentrated extract was then loaded onto a gel filtration column which was run in buffered 6 M guanidine-HCl. As guanidine precipitates during SDS polyacrylamide gel electrophoresis the chromatographic fractions were dialysed and the precipitate resuspended in buffered 8 M urea. The protein-containing fractions were

then analysed by polyacrylamide gel electrophoresis. On examination of the gels which had been stained with Coomassie Brilliant Blue, it was determined that in the majority of cases only about 40% of the eluted trichohyalin was 90% pure after a single gel filtration chromatographic run (Fig. 3.1a). Therefore, fractions which contained partially purified trichohyalin were passed through a second round of gel filtration chromatography. Although an additional proportion of the extracted trichohyalin was purified (Fig. 3.1b), the yield of trichohyalin was lower than expected. As the fractions from the first chromatographic run were dialysed prior to their re-chromatography it is possible that a proportion of the trichohyalin may bind to the surface of the dialysis membrane and be lost during dialysis. To examine this possibility fractions collected from the chromatographic separation of a sheep follicle extract (Fig. 3.2a) were passed through two rounds of dialysis and the fractions containing trichohyalin were re-chromatographed. The resultant transmission profile (Fig. 3.2b) showed that, in comparison to the original profile, very little protein was eluted in the position expected for the trichohyalin-containing fractions thus indicating that most of the protein had been lost between the two chromatographic runs. It is likely that this has occurred during the two rounds of dialysis.

In order to decrease the losses of trichohyalin, the gel filtrates were subsequently concentrated using the stirred ultrafiltration cell which had previously been used for concentrating the initial follicle extract (see above). This apparatus allowed the trichohyalin to be maintained in a chaotropic solution during the concentration of the fractions which would hopefully minimise any binding of trichohyalin to the filtration membrane. By performing a number of serial concentration steps, which involved the addition of buffered 8 M urea, the majority of the guanidine in the fractions was removed and replaced with urea, thus enabling subsequent electrophoretic analysis. Unexpectedly, upon the filtration of the fractions, the filtration rate was found to slowly decrease with time due to the formation of a "skin" on the surface of the filter. This "skin", which was not completely soluble in buffered 8 M urea, contained both precipitated trichohyalin and keratin proteins (Fig. 3.3). It is possible that the large surface area available to the concentrated protein solutions facilitated the precipitation of protein onto the filter. The level of protein loss was then compared with another form of ultrafiltration which has a smaller surface area available to the concentrated samples, namely centrifugal ultrafiltration through Amicon filtration cones. As seen in Fig. 3.4 considerably more

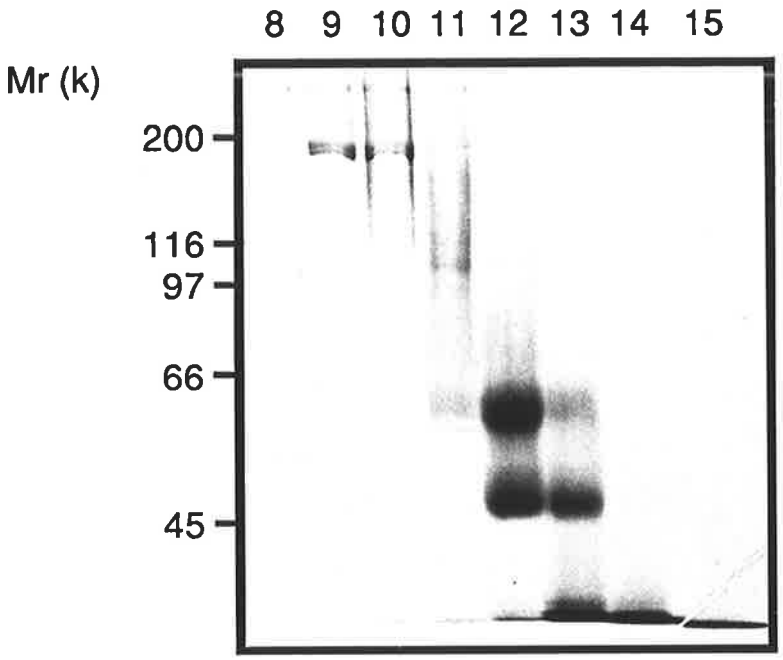
Figure 3.1 Electrophoretic analysis of fractions eluted from the CL-4B gel filtration column.

(a) Fractions 8 to 15, collected from the fractionation of a sheep follicle extract, were electrophoresed on a 7.5% SDS polyacrylamide gel. The separation of the proteins was of the highest standard obtained, with the majority of the trichohyalin, which appears as a doublet at a molecular weight of approximately 190 kdal, eluting in fractions 9 and 10. Fraction 9, in which the trichohyalin appears to be approximately 90% pure, contains about 60% of the eluted trichohyalin.

(b) Column fractions, which were derived from a number of CL-4B chromatographic runs of sheep follicle extract and contained impure trichohyalin, were combined and re-chromatographed on the CL-4B gel filtration column. Samples from fractions 8 to 15 were then electrophoresed on a 7.5% polyacrylamide gel. Almost all of the eluted trichohyalin (fractions 9 and 10) now appears to be more than 90% pure.

Proteins were detected with Coomassie Brilliant Blue and the relative levels of protein were judged by eye. The positions of the molecular weight standards are marked at the side of both gels.

a.


8 9 10 11 12 13 14 15
Mr (k)
200
116
97
66
45

b.

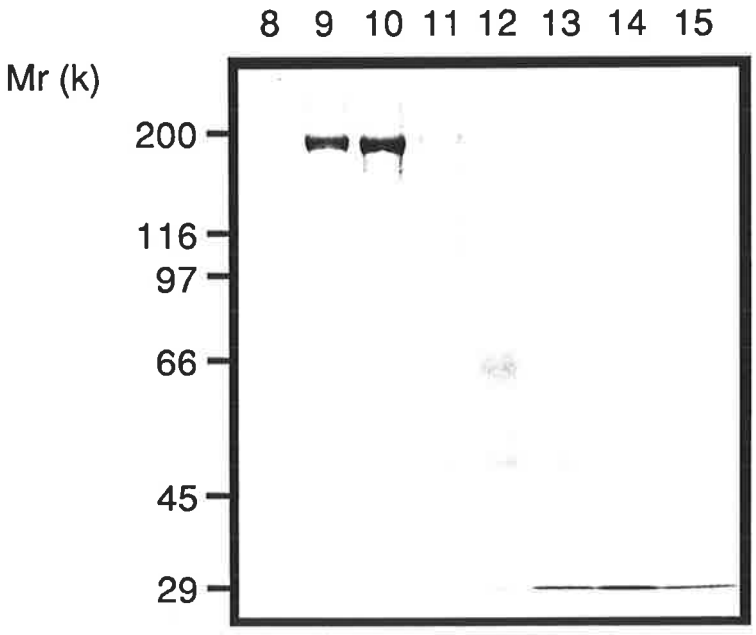
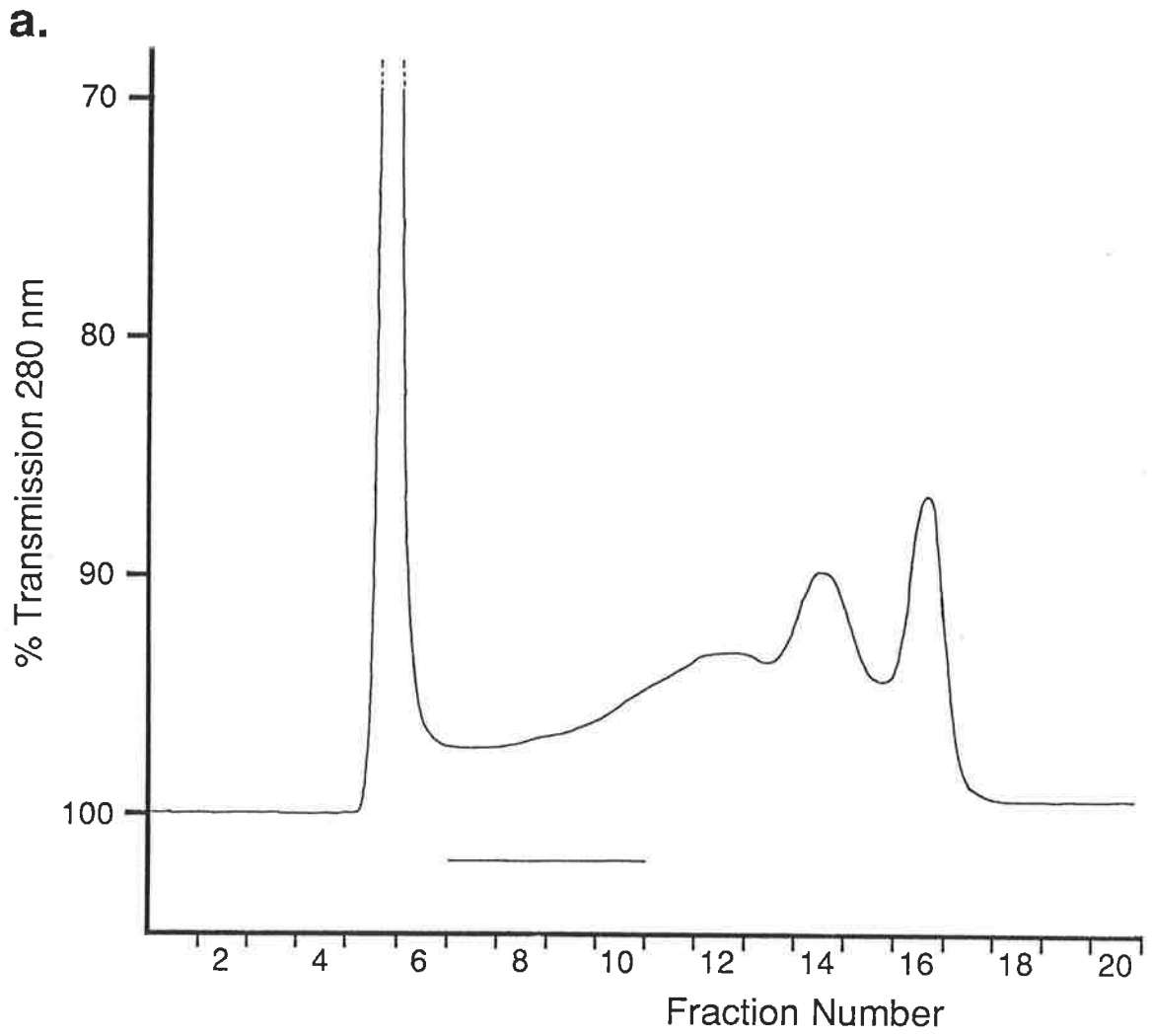

8 9 10 11 12 13 14 15
Mr (k)
200
116
97
66
45
29

Figure 3.2 The 280 nm transmission profiles of CL-4B gel filtration chromatographic runs.

(a) Chromatography of 2 ml of sheep follicle extract. The bar indicates the fractions which were determined by polyacrylamide gel electrophoresis to contain trichohyalin.

(b) The trichohyalin-containing fractions from (a) were passed through two rounds of dialysis and resuspension. They were then re-chromatographed on the CL-4B gel filtration column. The bar indicates the fractions which would normally be expected to contain trichohyalin. Almost no decrease in transmission occurs over this region, indicating that these fractions contain very little protein.


a.
70
80
90
100
% Transmission 280 nm
2
4
6
8
10
12
14
16
18
20
Fraction Number

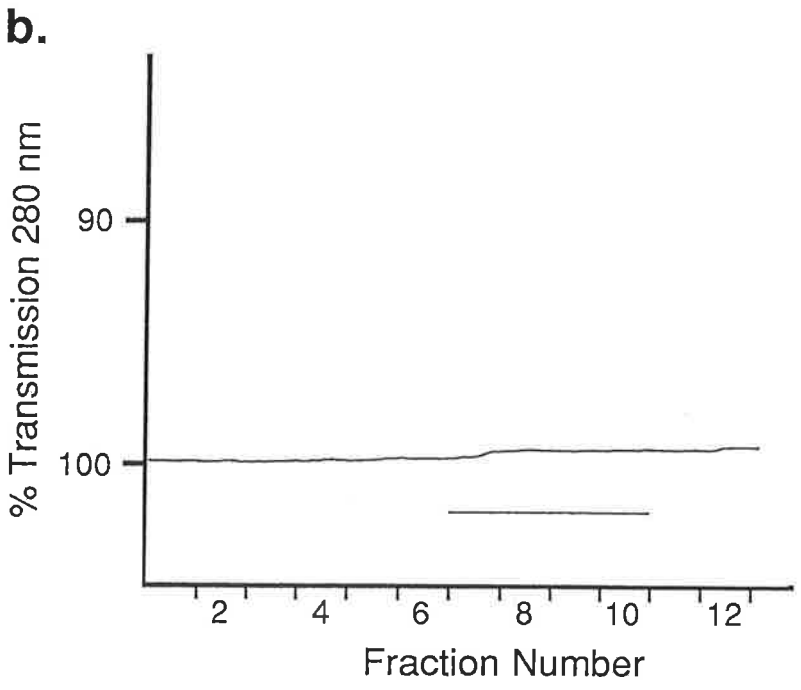

b.
90
100
% Transmission 280 nm
2
4
6
8
10
12
Fraction Number

Figure 3.3 Analysis of the protein content of the "skins" formed during fraction concentration in the stirred ultrafiltration cell.

Protein samples were prepared by attempted dissolution in buffered 8 M urea of three "skins" (1,2,3) formed during the concentration of a series of sheep follicle extract chromatographic fractions in the stirred ultrafiltration cell. These samples were electrophoresed on a 7.5% polyacrylamide gel and the proteins detected with Coomassie Brilliant Blue. Note that each "skin" contains both trichohyalin (190 kdal) and keratin proteins (40 kdal - 60 kdal). The positions of the molecular weight standards are marked at the side of the gel.

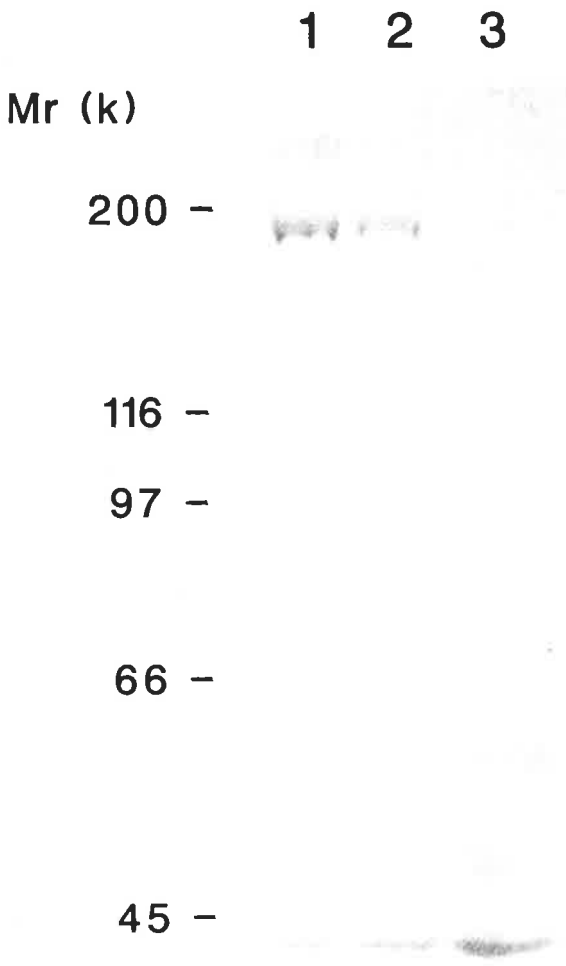
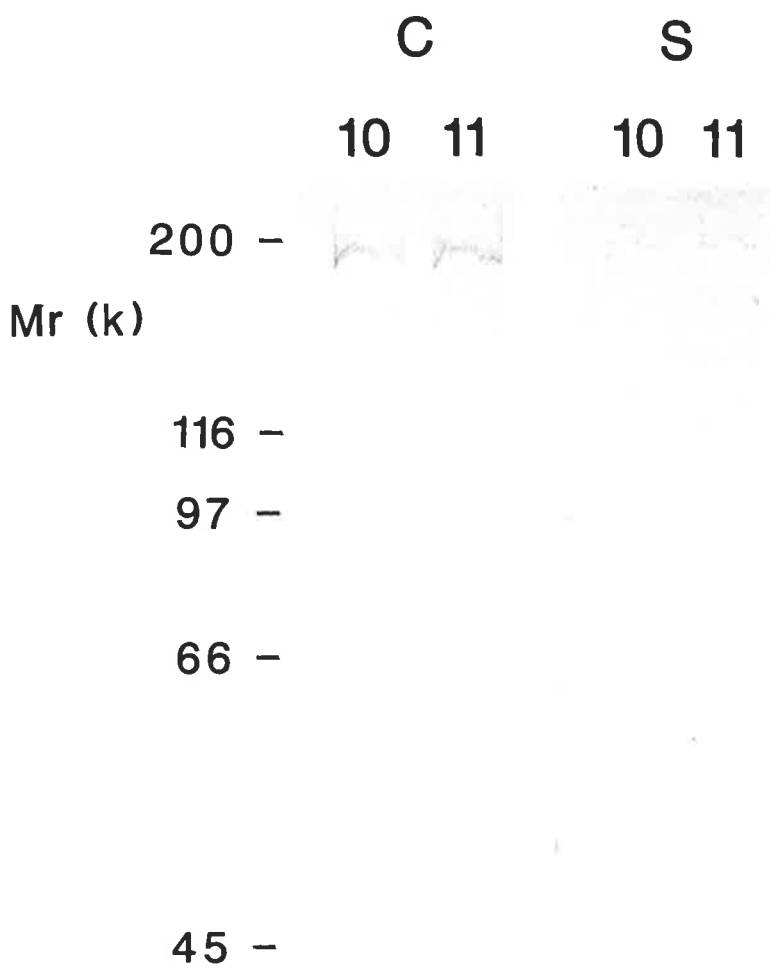

1 2 3
Mr (k)
200 -
116 -
97 -
66 -
45 -

Figure 3.4 Comparison of the trichohyalin recovery obtained after concentration of chromatographic fractions with either the stirred ultrafiltration cell or by centrifugation through the ultrafiltration cones.

Two trichohyalin-containing column fractions (10,11) were split and one half of each was concentrated in the stirred ultrafiltration cell (S) and the other by centrifugation in the ultrafiltration cones (C). Equal proportions of each sample were separated on a 7.5% polyacrylamide gel. Proteins were detected with Coomassie Brilliant Blue. The positions of the molecular weight standards are marked at the side of the gel.


C
S
10
11
10
11
200 -
Mr (k)
116 -
97 -
66 -
45 -

protein is recovered using the ultrafiltration cones than is the case with the stirred cell. All subsequent concentration steps were therefore performed in the ultrafiltration cones.

As stated above, the proteinaceous "skins" formed during filtration were not completely soluble in urea. The "skins" were found to dissolve upon the introduction of 1% 2-mercaptoethanol to the urea solution. This indicates that disulphide bond formation is involved in the precipitation of trichohyalin and the keratin proteins onto the filter. Thus 1% 2-mercaptoethanol was also included in all further concentration steps to decrease the rate of protein precipitation.

3.2.2 Cleavage of Sheep Trichohyalin

a. N-terminal Sequence Analysis

The simplest means of obtaining partial amino acid sequence from a protein is to determine the sequence at the N-terminus of the protein. Approximately 40μg of sheep trichohyalin, which had been purified to greater than 95% using the above procedure, was subjected to automated Edman analysis. Unfortunately, no significant signals were obtained, indicating that the N-terminal amino acid of trichohyalin is blocked thus stopping the first round of Edman degradation.

b. Cyanogen Bromide Cleavage

In order to obtain internal amino acid sequence, trichohyalin must be cleaved either chemically or enzymatically and some of the resultant peptides purified and sequenced. One possible chemical method is cleavage with cyanogen bromide. Cyanogen bromide has been found to cleave proteins with a high specificity on the carboxyl side of methionine residues (Gross and Witkop, 1962). According to the amino acid composition obtained by Rothnagel and Rogers (1986) sheep trichohyalin contains 1.0 moles% methionine (Table 3.1). On the basis of its molecular weight and amino acid composition, sheep trichohyalin can be calculated to contain approximately 1600 amino acids which implies that it contains about 16 methionine residues. Cleavage with cyanogen bromide should thus produce approximately 17 different peptides, a number of which could hopefully be purified for sequence analysis.

The cleavage of trichohyalin with cyanogen bromide was attempted with lyophilised sheep trichohyalin dissolved in either acetic acid or formic acid. Little cleavage was obtained when the normally required level of cyanogen bromide was added (Walker and Mayes, 1983). On the addition of increased amounts of cyanogen bromide

Table 3.1 Amino acid analysis of wool follicle trichohyalin.

Data from Rothnagel and Rogers (1986)

Amino Acid	Wool Follicle Trichohyalin
	mole percent
Asp/Asn	6.1
Thr	3.0
Ser	5.5
Glu/Gln	17.0
Pro	4.0
Gly	6.4
Ala	5.3
1/2-Cys	0.7
Val	4.1
Met	1.0
Ile	2.9
Leu	8.5
Tyr	2.1
Phe	2.7
Lys	11.6
His	2.7
Arg	16.5

significant levels of peptides were occasionally produced (Fig. 3.5). Unfortunately the level of cleavage was extremely variable and always incomplete, precluding the production of sufficient amounts of peptides for purification and sequencing.

To determine whether the unreliable cleavage was due to problems with the cleavage conditions, cyanogen bromide cleavage of carbonic anhydrase was attempted under identical conditions to those used for the sheep trichohyalin. Electrophoretic analysis of the cleavage products indicated that the native carbonic anhydrase had been completely cleaved by the cyanogen bromide to produce two peptides (Fig. 3.6). Thus the apparent inability of cyanogen bromide to completely cleave sheep trichohyalin is probably due to either an incorrect estimation of the number of methionine residues within trichohyalin, the inaccessibility of the methionines to cyanogen bromide in the tested solvents, or the insolubility of trichohyalin in either formic or acetic acid.

c. Proteolytic Cleavage

There are four narrow specificity reliable endoproteases which are commonly used for the production of peptides for subsequent sequence analysis. These are endoproteinase arginine-C (cleaves on the carboxyl side of arginine residues), endoproteinase glutamate-C (carboxyl side of glutamic acid residues), endoproteinase lysine-C (carboxyl side of lysine residues) and trypsin (carboxyl side of arginine and lysine residues). To examine the frequency of cleavage of these proteases the amino acid composition of trichohyalin was examined (Table 3.1). Although the respective levels of glutamic acid and glutamine cannot be conclusively determined due to the deamination of glutamine residues during the protein hydrolysis required for amino acid analysis, it would appear that the recognition amino acids for each of the four enzymes constitute greater than 10% of the trichohyalin residues. As trichohyalin contains approximately 1600 amino acids, cleavage with any of the four proteases would be expected to produce greater than 160 peptides with an average length of less than 10 amino acids. Although such a large number of peptides may make the purification of individual peptides difficult it was decided to cleave sheep trichohyalin with endoproteinase lysine-C which, of the four narrow specificity endoproteinases, should produce the smallest number of resultant peptides.

In order to maximise the efficiency of cleavage with endoproteinase lysine-C, the urea concentration was decreased to 2 M. Approximately 500 µg of pure sheep trichohyalin was digested at 37°C with a total of 10 µg of endoproteinase lysine-C. As

Figure 3.5 Analysis of cyanogen bromide cleaved sheep trichohyalin.

Purified sheep trichohyalin was cleaved with cyanogen bromide and electrophoresed on a 10% polyacrylamide gel (Track C). Although the level of cleavage was the best obtained, uncut trichohyalin is still visible and the cumulative mass of the five prominent peptides (ranging in molecular weight from 70 kdal to 95 kdal) is greater than 300 kdal indicating that their cleavage is incomplete. Track U contains uncut sheep trichohyalin. Proteins were detected with Coomassie Brilliant Blue. The positions of the molecular weight standards are marked at the side of the gel.

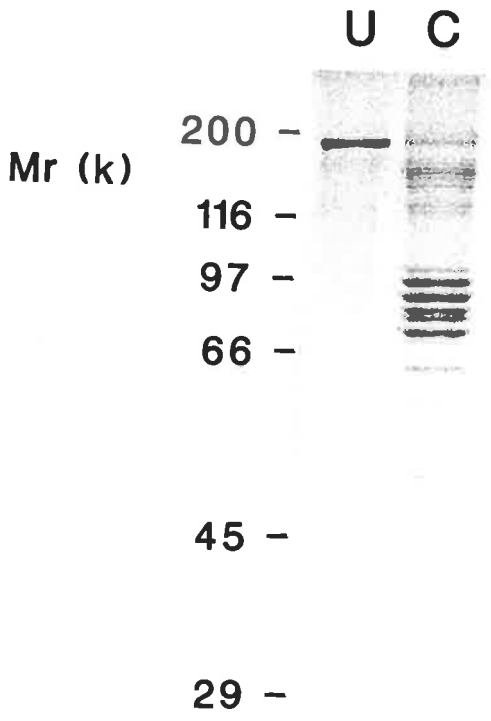
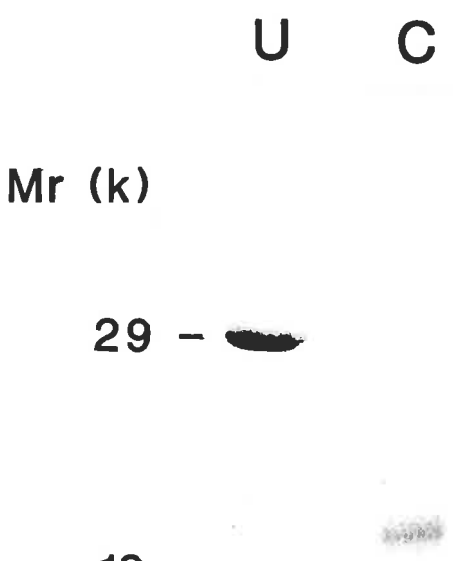

U
C
Mr (k)
200 –
116 –
97 –
66 –
45 –
29 –

Figure 3.6 Examination of the efficiency of cyanogen bromide cleavage.

Carbonic anhydrase (Mr 29,000) was digested with cyanogen bromide and both cut (C) and uncut (U) carbonic anhydrase were separated on a 15% polyacrylamide gel. Proteins were detected with Coomassie Brilliant Blue. Note that carbonic anhydrase has been completely cleaved by cyanogen bromide yielding two distinct peptide products which appear to have molecular weights of approximately 20,000 and 9,000. The faint bands in both tracks are caused by contaminants in the carbonic anhydrase sample. The positions of the molecular weight standards are marked at the side of the gel.


U
C
Mr (k)
29 –
18 –
16 –
6 –

determined by polyacrylamide gel electrophoresis, the proteolysis appeared to go to completion in 16 hours.

To examine the complexity of the peptide mix, the digest was chromatographed on a butyl reverse-phase cartridge equilibrated with 0.1% trifluoroacetic acid. The separation of the peptides was examined using various gradients of acetonitrile (Fig. 3.7). With each of the gradients used the separation profile remained complex and it was believed at the time that it would be extremely difficult to purify individual peptides from the mix.

3.2.3 Proteolysis of Guinea Pig Trichohyalin

a. Comparison of Sheep and Guinea Pig Trichohyalin

As purified peptides could not be obtained from cleaved sheep trichohyalin it was decided to attempt cleavage of trichohyalin from a different species. The guinea pig was chosen, partly because guinea pig hairs contain a medulla and are therefore more enriched in trichohyalin, although it should be noted that the IRS and medulla trichohyalin may be different (see Section 1.3.1). It was also hoped that the guinea pig trichohyalin amino acid sequence may differ sufficiently from that of sheep trichohyalin to alter the overall cleavage of trichohyalin, such that individual peptides could be purified and sequenced, and yet be similar enough to allow the identification or confirmation of sheep cDNA clones. To examine the similarity of the overall amino acid compositions, amino acid analyses were performed on both sheep and guinea pig trichohyalin (Table 3.2). Although not identical, the moles% of most amino acids is very similar in both proteins.

Further analysis of the sequence similarities was obtained by examining the immunocross-reactivity of the two trichohyalin species. Previous work had shown that guinea pig trichohyalin is highly immunoreactive with polyclonal antibodies raised against sheep trichohyalin (Rothnagel and Rogers, 1986). It was decided to further this examination by testing the immunoreactivity of proteolytically produced peptides. Sheep and guinea pig trichohyalin were cleaved with endoproteinase lysine-C and time samples taken. The time samples were blotted onto nitrocellulose and probed with a polyclonal antibody which had been raised against purified sheep trichohyalin (Fig. 3.8). Although the two protein digests cannot be directly compared, it can be seen that the antibody binds to a range of similarly sized peptides in both species and that even small guinea pig peptides (30 kdal) bind the anti-sheep trichohyalin antibody. It therefore appears that the

Figure 3.7 Separation of endoproteinase lysine-C digested sheep trichohyalin by reverse phase HPLC.

Sheep trichohyalin was digested with endoproteinase lysine-C, loaded onto a butyl reverse phase cartridge and separated by numerous different gradients of acetonitrile. The depicted profile has the best obtainable separation but was still unable to yield separate pure peptide peaks. Buffer A; 0.11% trifluoroacetic acid (TFA). Buffer B; acetonitrile, 0.1% TFA.

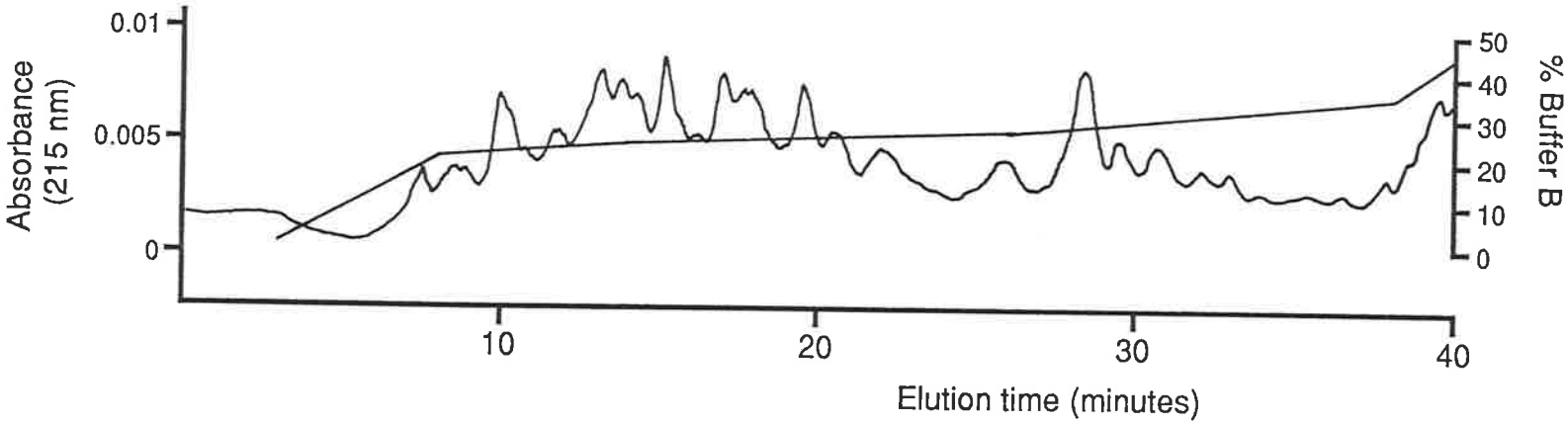

Absorbance
(215 nm)
0.01
0.005
0
50
40
30
20
10
0
% Buffer B
10
20
30
40
Elution time (minutes)

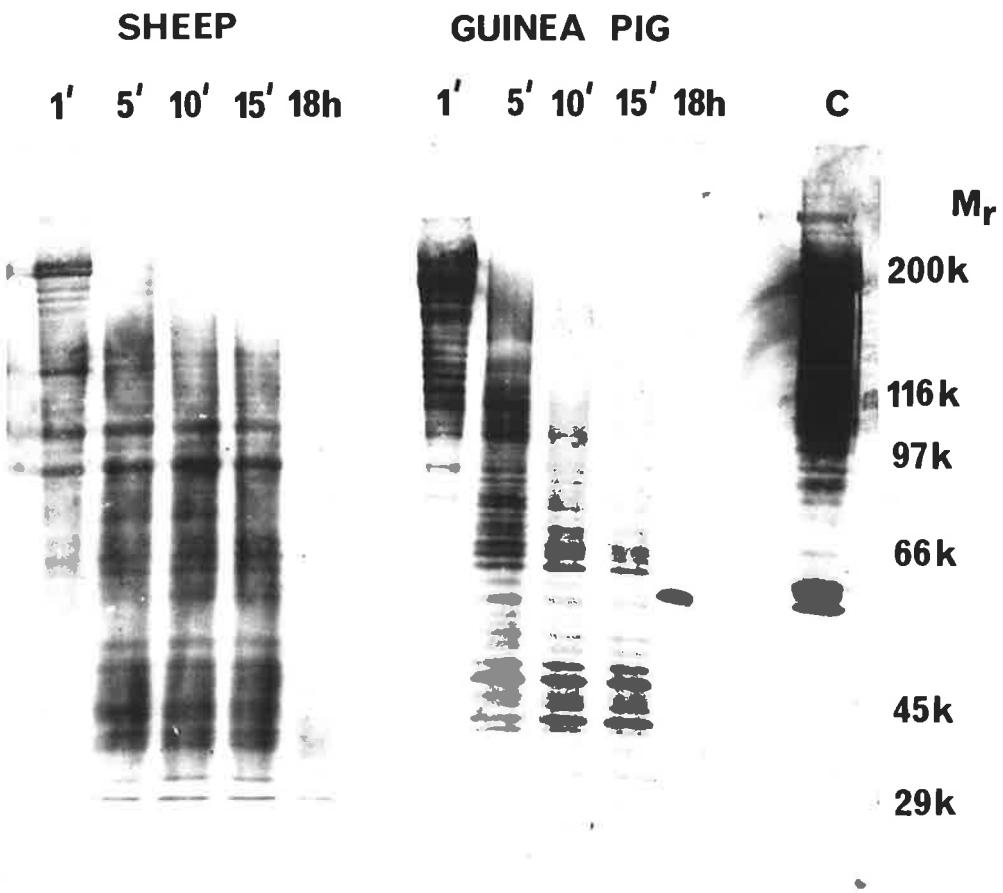
Table 3.2 Amino acid analyses of sheep trichohyalin and guinea pig trichohyalin.

Analyses were kindly performed by Dr. R. Marshall, Division of Biotechnology, Commonwealth Scientific and Industrial Research Organisation, Melbourne.

Amino Acid	Sheep Trichohyalin	Guinea Pig Trichohyalin
	mole percent	*mole percent*
Asp/Asn	6.4	4.8
Thr	2.9	1.9
Ser	5.4	2.9
Glu/Gln	28.0	35.7
Pro	3.3	1.9
Gly	5.3	3.2
Ala	4.7	2.9
1/2-Cys	0.6	0.7
Val	4.0	2.4
Met	0.1	0.3
Ile	2.5	1.6
Leu	10.0	10.4
Tyr	2.1	1.2
Phe	2.4	2.6
Lys	6.7	4.9
His	1.7	1.2
Arg	13.7	21.0
Trp	0.2	0.3

Figure 3.8 Western blot analysis of time samples taken from endoproteinase lysine-C digests of sheep and guinea pig trichohyalin.

Purified sheep and guinea pig trichohyalin were digested with endoproteinase lysine-C and time samples taken at 1', 5', 10', 15' and 18 hours. These samples were electrophoresed on a 7.5% polyacrylamide gel, transferred to nitrocellulose and reacted with antiserum raised against sheep trichohyalin. Track C contains a mixture of sheep trichohyalin and keratin proteins. Note that the wide range of bands detected in track C are believed to be produced by proteolytic breakdown of trichohyalin during its extraction from the follicle. The positions of the molecular weight standards are marked at the side of the gel.


SHEEP
GUINEA PIG
1′ 5′ 10′ 15′ 18h
1′ 5′ 10′ 15′ 18h
C
Mr
200k
116k
97k
66k
45k
29k

two proteins contain similar antigenic sites, suggesting that the amino acid sequences of sheep and guinea pig trichohyalin are highly homologous.

b. N-terminal Sequence Analysis

Approximately 40 μg of guinea pig trichohyalin, which had been purified by the modified purification procedure described in Section 3.2.1, was subjected to automated Edman degradation. As was the case for sheep trichohyalin, insignificant signals were again obtained indicating that the guinea pig trichohyalin N-terminus is also blocked.

c. Endoproteinase Lysine-C Cleavage

500 μg of pure guinea pig trichohyalin was digested with a total of 10 μg of endoproteinase lysine-C under conditions identical to those used for sheep trichohyalin (Section 3.2.2c). The digestion was examined by polyacrylamide gel electrophoresis (Fig. 3.9) and found to produce peptides ranging in size from 3 kdal to 40 kdal.

d. Purification and Sequencing of Proteolytic Peptides

The initial separation of the guinea pig trichohyalin proteolytic peptides was performed on a Superose 12 gel filtration column (Fig. 3.10a). 0.5 ml fractions were collected and the protein-containing fractions, as determined by SDS polyacrylamide gel electrophoresis, were loaded onto a butyl reverse-phase HPLC cartridge and separated by various gradients of acetonitrile (Fig. 3.10b). Individual peaks were collected and identical peaks from adjacent runs were combined. Five peptides were then repurified by reverse-phase HPLC (Fig. 3.10c), examined by polyacrylamide gel electrophoresis (Fig. 3.10d) and sequenced by automated Edman degradation. The resultant peptide sequences are shown in Fig. 3.11. Three of the peptides, B, D and F1, show considerable homology, with each containing the sequence -QL-, surrounded by charged amino acids. In addition, peptides B and D also begin with the sequence -FR-, indicating an association with the lysine residue immediately prior to the cleavage site.

3.3 Discussion

The hardened cells of the hair follicle IRS and medulla have a peculiar protein chemistry; the structural proteins contain the amino acid citrulline (Rogers, 1958a, 1963; Rogers and Simmonds, 1958) and are cross-linked by ε-(γ-glutamyl)lysine iso-peptide bonds (Harding and Rogers 1971, 1972a). Since the iso-peptide bonds are present, the intact hardened proteins cannot be purified and analysed and thus the precursor forms of the structural proteins must be examined. One of these precursor proteins, trichohyalin,

Figure 3.9 Analysis of an endoproteinase lysine-C digest of guinea pig trichohyalin.

Purified guinea pig trichohyalin was digested with endoproteinase lysine-C and electrophoresed on a 15% polyacrylamide gel (Track C). Track U contains undigested guinea pig trichohyalin. Proteins were detected with Coomassie Brilliant Blue. The positions of the molecular weight standards are marked at the side of the gel.

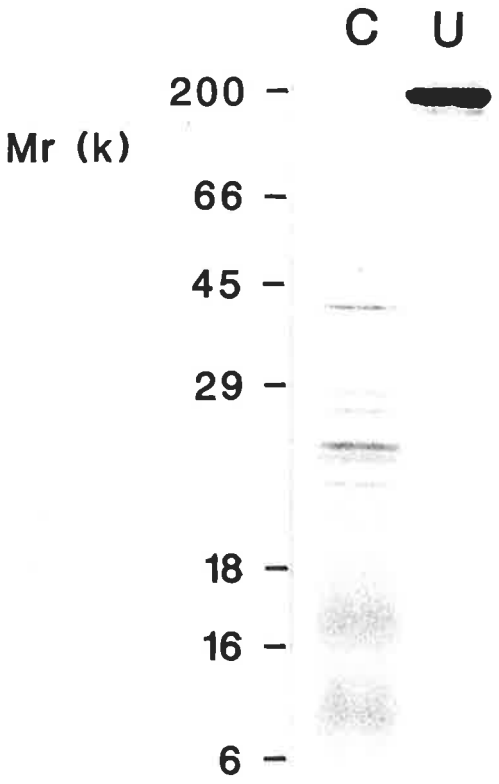

C
U
200
Mr (k)
66
45
29
18
16
6

Figure 3.10 Purification of guinea pig trichohyalin proteolytic peptide D.

(a) Guinea pig trichohyalin which had been digested with endoproteinase lysine-C was initially loaded onto a Superose 12 gel filtration column and eluted in 2 M urea, 50 mM Tris-HCl, 100 mM NaCl. The bar indicates the fractions which were subsequently found to contain peptide D.

Continued.....

a.

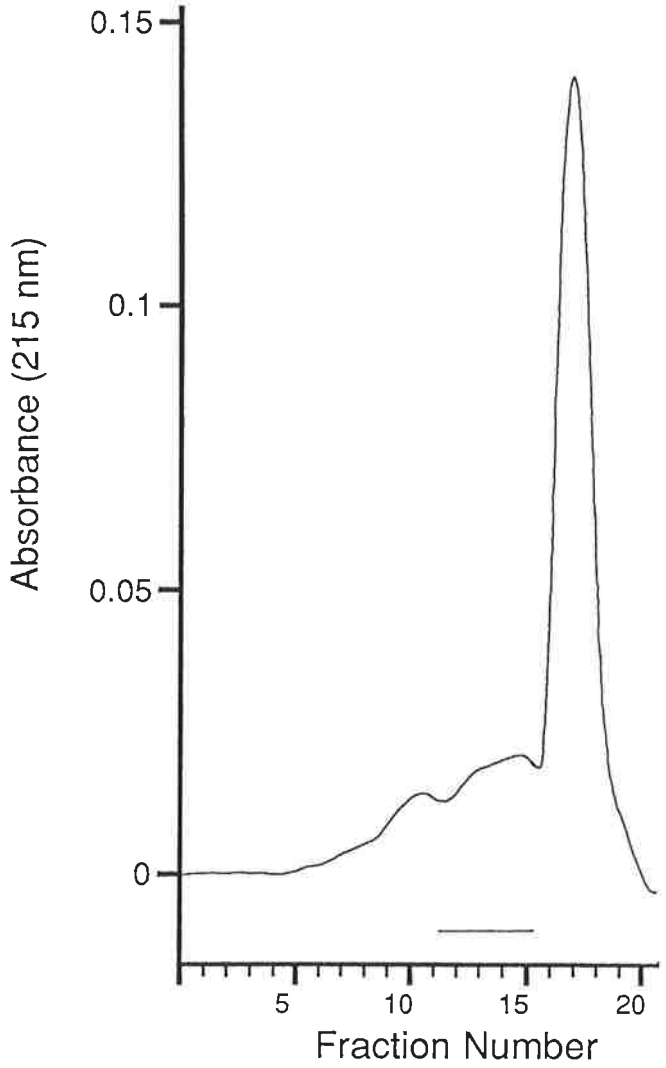


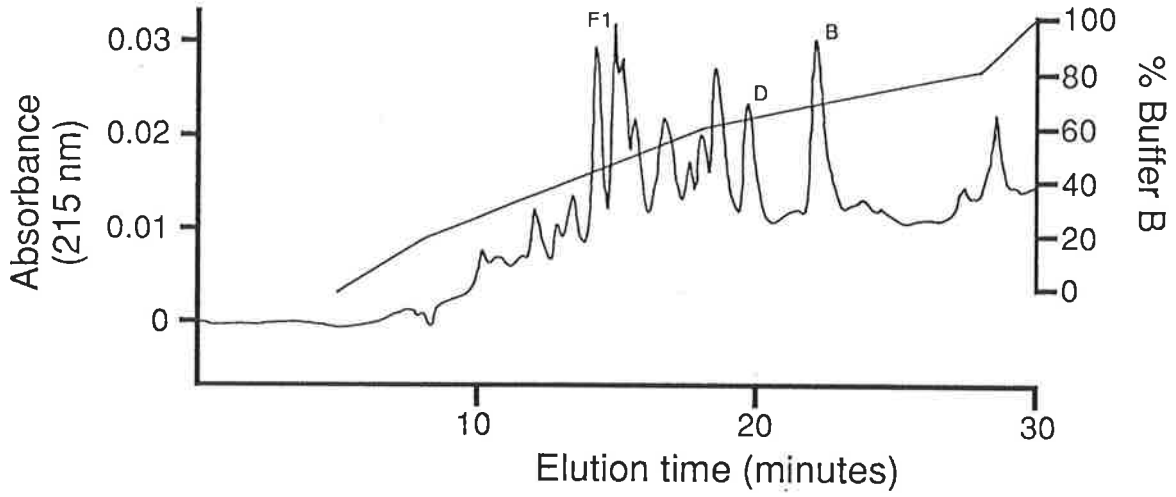
Figure 3.10 (Cont.)

(b) Fractions 5 to 18 from (a), which had been shown by polyacrylamide gel electrophoresis to contain peptides, were chromatographed separately on a butyl reverse phase HPLC cartridge and the individual peaks were collected. The depicted profile is of the Superose 12 fraction 14 and the peaks containing peptides B, D, and F1 are marked. Peptide A eluted off in earlier fractions and peptide M in later fractions. Buffer A; 10% acetonitrile, 0.11% TFA. Buffer B; 35% acetonitrile, 0.1% TFA.

(c) On the basis of the absorbance profiles a number of the relatively pure peptides were selected and the corresponding peaks from separate runs were combined. The peptides were then rechromatographed on the reverse phase cartridge to remove any remaining contamination prior to sequencing. The profile for peptide D is shown. Buffer A; 10% acetonitrile, 0.11% TFA. Buffer B; 35% acetonitrile, 0.1% TFA.

Continued.....

b.


F1
B
D
0.03
0.02
0.01
0
Absorbance (215 nm)
100
80
60
40
20
0
% Buffer B
10
20
30
Elution time (minutes)

c.

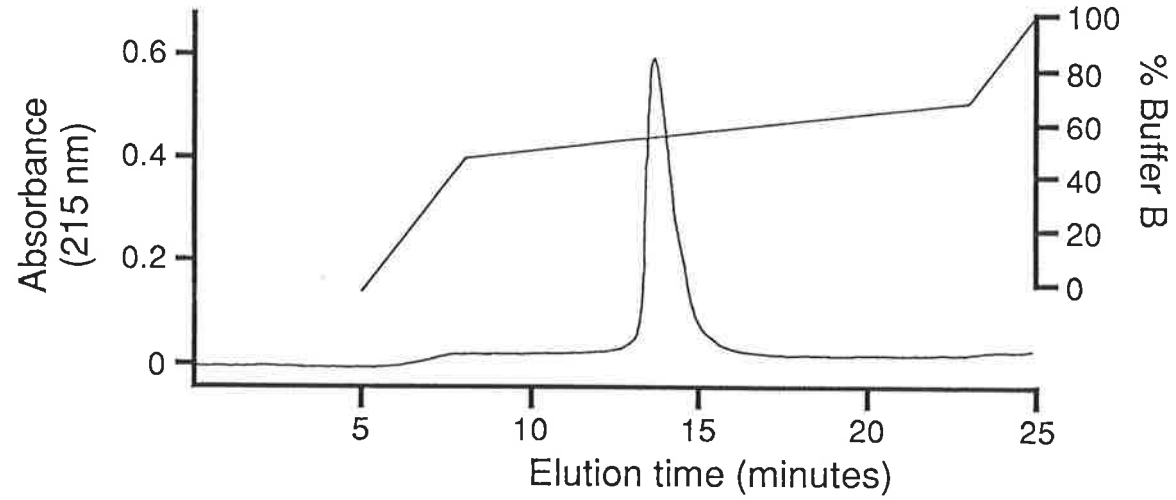

0.6
0.4
0.2
0
Absorbance (215 nm)
100
80
60
40
20
0
% Buffer B
5
10
15
20
25
Elution time (minutes)

Figure 3.10 (Cont.)

(d) The pure peptides were electrophoresed on a 15% polyacrylamide gel to ensure that the peaks did contain protein. The analysis of peptide D (M_r 10,000) is shown. Track T contains guinea pig trichohyalin digested with endoproteinase lysine-C. Proteins within track T were stained with Coomassie Brilliant Blue whilst peptide D was detected with a positive-image silver stain (Section 2.2.2f). The positions of the molecular weight standards are marked at the side of the gel.

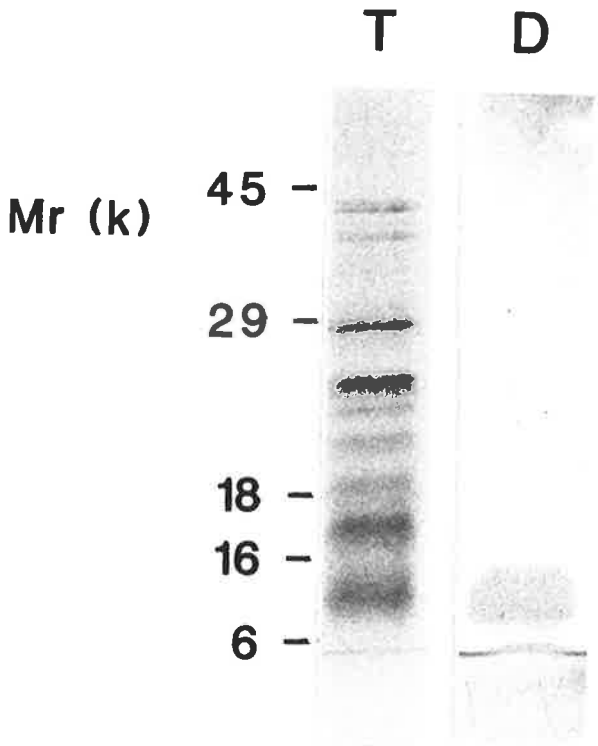
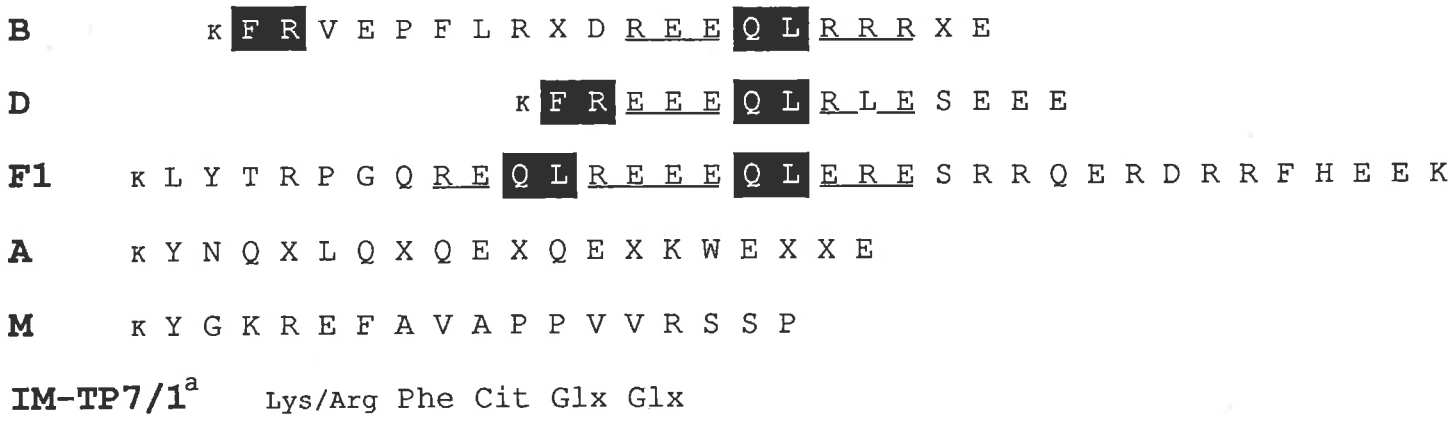

T
D
Mr (k)
45
29
18
16
6

Figure 3.11 Amino-terminal sequences of five guinea pig trichohyalin peptides.

Five peptides (B, D, F1, A and M) were purified from the digest of guinea pig trichohyalin with endoproteinase lysine-C and were sequenced by the gas-phase method. The resultant sequences are shown. Unassigned residues are indicated by the letter X. A lysine residue (small letters) has been placed at the amino end of each peptide because of the site-specific cleavage of endoproteinase lysine-C. The peptides B, D and F1 have been aligned to emphasise their similarity. Each of the three peptides contains at least one QL combination (boxed) which is surrounded by a highly charged region (underlined). Peptides B and D also begin with FR (boxed) indicating an association of these residues with lysine, a substrate amino acid for transglutaminase. Also shown is the amino-terminal sequence of the peptide IM-TP7/1, a citrulline (Cit) containing peptide purified from a tryptic digest of porcupine quill medulla (Steinert *et al.*, 1969; see Table 1.3). Glx indicates that the residue could be either glutamic acid or glutamine. As citrulline is formed by the post-translational deimination of arginine the sequence of IM-TP7/1 could be identical to the N-terminal sequence of peptide D.


B K F R V E P F L R X D R E E Q L R R R X E
D K F R E E E Q L R L E S E E E
F1 K L Y T R P G Q R E Q L R E E E Q L E R E S R R Q E R D R R F H E E K
A K Y N Q X L Q X Q E X Q E X K W E X X E
M K Y G K R E F A V A P P V V R S S P
IM-TP7/1[a] Lys/Arg Phe Cit Glx Glx

had previously been purified (Rothnagel and Rogers, 1986) and found to have an unusual chemistry. Over 55% of the amino acids are charged or highly polar (Table 3.1) yet the protein is insoluble in aqueous solutions, as shown by its ability to aggregate into non-membrane bound granules in the developing IRS and medulla cells. The solubility of trichohyalin in guanidine without the presence of reducing agents (Rothnagel and Rogers, 1986) indicates that intermolecular disulphide bonds are not present within the trichohyalin granules and that trichohyalin must contain surface hydrophobic residues which interact with those of adjacent trichohyalin molecules to form the granular arrangement. It would appear that during the dialysis of the trichohyalin-containing fractions in the purification procedure (Rothnagel and Rogers, 1986), these hydrophobic interactions not only allow the trichohyalin to aggregate with itself but also enable it to bind to the inner surface of the dialysis membrane leading to significant losses of trichohyalin.

To decrease these losses, the fractions were initially concentrated in a stirred cell ultrafiltration apparatus in which the proteins can be maintained in highly chaotropic conditions while the solvent volume is decreased by filtration. Nevertheless, significant amounts of protein were still found to precipitate onto the filter, further demonstrating the low solubility of even denatured trichohyalin (Fig. 3.3). A reduction in the amount of precipitated trichohyalin was achieved by using ultrafiltration cones which had a smaller available filtration surface area. The addition of reducing agents also increased the solubility of the precipitated trichohyalin and keratin proteins. This suggests that intramolecular disulphide bonds are present within the native trichohyalin and that during the denaturation of trichohyalin these bonds are broken and the free cysteine residues are then able to form disulphide cross-links with both other trichohyalin molecules and the extracted keratins. On the basis of these findings all concentration steps used in the final purification protocol were performed in the ultrafiltration cones and in the presence of 1% 2-mercaptoethanol.

The major aim of the work described in this chapter was the acquisition of a partial amino acid sequence for sheep trichohyalin. This sequence could then be used in either the purification of cDNA clones via synthesis of an oligodeoxynucleotide probe or in the confirmation of the identity of trichohyalin cDNA clones isolated by other methods, e.g., by screening of an expression cDNA library with an antibody raised against trichohyalin. It was initially attempted to obtain the N-terminal sequence of the complete trichohyalin

molecule. For proteins the size of trichohyalin (190 kdal) oligo-dT primed cDNA clones usually do not extend to the coding region of the N-terminal portion of the protein and thus the N-terminal sequence would not be useful for identifying or analysing cDNA clones. Yet this sequence could be important for the precise localisation of the initiation codon within the gene or cDNA sequence. Unfortunately no sequence was obtained from the N-terminus, probably due to the blockage of the terminal amino acid.

Although the desired source for the amino acid sequence was sheep trichohyalin, no suitable method was found for its cleavage and the purification of the resultant peptides. According to the amino acid analysis of Rothnagel and Rogers (Table 3.1) sheep trichohyalin contained approximately 16 methionine residues suggesting that cleavage with cyanogen bromide should yield about 17 distinct peptides from which individual peptides might be purified. Nevertheless, cyanogen bromide was unable to cut trichohyalin to completion (Fig. 3.5). Subsequent amino acid analysis (Table 3.2) indicated that there may be only one methionine residue within sheep trichohyalin, and it is possible that this could be positioned at or near one of the termini. Complete cleavage with cyanogen bromide may therefore not have been detected by polyacrylamide gel electrophoresis. The cleavage products seen in Fig. 3.5 may be due to partial cleavage at tryptophan residues which can occur if the protein has been partially oxidised (Blumenthal *et al.*, 1975).

Attempts were also made to purify peptides produced by cleavage of sheep trichohyalin with endoproteinase lysine-C. Due to the complexity of the peptide mix individual peptides were not separated by single stage reverse phase HPLC. It is possible that the protocol used for the eventual purification of proteolytic peptides produced from guinea pig trichohyalin may also have been successful if attempted on the sheep trichohyalin digest.

It is interesting to note that the levels of methionine, lysine and glutamic acid/ glutamine in the amino acid analysis of sheep trichohyalin performed in the work for this thesis (Table 3.2) differ significantly from those of Rothnagel and Rogers (Table 3.1). These differences may indicate that the trichohyalin sequence and thus amino acid composition can vary markedly between different sheep breeds (Merino-Dorset Horn x Border Leicester sheep were used for the work described in this thesis whereas Border Leicester x Dorset Horn sheep were used by Rothnagel and Rogers). Another possibility is that the determination of amino acid composition may tend to be unreliable, especially

with amino acids present at very high (glutamic acid/glutamine) or very low (methionine) levels.

As individual peptides could not be purified from sheep trichohyalin, the trichohyalin of another species, namely the guinea pig, was purified. Guinea pig trichohyalin and its proteolytic peptides are immunocross-reactive with sheep trichohyalin (Rothnagel and Rogers, 1986; Fig. 3.8). This suggests that the antigenic determinants are conserved, and that the respective amino acid sequences may be very similar, enabling sheep cDNA clones to be identified or confirmed using guinea pig peptide sequences. Additionally, although the amino acid composition of guinea pig trichohyalin is comparable to that of sheep trichohyalin, the proportion of lysine residues is significantly lower than in sheep trichohyalin, e.g., 80 vs. 110 for the wool trichohyalin fraction used earlier (see Table 3.2). Therefore guinea pig trichohyalin was digested with endoproteinase lysine-C and five of the resultant peptides were purified and sequenced (Figure 3.11). Interestingly, three of them showed considerable homology (B, D and F1), containing the sequence -QL- surrounded by highly charged residues. This environment may be required for the glutamine residue to be cross-linked by follicle transglutaminase. Additionally, two peptides (B, D) began with -FR-, which would be positioned immediately after the lysine residue at the endoproteinase lysine-C site. This sequence, together with the subsequent glutamic acid residues of peptide D, was also found in a peptide which had previously been purified from a tryptic digest of hardened IRS cells (Steinert *et al.*, 1969; Fig. 3.11). This conserved sequence may be important for the cross-linking of the lysine residue. The homology of the peptide sequences also suggests that peptide repeats are present within trichohyalin, a proposal which had previously been put forward by Rothnagel and Rogers (1986) on the basis of two-dimensional gel analysis of trichohyalin breakdown products.

Chapter 4

Analysis of a Sheep Trichohyalin cDNA Clone

4.1 Introduction

The purification of a cDNA clone for trichohyalin is a critical step in the determination of the complete trichohyalin protein sequence. For a 190 kdal protein, the likelihood of purifying a full-length cDNA clone is very low, yet the information obtained from even a partial clone will be essential in determining the best method for procuring the remainder of the coding sequence, i.e. by extending along the mRNA from specific primers to obtain a full-length sequence or by purifying the trichohyalin gene. It will also allow an initial examination of the coding sequence which may determine whether the core IF α-helical region is present in trichohyalin.

This chapter describes the characterisation and sequencing of λsTr1, a partial sheep follicle cDNA clone to trichohyalin. The encoded protein sequence is determined and its structure is analysed and discussed in relation to the role of trichohyalin. Much of the work presented in this chapter has already been published (Fietz *et al.*, 1990; see Appendix).

4.2 Results

4.2.1 Characterisation of λsTr1

A sheep follicle cDNA library was constructed in this laboratory in collaboration with Dr. R. Presland (Fietz *et al.*, 1990). cDNA was prepared by oligo-dT priming of sheep follicle poly A(+) RNA and species of greater than 1 kb in size were selected by sucrose gradient centrifugation. After addition of EcoR I linkers the cDNA was ligated into the expression vector λgt11 (Huynh *et al.*, 1985)

The library was screened with a polyclonal antibody raised against sheep trichohyalin (see Section 2.2.2j). Antibody-bound plaques were purified, the phage DNA prepared and provided for analysis. The longest cDNA clone, λsTr1, was 2.4 kb long and, upon digestion with EcoR I, was resected from the λ vector, producing fragments of 1.9 kb and 0.47 kb. As stated earlier (Section 3.2.2b) trichohyalin is expected to contain approximately 1600 amino acids and should thus be encoded by an mRNA species of greater than 5 kb in length. Therefore λsTr1 is less than half the expected size for a full-length trichohyalin cDNA clone.

For further analysis, the two EcoR I fragments were subcloned into pGEM-2, producing the clones pGEM-1.9 and pGEM-0.47. Restriction analysis of these clones determined that the cDNA insert of λsTr1 contained unique Hind III, BamH I and Sac I restriction sites in addition to the unique internal EcoR I site. Additionally, all four unique sites were located within one half of the insert, whilst the other half was found to contain multiple Pst I restriction sites (Fig. 4.1).

4.2.2 Nucleotide Sequencing of λsTr1

Using the basic restriction map for λsTr1 (Fig 4.1), DNA fragments were selected and prepared by restriction digestion of pGEM-1.9 or pGEM-0.47 for cloning into the appropriate M13 vectors (Messing and Vieira, 1982; Norrander *et al.*, 1983). Further fragments were prepared by progressive deletion of linear pGEM-1.9 with Bal 31 exonuclease (Section 2.2.3j) and cloning of the desired fragments into the required M13 vector. Single stranded DNA prepared from these clones was sequenced by the dideoxy chain termination method (Section 2.2.3n). The location and extent of sequenced clones is shown in Fig. 4.2. Additionally, to check that the two EcoR I fragments were contiguous, the 0.7 kb BamH I/Hind III fragment of λsTr1 was subcloned into M13 and the nucleotide sequence spanning the EcoR I site was obtained.

The nucleotide sequence of λsTr1, together with the deduced amino acid sequence, is shown in Fig. 4.3.

λsTr1 is 2,408 bp long and contains an open reading frame which spans the first 1,375 bp at the 5' end of the clone. The remainder of the insert consists of the 1030 bp 3' non-coding region of which the last 5 bases are adenines. A putative polyadenylation signal (underlined in Fig. 4.3) begins 19 bases upstream from the first of the 5 adenines, suggesting that the adenine residues constitute the start of the poly A tail. This was shown by the subsequent sequencing of the 3' end of two other positive cDNA clones, λsTr4 and λsTr16, both of which also terminate with short adenine stretches positioned at the same site.

4.2.3 Confirmation of λsTr1 Identity

As stated above (Section 4.2.1), the mRNA for trichohyalin is expected to be greater than 5 kb in length. Northern analysis determined that the 1.9 kb EcoR I fragment hybridised to a sheep follicle mRNA species of approximately 6 kb in length (Fig. 4.4)

Figure 4.1 Initial restriction map of λsTr1.

A basic restriction map of the cDNA insert from λsTr1 was constructed using restriction analysis of pGEM-1.9, pGEM-0.47 and λsTr1 itself. The first 600 bp of the insert contains numerous PstI sites, the location of which is uncertain (P(?)). Note that one half of the insert contains unique restriction sites for EcoR I (E), Hind III (H), BamH I (B) and Sac I (S) whilst the other half contains multiple Pst I (P) sites. E' indicates terminal EcoR I linkers added during cDNA clone formation.


E'
P(?)
P
H
S
E
B
E'

Figure 4.2 Detailed restriction map of λsTr1 which also depicts the sequencing strategy.

The restriction map of λsTr1 is shown and includes all the restriction sites used for subcloning and sequencing. The sequencing strategy is also depicted, with arrows indicating the extent and direction of sequencing reactions. Hatched and open bars show the coding and 3' non-coding regions, respectively. B, BamH I; D, Dra I; E, EcoR I; H, Hind III; P, Pst I; S, Sac I. E' indicates the terminal EcoR I linkers added during cDNA clone formation.

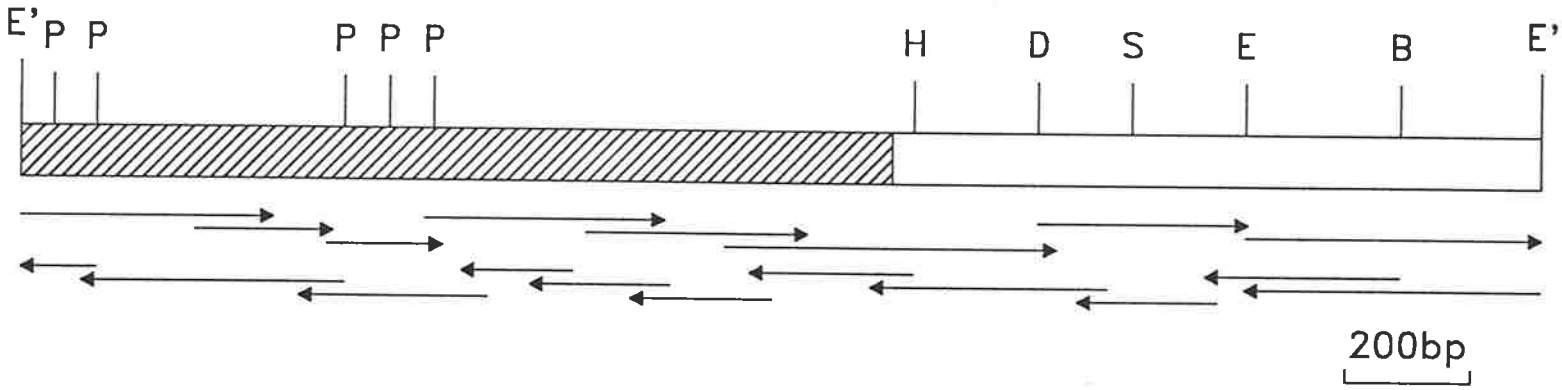

E'
P
P
P
P
P
H
D
S
E
B
E'
200bp

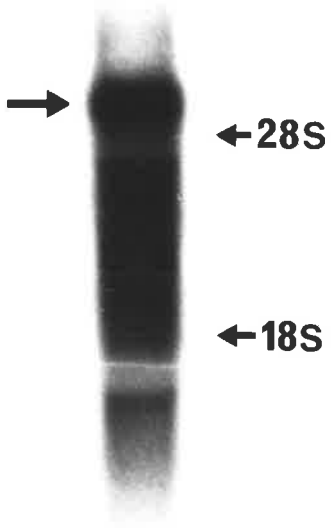
Figure 4.3 The nucleotide and predicted amino acid sequence of λsTr1.

The complete nucleotide sequence and the deduced amino acid sequence are shown. The restriction endonuclease sites, depicted in Figure 4.2, and the likely polyadenylation signal, which is present within the 3' non-coding region, are underlined.

```
1   L R R Q E R D R K F R E E E Q L L Q E R E E Q L R R Q E R D
GCTGCGCCGCCAAGAACGCGACAGAAAGTTCCGTGAGGAGGAACAGCTCCTGCAGGAAAGGGAAGAACAGCTGCGCCGCCAAGAACGCGA 90
                                                  Pst I

31  R K F R E E E Q L L Q E R E E Q L R R Q E R D R K F R E E E
CAGAAAGTTCCGTGAGGAGGAACAGCTCCTGCAGGAAAGGGAAGAACAGCTGCGCCGCCAGGAACGCGACAGAAAGTTCCGCGAAGAGGA 180
                              Pst I

61  Q Q L R L L E R E Q Q L R Q E R N R K F R E E Q L L R E R E
ACAGCAGCTGCGCCTCCTGGAACGCGAGCAACAGCTACGCCAAGAACGAAATAGAAAATTCCGCGAAGAACAGCTCCTGCGAGAAAGGGA 270

91  E Q L R L Q E G E P Q L R Q K R D R K F H E E E Q L L Q E R
AGAACAGCTGCGCCTCCAGGAGGGCGAGCCGCAGCTTCGCCAGAAGCGCGATAGAAAGTTCCATGAGGAGGAACAGCTCCTGCAAGAAAG 360

121 E E Q L R R Q E R D R K F R E E E Q L L Q E R E K L R R Q E
AGAAGAACAGCTGCGCCGCCAGGAACGCGACAGAAAGTTCCGTGAGGAGGAACAGCTCCTGCAAGAAAGAGAAAAGCTGCGCCGCCAGGA 450

151 R E P Q L R Q E R D R K F H E E E Q L L Q E R E E Q L R R Q
GCGCGAGCCACAACTTCGCCAGGAACGCGACAGAAAGTTCCATGAGGAGGAACAGCTCCTGCAGGAAAGGGAAGAACAGCTGCGCCGCCA 540
                                                  Pst I

181 E R D R K F R E E E Q L L Q E R E E Q L R R Q E R D R K F R
GGAACGCGACAGAAAGTTCCGTGAGGAGGAACAGCTCCTGCAGGAAAGGGAAGAACAGCTGCGCCGCCAGGAGCGCGACAGAAAGTTCCG 630
                                  Pst I

211 E E E Q L L Q E R E E Q L R R Q E R D R K F R E E E Q L L K
TGAGGAGGAACAGCTCCTGCAGGAAAGGGAAGAACAGCTGCGCCGCCAGGAACGCGACAGAAAGTTCCGTGAGGAGGAACAGCTCCTGAA 720
                  Pst I

241 E S E E Q L R R Q E R D R K F H E K E H L L R E R E E Q Q L
AGAAAGCGAAGAACAGCTCCGCCGCCAGGAGCGCGACAGGAAGTTCCATGAGAAAGAACACCTCCTGCGAGAAAGGGAGGAACAGCAGCT 810

271 R R Q E L E.G V F S Q E E Q L R R A E Q E E E Q R R Q R Q R
GCGCCGTCAGGAACTTGAGGGGGTCTTCTCCCAGGAGGAACAGCTGAGGCGCGCCGAGCAAGAGGAAGAACAGCGACGTCAGAGGCAGAG 900

301 D R K F L E E G Q S L Q R E R E E E K R R V Q E Q D R K F L
AGACAGGAAATTCCTCGAGGAAGGGCAGAGCCTCCAGCGCGAGCGAGAGGAAGAAAAGCGCCGCGTCCAGGAGCAGGACAGGAAGTTCCT 990

331 E Q E E Q L H R E E Q E E L R R R Q Q L D Q Q Y R A E E Q F
CGAGCAGGAAGAGCAGCTGCACCGCGAGGAGCAGGAAGAGCTGAGGCGCCGGCAGCAGCTAGACCAGCAGTACCGGGCGGAGGAGCAGTT 1080

361 A R E E K R R R Q E Q E L R Q E E Q R R R Q E R E R K F R E
TGCTAGGGAGGAGAAGAGGCGTCGTCAGGAACAAGAATTGAGGCAAGAAGAGCAGAGACGCCGCCAGGAGCGGGAGAGGAAATTCCGGGA 1170

391 E E Q L R R Q Q Q E E Q K R R Q E R D V Q Q S R R Q V W E E
AGAAGAACAGCTCCGCCGCCAGCAGCAGGAGGAGCAGAAGCGTCGCCAAGAGCGCGACGTGCAGCAGAGCCGCCGCCAAGTGTGGGAGGA 1260

421 D K G R R Q V L E A G K R Q F A S A P V R S S P L Y E Y I Q
AGACAAGGGCCGCCGGCAGGTCCTGGAGGCTGGCAAGCGGCAGTTTGCCAGTGCCCCAGTGCGCTCCAGTCCGCTCTACGAGTACATCCA 1350

451 E Q R S Q Y R P *
AGAGCAGAGGTCTCAGTACCGCCCTTAGGAGATGCTGCCAAATCCCGACATCTGCCGAAGCTTCGAGCAAAGGAAAATGAGAATCACTGA 1440
                                                  Hind III

GTACCAAATGACTCTGGTTGTGGGAAAACTCTGGTGTTAGACTAACTCTTTTTTACAAAATCTTTAATCTATACTTTTTCATGTGCTTTG 1530

TACTTCTGCCTTTTATTCTTCCTTAAATAGTTCTTTAGGATGTCTTTGCTCTTTGGTGCAGATTTGGTGTGCATTTTTAAAAAACATAAA 1620
                                                  Dra I

AGCCATTTAATTTGTTTAAGGAATTTTGTTTGGGAAACACGTTCATTCATTGCCTTCAGAAGTAACAAAAATATTGTGTCCATTTGAGAT 1710

TCAAAAGAATGGGTCAGCTTTTTTATTGTTGATCCATCTTATAGAGAGCTCAGATATTTTTTATGTTCAAGTTGATATTTCTTCCTGGGC 1800
                                  Sac I

CTAAATTATGTTAATATTTATCTCCAAATAGCCTCCCACTTTTTGTGGCATAATTAGCACAGATTCTGCAAGGGGACCGAATTTTTCCAA 1890

GAACCCCTGAATAGTGCAGGAAGAATGGCTTCTCCAGAAAAAGTCTCTGAATTCAGCCCATAATTGGAAGAATTATCTTTAAGACTTGGA 1980
                                  EcoR I

ATTAGATTTCTTTTCCATTTAATTCCATAAATACTTAAAATATCATGAAGCAAAAAGCAAGGCGTTTCTCAAACAACTCCAGAGTTCAAA 2070

CTTTAAAACCATGTCTGCTGTAGTCTGTAGCATTCAGTTCCTTTCCCCCAGTCTTGGATGAGCTTGAGAATATTACTGCCTTTGTTAATT 2160
Dra I

TTAGCTGAGAAGGGACCTGCTCAGGATCCTGTAAACATCTGTCTTGCTCTAGAGCCAACAAGAGATCACAGAGCATTTGGGGGTGGGGAA 2250
                  BamH I

AAGGGAAGTTTTGTGGACAGAAGGGCAAGTCCCTTGAGGACCTTTGACAAGCCCTGTCAGCCCACAATCCTTTGAGCCCTACTGATACTA 2340

CCTTGGAGACATGTTAAAATAATTGGACTGTGTAAAAAATTAATAATAAATATTTTGGCAATTAAAAA 2408
```

Figure 4.4 Northern blot analysis of sheep follicle RNA.

A sample of total sheep follicle RNA (5 μg) was denatured with glyoxal and fractionated on a 1% agarose gel. The resultant filter was probed with the 1.9 kb EcoR I fragment of λsTr1. Full-length trichohyalin mRNA (approx. 6 kb) is detected (arrow) together with partially degraded message. Ribosomal RNA marker positions are indicated.


28S
18S

showing that λsTr1 is derived from an mRNA species large enough to encode trichohyalin.

Comparison of the guinea pig trichohyalin peptide sequences with the deduced amino acid sequence of λsTr1 showed that, although there are no identical matches, four of the five guinea pig sequences are homologous to sequences from λsTr1 (Fig. 4.5). Taking into account the probable cross-species sequence differences, the high degree of similarity indicates that λsTr1 does indeed code for sheep trichohyalin.

4.2.4 Analysis of the Predicted Amino Acid Sequence

a. Deduced Protein Size

The open reading frame of λsTr1 encodes the 458 C-terminal amino acids of sheep trichohyalin, which have a predicted molecular weight of approximately 60,000. Thus λsTr1 encodes approximately 30% of the native 190 kdal protein.

b. Amino Acid Composition

The amino acid composition of the deduced partial trichohyalin sequence is shown in Table 4.1 and compared with that of the complete trichohyalin protein (see Section 3.2.3a). The encoded protein is extremely hydrophilic, with over 59% of the residues being charged and a further 20% being polar. The most abundant amino acids are, in decreasing order, glutamic acid/glutamine, arginine, leucine and lysine, which corresponds with the analysis of total wool trichohyalin (Table 4.1). However, the absolute levels of glutamic acid/glutamine and arginine are considerably higher in the deduced protein sequence than in the total protein which suggests that they are enriched within the C-terminal third of trichohyalin. Interestingly, there are no sulphur-containing amino acids in the deduced sequence, which correlates with their low level in total trichohyalin (Table 4.1) and also indicates that this portion of trichohyalin is unable to form any of the intermolecular disulphide cross-links present in the protein precipitate formed during protein concentration (Section 3.2.1)

c. Protein Sequence Comparisons

The deduced amino acid sequence of λsTr1 was compared with the available IF amino acid sequences (Conway and Parry, 1988). No significant homology with the IF sequences was detected and, significantly, no region comparable to the IF α-helical core or heptad repeat structure was found within the deduced sequence. Comparisons were

Figure 4.5 Comparison of guinea pig trichohyalin peptide sequences with regions of the deduced cDNA amino acid sequence.

The sequences of four of the guinea pig trichohyalin peptides (B, D, F1 and M; Section 3.2.3d) are shown together with the most comparable regions of the trichohyalin amino acid sequence derived from λsTr1. Residues within the peptides which are identical to the corresponding amino acid in the cDNA deduced sequence are shown in bold. The position of the first and last residue within each sequence is also shown.

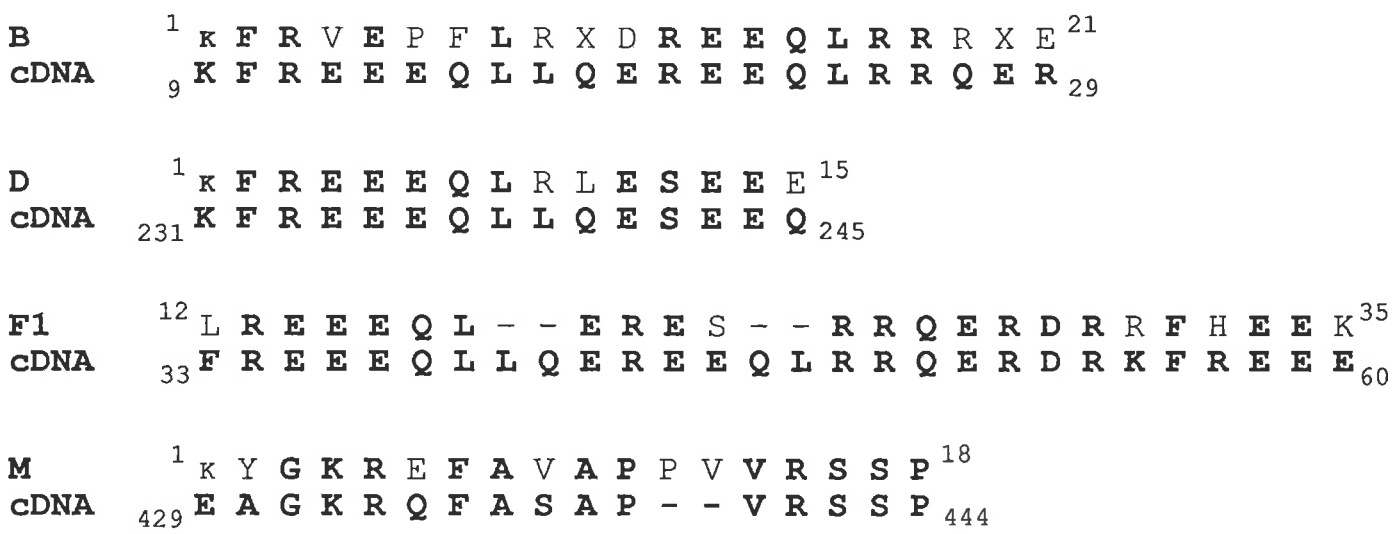

B 1 K F R V E P F L R X D R E E Q L R R R X E 21
cDNA 9 K F R E E E Q L L Q E R E E Q L R R Q E R 29
D 1 K F R E E E Q L R L E S E E E 15
cDNA 231 K F R E E E Q L L Q E S E E Q 245
F1 12 L R E E E Q L - - E R E S - - R R Q E R D R R F H E E K 35
cDNA 33 F R E E E Q L L Q E R E E Q L R R Q E R D R K F R E E E 60
M 1 K Y G K R E F A V A P P V V R S S P 18
cDNA 429 E A G K R Q F A S A P - - V R S S P 444

Table 4.1 Comparison of the amino acid content of the deduced cDNA sequence with that of wool follicle trichohyalin (see Table 3.2).

Amino Acid	Deduced cDNA Sequence	Wool Follicle Trichohyalin
	mole percent	*mole percent*
Asp/Asn	3.3/0.2	6.4
Thr	0.0	2.9
Ser	1.7	5.4
Glu/Gln	26.0/18.1	28.0
Pro	1.1	3.3
Gly	1.1	5.3
Ala	1.3	4.7
1/2-Cys	0.0	0.6
Val	1.3	4.0
Met	0.0	0.1
Ile	0.2	2.5
Leu	10.9	10.0
Tyr	0.9	2.1
Phe	3.7	2.4
Lys	5.0	6.7
His	1.1	1.7
Arg	23.8	13.7
Trp	0.2	0.2

made with the sequences within the Genbank and NBRF databases but no homologous sequences were found.

d. Secondary Structure Analysis

Although no IF-like sequences or heptad repeats were found within the deduced trichohyalin sequence, secondary structure analysis predicted that the majority of the protein has the propensity to form α-helix (Fig. 4.6a). Amidst the predicted α-helical regions, regularly spaced short sequences are predicted to form random coil or turn. No region within the deduced sequence was predicted to be involved in the formation of ß-sheets (data not shown).

e. Repetitive Protein Structure

The regular spacing of the predicted random coil regions (Fig. 4.6a) has suggested the presence of a repeated region. This correlates with evidence from the guinea pig peptide sequences (see Section 3.2.3d). A dot matrix plot was made of the λsTr1 amino acid compared with itself (Fig. 4.7). This plot revealed a large number of diagonals spaced in the main by about 25 amino acids which cover all but the C-terminal 50 residues. The presence of these diagonals suggests that a sequence within the clone is continually repeated along the length of the clone. More detailed analysis determined that the deduced λsTr1 coding sequence contains a series of tandem repeats which are based on a 23 amino acid consensus sequence shown at the top of Fig. 4.8. When the sequence is aligned with respect to the repeat, it is seen to contain 25 full or partial length repeats which extend from the beginning of the sequence to 29 amino acids from the C-terminus. The first 14 repeats are highly conserved with respect to the consensus whereas the last 11 repeats show considerably less conservation.

Analysis of the consensus sequence indicates that although 15 of the 23 amino acids are charged it has a net charge of only -1. The consensus contains all of the substrate residues for peptidylarginine deiminase and transglutaminase, i.e. arginine, glutamine and lysine, with multiple arginine and glutamine residues. As stated earlier (Section 4.2.4c), coiled-coil heptad repeats are not present in λsTr1, yet a differing heptad structure is seen within the consensus with a series of three hydrophobic amino acids positioned 7 residues apart, i.e. phenylalanine at position 4, leucine at 11 and leucine at 18 (Fig. 4.8). Secondary structure analysis of the consensus sequence indicates that the whole region is predicted to have the ability to form α-helix although the

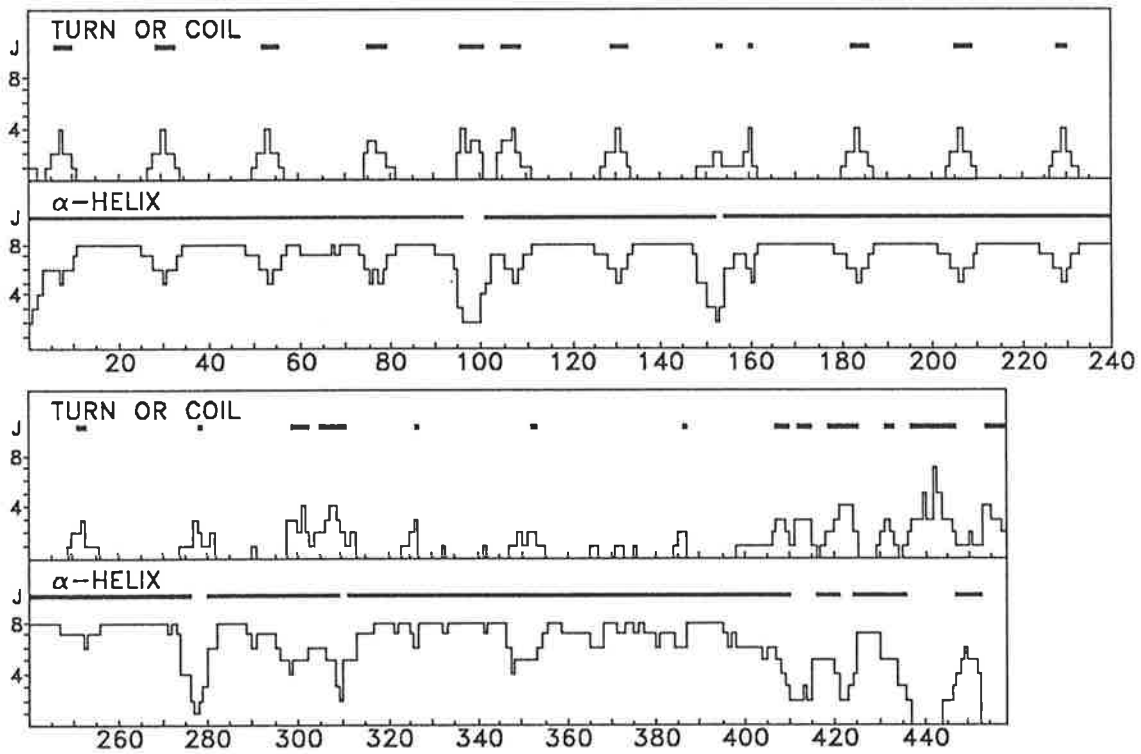
Figure 4.6 The predicted secondary structure of the deduced amino acid sequence.

The secondary structure of the deduced amino acid sequence was analyzed by the program PREDICT (see Section 2.2.5c), and the number of predictions for α-helix and for turn or coil are graphed. Predicted regions of the given secondary structure are indicated by the solid bars at the top of each section (labelled J).

(a) Secondary structure predictions for the total deduced protein sequence. Note that α-helix is predicted for the vast majority of the sequence and that small regions of coil, which predominantly overlap regions of α-helix, are also predicted.

(b) The predicted structure for two consecutive repeats is enlarged (e.g., from residue 183 to 208). Note that the region from asp(1) to phe(4) (positions 1-4 and 24-27) could be either α-helix or random coil.

a.


TURN OR COIL
α-HELIX
20 40 60 80 100 120 140 160 180 200 220 240
TURN OR COIL
α-HELIX
260 280 300 320 340 360 380 400 420 440

b.

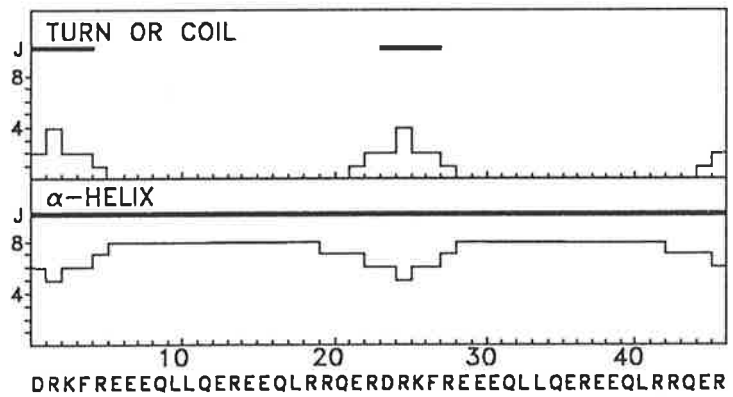

TURN OR COIL
α-HELIX
10 20 30 40
DRKFREEEQLLQEREEQLRRQERDRKFREEEQLLQEREEQLRRQER

Figure 4.7 Analysis of peptide repeats within the deduced amino acid sequence.

The predicted amino acid sequence of λsTr1 was analyzed for similar internal sequences using the computer program DIAGON (Staden, 1982) with a window length of 23. The output of this comparison was then plotted. Internal similarities are represented by the lines parallel to the central diagonal. The axes are labelled in residue numbers.

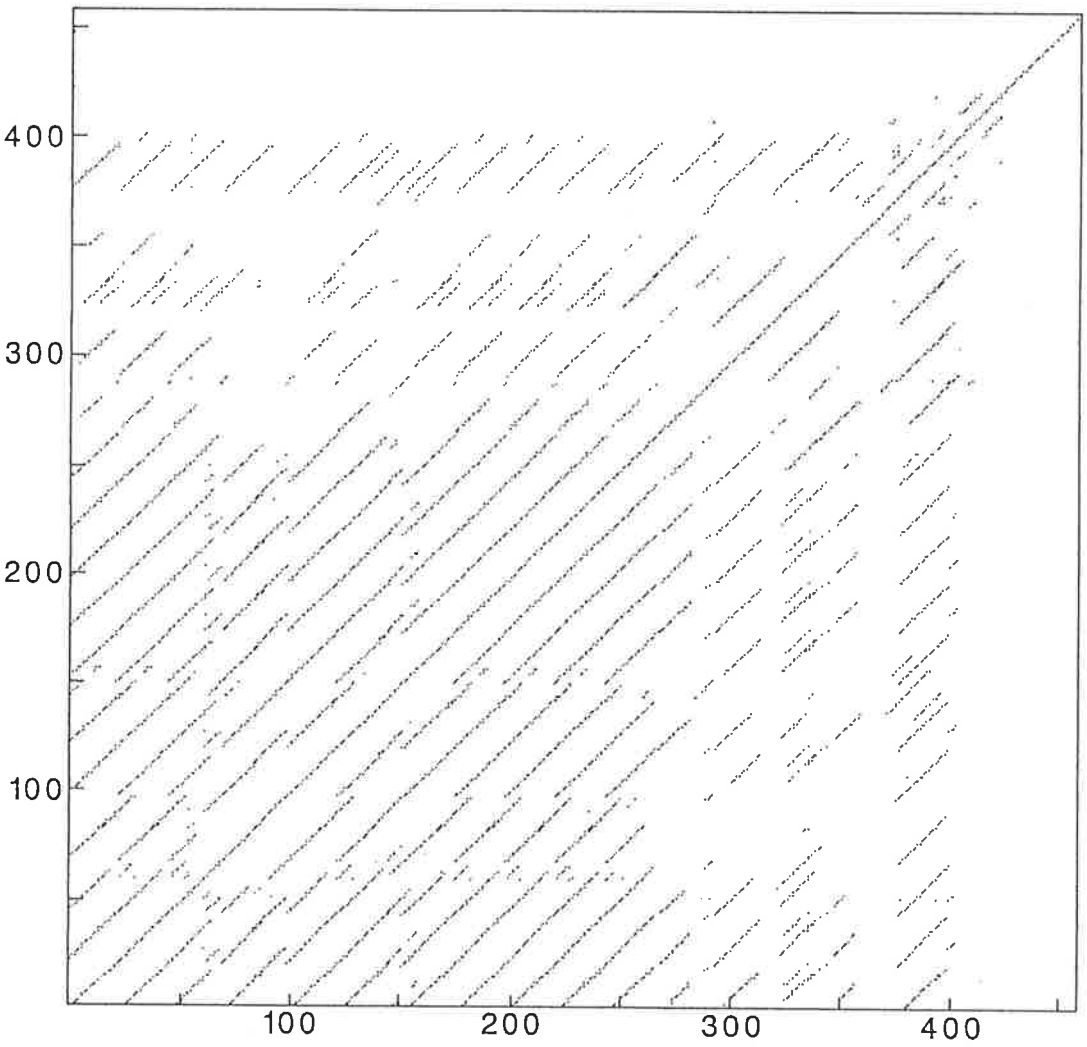
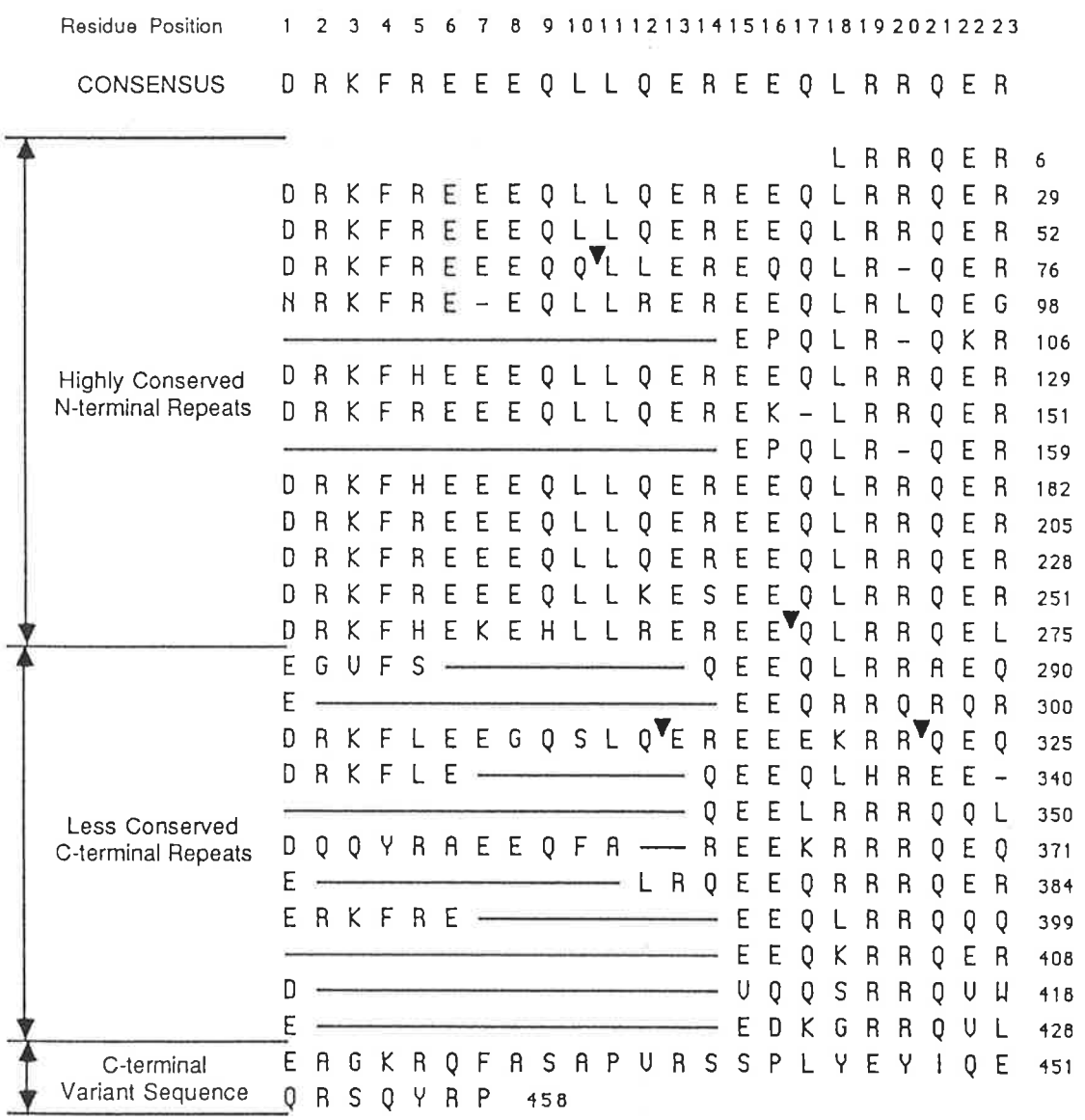

400
300
200
100
100
200
300
400

Figure 4.8 The amino acid sequence deduced from λsTr1.

The predicted amino acid sequence is aligned with respect to the 23 amino acid consensus sequence (top) which was derived from the N-terminal segment of highly conserved repeats. Dashes and arrow heads indicate space insertions or sequence deletions that have been introduced for optimal alignment. Each arrowhead indicates the removal of only one or two amino acids.


Residue Position 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
CONSENSUS D R K F R E E E Q L L Q E R E E Q L R R Q E R
L R R Q E R 6
D R K F R E E E Q L L Q E R E E Q L R R Q E R 29
D R K F R E E E Q L L Q E R E E Q L R R Q E R 52
D R K F R E E E Q Q L L E R E Q Q L R - Q E R 76
N R K F R E - E Q L L R E R E E Q L R L Q E G 98
E P Q L R - Q K R 106
D R K F H E E E Q L L Q E R E E Q L R R Q E R 129
D R K F R E E E Q L L Q E R E K - L R R Q E R 151
E P Q L R - Q E R 159
D R K F H E E E Q L L Q E R E E Q L R R Q E R 182
D R K F R E E E Q L L Q E R E E Q L R R Q E R 205
D R K F R E E E Q L L Q E R E E Q L R R Q E R 228
D R K F R E E E Q L L K E S E E Q L R R Q E R 251
D R K F H E K E H L L R E R E E Q L R R Q E L 275
E G V F S Q E E Q L R R A E Q 290
E E E Q R R Q R Q R 300
D R K F L E E G Q S L Q E R E E E K R R Q E Q 325
D R K F L E Q E E Q L H R E E - 340
Q E E L R R R Q Q L 350
D Q Q Y R A E E Q F A R E E K R R R Q E Q 371
E L R Q E E Q R R R Q E R 384
E R K F R E E E Q L R R Q Q Q 399
E E Q K R R Q E R 408
D V Q Q S R R Q V W 418
E E D K G R R Q V L 428
E A G K R Q F A S A P V R S S P L Y E Y I Q E 451
Q R S Q Y R P 458
Highly Conserved N-terminal Repeats
Less Conserved C-terminal Repeats
C-terminal Variant Sequence

first four residues of the consensus (asp(1) to phe(4)) are also predicted to be able to form a short random coil (Fig. 4.6b).

4.2.5 Detection of Sheep Trichohyalin Gene Sequences

Sheep genomic DNA samples were digested with BamH I, EcoR I and Hind III, electrophoresed, blotted and hybridised separately with the two EcoR I fragments of λsTr1. The 1.9 kb EcoR I fragment contains the repetitive coding region of λsTr1 and will detect any genes containing the repetitive sequence, thus determining whether trichohyalin is present as a single gene or as a family of genes.

The 1.9 kb EcoR I probe detected a single band within all 3 tracks (Fig 4.9) indicating that the sheep genome does indeed contain only a single trichohyalin gene. Additionally, the hybridisation to one rather than two Hind III fragments indicates that the genomic sheep sequence appears to differ from that of the cDNA clone as the Hind III site within the 1.9 kb EcoR I fragment of λsTr1 does not appear within the genomic DNA.

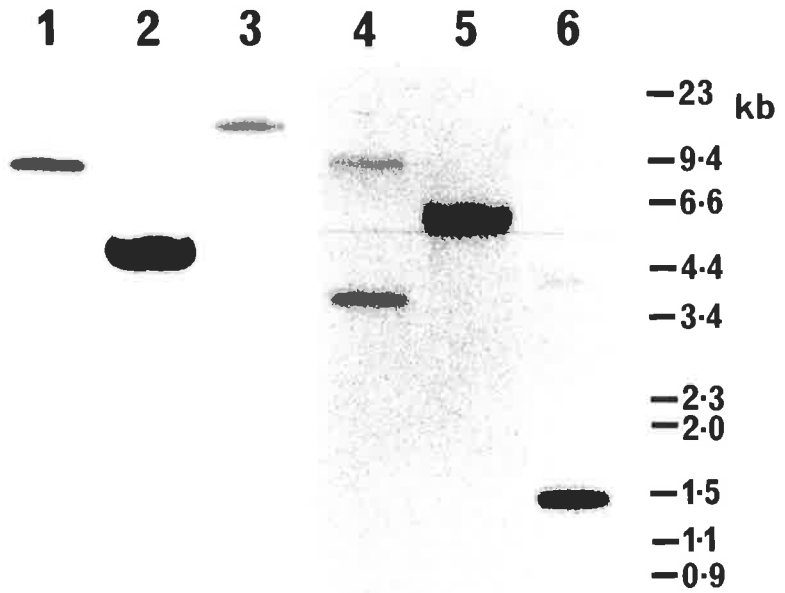
The hybridisation pattern of the 0.47 kb EcoR I fragment from the 3' non-coding region of λsTr1 also demonstrates differences between the cDNA and genomic sequences. As expected, a single band, differing in size to that detected by the coding probe, was detected in the EcoR I track and two approximately equal intensity bands were detected within the BamH I track. Yet, within the Hind III track, in addition to the 13 kb band bound by the 1.9 kb probe (not seen in Fig. 4.9 but visible after prolonged exposure), two other bands, both larger than 0.47 kb, are also detected. This indicates that a Hind III site which is not present within the 0.47 kb EcoR I fragment of the cDNA sequence is found within the gene. Also, the hybridisation of the 0.47 kb probe to three bands larger than itself suggests that there is an intron present within the 3' non-coding region or that there may have been only partial digestion of the genomic DNA with Hind III. The second possibility is unlikely due to the presence of only a single band when the genomic DNA was probed with the 1.9 kb EcoR I fragment (see Fig. 4.9, Track 3).

4.2.6 Examination of Cross-species Nucleotide Sequence Homology

Immunological studies had previously shown that trichohyalin purified from various mammalian species was highly immunocross-reactive (Rothnagel and Rogers, 1986). To examine the sequence homology of different species, a filter containing sheep, human and mouse genomic DNA digested with EcoR I was hybridised with the 1.9 kb

Figure 4.9 Southern blot analysis of sheep genomic DNA.

Total sheep genomic DNA (4 μg/lane) was digested with BamH I (lanes 1 and 4), EcoR I (2 and 5), and Hind III (3 and 6). Lanes 1-3 were probed with the 1.9 kb EcoR I fragment of λsTr1 (coding) and lanes 4-6 were probed with the 0.47 kb EcoR I fragment (3' non-coding). After prolonged exposure of lane 6, an additional band is seen and is the same size as the band present in lane 3 (13 kb). The filters were washed in 2x SSC, 0.1% SDS at 65°C. The autoradiograph of lanes 1-3 was exposed for 20 h and that of lanes 4-6 for 6 d. Size markers are shown (in kilobase pairs).


1
2
3
4
5
6
23
kb
9·4
6·6
4·4
3·4
2·3
2·0
1·5
1·1
0·9

EcoR I fragment of λsTr1. Figure 4.10 shows an autoradiograph of the resultant filter after washing at low stringency. Although the probe hybridised to a single band within the DNA of all 3 species, the signal intensity varied greatly. The hybridisation to the human DNA was much weaker than to the sheep DNA but was much stronger than to the mouse DNA, where only a very weak signal was obtained.

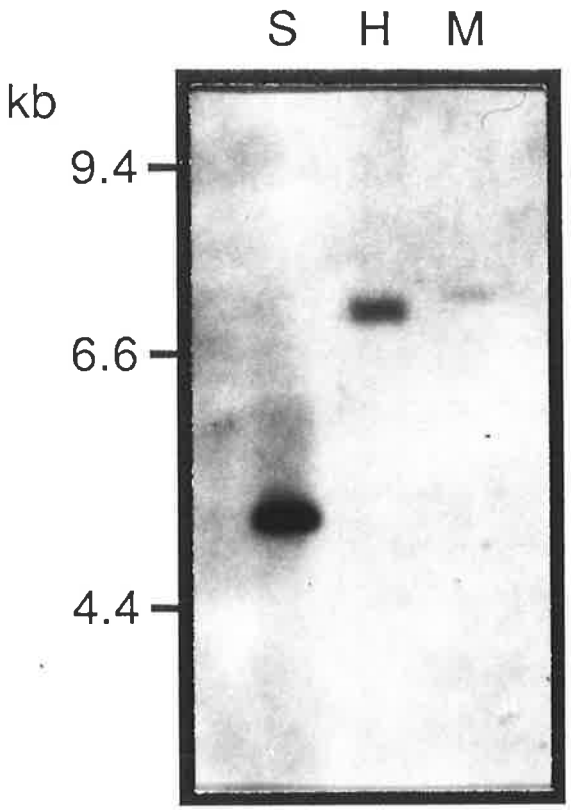
4.3 Discussion

This chapter has described the characterisation and sequencing of a partial sheep trichohyalin cDNA clone, λsTr1. The clone encodes 2,408 bp of the approximately 6 kb sheep wool follicle trichohyalin mRNA (Fig. 4.4), consisting of 1,375 bp of the coding sequence, the complete 3' non-coding region of 1025 bp and a very short poly A tail. The coding region encodes 458 amino acids which would form a partial protein with a predicted mass of 60 kdal, i.e., about 30% of the intact trichohyalin molecule. The deduced protein sequence has a similar amino acid composition to purified wool follicle trichohyalin (Table 4.1) and contains regions which are nearly identical in sequence to the purified guinea pig trichohyalin peptides (Fig. 4.5) confirming that λsTr1 does encode trichohyalin.

Analysis of the deduced protein sequence has shown that 95% of the sequence consists of full or partial length repeats of a 23 amino acid sequence (Fig. 4.8). This confirms predictions made by Rothnagel and Rogers (1986) based on the two-dimensional gel analysis of trichohyalin breakdown products and also those made from the homology of the guinea pig peptide sequences (Section 3.3). The consensus sequence is characterised by a high proportion of charged residues (15/23) and the presence of the substrate amino acids for peptidylarginine deiminase and transglutaminase. As there is only a single lysine residue within each repeat it is highly likely that it is a substrate for follicle transglutaminase and that the adjacent phenylalanine and surrounding charged residues may be required for its utilisation by transglutaminase. In addition, the glutamine at position 17 of the consensus sequence (Fig. 4.8) is in a charged environment (EEQLRR) which is similar to that of the cross-linked glutamine residue in fibrin (EGQQHH), another transglutaminase substrate (Chen and Doolittle, 1971). Thus the glutamine at position 17 may be cross-linked by follicle transglutaminase. Interestingly, the proposed glutamine and lysine substrate sites for

Figure 4.10 Hybridisation analysis of a genomic Zoo blot .

Sheep, human and mouse total genomic DNA (4 μg/lane) were digested with EcoR I, separated on a 1% agarose gel, transferred to Zetaprobe membrane and probed with the 1.9 kb EcoR I fragment of λsTr1. The filter was washed in 2x SSC, 0.1% SDS at 65°C. The autoradiograph was exposed for 11 d. Size markers are shown (in kilobase pairs).


S
H
M
kb
9.4
6.6
4.4

follicle transglutaminase are nearly identical to the two common sequences present in the guinea pig peptides (Fig. 3.11).

The repeat region of the protein can be divided on the basis of the length of each repeat and the level of amino acid conservation into amino (N)- and carboxy (C)-terminal segments. The 14 repeats within the N-terminal segment are mainly full length and are highly conserved, with 75-100% of the amino acids within each repeat identical to those of the consensus sequence (Fig. 4.8). Of the 11 repeats within the C-terminal segment, all but one are of partial length and together they show a lower degree of conservation with consensus sequence (40-89%), although each does retain an eight amino acid stretch (residues 15-22; see Fig. 4.8). This overall structure suggests that within the N-terminal segment of the deduced sequence the full 23 amino acid repeat is functionally required, whereas at the carboxy terminus only a smaller region is necessary. The high levels of arginine and glutamic acid/glutamine in the deduced sequence with respect to the total wool trichohyalin (Table 4.1), suggests that the 23 amino acid repeat is not present throughout the complete trichohyalin molecule and that the remaining two-thirds of the protein will contain sequences distinct from the repeat with reduced levels of arginine and glutamic acid/glutamine.

As stated earlier, two epidermal structural proteins, involucrin and filaggrin, are substrates for epidermal transglutaminase and peptidylarginine deiminase respectively (Section 1.4). Like trichohyalin, both of these proteins also consist of tandem peptide repeats (Eckert and Green, 1986; Tseng and Green; 1988; Rothnagel *et al.*, 1987; Haydock and Dale, 1990). Interestingly, although the sequences are not homologous the 10 amino acid human involucrin repeat (QEGQLKHLEQ) does have a number of similar characteristics to the trichohyalin repeat (DRKFREEEQLLQEREEQLRRQER). Both repeats are extremely hydrophilic with 70% of the residues in the involucrin consensus sequence and 83% of the trichohyalin consensus sequence residues being polar or charged. Of these, glutamic acid and glutamine are predominant, constituting 50% of the involucrin repeat and 48% of the trichohyalin repeat. The involucrin repeat also contains the sequence EGQLKH in which the glutamine residue is in a charged environment similar to that seen in trichohyalin and fibrin, and has also been suggested to be a substrate for transglutaminase (Eckert and Green, 1986). Both the trichohyalin and involucrin repeats also have an unusual nucleotide composition, in that the T content is

very low. Only 9% of the nucleotides in the trichohyalin consensus nucleotide sequence and 8% of those in the involucrin consensus (Eckert and Green, 1986) are thymidines.

Analysis of the deduced amino acid sequence has shown that it contains no regions with homology to the conserved IF α-helical core. Although the heptad repeat structure, typical of proteins involved in coiled-coil formation, is not present in the deduced sequence, a differing heptad repeat appears to be present within the repeat consensus sequence. In the case of the trichohyalin repeats only the first residue of the heptad (phe at position 4, leu(11) and leu(18)) is hydrophobic (Fig. 4.8) and the heptad structure is not continued between the repeats, i.e. leu(18) and phe(4) of the subsequent repeat are 9 residues apart. It is possible that this altered heptad structure may be able to form a differing elongated α-helical conformation which can still produce filaments with a diameter of 8-10 nm. Alternatively, if a short region of random coil is present within each repeat (Fig. 4.6b) this may allow the short α-helical regions within a trichohyalin molecule to interact and form a compact structure which may be able to act as an IFAP within the IRS.

Further indications with regard to the function of trichohyalin within the IRS may be gained by obtaining the remainder of the protein sequence. As stated above, the 23 amino acid repeat is unlikely to persist throughout all of the remaining two-thirds of the trichohyalin protein and it is thus possible that a region homologous to the α-helical core of the IF proteins could be present. This will be examined in Chapter 5.

Hybridisation of the repeat-containing 1.9 kb EcoR I probe to a Zoo blot containing sheep, human and mouse genomic DNA indicated that the sequence of the trichohyalin gene shows considerable divergence between species (Fig. 4.10). The hybridisation signal obtained with the hybridisation to the human and mouse DNA was considerably weaker than that obtained with the sheep DNA. This correlates with the differences between the sequences of the guinea pig proteolytic peptides and the protein sequence deduced from the cDNA clone (Fig. 4.5). Interestingly, although there are significant sequence differences between the trichohyalin genes of various species, the antigenic determinants appear to be relatively unaffected as is shown by the strong immunocross-reactivity of human, guinea pig, rat and sheep trichohyalin.

Chapter 5

Purification and Analysis of Sheep Genomic Trichohyalin Clones

5.1 Introduction

The purification and sequencing of λsTr1, a partial cDNA clone for trichohyalin, has yielded the sequence of the C-terminal 30% of the trichohyalin protein. This has provided incomplete information with regard to the overall structure and function of trichohyalin. In order to perform more detailed analysis and hopefully determine the function of trichohyalin the complete protein sequence had to be obtained. It was necessary therefore to determine whether to purify the trichohyalin gene or to obtain cDNA sequence covering the complete protein. To opt for the purification of the gene it was important to determine that the trichohyalin gene was unlikely to contain introns as the presence of introns could severely hamper the acquisition of the complete coding sequence. Genomic Southern data suggests that the trichohyalin gene may not contain introns. The 1.9 kb EcoR I fragment of λsTr1 hybridised to a 4.7 kb sheep genomic EcoR I fragment (Fig. 4.9). As the 23 amino acid repeat encoded by the 1.9 kb EcoR I fragment probably extends into the N-terminal two-thirds of the trichohyalin protein, the hybridisation to the 4.7 kb fragment would appear to limit the number and size of any introns present within the portion of the gene encoding the repeat. In addition, the genes for the epidermal transglutaminase and peptidylarginine deiminase substrates, respectively involucrin and filaggrin, do not contain any introns within their respective coding regions (Eckert and Green, 1986; R. Presland, personal communication) suggesting that this may also be the case for the trichohyalin gene. Thus it was decided to isolate, sequence and analyse the trichohyalin gene.

This chapter details the purification and analysis of clones containing all or part of the trichohyalin gene, the sequencing of the gene and the analysis of the resultant DNA and deduced protein sequences.

5.2 Results

5.2.1 Detection and Purification of Sheep Genomic Trichohyalin Clones

The isolation of the sheep trichohyalin gene was initially attempted using both cosmid and λ libraries which had been prepared using genomic DNA obtained from Merino-derived sheep breeds. In each case 2 to 5 genome equivalents of clones were plated and screened with the 1.9 kb EcoR I fragment of λsTr1. No positive clones were

detected in any of the libraries. Thus a commercially produced sheep genomic library, containing genomic DNA inserted into λEMBL3, was obtained from Clontech. Six genome equivalents of clones were screened with the 1.9 kb EcoR I cDNA fragment and 13 positive clones were detected. Of these, eight were fully purified and the λ DNA prepared.

5.2.2 Analysis and Mapping of λsGT1b

The eight purified λ clones were cleaved with EcoR I and the digests analysed by Southern hybridisation using both the 0.47 kb and 1.9 kb EcoR I trichohyalin cDNA fragments as probes (clones λsGT1b, 2, 5 and 6 are shown in Fig. 5.1). Both probes hybridised to either a single insert fragment of approximately 5.5 kb (clones λsGT1a, 5, 6, 7, 8a, 8b, ll) or to a fragment containing both insert and λ DNA (λsGT2). Resection of the insert from the EMBL-3 DNA using Sal I showed that λsGT1b contained the largest genomic DNA insert (17 kb) and this, together with the clone containing the complete 5.5 kb repeat-containing EcoR I fragment (see above), led to λsGT1b being chosen for detailed mapping and analysis.

Digestion of λsGT1b with EcoR I yielded three fragments positioned within the insert DNA (4.5 kb, 5.2 kb and 5.5 kb). These fragments were sub-cloned into the phagemid pGEM-7Zf(+) producing the clones pGEM-4.5, pGEM-5.2 and pGEM-5.5. Restriction analysis of these clones, together with Southern analysis of Sal I cut λsGT1b which had been subjected to single and double restriction digests with EcoR I, Hind III, BamH I and Kpn I, produced a restriction map of λsGT1b (Fig. 5.2). The hybridisation of the 1.9 kb EcoR I fragment to two Kpn I fragments (data not shown) allowed the orientation of the trichohyalin gene to be determined by Southern analysis of Kpn I cut λsGT1b using the 0.47 kb EcoR I fragment as a probe (Fig. 5.3).

Interestingly, on comparison of the digests of the remaining 7 clones with λsGT1b at least 3 of the clones (λsGT6, 8b and 11) contain a 300 bp insert within the 1.7 kb Kpn I/Hind III fragment which is present in the 5.2 kb EcoR I fragment of λsGT1b. It therefore appears that the two different sets of clones contain DNA representing the two trichohyalin alleles present within the genomic DNA.

Figure 5.1 Southern blot analysis of positive genomic clones.

Purified genomic clones λsGT1b, 2, 5 and 6 were digested with EcoR I, electrophoresed on a 1% agarose gel and transferred to Zetaprobe membrane. The filter was probed separately with the 1.9 kb and the 0.47 kb EcoR I fragments of λstr1. Both probes hybridised to identical sized bands within each digest. The fragment detected in the digests of λsGT1b, 5 and 6 are all of the same size; the small differences in position were caused by the "smiling" of the gel during electrophoresis. Both hybridisations were washed in 2x SSC, 0.1% SDS at 65°C. The autoradiograph of the 1.9 kb fragment hybridisation was exposed for 2 h and that of the 0.47 kb hybridisation was exposed for 6 d. Size markers are shown (in kb).

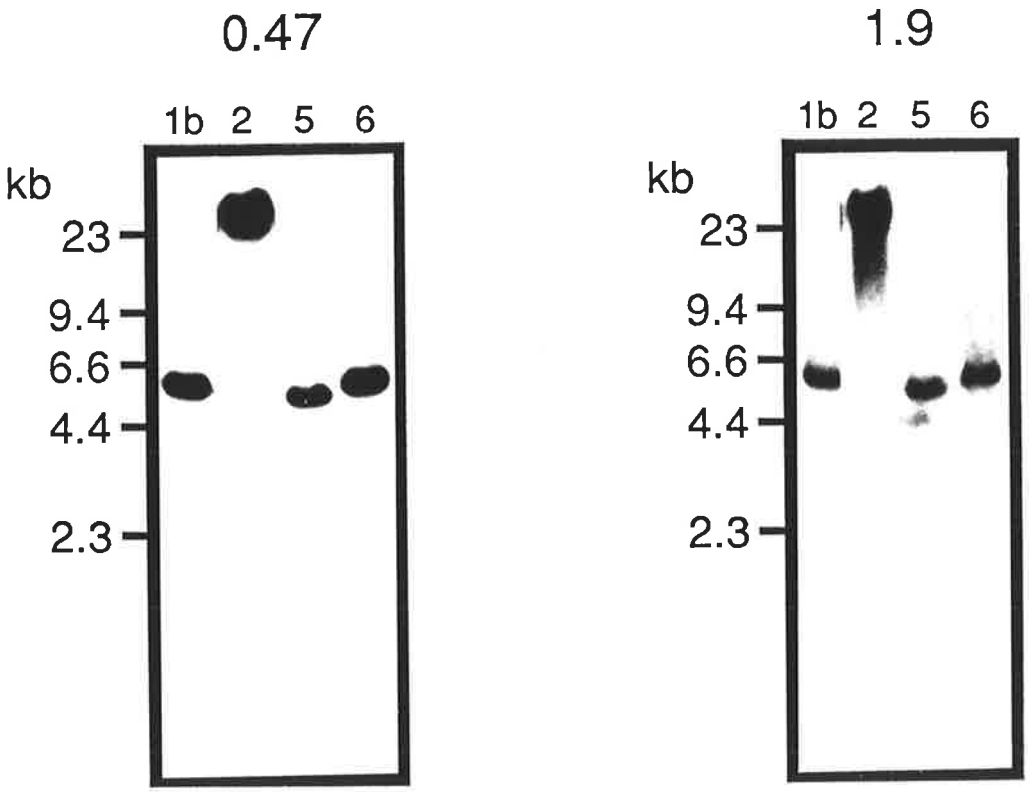

0.47
1b 2 5 6
kb
23
9.4
6.6
4.4
2.3
1.9
1b 2 5 6
kb
23
9.4
6.6
4.4
2.3

Figure 5.2 Preliminary restriction mapping of λsGT1b.

The genomic insert of λsGT1b (blocked) was mapped using the restriction enzymes BamH I (B), EcoR I (E), Hind III (H) and Kpn I (K). The Sal I restriction sites (S), which immediately flank the genomic insert, are also shown. Note that there are no internal Sal I sites.

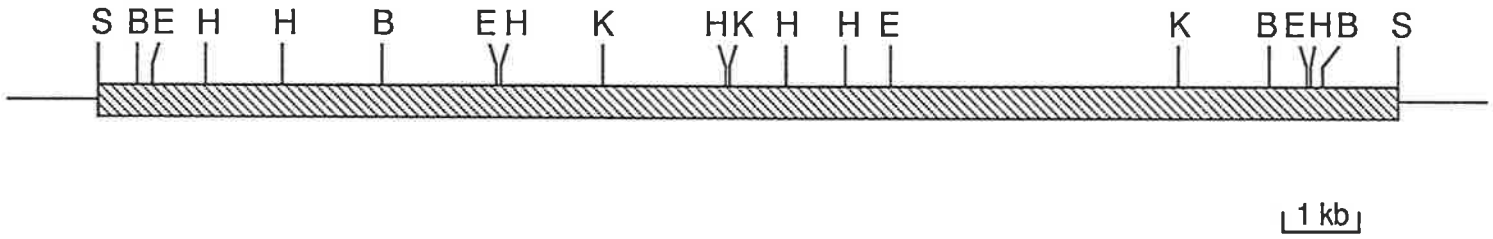
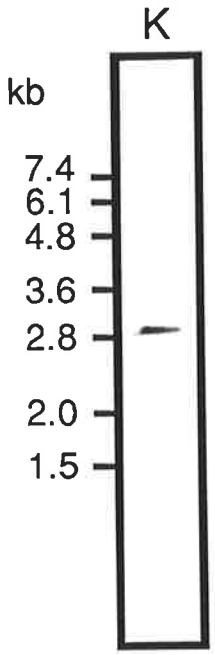

S B E H H B E H K H K H H E K B E H B S
1 kb

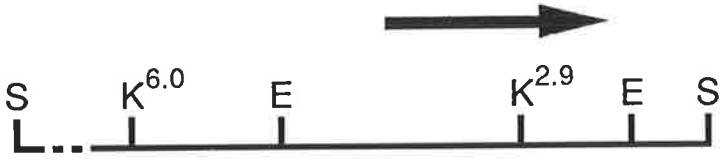
Figure 5.3 Determination of the orientation of the trichohyalin gene within λsGT1b.

The 1.9 kb EcoR I fragment of λsTr1 hybridised to two fragments produced by digestion of the λsGT1b insert with Kpn I (6 kb and 2.9 kb, data not shown). Therefore a Southern blot containing λsGT1b digested with Kpn I and Sal I was hybridised with the 0.47 kb EcoR I fragment from the 3' end of the cDNA clone, in order to determine the orientation of the trichohyalin gene. Hybridisation to the 2.9 kb fragment (a) indicated that the trichohyalin gene is transcribed in the direction shown in (b). E, EcoR I; K, Kpn I; S, Sal I. Note that the two EcoR I restriction sites flank the 5.5 kb EcoR I fragment of λsGT1b.

a.



b.



5.2.3 Sequencing the Trichohyalin Gene

From Northern analysis of total follicle RNA the trichohyalin mRNA species has been predicted to be about 6 kb long (Section 4.2.3). If the coding region of the trichohyalin gene does not contain any introns, as was proposed above (Section 5.1), then the complete coding and 3' non-coding regions should be contained within the 5.5 kb EcoR I fragment, which hybridised to both trichohyalin cDNA probes, and the upstream 5.2 kb EcoR I fragment. Thus these two EcoR I fragments of λsGT1b were mapped in greater detail using the restriction enzymes Xho I, Pst I and Xba I (Fig. 5.4).

Attempts were made to map the extent of the trichohyalin gene by probing follicle RNA with various fragments from λsGT1b (see Fig. 5.4 for fragment locations). As expected, positive signals were given by the 1.2 kb EcoR I/Pst I and 1.6 kb Xho I fragments contained within the 5.5 kb EcoR I fragment (Fig. 5.5). Unexpectedly, no signal was obtained when follicle RNA was probed with either the 5.2 kb EcoR I fragment or any of its numerous subfragments, including the 0.6 kb Hind III/EcoR I fragment which is adjacent to the 5.5 kb EcoR I fragment (Fig. 5.5). Additionally the 4.5 kb EcoR I fragment, which is the next adjacent upstream EcoR I fragment, also did not hybridise to the trichohyalin mRNA (data not shown). There appear to be three possible explanations for the Northern results: (1) the transcribed region of the trichohyalin gene is completely contained within the 5.5 kb EcoR I fragment, although this is unlikely as the trichohyalin mRNA has a predicted size of 6 kb; (2) the gene contains a large intron covering both the 5.2 kb and 4.5 kb EcoR I fragments, or; (3) there is only a short stretch of the mRNA which is transcribed from the 5.2 kb EcoR I fragment and it is not detected by Northern hybridisation. In order to resolve these alternatives the 5.5 kb EcoR I fragment and the adjacent 2 kb of the 5.2 kb EcoR I fragment were sequenced.

It was initially attempted to obtain clones for sequencing the trichohyalin gene using the Erase-a-Base nested deletion kit. This kit incorporates the enzyme Exonuclease III which, when used on correctly cut phagemid DNA, can allow the mono-directional deletion of the insert DNA and then the production of single stranded DNA for sequencing without the need for subcloning of the insert DNA. Although a number of attempts were made at producing deleted clones with the Erase-a-Base kit only 3 of the 17 non-parental clones prepared contained the desired deletion.

Figure 5.4 Detailed restriction map of the 5.2 kb and 5.5 kb EcoR I fragments of λsGT1b.

The 5.2 kb and 5.5 kb EcoR I fragments of λsGT1b were subjected to more detailed restriction mapping using the enzymes Pst I (P), Xba I (Xb), and Xho I (Xh). The orientation of the trichohyalin gene is shown. Also indicated are various subfragments which were used for subsequent analysis; these will be referred to later. B, BamH I; E, EcoR I; H, Hind III; K, Kpn I.

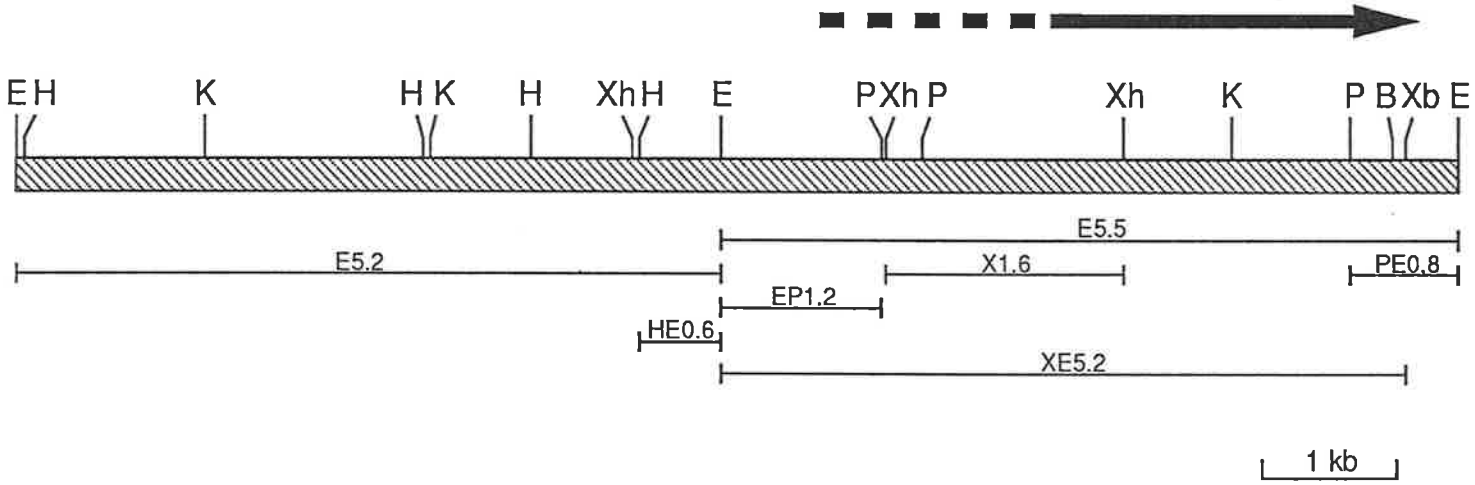
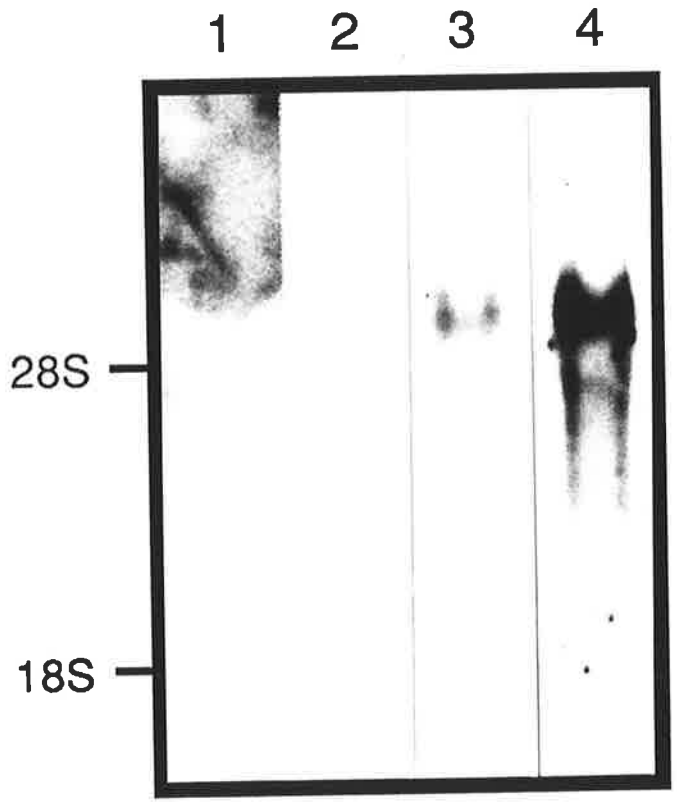

E H
K
H K
H
Xh H
E
P Xh P
Xh
K
P
B
Xb
E
E5.5
E5.2
X1.6
PE0.8
EP1.2
HE0.6
XE5.2
1 kb

Figure 5.5 Attempted location of the 5' end of the trichohyalin gene by Northern analysis.

Identical tracks, each containing 5μg of total wool follicle RNA which had been separated on a 1% agarose/formaldehyde gel, were transferred to Nytran membrane. The respective filters were probed with various fragments in an attempt to approximately locate the 5' end of the trichohyalin gene (see Fig. 5.4 for the location of the probe fragments). Track 1 was probed with the fragment E5.2, track 2 with HE0.6, track 3 with EP1.2 and track 4 with X1.6. Each of the filters were washed in 2x SSC, 0.1% SDS at 65°C. The autoradiograph of tracks 1, 2 and 4 were exposed for 16 h whilst that of track 3 was exposed for 4 d. Ribosomal marker positions are indicated.


1
2
3
4
28S
18S

Therefore the majority of the sequence was obtained using fragments prepared by deletion with Bal 31 exonuclease. Subfragments of the 5.5 kb and 5.2 kb EcoR I fragments of λsGT1b were subcloned into pGEM-7Zf(+) or pGEM-3Zf(+), the desired deletions performed and the resultant fragments cloned into the suitable M13 vectors (Messing and Vieira, 1982; Norrander *et al.*, 1983). These clones were sequenced by the dideoxy chain termination method using the Klenow fragment of E. coli DNA Polymerase I (Section 2.2.3n). The complete 5.5 kb EcoR I fragment and the adjoining 1.2 kb of the 5.2 kb EcoR I fragment were sequenced in both orientations whilst much of the 900 bp immediately upstream of this was sequenced in only one orientation. The complete sequencing protocol is shown in Figure 5.6.

To obtain sequence spanning the EcoR I restriction site situated between the 5.5 kb and 5.2 kb fragments the 2.0 kb Xho I fragment containing this site (see Fig. 5.4) was subcloned into pGEM-7Zf(+), deleted with Bal 31 exonuclease and the appropriate sized clones sequenced. This indicated that the two EcoR I fragments are immediately adjacent.

Numerous ambiguities remained within the determined nucleotide sequence even after both strands had been sequenced using the Klenow fragment. Although most of these ambiguities were overcome when the respective clones were sequenced using extension with Taq I polymerase at 70°C, eight ambiguities remained. Each of these ambiguities (nucleotide positions 3891, 3957, 4095, 4539, 4605, 4671, 5583 and 5649) involved the orientation of a C/G pair and were positioned within identical environments (see Fig. 5.7). Analysis of the surrounding 6 base sequence determined that the actual sequence would be either GACGTC or GAGCTC. Both of these sequences are palindromic and are cut by the restriction enzymes Aat II and Sac I respectively. As both the surrounding sequence and the electrophoretic patterns of the dideoxy chain termination reactions are almost identical for each ambiguity it is highly probable that all of them contain the same central sequence. All of the ambiguities occur within a 3.0 kb Pst I fragment and the clone pGP3.0A, which contains the 3.0 kb Pst I fragment subcloned into pGEM-3Zf(+), was digested with both Aat II and Sac I. Analysis of the resultant restriction patterns determined that only Sac I cut within the insert (Fig. 5.7) and therefore the actual sequence is in each case GAGCTC.

The final DNA sequence, spanning 7456 bp, is shown in Figure 5.8.

Figure 5.6 The protocol used for sequencing the 5.5 kb and part of the 5.2 kb EcoR I fragments of λsGT1b.

The restriction map of the 5.2 kb and 5.5 kb EcoR I fragments is shown (reproduced from Fig. 5.4) together with the sequencing strategy. Arrows indicate the extent and direction of sequencing reactions. B, BamH I; E, EcoR I; H, Hind III; K, Kpn I; P, Pst I; Xb, Xba I; Xh, Xho I.

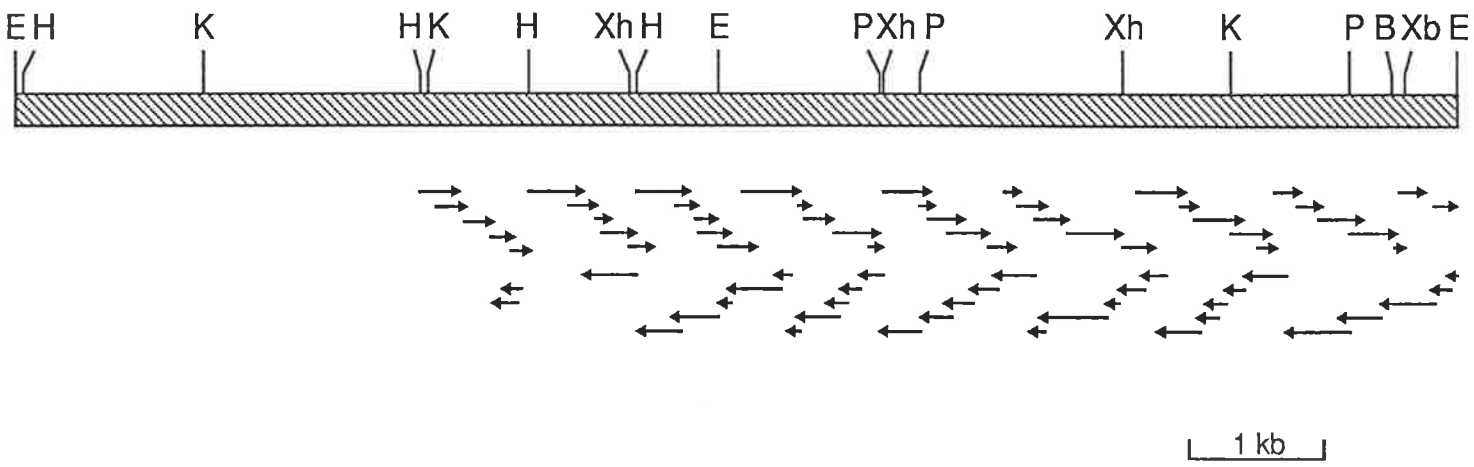
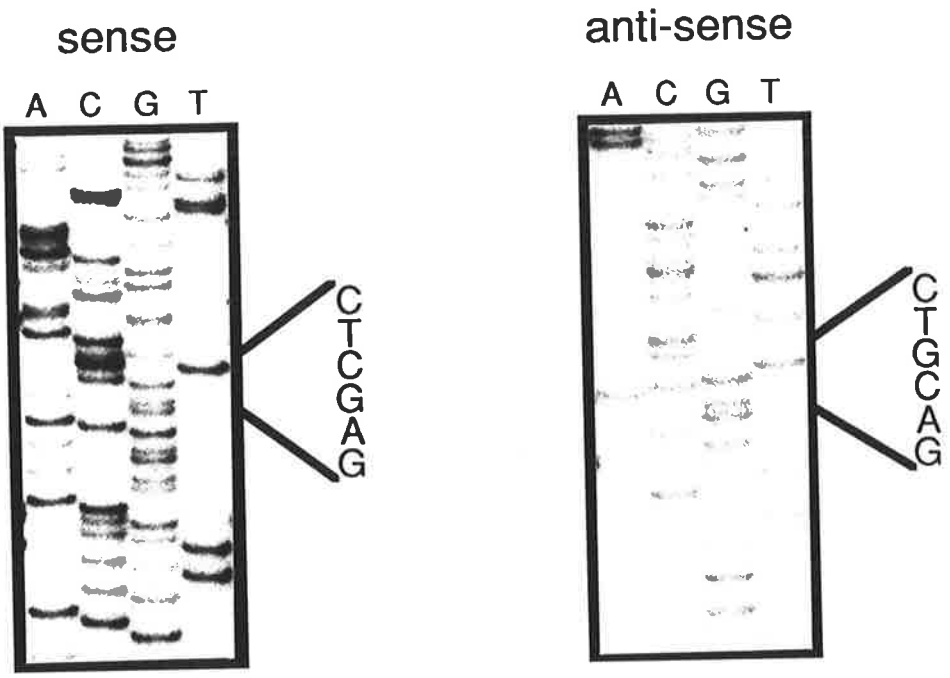

E H
K
H K
H
Xh H
E
P Xh P
Xh
K
P B Xb E
1 kb

Figure 5.7 Resolution of ambiguous sequences using restriction analysis.

After the completion of the sequencing reactions outlined in Figure 5.6, eight sequences within the presumed coding region, each of which is centred on an identical 20 base sequence (GGAGCAGGA(GC/CG)TCCGCCAGG), remained ambiguous. Shown in (a) are autoradiographs of polyacrylamide gels containing the sense and anti-sense sequencing reactions spanning one of these ambiguities. Note that the ambiguity is at the centre of the denoted sequences, i.e., GAGCTC (sense) and GACGTC (anti-sense). As these two six base sequences are the recognition sites for the enzymes Sac I and Aat II respectively, the clone pGP3.0 containing the 3.0 kb Pst I fragment of λsGT1b which encodes all eight ambiguities, was digested separately with Aat II (A) and Sac I (S). The digests were then separated on a 1% agarose gel. A photograph of the ethidium bromide stained gel is shown (b) and indicates that pGP3.0 contains only the single Aat II site present within the vector sequence but produced all of the fragments expected upon the digestion of each of the ambiguous sites with Sac I (4196, 774, 444, 246, 138, 138, 66, 66, 66, 66 and 66 bp).

a.


sense
A C G T
CTCGAG
anti-sense
A C G T
CTGCAG

b.

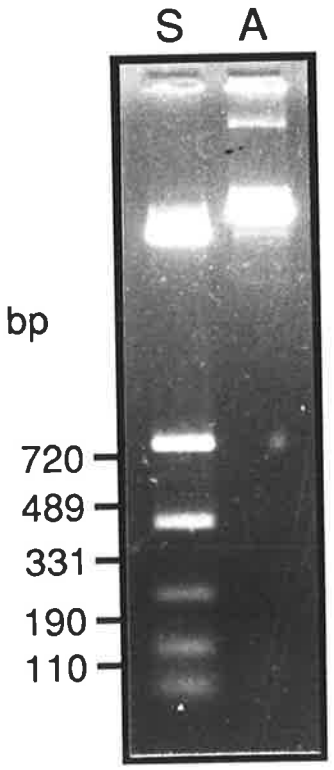

S A
bp
720
489
331
190
110

Figure 5.8 The nucleotide sequence and predicted amino acid sequence of the trichohyalin gene.

The 5.5 kb EcoR I fragment and the adjacent 2.2 kb of the 5.2 kb EcoR I fragment of λsGT1b were sequenced as outlined in Figure 5.6; the resultant sequence is shown. The complete predicted coding sequence and most of the non-coding sequence is unambiguous although the non-coding sequences from bases 1 to 526 and 755 to 941 requires confirmation by sequencing in the opposite orientation to the presently available clones. The restriction sites shown in Figure 5.4 are denoted and the proposed polyadenylation signal is underlined (position 7220). Also shown is the complete predicted amino acid sequence of trichohyalin.

Continued.....

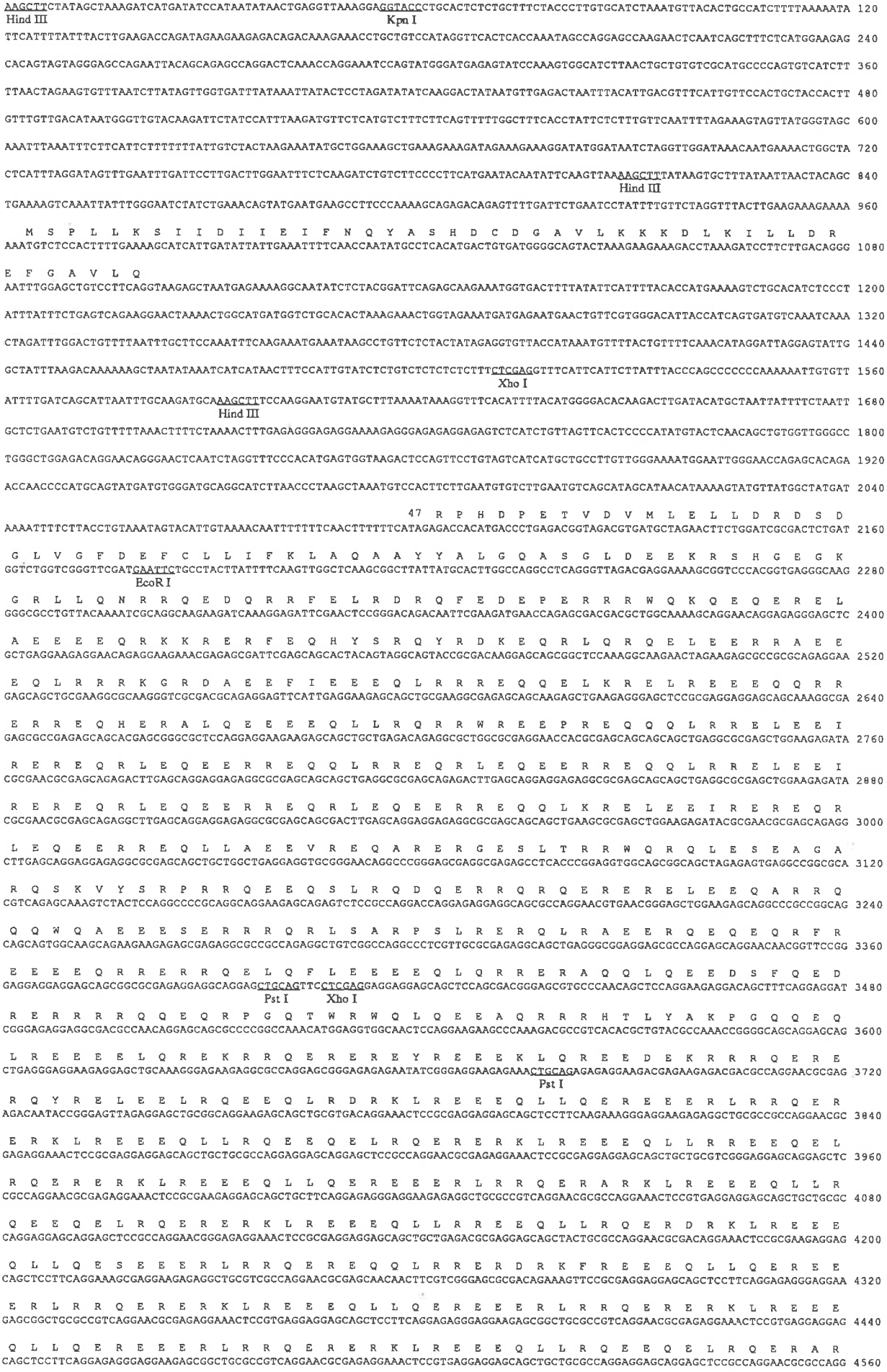

AAGCTTCTATAGCTAAAGATCATGATATCCATAATATAACTGAGGTTAAAGGAGGTACCCTGCACTCTCTGCTTTCTACCCTTGTGCATCTAAATGTTACACTGCCATCTTTTAAAAATA 120
Hind III
Kpn I
TTCATTTTATTTACTTGAAGACCAGATAGAAGAAGAGACAGACAAAGAAACCTGCTGTCCATAGGTTCACTCACCAAATAGCCAGGAGCCAAGAACTCAATCAGCTTTCTCATGGAAGAG 240
CACAGTAGTAGGGAGCCAGAATTACAGCAGAGCCAGGACTCAAACCAGGAAATCCAGTATGGGATGAGAGTATCCAAAGTGGCATCTTAACTGCTGTGTCGCATGCCCCAGTGTCATCTT 360
TTAACTAGAAGTGTTTAATCTTATAGTTGGTGATTTATAAATTATACTCCTAGATATATCAAGGACTATAATGTTGAGACTAATTTACATTGACGTTTCATTGTTCCACTGCTACCACTT 480
GTTTGTTGACATAATGGGTTGTACAAGATTCTATCCATTTAAGATGTTCTCATGTCTTTCTTCAGTTTTTGGCTTTCACCTATTCTCTTTGTTCAATTTTAGAAAGTAGTTATGGGTAGC 600
AAATTTAAATTTCTTCATTCTTTTTTTATTGTCTACTAAGAAATATGCTGGAAAGCTGAAAGAAAGATAGAAAGAAAGGATATGGATAATCTAGGTTGGATAAACAATGAAAACTGGCTA 720
CTCATTTAGGATAGTTTGAATTTGATTCCTTGACTTGGAATTTCTCAAGATCTGTCTTCCCCTTCATGAATACAATATTCAAGTTAAAAGCTTTATAAGTGCTTTATAATTAACTACAGC 840
Hind III
TGAAAAGTCAAATTATTTGGGAATCTATCTGAAACAGTATGAATGAAGCCTTCCCAAAAGCAGAGACAGAGTTTTGATTCTGAATCCTATTTTGTTCTAGGTTTACTTGAAGAAAGAAAA 960
1 M S P L L K S I I D I I E I F N Q Y A S H D C D G A V L K K K D L K I L L D R
AAATGTCTCCACTTTTGAAAAGCATCATTGATATTATTGAAATTTTCAACCAATATGCCTCACATGACTGTGATGGGGCAGTACTAAAGAAGAAAGACCTAAAGATCCTTCTTGACAGGG 1080
40 E F G A V L Q
AATTTGGAGCTGTCCTTCAGGTAAGAGCTAATGAGAAAAGGCAATATCTCTACGGATTCAGAGCAAGAAATGGTGACTTTTATATTCATTTTACACCATGAAAAGTCTGCACATCTCCCT 1200
ATTTATTTCTGAGTCAGAAGGAACTAAAACTGGCATGATGGTCTGCACACTAAAGAAACTGGTAGAAATGATGAGAATGAACTGTTCGTGGGACATTACCATCAGTGATGTCAAATCAAA 1320
CTAGATTTGGACTGTTTTAATTTGCTTCCAAATTTCAAGAAATGAAATAAGCCTGTTCTCTACTATAGAGGTGTTACCATAAATGTTTTACTGTTTTCAAACATAGGATTAGGAGTATTG 1440
GCTATTTAAGACAAAAAAGCTAATATAAATCATCATAACTTTCCATTGTATCTCTGTCTCTCTCTTTCTCGAGGTTTCATTCATTCTTATTTACCCAGCCCCCCCAAAAAATTGTGTT 1560
Xho I
ATTTTGATCAGCATTAATTTGCAAGATGCAAAGCTTTCCAAGGAATGTATGCTTTAAAATAAAGGTTTCACATTTTACATGGGGACACAAGACTTGATACATGCTAATTATTTTCTAATT 1680
Hind III
GCTCTGAATGTCTGTTTTTAAACTTTTCTAAAACTTTGAGAGGGAGAGGAAAAGAGGGAGAGAGGAGAGTCTCATCTGTTAGTTCACTCCCCATATGTACTCAACAGCTGTGGTTGGGCC 1800
TGGGCTGGAGACAGGAACAGGGAACTCAATCTAGGTTTCCCACATGAGTGGTAAGACTCCAGTTCCTGTAGTCATCATGCTGCCTTGTTGGGAAAATGGAATTGGGAACCAGAGCACAGA 1920
ACCAACCCCATGCAGTATGATGTGGGATGCAGGCATCTTAACCCTAAGCTAAATGTCCACTTCTTGAATGTGTCTTGAATGTCAGCATAGCATAACATAAAAGTATGTTATGGCTATGAT 2040
47 R P H D P E T V D V M L E L L D R D S D
AAAATTTTCTTACCTGTAAATAGTACATTGTAAAACAATTTTTTTCAACTTTTTTCATAGAGACCACATGACCCTGAGACGGTAGACGTGATGCTAGAACTTCTGGATCGCGACTCTGAT 2160
67 G L V G F D E F C L L I F K L A Q A A Y Y A L G Q A S G L D E E K R S H G E G K
GGTCTGGTCGGGTTCGATGAATTCTGCCTACTTATTTTCAAGTTGGCTCAAGCGGCTTATTATGCACTTGGCCAGGCCTCAGGGTTAGACGAGGAAAAGCGGTCCCACGGTGAGGGCAAG 2280
EcoR I
107 G R L L Q N R R Q E D Q R R F E L R D R Q F E D E P E R R R W Q K Q E Q E R E L
GGGCGCCTGTTACAAAATCGCAGGCAAGAAGATCAAAGGAGATTCGAACTCCGGGACAGACAATTCGAAGATGAACCAGAGCGACGACGCTGGCAAAAGCAGGAACAGGAGAGGGAGCTC 2400
147 A E E E E Q R K K R E R F E Q H Y S R Q Y R D K E Q R L Q R Q E L E E R R A E E
GCTGAGGAAGAGGAACAGAGGAAGAAACGAGAGCGATTCGAGCAGCACTACAGTAGGCAGTACCGCGACAAGGAGCAGCGGCTCCAAAGGCAAGAACTAGAAGAGCGCCGCGCAGAGGAA 2520
187 E Q L R R R K G R D A E E F I E E E Q L R R R E Q Q E L K R E L R E E E Q Q R R
GAGCAGCTGCGAAGGCGCAAGGGTCGCGACGCAGAGGAGTTCATTGAGGAAGAGCAGCTGCGAAGGCGAGAGCAGCAAGAGCTGAAGAGGGAGCTCCGCGAGGAGGAGCAGCAAAGGCGA 2640
227 E R R E Q H E R A L Q E E E E Q L L R Q R R W R E E P R E Q Q Q L R R E L E E I
GAGCGCCGAGAGCAGCACGAGCGGGCGCTCCAGGAGGAAGAAGAGCAGCTGCTGAGACAGAGGCGCTGGCGCGAGGAACCACGCGAGCAGCAGCAGCTGAGGCGCGAGCTGGAAGAGATA 2760
267 R E R E Q R L E Q E E R R E Q Q L R R E Q R L E Q E E R R E Q Q L R R E L E E I
CGCGAACGCGAGCAGAGACTTGAGCAGGAGGAGAGGCGCGAGCAGCAGCTGAGGCGCGAGCAGAGACTTGAGCAGGAGGAGAGGCGCGAGCAGCAGCTGAGGCGCGAGCTGGAAGAGATA 2880
307 R E R E Q R L E Q E E R R E Q R L E Q E E R R E Q Q L K R E L E E I R E R E Q R
CGCGAACGCGAGCAGAGGCTTGAGCAGGAGGAGAGGCGCGAGCAGCGACTTGAGCAGGAGGAGAGGCGCGAGCAGCAGCTGAAGCGCGAGCTGGAAGAGATACGCGAACGCGAGCAGAGG 3000
347 L E Q E E R R E Q L L A E E V R E Q A R E R G E S L T R R W Q R Q L E S E A G A
CTTGAGCAGGAGGAGAGGCGCGAGCAGCTGCTGGCTGAGGAGGTGCGGGAACAGGCCCGGGAGCGAGGCGAGAGCCTCACCCGGAGGTGGCAGCGGCAGCTAGAGAGTGAGGCCGGCGCA 3120
387 R Q S K V Y S R P R R Q E E Q S L R Q D Q E R R Q R Q E R E R E L E E Q A R R Q
CGTCAGAGCAAAGTCTACTCCAGGCCCCGCAGGCAGGAAGAGCAGAGTCTCCGCCAGGACCAGGAGAGGAGGCAGCGCCAGGAACGTGAACGGGAGCTGGAAGAGCAGGCCCGCCGGCAG 3240
427 Q Q W Q A E E E S E R R R Q R L S A R P S L R E R Q L R A E E R Q E Q E Q R F R
CAGCAGTGGCAAGCAGAAGAAGAGAGCGAGAGGCGCCGCCAGAGGCTGTCGGCCAGGCCCTCGTTGCGCGAGAGGCAGCTGAGGGCGGAGGAGCGCCAGGAGCAGGAACAACGGTTCCGG 3360
467 E E E E Q R R E R R Q E L Q F L E E E E Q L Q R R E R A Q Q L Q E E D S F Q E D
GAGGAGGAGGAGCAGCGGCGCGAGAGGAGGCAGGAGCTGCAGTTCCTCGAGGAGGAGGAGCAGCTCCAGCGACGGGAGCGTGCCCAACAGCTCCAGGAAGAGGACAGCTTTCAGGAGGAT 3480
Pst I
Xho I
507 R E R R R R Q Q E Q R P G Q T W R W Q L Q E E A Q R R R H T L Y A K P G Q Q E Q
CGGGAGAGGAGGCGACGCCAACAGGAGCAGCGCCCCGGCCAAACATGGAGGTGGCAACTCCAGGAAGAAGCCCAAAGACGCCGTCACACGCTGTACGCCAAACCGGGGCAGCAGGAGCAG 3600
547 L R E E E E L Q R E K R R Q E R E R E Y R E E E K L Q R E E D E K R R R Q E R E
CTGAGGGAGGAAGAGGAGCTGCAAAGGGAGAAGAGGCGCCAGGAGCGGGAGAGAGAATATCGGGAGGAAGAGAAACTGCAGAGAGAGGAAGACGAGAAGAGACGACGCCAGGAACGCGAG 3720
Pst I
587 R Q Y R E L E E L R Q E E Q L R D R K L R E E E Q L L Q E R E E E R L R R Q E R
AGACAATACCGGGAGTTAGAGGAGCTGCGGCAGGAAGAGCAGCTGCGTGACAGGAAACTCCGCGAGGAGGAGCAGCTCCTTCAAGAAAGGGAGGAAGAGAGGCTGCGCCGCCAGGAACGC 3840
627 E R K L R E E E E Q L L R Q E E Q E L R Q E R E R K L R E E E Q L L R R E E Q E L
GAGAGGAAACTCCGCGAGGAGGAGCAGCTGCTGCGCCAGGAGGAGCAGGAGCTCCGCCAGGAACGCGAGAGGAAACTCCGCGAGGAGGAGCAGCTGCTGCGTCGGGAGGAGCAGGAGCTC 3960
667 R Q E R E R K L R E E E Q L L Q E R E E E R L R R Q E R A R K L R E E E Q L L R
CGCCAGGAACGCGAGAGGAAACTCCGCGAAGAGGAGCAGCTGCTTCAGGAGAGGGAGGAAGAGAGGCTGCGCCGTCAGGAACGCGCCAGGAAACTCCGTGAGGAGGAGCAGCTGCTGCGC 4080
707 Q E E Q E L R Q E R E R K L R E E E Q L L R R E E Q L L R Q E R D R K L R E E E
CAGGAGGAGCAGGAGCTCCGCCAGGAACGGGAGAGGAAACTCCGCGAGGAGGAGCAGCTGCTGAGACGCGAGGAGCAGCTACTGCGCCAGGAACGCGACAGGAAACTCCGCGAAGAGGAG 4200
747 Q L L Q E S E E E R L R R Q E R E Q Q L R R E R D R K F R E E E Q L L Q E R E E
CAGCTCCTTCAGGAAAGCGAGGAAGAGAGGCTGCGTCGCCAGGAACGCGAGCAACAACTTCGTCGGGAGCGCGACAGAAAGTTCCGCGAGGAGGAGCAGCTCCTTCAGGAGAGGGAGGAA 4320
787 E R L R R Q E R E R K L R E E E Q L L Q E R E E E R L R R Q E R E R K L R E E E
GAGCGGCTGCGCCGTCAGGAACGCGAGAGGAAACTCCGTGAGGAGGAGCAGCTCCTTCAGGAGAGGGAGGAAGAGCGGCTGCGCCGTCAGGAACGCGAGAGGAAACTCCGTGAGGAGGAG 4440
827 Q L L Q E R E E E R L R R Q E R E R K L R E E E Q L L R Q E E Q E L R Q E R A R
CAGCTCCTTCAGGAGAGGGAGGAAGAGCGGCTGCGCCGTCAGGAACGCGAGAGGAAACTCCGTGAGGAGGAGCAGCTGCTGCGCCAGGAGGAGCAGGAGCTCCGCCAGGAACGCGCCAGG 4560

Figure 5.8 (Cont.)

```
867  K  L  R  E  E  E  Q  L  L  R  Q  E  E  Q  E  L  R  Q  E  R  D  R  K  L  R  E  E  E  Q  L  L  R  Q  E  E  Q  E  L  R  Q
     AAACTCCGTGAGGAGGAGCAGCTGCTGCGCCAGGAGGAGCAGGAGCTCCGCCAGGAACGCGACAGGAAACTCCGCGAGGAGGAGCAGCTGCTGCGCCAGGAGGAGCAGGAGCTCCGCCAG 4680

907  E  R  D  R  K  L  R  E  E  E  Q  L  L  Q  E  S  E  E  E  R  L  R  R  Q  E  R  E  R  K  L  R  E  E  E  Q  L  L  R  R  E
     GAACGCGACAGGAAACTCCGCGAAGAGGAGCAGCTCCTTCAGGAAAGCGAGGAAGAGAGGCTGCGCCGTCAGGAACGCGAGAGGAAACTCCGTGAAGAGGAGCAGCTGCTGCGTCGGGAG 4800

947  E  Q  E  L  R  R  E  R  A  R  K  L  R  E  E  E  Q  L  L  Q  E  R  E  E  E  R  L  R  R  Q  E  R  A  R  K  L  R  E  E  E
     GAGCAGGAGCTCCGTCGGGAACGCGCCAGGAAACTCCGTGAGGAGGAGCAGCTGCTTCAGGAGAGGGAGGAAGAGAGGCTGCGCCGTCAGGAGCGCGCCAGGAAACTCCGTGAGGAGGAG 4920

987  Q  L  L  R  R  E  E  Q  E  L  R  Q  E  R  D  R  K  F  R  E  E  E  Q  L  L  Q  E  R  E  E  E  R  L  R  R  Q  E  R  D  R
     CAGCTGCTGCGCCGGGAGGAGCAGGAGCTTCGTCAGGAGCGCGACAGAAAGTTCCGCGAAGAGGAGCAGCTGCTTCAGGAGAGGGAGGAAGAGAGGCTGCGCCGTCAGGAACGCGACAGA 5040

1027 K  F  R  E  E  E  R  Q  L  R  R  Q  E  L  E  E  Q  F  R  Q  E  R  D  R  K  F  R  L  E  E  Q  I  R  Q  E  K  E  E  K  Q
     AAGTTCCGCGAGGAAGAAAGACAGCTTCGTCGCCAGGAACTCGAGGAACAGTTTCGTCAGGAGCGCGATAGAAAATTCCGCTTAGAGGAACAGATCCGCCAGGAGAAGGAGGAAAAGCAG 5160
                                           Xho I
1067 L  R  R  Q  E  R  D  R  K  F  R  E  E  E  Q  Q  R  R  R  Q  E  R  E  Q  Q  L  R  R  E  R  D  R  K  F  R  E  E  E  Q  L
     CTGCGCCGTCAGGAGCGCGACAGGAAATTCCGCGAGGAGGAGCAGCAGCGACGCCGCCAGGAACGCGAGCAACAACTTCGTCGGGAGCGCGACAGAAAGTTCCGCGAGGAGGAGCAGCTC 5280

1107 L  Q  E  R  E  E  E  R  L  R  R  Q  E  R  A  R  K  L  R  E  E  E  Q  L  L  R  R  E  E  Q  L  L  R  Q  E  R  D  R  K  F
     CTTCAGGAGAGGGAGGAAGAGCGGCTGCGCCGTCAGGAGCGCGCCAGGAAACTCCGCGAGGAGGAGCAGCTGCTGAGACGCGAGGAGCAGCTACTGCGCCAGGAACGCGACAGAAAGTTC 5400

1147 R  E  E  E  Q  L  L  Q  E  S  E  E  E  R  L  R  R  Q  E  R  E  R  K  L  R  E  E  E  Q  L  L  Q  E  R  E  E  E  R  L  R
     CGCGAGGAGGAGCAGCTGCTTCAGGAAAGCGAGGAAGAGAGGCTGCGCCGCCAGGAACGCGAGAGGAAACTCCGCGAGGAGGAGCAGCTGCTTCAGGAAAGGGAGGAAGAGCGGCTGCGC 5520

1187 R  Q  E  R  A  R  K  L  R  E  E  E  Q  L  L  R  Q  E  E  Q  E  L  R  Q  E  R  A  R  K  L  R  E  E  E  Q  L  L  R  Q  E
     CGTCAGGAGCGCGCCAGGAAACTCCGCGAGGAGGAGCAGCTGCTGCGCCAGGAGGAGCAGGAGCTCCGCCAGGAACGCGCCAGGAAACTCCGCGAGGAGGAGCAGCTGCTGCGCCAGGAG 5640

1227 E  Q  E  L  R  Q  E  R  D  R  K  F  R  E  E  E  Q  L  L  R  R  E  E  Q  E  L  R  R  E  R  D  R  K  F  R  E  E  E  Q  L
     GAGCAGGAGCTCCGCCAGGAACGCGACAGAAAGTTCCGTGAGGAGGAGCAGCTGCTGCGTCGGGAGGAGCAGGAGCTCCGTCGGGAGCGCGACAGAAAGTTCCGCGAGGAGGAGCAGCTG 5760

1267 L  Q  E  R  E  E  E  R  L  R  R  Q  E  R  A  R  K  L  R  E  E  E  E  Q  L  L  F  E  E  Q  E  E  Q  R  L  R  Q  E  R  D
     CTTCAGGAAAGGGAGGAAGAGCGGCTGCGCCGTCAGGAGCGCGCCAGGAAACTCCGCGAGGAAGAAGAGCAACTGCTGTTTGAGGAGCAGGAAGAGCAGAGGCTCCGGCAGGAGCGAGAC 5880

1307 R  R  Y  R  A  E  E  Q  F  A  R  E  E  K  S  R  R  L  E  R  E  L  R  Q  E  E  E  Q  R  R  R  R  E  R  E  R  K  F  R  E
     CGGCGGTACCGAGCAGAGGAGCAGTTTGCCAGAGAGGAGAAGAGCCGTCGTCTGGAAAGAGAGCTGCGTCAAGAAGAAGAGCAGAGACGCCGCCGGGAACGGGAGAGGAAATTCCGGGAA 6000
       Kpn I
1347 E  Q  L  R  R  Q  Q  E  E  E  Q  R  R  R  Q  L  R  E  R  Q  F  R  E  D  Q  S  R  R  Q  V  L  E  P  G  T  R  Q  F  A  R
     GAGCAGCTCCGCCGCCAGCAGGAGGAGGAGCAGAGGCGCCGCCAACTGCGGGAGAGACAATTCCGAGAAGATCAGAGCCGTCGGCAAGTTCTGGAGCCCGGTACACGTCAGTTTGCTAGG 6120

1387 V  P  V  R  S  S  P  L  Y  E  Y  I  Q  E  Q  R  S  Q  Y  R  P  *  1407
     GTCCCGGTGCGCTCCAGCCCTCTTTATGAATACATCCAAGAGCAGAGATCTCAGTACCGCCCCTAAGAGATGTTGCCGGTGTCCTGACACCTGCCAAAGCCTCAAGCAAAAGAAAGTGAG 6240

     AAACAGTGCGTACCAAACGATAACGCAAATGTTTCTGGTTGTGGGAAAATTCTCTGATGTTAGAATGTGTCTTTTCTTCCAAAATCTTTACACTCATTTCATGTACTTTGTACTTCTACC 6360

     TTTTATTCTTTGTCAAGTAGTTCTTTACTACAATATATCTTTGCTCTTTGGTGCAGATATAGTGAGCATTTTTTAAAACAACAACCCTTTAATTTGTTTAAGGAGTTTTGTTTGAGGAAC 6480

     ACGTTGATTTATTGCCTTAAAAGTGGCAAGAATAATAGGACAATTTGATATTGAGAAAAATTGGCTCTACATCATTAACAATTGAACCATCTTGTGGAGTGTTCAAATGTTTTTAATGTT 6600

     CAAGTTGATATTTCGTCTTAGGCCTAAATTTAATTCACCTTTACTTCCAAATAGCTTCAGATCTTTTGTGGCATAGTTAACACAGATTCTTCCAGGTGACTGAGTTTTTACAATCCTTGA 6720

     ATAGTACAAGGGGAAGTGACTTCTGCAGTAAAACTCCCTATGAAGTTAAGTCCTTAACTGGAAGAATACCACTTAAGAATAAGGTTTAGCTTGGAGACAAAATTCGGAATATGTTTTCCT 6840
                           Pst I
     ACACAACTCCACTTGAACACAAACACTTAATCTCTTGAAGCAAAAATAAGATGTTTCTCAAACAATTTCAGAGCTCATATATTAAAACTATCTCCTGTAGTTTCTAGTGCTGTGTTCTGT 6960
                                                                            Sac I
     TTCCCCCAAGTTTTGATGGGTTTGGGAATATACCTTTGTGAACTTTGTCTTAGAGGGAACCTGACCAGGATCCTATCTTGCTGTAGGACCCATGAGATCACAGAACGTGTTGGGGGATGG 7080
                                                                 BamH I
     AAACGGGGAAAATGTGTGAACAGAGCGGCAAGTTTCTAGATGCTCTTTAACAAGATTCATCAGCCCACAATCCTTTGAGGCCTATTGATACTACCTTGGAGACAGTCTTGTTGAAATAAT 7200
                                  Xba I
     TGCACTGTGAGACATCAATAATAAATGTTTGGCAAGTAATTATTTGTGTCTGTTTTAATTACATCTTGGTCATAACTCAAGATCTAGTTGTCTGAATATTGGCATAAGTTCCCAATCACT 7320

     TAAAAATACTTTGTTACTCTTATCTATGACCTAATAAGAAAAGGGGAAGACCCTCTTTTTCCCCACCTTTTTTCCCAGTTAAATTATTTCCATCCTCACACTGATTTCTCTTTCCAAGTG 7440

     TTTTTGTTCATGACCCCAGAGTTTCCTATGCCTTTTCTTGCATTGACTCCCTAGTCATTACAAATACTCAGTGCTCTCAGTGTTTAATATTTCCTTCTGAGAATTC 7546
                                                                                                          EcoR I
```

5.2.4 Analysis of the Trichohyalin Gene Sequence

a. Definition of the Gene Structure

Comparison of the genomic sequence with that of λsTr1 allowed the stop codon of the trichohyalin gene to be located at position 6184 (Fig. 5.8). Analysis of the upstream sequence indicated that the stop codon terminates a 4083 bp open reading frame (ORF) which extends 84 bases into the 5.2 kb EcoR I fragment. The ORF was then examined for in-frame methionine codons or 3' (donor) intron splice sites, either of which could form the 5' end of the actual ORF. The first 105 bases of the ORF were found to contain the only methionine (position 2131) and the only two possible intron donor sites which are at positions 2100 (immediately preceding the ORF) and 2202. It therefore still remains possible that the ORF contains either the complete trichohyalin coding region, beginning at the methionine codon, or is only a single large exon with the remainder of the coding region present on an upstream exon. These possibilities are summarised in Figure 5.9.

The first method used to resolve this problem was Northern analysis. As both the first intron splice site and the methionine codon are positioned at least 50 residues into the 5.2 kb EcoR I fragment, and thus also into the 0.6 kb Hind III/EcoR I fragment, Northern hybridisation with the 0.6 kb Hind III/EcoR I fragment should determine whether either of these sites forms the 5' end of the reading frame. As shown earlier (Fig. 5.5), the 0.6 kb Hind III/EcoR I fragment does not hybridise to the trichohyalin mRNA suggesting that this fragment is contained within an intron whose donor site is at position 2202.

However, this interpretation of the Northern analysis was subsequently questioned by a comparison of the genomic 3' non-coding sequence with that of the cDNA clone. This comparison indicated that the two sequences have considerable differences, suggesting that the lack of hybridisation of the 0.6 kb Hind III/EcoR I fragment to the trichohyalin mRNA could be due to marked differences in the corresponding genomic and cDNA sequences. This was corroborated by additional Northern analysis where the 1.2 kb EcoR I/Pst I genomic fragment, which contains at most 20 bp of non-coding sequence, gave a much weaker hybridisation signal to follicle RNA than the 0.47 kb EcoR I fragment from the 3' non-coding region of the cDNA clone when the two were hybridised under identical conditions (Fig. 5.10). Thus the

Figure 5.9 Possible arrangements for the initiation of the coding sequence at the 5' end of the large ORF present in the trichohyalin gene.

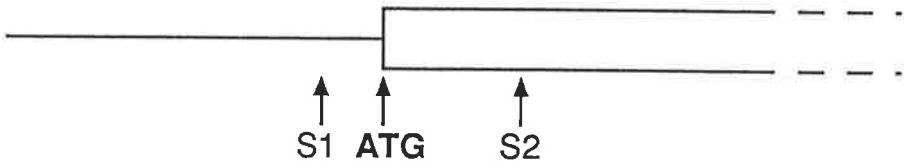
Analysis of the obtained nucleotide sequence has shown that it contains a 4083 bp ORF (nucleotides 2101 to 6183). The 105 bp at the 5' end of the ORF contain the only possible 3' intron splice sites (positions 2100 and 2202) and initiating methionine codons (position 2131). Represented are the three possible arrangements for the initiation of the coding sequence at the 5' end of the large ORF.

(a) The coding region for the trichohyalin gene begins at the methionine codon at position 2131 (ATG), i.e., trichohyalin is encoded by a single exon.

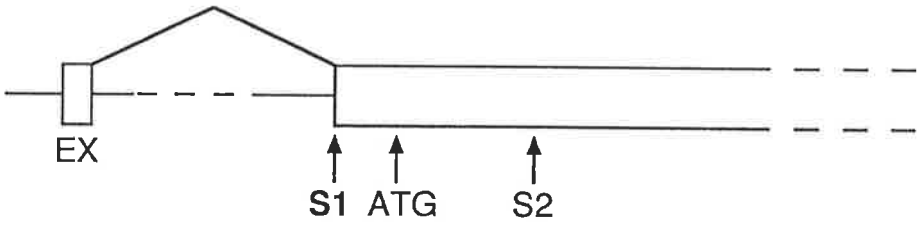
(b) The coding sequence begins at position 2101 (S1) and is immediately preceded by an intron present within the coding region. The remainder of the trichohyalin coding sequence is located on one or more upstream exons (EX).

(c) The coding sequence begins at position 2203 (S2) and is immediately preceded by an intron present within the coding region. The remainder of the trichohyalin coding sequence is located on one or more upstream exons (EX).

a.



b.



c.

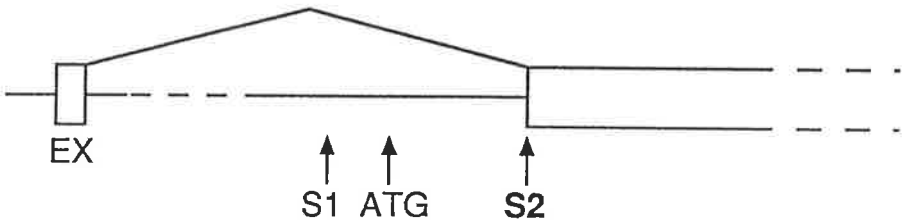
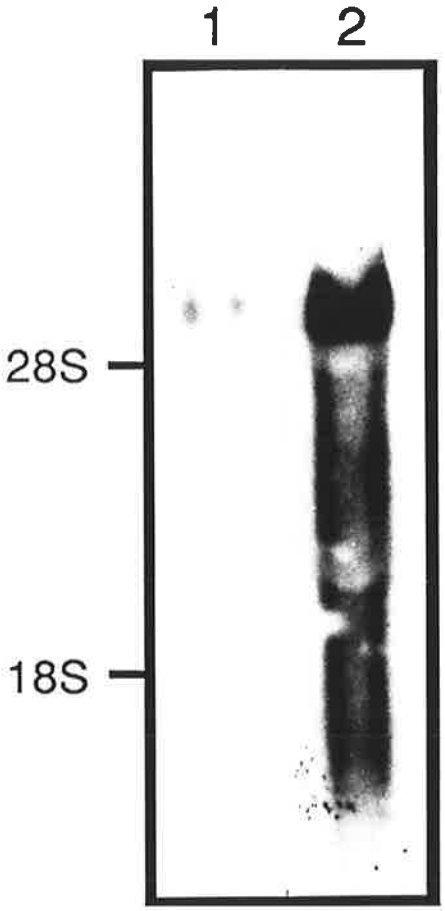


Figure 5.10 Examination of the homology of genomic and cDNA sequences with wool follicle RNA.

Two Nytran filters, each containing 5 μg of total wool follicle RNA which had been separated on a 1% agarose/formaldehyde gel, were probed separately with a coding fragment from the genomic clone (EP1.2, track 1) and the 0.47 kb EcoR I fragment from the 3' non-coding region of the cDNA clone (track 2). The two filters were hybridised and washed under identical conditions. The autoradiograph of both tracks was exposed for 4 d.


1
2
28S
18S

nucleotide sequences within the 1.2 kb EcoR I/Pst I fragment which is present at the 5' end of the ORF appear to be considerably different from those in the Merino mRNA which suggests that the negative hybridisation result obtained with the 0.6 kb Hind III/EcoR I fragment could be due to considerable sequence differences together with the presence of only a small target sequence.

Not only do the differences between the follicle RNA and genomic sequences prevent deduction of the occurrence of splice sites by Northern analysis but they also stop analysis by methods such as RNA protection and S1 nuclease protection where the results are dependent upon the complete hybridisation of the genomic and cognate mRNA sequences in order to prevent degradation of single stranded nucleic acid. Thus the 5' end of the reading frame could not be determined experimentally.

Fortuitously, when the first 45 amino acid sequence encoded by the 4083 bp ORF was used to search the Genbank database it was found to have a high degree of homology with the calcium-binding domains of a number of different calcium-binding proteins. Analysis of the trichohyalin sequence showed that it contained a complete calcium-binding domain. When the trichohyalin gene sequence was compared with those for a group of small calcium-binding proteins, which are characterised by the presence of two calcium-binding domains (EF hands, see Kretsinger, 1980) encoded by two separate exons (Lagasse and Clerc, 1988; Krisinger *et al.*, 1988), it was found that the splice site at position 2100 corresponds exactly with the location of the conserved splice site within the genes for the small calcium-binding proteins. Therefore the upstream sequence was searched for a 5' intron splice site (acceptor site) and one with good homology to the consensus sequence (Mount, 1982) was found at position 1100. An in-frame methionine codon is present upstream of the acceptor site (position 963) and is preceded by an in-frame termination codon (position 948). No likely 3' intron splice sites are present within this region suggesting that the initiating methionine codon for trichohyalin gene is at position 963. Note that the sequence encoded by the first coding exon is homologous to the first EF hand present within the small calcium binding proteins (this will be detailed later).

The genes encoding the small calcium binding proteins not only contain an intron positioned between the coding regions of the two EF hands but also contain an intron within the 5' non-coding region. Corresponding with this, a sequence homologous to an intron donor site is present 22 bases upstream of the ATG codon (position 940). No

sequence with strong homology to the 5' intron splice site consensus sequence is present within the available upstream sequence suggesting that the first exon of the trichohyalin gene is located further upstream.

The trichohyalin gene would therefore appear to consist of three exons, the first containing a portion of the 5' non-coding region, the second (160 bp) containing the remainder of the 5' non-coding region and part of the coding region, and the third (5137 bp) containing the remainder of the coding region and the complete 3' non-coding region (Fig. 5.11). An intron of exactly 1 kb splits the second and third exons. The complete nucleotide and deduced amino acid sequence are shown in Figure 5.8. The deduced sequence is 1407 amino acids long and has a predicted molecular weight of 183,780.

b. Amino Acid Composition

The total amino acid composition of the deduced trichohyalin protein sequence is shown in Table 5.1 and compared with that of the sheep cDNA sequence (see Section 4.2.4b) and wool follicle trichohyalin (see Section 3.2.3a). The overall amino acid composition is very similar to that seen within the partial cDNA sequence, particularly with respect to the high levels of glutamic acid, arginine, glutamine and leucine which together total 81.4% of the protein's residues. These are markedly higher than seen in the amino acid composition for wool follicle trichohyalin, especially in regard to glutamic acid/glutamine and arginine.

The deduced trichohyalin protein sequence is, like the cDNA deduced sequence, extremely hydrophilic with 61% of the residues charged and a further 18% being polar. Computer analysis of the overall charge ratio, using the computer program ISOELECTRIC, has indicated that the pI of trichohyalin is approximately 5.44.

On the basis of amino acid composition, the 95 amino acid sequence at the N-terminus of trichohyalin is quite distinct from the remainder of the protein (Table 5.2). This N-terminal region, due to the amino acid requirements for the contained calcium-binding domains, has a higher content of hydrophobic amino acids. In addition, glutamine, glutamic acid, arginine and leucine total only 30% of the residues and less than half of the amino acids within the N-terminal region are charged or polar (Table 5.2).

The content of sulphur-containing amino acids within trichohyalin is extremely low. The deduced protein sequence contains only two methionine and two cysteine residues. All four of these residues occur within the N-terminal "hydrophobic" portion.

Figure 5.11 Structure of the trichohyalin gene.

The predicted structure of the trichohyalin gene is shown. It consists of three exons, the first of which is believed to be located upstream of the available sequence (EXON 1?). The predicted start (ATG) and stop (TAA) codons are shown as is the position of the putative polyadenylation signal. Also shown are a number of the restriction sites present within the available sequence. Boxed regions indicate the exons and the hatched area indicates the predicted coding region of the trichohyalin gene. E, EcoR I; H, Hind III; K, Kpn I; P, Pst I; X, Xho I.

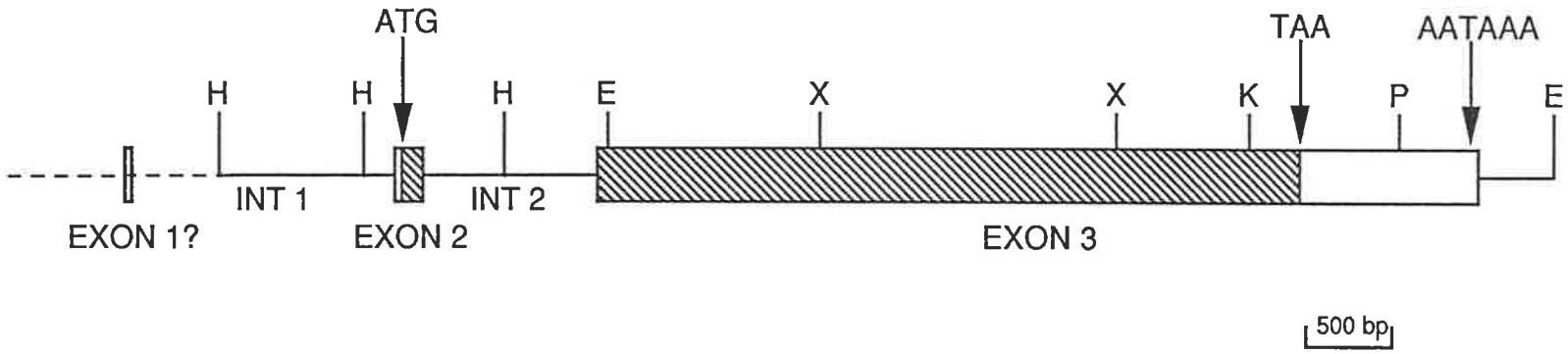

ATG
TAA
AATAAA
H
H
H
E
X
X
K
P
E
INT 1
INT 2
EXON 1?
EXON 2
EXON 3
500 bp

Table 5.1 Comparison of the amino acid content of the predicted trichohyalin gene sequence with those of the deduced cDNA sequence (see Table 4.1) and wool follicle trichohyalin (see Table 3.2).

Amino Acid	Wool Follicle Trichohyalin	Deduced cDNA Sequence	Deduced Gene Sequence
	mole percent	*mole percent*	*mole percent*
Asp/Asn	6.4	3.3/0.2	2.6/0.1
Thr	2.9	0.0	0.4
Ser	5.4	1.7	1.7
Glu/Gln	28.0	26.0/18.1	28.6/14.9
Pro	3.3	1.1	0.9
Gly	5.3	1.1	1.1
Ala	4.7	1.3	2.4
Cys	0.6	0.0	0.1
Val	4.0	1.3	0.7
Met	0.1	0.0	0.1
Ile	2.5	0.2	0.9
Leu	10.0	10.9	12.7
Tyr	2.1	0.9	0.9
Phe	2.4	3.7	1.9
Lys	6.7	5.0	3.8
His	1.7	1.1	0.4
Trp	0.2	0.2	0.4
Arg	13.7	23.8	25.2

Table 5.2 Comparison of the amino acid content of the first 95 amino acids of the predicted trichohyalin amino acid sequence with that of the remainder of the protein.

Amino Acid	Trichohyalin (residues 1-95)	Trichohyalin (residues 96-1407)
	mole percent	*mole percent*
Asp	11.6	1.9
Asn	1.1	0.1
Thr	1.1	0.3
Ser	5.3	1.4
Glu	5.3	30.3
Gln	4.2	15.7
Pro	3.2	0.8
Gly	6.3	0.7
Ala	8.4	2.0
Cys	2.1	0.0
Val	5.3	0.4
Met	2.1	0.0
Ile	7.4	0.5
Leu	16.8	12.3
Tyr	3.2	0.8
Phe	5.3	1.7
Lys	6.3	3.7
His	2.1	0.3
Trp	0.0	0.5
Arg	3.2	26.8

c. Repetitive Protein Structure

The deduced amino acid sequence of trichohyalin was compared with itself and the results were displayed by a dot matrix plot (Fig. 5.12). The C-terminal repeat structure present within the cDNA clone appears to cover almost half of the trichohyalin protein. Analysis of the amino acid sequence has shown that this repeat region consists of a central highly conserved domain and two less-conserved flanking domains (see Fig. 5.13). The total C-terminal repetitive region covers almost 60% of the protein.

In addition to the large C-terminal repeat the dot matrix plot shows that there is also a small repetitive region (Fig. 5.12) which consists of three full or partial copies of a forty amino acid repeat (Fig. 5.13) and extends from residue 242 to 355.

Analysis of the large C-terminal repeat has shown that although the repeat size and sequence are very similar to that of the repeats in the sheep cDNA clone the two repetitive regions differ significantly. The highly conserved region of the complete protein, which consists of 30 full or partial length repeats, is based not on one but two consensus sequences, one of 24 amino acids (repeat A) and the other of 22 amino acids (repeat B) (Fig. 5.14b). Within the repetitive region there are 15 copies of each repeat in an apparently random arrangement. The two consensus sequences are nearly identical over 15 of the amino acids, viz., the first eleven and last four amino acids, but contain an intervening variable region (positions 12 to 20 (repeat A)/18 (repeat B)) in which the number and arrangement of amino acids within the two consensus sequences differ . In addition the two consensus sequences vary in the predominance of aspartic acid and alanine at position 1 and phenylalanine and leucine at position 4. Although the two repeat sequences differ they still contain the substrate amino acids for follicle transglutaminase and peptidylarginine deiminase within the highly conserved portions of the consensus sequences, namely lysine at position 3, glutamine at positions 9 and 22/20 and arginine at positions 2, 5, 21/19 and 24/22.

Note that the highly conserved repetitive region also appears to contain two identical insertions of eight amino acids which occur between repeats 7 and 8 and between repeats 21 and 22 (Fig. 5.13). These are nearly identical in sequence to the eight amino acid insertions seen in the cDNA deduced amino acid sequence (Fig. 5.15).

The C-terminal repetitive region also contains two less-conserved flanking domains (Fig. 5.13). The less-conserved N-domain contains seven repeats and the C-domain four repeats. Most of these repeats contain large deletions and the repeat

Figure 5.12 Dot plot of the trichohyalin amino acid sequence compared with itself.

The predicted amino acid sequence of trichohyalin was analysed for similar internal sequences with the UWGCG computer programme COMPARE (Devereux *et al.*, 1984) using a word length of 6. Output from this programme was then displayed by the programme DOTPLOT (Devereux *et al.*, 1984). Lines parallel to the diagonal indicate internal similarities, i.e., nucleotide repeats. The axes are labelled in residue numbers.

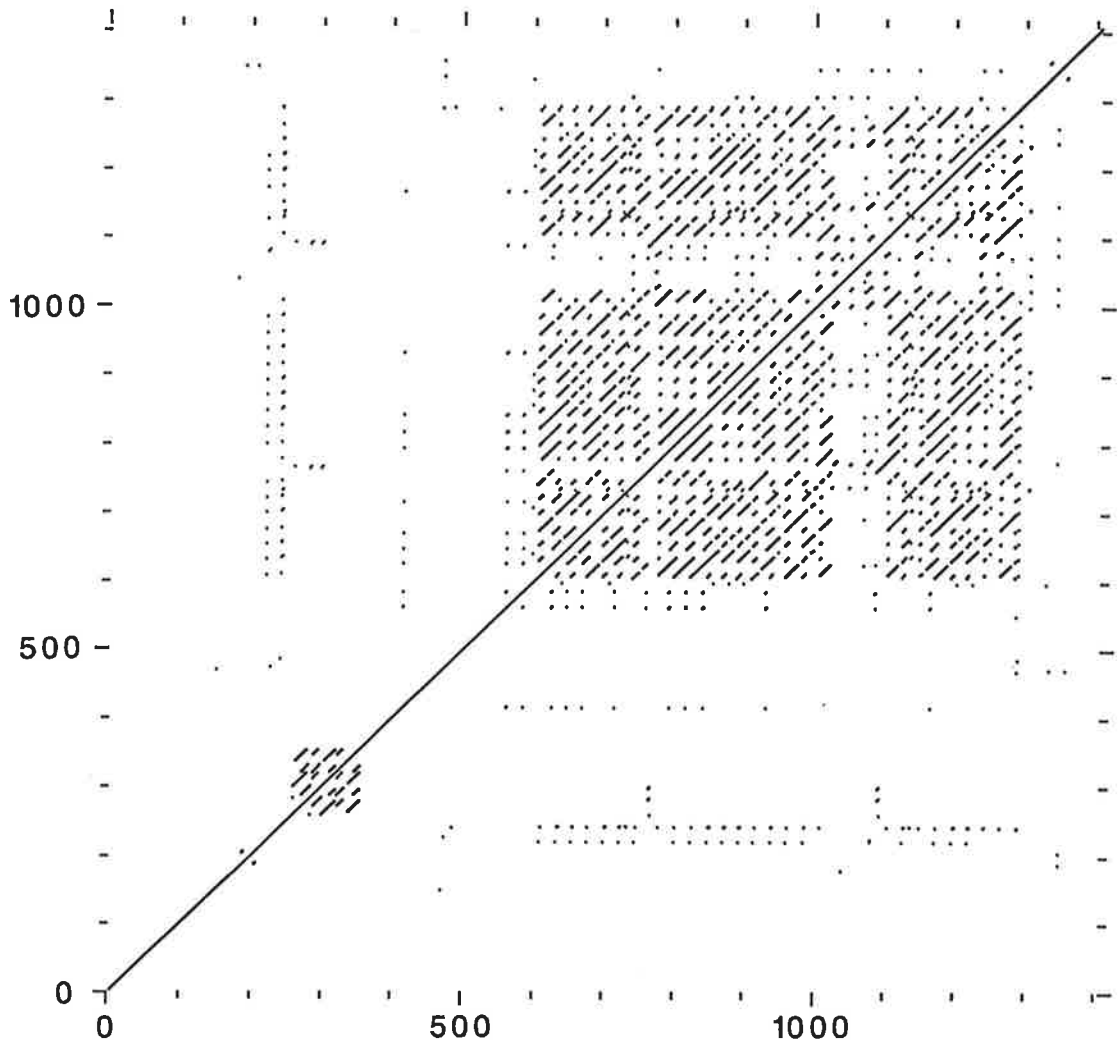

1000
500
0
0
500
1000

Figure 5.13 Amino acid sequence of trichohyalin.

The predicted amino acid sequence of trichohyalin is aligned with respect to the consensus sequences for the N-terminal and C-terminal repetitive regions (see Fig. 5.14). Both repetitive regions are marked. The C-terminal repetitive region has been subdivided into a central conserved domain and less-conserved N- and C-flanking domains. The repeats within the central domain are each denoted as being of either type A or type B. To optimise alignment of the sequence with the consensus sequences spaces have been introduced into the sequence (dashes) and single amino acids removed (arrowheads). Also shown are the helix (open box) and turn (hatched box) regions of the proposed EF hands.

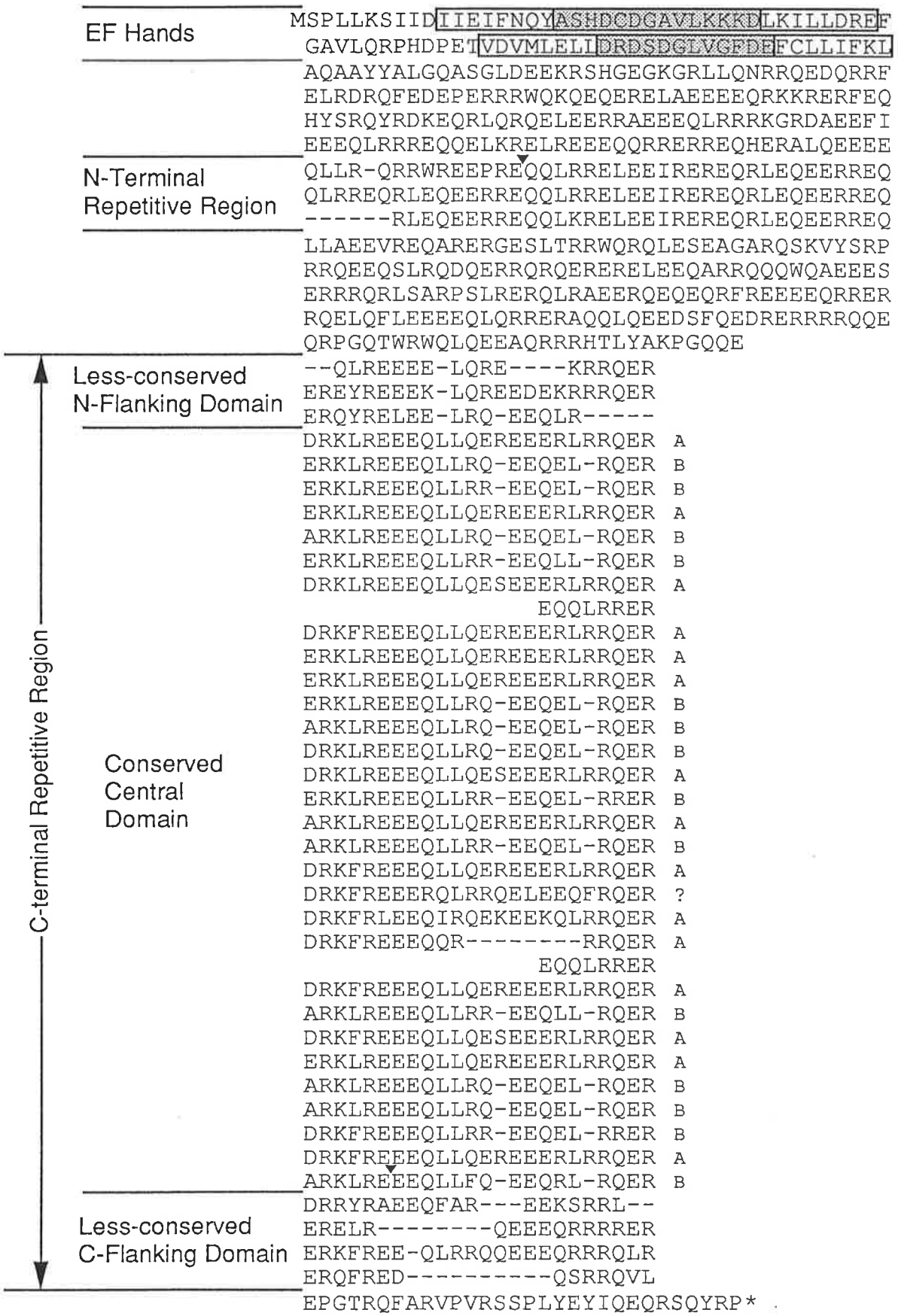

EF Hands
N-Terminal Repetitive Region
C-terminal Repetitive Region
Less-conserved N-Flanking Domain
Conserved Central Domain
Less-conserved C-Flanking Domain
MSPLLKSIIDIIEIFNQYASHDCDGAVLKKKDLKILLDREF
GAVLQRPHDPETVDVMLELIDRDSDGLVGFDEFCLLIFKL
AQAAYYALGQASGLDEEKRSHGEGKGRLLQNRRQEDQRRF
ELRDRQFEDEPERRRWQKQEQERELAEEEEQRKKRERFEQ
HYSRQYRDKEQRLQRQELEERRAEEEQLRRRKGRDAEEFI
EEEQLRRREQQELKRELREEEQQRRERREQHERALQEEEE
QLLR-QRRWREEPREQQLRRELEEIREREQRLEQEERREQ
QLRREQRLEQEERREQQLRRELEEIREREQRLEQEERREQ
------RLEQEERREQQLKRELEEIREREQRLEQEERREQ
LLAEEVREQARERGESLTRRWQRQLESEAGARQSKVYSRP
RRQEEQSLRQDQERRQRQERERELEEQARRQQQWQAEEES
ERRRQRLSARPSLRERQLRAEERQEQEQRFREEEEQRRER
RQELQFLEEEEQLQRRERAQQLQEEDSFQEDRERRRRQQE
QRPGQTWRWQLQEEAQRRRHTLYAKPGQQE
--QLREEEE-LQRE----KRRQER
EREYREEEK-LQREEDEKRRRQER
ERQYRELEE-LRQ-EEQLR-----
DRKLREEEQLLQEREEERLRRQER A
ERKLREEEQLLRQ-EEQEL-RQER B
ERKLREEEQLLRR-EEQEL-RQER B
ERKLREEEQLLQEREEERLRRQER A
ARKLREEEQLLRQ-EEQEL-RQER B
ERKLREEEQLLRR-EEQLL-RQER B
DRKLREEEQLLQESEEERLRRQER A
EQQLRRER
DRKFREEEQLLQEREEERLRRQER A
ERKLREEEQLLQEREEERLRRQER A
ERKLREEEQLLQEREEERLRRQER A
ERKLREEEQLLRQ-EEQEL-RQER B
ARKLREEEQLLRQ-EEQEL-RQER B
DRKLREEEQLLRQ-EEQEL-RQER B
DRKLREEEQLLQESEEERLRRQER A
ERKLREEEQLLRR-EEQEL-RRER B
ARKLREEEQLLQEREEERLRRQER A
ARKLREEEQLLRR-EEQEL-RQER B
DRKFREEEQLLQEREEERLRRQER A
DRKFREEERQLRRQELEEQFRQER ?
DRKFRLEEQIRQEKEEKQLRRQER A
DRKFREEEQQR--------RRQER A
EQQLRRER
DRKFREEEQLLQEREEERLRRQER A
ARKLREEEQLLRR-EEQLL-RQER B
DRKFREEEQLLQESEEERLRRQER A
ERKLREEEQLLQEREEERLRRQER A
ARKLREEEQLLRQ-EEQEL-RQER B
ARKLREEEQLLRQ-EEQEL-RQER B
DRKFREEEQLLRR-EEQEL-RRER B
DRKFREEEQLLQEREEERLRRQER A
ARKLREEEQLLFQ-EEQRL-RQER B
DRRYRAEEQFAR---EEKSRRL--
ERELR--------QEEEQRRRRER
ERKFREE-QLRRQQEEEQRRRQLR
ERQFRED----------QSRRQVL
EPGTRQFARVPVRSSPLYEYIQEQRSQYRP*

Figure 5.14 Consensus sequences of the N-terminal and C-terminal repetitive regions.

(a) The consensus sequence of the N-terminal repetitive region was derived from the three repeats marked in Figure 5.13.

(b) The repeat A and repeat B consensus sequences of the C-terminal repetitive region were derived respectively from the type A and type B repeats marked in Figure 5.13. If two residues occurred in more than 30% of the repeats then both are denoted, with the first of the two occurring at a greater frequency.

a.

```
          5         10        15        20        25        30        35        40
Q L R R E Q R L E Q E E R R E Q Q L R R E L E E I R E R E Q R L E Q E E R R E Q
```

b.

```
                  5         10        15        20        24
Repeat A  D R K L/F R E E E Q L L Q E R E E E R L R R Q E R

                  5         10        15        20  22
Repeat B  A/E R K L R E E E Q L L R Q/R E E Q E L R Q E R
```

Figure 5.15 Shown are the eight amino acid inserts present at residues 99-106 and 152-159 in the deduced cDNA sequence and residues 762-769 and 1088-1095 in the deduced genomic protein sequence.

cDNA	E	P	Q	L	R	Q	K	R
	E	P	Q	L	R	Q	E	R
Genomic	E	Q	Q	L	R	R	E	R
	E	Q	Q	L	R	R	E	R

sequences show only partial conservation with either of the two C-terminal consensus sequences.

The short N-terminal repetitive region, although containing high levels of glutamic acid, arginine, glutamine and leucine as is the case for the C-terminal repeat, does not appear to have any homology with the consensus sequences of the C-terminal repeat. Although there are only 3 copies of the repeat within this repetitive region the sequence is highly conserved with the consensus sequence shown in Figure 5.14a. It should be noted that there are no lysine residues within the consensus sequence. Additionally the sequence REQQLRRE, which occurs at positions 14 to 21 within the consensus sequence, would, if inserted between the last two residues of a C-terminal repeat, yield an eight amino acid insert identical to the two eight amino acid sequences which interrupt the conserved section of the C-terminal repeats (residues 762-769 and 1088-1095).

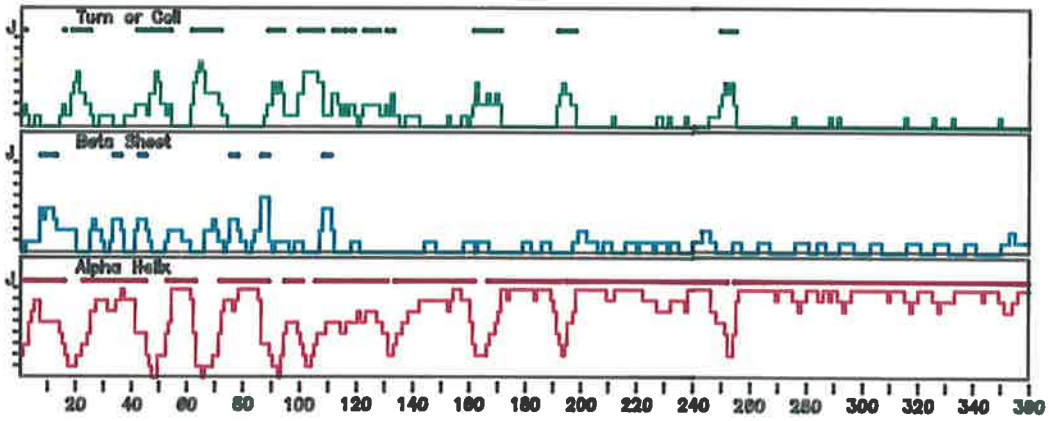
d. Secondary Structure Analysis

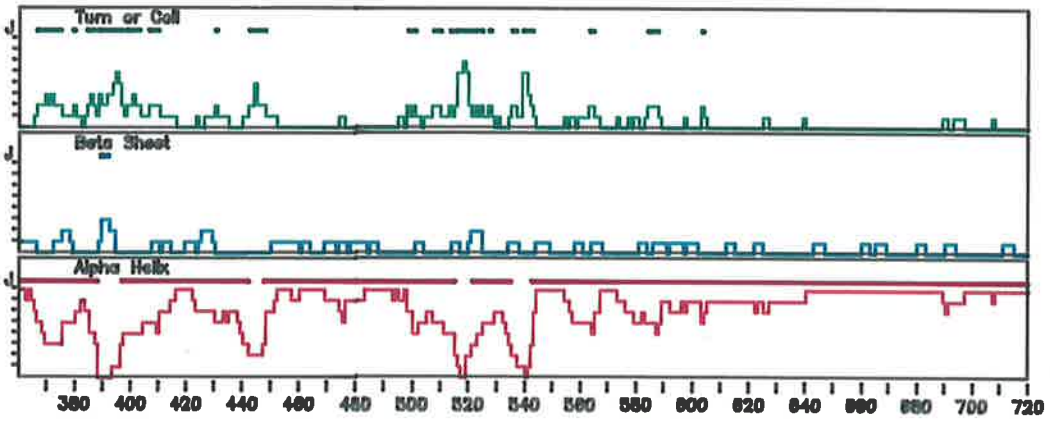
The secondary structure for the complete protein was analysed by the computer program PREDICT of Eliopoulos *et al.* (1982) (Fig. 5.16). As for the cDNA deduced sequence α-helix is the predominant secondary structure and is predicted to be able to form over almost the complete trichohyalin protein. The majority of the predicted turn or coil exists within the non-repetitive regions of trichohyalin e.g. residues 20-135, 365-410, 500-545, 1360-1407. Note the oscillating predictions of helix and turn within the first 90 amino acids of trichohyalin. This corresponds to the helix-turn-helix structures of the two EF hands (residues 11-40 and 54-81). The regular spacing of turn or coil seen within the repetitive portion of the cDNA deduced sequence (Fig. 4.6) does not occur throughout the C-terminal repeat of the total protein with this region containing long stretches with no predicted turn or coil (residues 605-740, 780-885, 910-1000, 1150-1230). Note that no turn or coil is predicted over the majority of the N-terminal repetitive region (residues 255-355) and almost no β-sheet is predicted throughout the trichohyalin molecule.

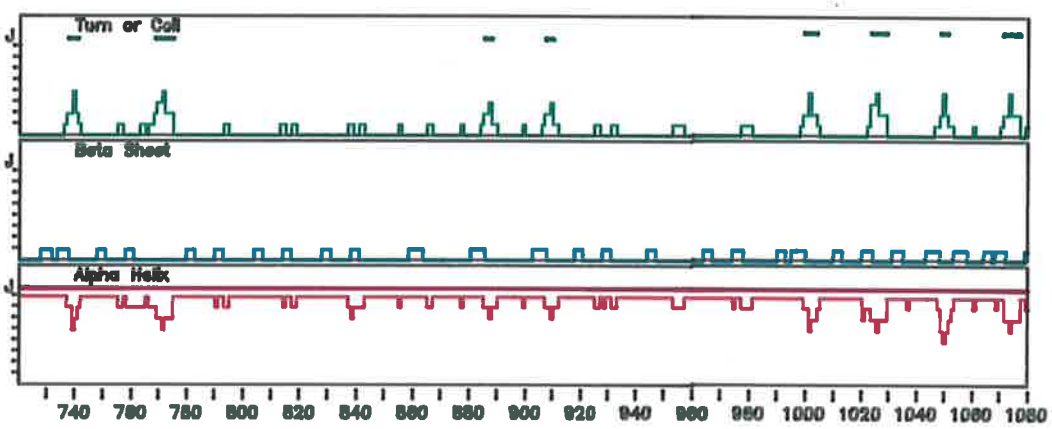
Examination of the predicted secondary structure for the individual consensus sequences has shown that the N-terminal repeat consensus sequence and the C-terminal B-repeat are predicted to form only α-helix (Fig. 5.17). Both of these consensus sequences also contain three hydrophobic residues spaced seven amino acids apart (leu(18), ile(25) and leu(32) in the N-terminal consensus sequence and leu(4, 11 and 18) for repeat B) as is seen in the heptad-like arrangement of the cDNA-deduced consensus

Figure 5.16 Secondary structure analysis of trichohyalin.

The secondary structure of the complete predicted amino acid sequence of trichohyalin was examined by the computer programme PREDICT (Eliopoulos *et al.*, 1982) and the number of predictions for turn or coil (green), ß-sheet (blue) and α-helix (red) are plotted. Predicted regions of the given secondary structure are indicated by the solid bars at the top of each section (J).


Turn or Coil
Beta Sheet
Alpha Helix
20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360


Turn or Coil
Beta Sheet
Alpha Helix
380 400 420 440 460 480 500 520 540 560 580 600 620 640 660 680 700 720


Turn or Coil
Beta Sheet
Alpha Helix
740 760 780 800 820 840 860 880 900 920 940 960 980 1000 1020 1040 1060 1080

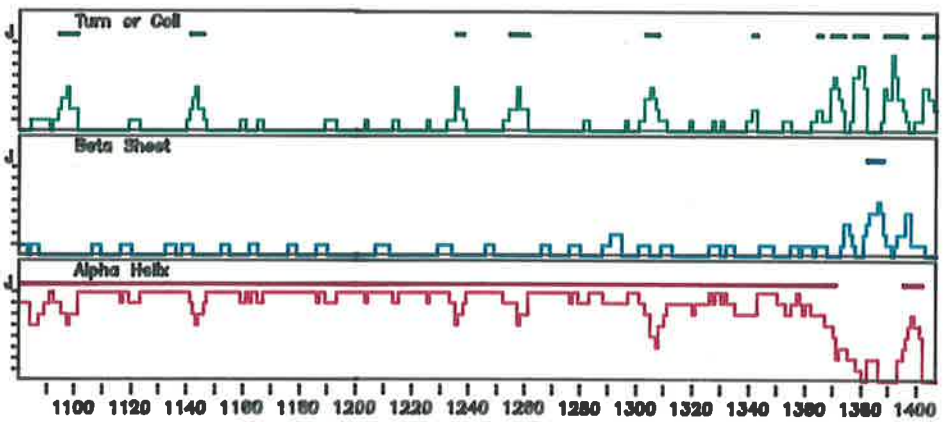
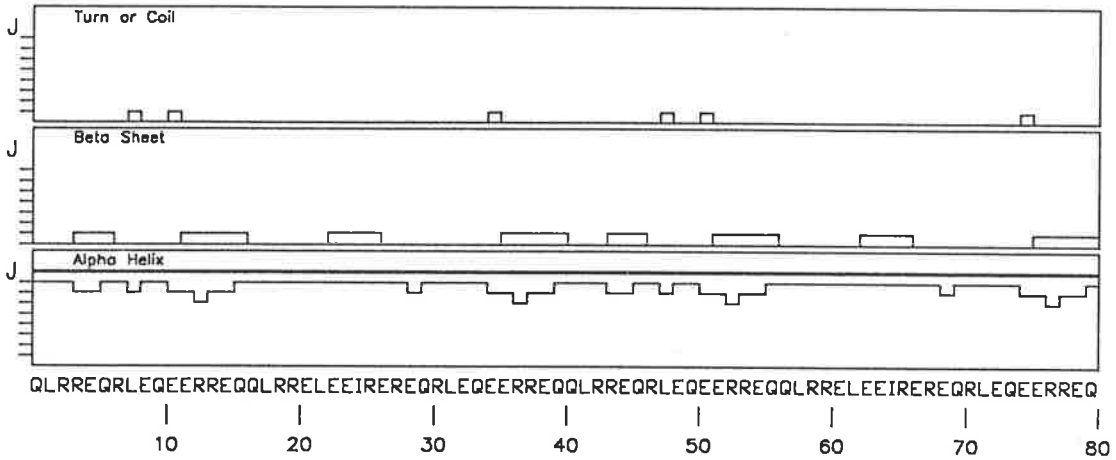

Turn or Coil
Beta Sheet
Alpha Helix
1100 1120 1140 1160 1180 1200 1220 1240 1260 1280 1300 1320 1340 1360 1380 1400

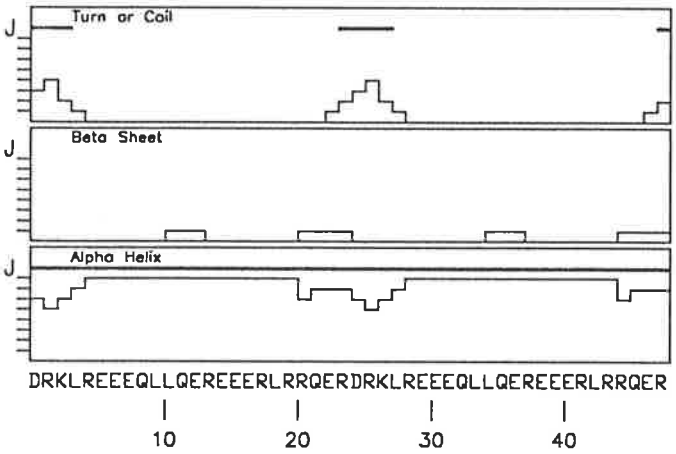
Figure 5.17 Secondary structure analysis of the consensus sequences of the N-terminal and C-terminal repetitive regions.

The secondary structures of two consecutive copies of the N-terminal consensus sequence (a), C-terminal repeat A consensus sequence (b) and C-terminal repeat B consensus sequence (c) were analysed by the computer programme PREDICT (Eliopoulos *et al.*, 1982) and plotted as described in Figure 5.16. Note that α-helix is the only predicted secondary structure in the N-terminal and the C-terminal repeat B consensus sequences.

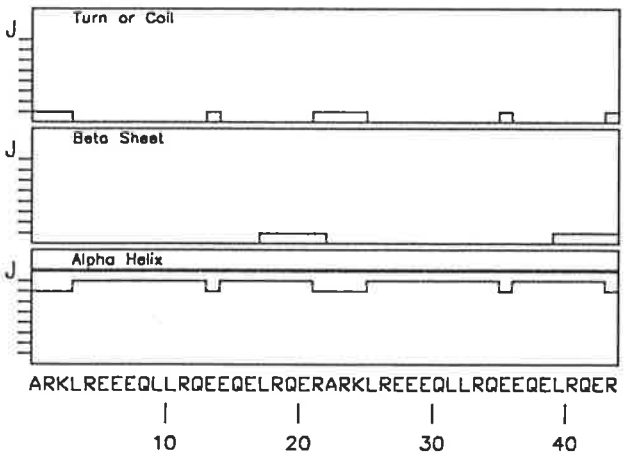
a.


Turn or Coil
Beta Sheet
Alpha Helix
QLRREQRLEQEERREQQLRRELEEIREREQRLEQEERREQQLRREQRLEQEERREQQLRRELEEIREREQRLEQEERREQ
10
20
30
40
50
60
70
80

b.


Turn or Coil
Beta Sheet
Alpha Helix
DRKLREEEQLLQEREEERLRRQERDRKLREEEQLLQEREEERLRRQER
10
20
30
40

c.


Turn or Coil
Beta Sheet
Alpha Helix
ARKLREEEQLLRQEEQELRQERARKLREEEQLLRQEEQELRQER
10
20
30
40

sequence. Also, the N-terminal repeat contains a leucine residue at position 22 and approximately half of the repeat B sequences contain an alanine at position 1. The positions of these residues would allow the formation of short α-helical coiled-coil heptad arrangements. Note that although heptad like structures are present in both consensus sequences, in neither case does any heptad arrangement spread between consecutive repeats.

Although α-helix is predicted over the complete C-terminal repeat A consensus sequence, turn or coil is also predicted over the region spanning residues 24, 1, 2 and 3 (Fig. 5.17). Repeat A also only contains two hydrophobic residues seven amino acids apart (leu/phe(4) and leu(11)) and does not contain any coiled-coil heptad arrangement.

Correlation of the consensus secondary structure predictions with that for the complete trichohyalin protein has shown that the prediction of a short region of turn or coil, as seen within the repeat A predicted structure, only occurs in those C-terminal repeats which contain aspartic acid at position 1, whether they be type A or B repeats.

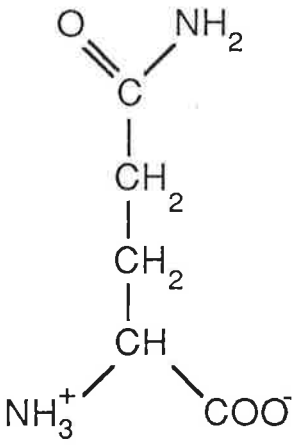
As all of the secondary structure predictions have been made on the amino acid sequence of the precursor trichohyalin molecule, i.e., before the post-translational modifications, it was decided to examine the effect of the replacement of arginine residues with citrulline residues within the C-terminal and N-terminal repetitive sequences. As the PREDICT computer program is based on the use of only the standard twenty amino acids glutamine was chosen as the most similar amino acid to citrulline (see Fig. 5.18) and thus used for subsequent replacements. The replacement of the arginine residues within the solely α-helical region of either the C-terminal or N-terminal repeats was found to have no effect on the predicted secondary structure (data not shown). However, within the C-terminal repeats conversion of the arginine residues on either side of an aspartic acid residue at position 1, i.e., the last arginine of the previous repeat and the position 2 arginine, was found to totally remove the turn or coil which had previously been predicted for that region (Fig. 5.19). Therefore the complete repeat sequence was now predicted to only be able to form α-helix.

e. Database Searches

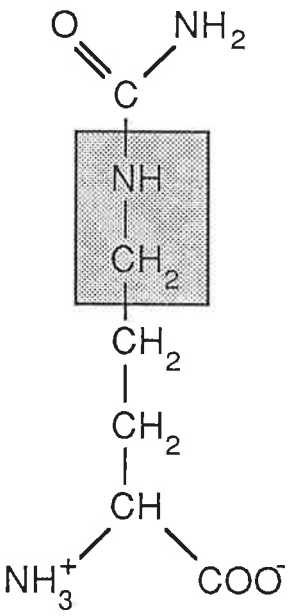
The deduced trichohyalin amino acid sequence was initially searched for regions with homology to published IF amino acid sequences (Conway and Parry, 1988). No regions were found with significant homology to either the central conserved IF α-helical region or any of the N- or C-terminal domains.

Figure 5.18 Comparison of the structures of glutamine and citrulline.

Note that both amino acids contain terminal amide groups but that citrulline has a longer side-chain than glutamine.



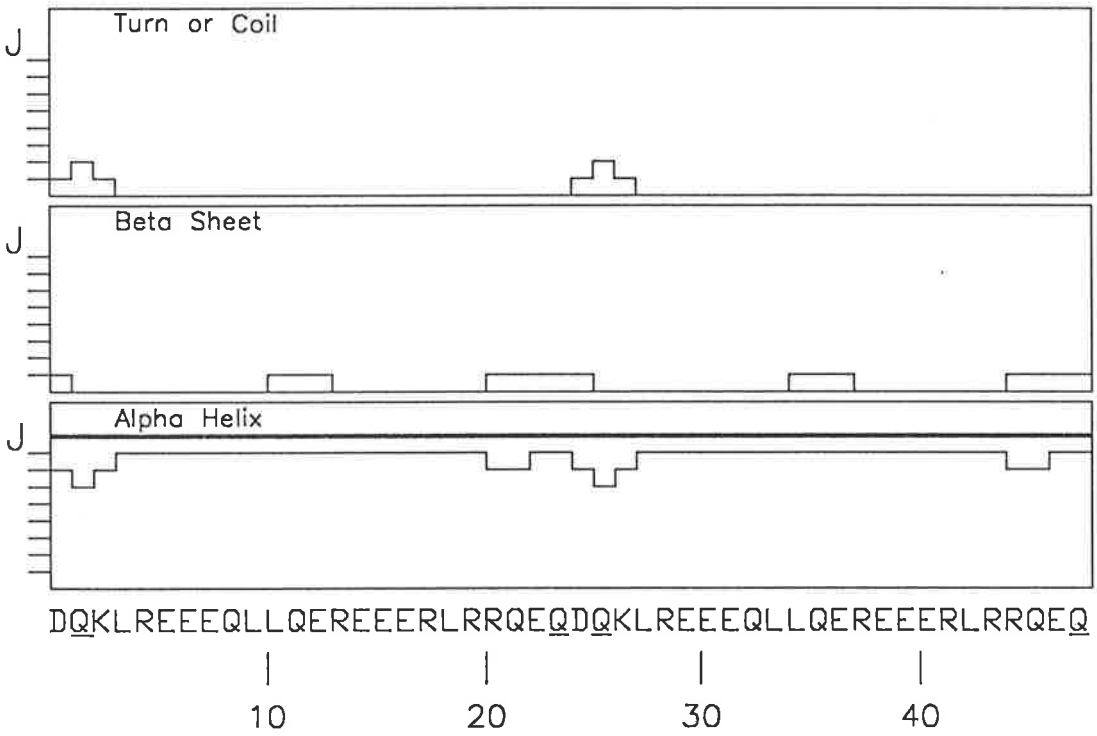
Glutamine



Citrulline

Figure 5.19 Alteration of the secondary structure of the C-terminal repeat A consensus sequence by the replacement of arginine residues with glutamine.

The arginine residues at positions 2 and 24 of the C-terminal repeat A consensus sequence were replaced with glutamine residues (underlined) and the secondary structure of two such consecutive repeats was analysed by the programme PREDICT (Eliopoulos *et al.*, 1982). The resultant plot is shown. Note that the alterations have removed the predicted regions of turn or coil seen with the normal repeat A sequence (see Fig. 5.17b).

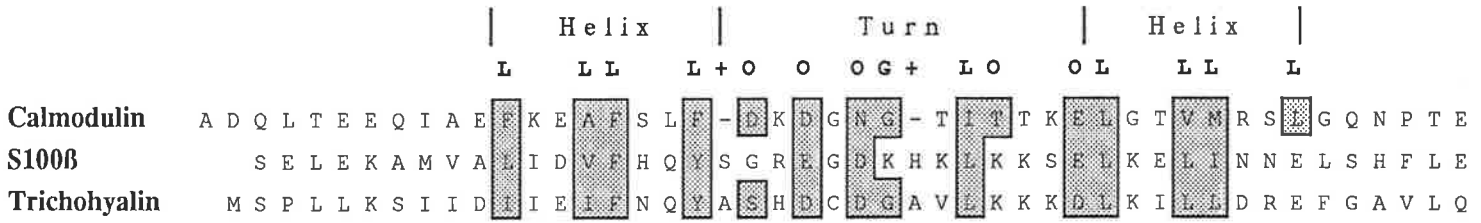

Turn or Coil
J
Beta Sheet
J
Alpha Helix
J
DQKLREEEQLLQEREEERLRRQEQDQKLREEEQLLQEREEERLRRQEQ
10
20
30
40

As stated earlier (Section 5.2.4a), a number of calcium binding proteins, namely, troponin C (Gahlmann *et al.*, 1988; Parmacek and Leiden, 1989), mrp8 (Lagasse and Clerc, 1988), intestinal vitamin D-dependent calcium binding protein (Desplan *et al.*, 1983; Darwish *et al.*, 1987; Krisinger *et al.*, 1988), S100β (Kuwano *et al.*, 1984; Dunn *et al.*, 1987) and sorcin (Van der Bliek *et al.*, 1986), were detected when the translated Genbank database was searched with the first 45 encoded amino acids of the second coding exon of the trichohyalin gene. On further analysis, the predicted N-terminal 85 amino acids of trichohyalin were found to show considerable homology with a family of calcium binding proteins, with the homology being centred on the calcium-binding domains of these proteins (Fig. 5.20). The calcium binding domains, or EF hands, consist of two characteristic α-helices which are separated by a turn (Kretsinger, 1980; see Fig. 5.21) with the calcium ion bound to the turn by the presence of oxygen-containing side-chains. Each of the proteins within the family of EF hand proteins contains between two and four EF hands (Kretsinger, 1980). The structure of the trichohyalin EF hands is almost identical to that of the proteins within the sub-family of small calcium-binding proteins or S100-like proteins. These proteins contain only two EF hands; the N-terminal hand is a variant EF hand containing 30 amino acids whilst the second is a normal EF hand of 28 amino acids (Szebenyi *et al.*, 1981). The greatest similarity of the trichohyalin EF hands is with those of bovine intestinal calcium-binding protein (Fullmer and Wasserman, 1981) with 28 of the amino acids within the region spanned by the two EF hands being identical. Although the variant EF hand of trichohyalin contains the two inserted amino acids seen within the variant EF hands of the small calcium binding proteins, its hand structure has a greater similarity to the normal EF hand, i.e., the glycine residue within the turn is located at the same position as within the normal hand and is surrounded by an almost normal array of oxygen-containing amino acids (see Fig. 5.20). The gene structure of trichohyalin and the S100-like proteins is also very similar; the intron within the trichohyalin gene which occurs between the two EF hands is at the same position as the introns within the small calcium binding proteins (Fig. 5.20). Within the other EF hand proteins, e.g., calmodulin, parvalbumin and troponin C, the gene structure is quite different with the various genes containing up to five introns which almost all splice at points within the regions coding for the individual EF hands.

Figure 5.20 Comparison of the EF hands of trichohyalin with those of calmodulin and S100ß.

The sequences of the first and second EF hands of trichohyalin were aligned with the corresponding regions of bovine calmodulin (Vanaman *et al.*, 1977) and human S100ß (Jensen *et al.*, 1985) which is a member of the family of S100-like proteins. The helix and turn regions of each hand are denoted as are the conserved residues of the EF hands: L, hydrophobic; O, oxygen-containing; G, glycine. Note that trichohyalin and S100ß contain a variant first hand which contains two single amino acid inserts (+).

FIRST EF HAND



SECOND EF HAND

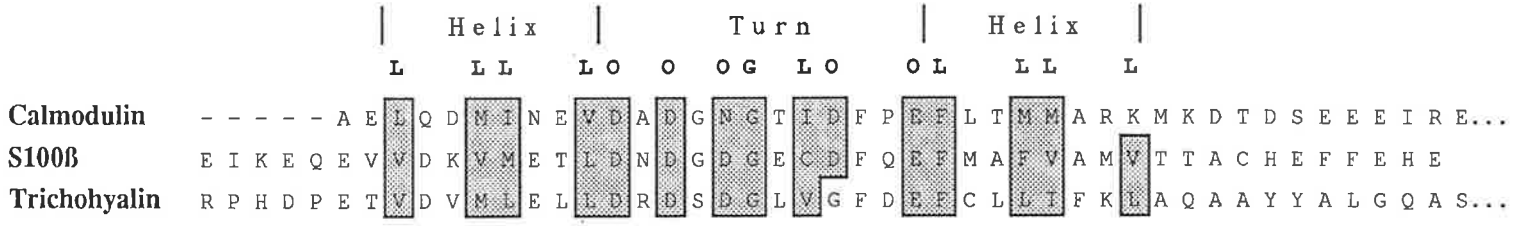
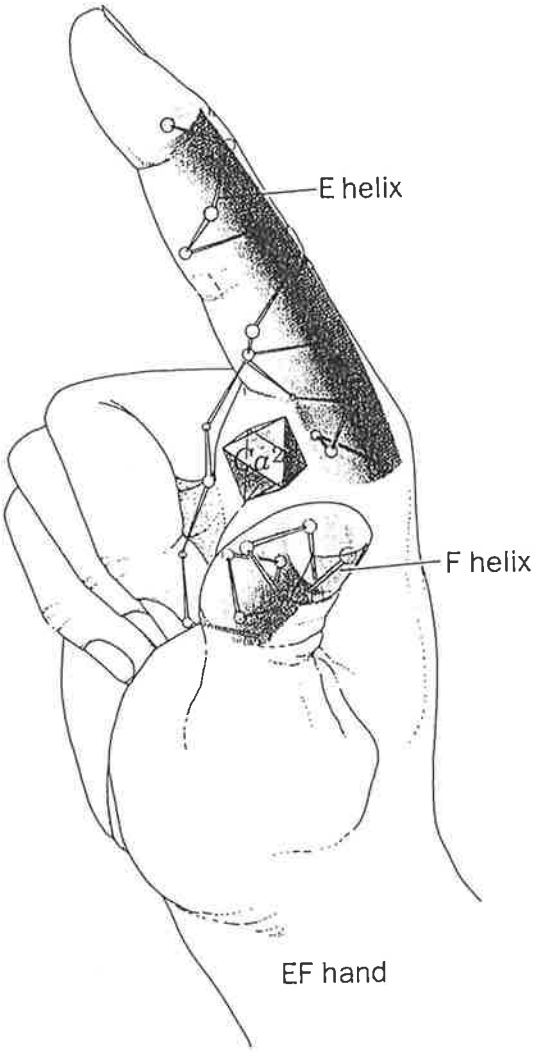


Figure 5.21 Schematic representation of an EF hand.

Shown are the arrangements of the two α-helices (E-helix and F-helix), the turn and the bound calcium ion in an EF hand. The relative positions of the E-helix and F-helix are similar to the positions of the extended forefinger and thumb of a right hand.

(Reproduced from Voet and Voet, 1990.)


E helix
F helix
EF hand

It should be noted that the each of the EF hands within trichohyalin contain a cysteine residue. Based on the crystal structures of the different calcium binding proteins (for a review see Strynadka and James, 1989) the two cysteine side-chains will probably occur at distant parts on the outer surface of the calcium binding domain. It is therefore unlikely that they will be able to form an intramolecular disulphide bond and thus they may be able to form disulphide linkages with either adjacent trichohyalin molecules or with an associated protein.

Both the Genbank and Swiss Protein Databases were then searched for sequences homologous to the remainder of the trichohyalin molecule. At the nucleotide level significant homology was found between the N-terminal third of the trichohyalin coding region and the coding region of human involucrin (Eckert and Green, 1986). Within a 1210 bp overlap which encodes amino acids 124 to 518 of trichohyalin, 55% of the nucleotides within the trichohyalin and involucrin gene sequences were identical. When examined at the amino acid level 23% of the residues were found to be identical. The nucleotide content and the amino acid content of the two genes within the overlap region is shown in Table 5.3. Both sequences contain correspondingly low levels of thymidine (less than 9%) and high levels of guanosine (greater than 40%). In addition both trichohyalin and involucrin have high levels of glutamic acid and glutamine which together total 48% of the trichohyalin residues and 51% of the involucrin residues. However, the levels of arginine, lysine, glycine, histidine and proline are significantly different, particularly in the case of arginine where it accounts for 26% of the trichohyalin residues but for less than 1% of those in involucrin.

No other proteins with significant homology to trichohyalin were found in either the Genbank or Swiss Protein databases.

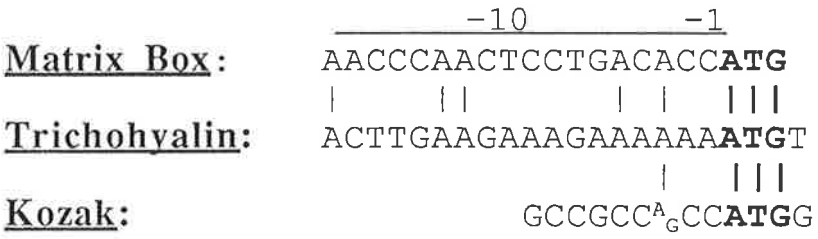
f. Analysis of the Non-coding and Flanking Regions

Although the first exon of the trichohyalin gene has not been located the proximal region of the 5' non-coding region together with the intron, 3' non-coding and flanking regions are defined and can be analysed. Within the 5' non-coding region the sequence immediately preceding the initiation codon bears little homology with the Kozak consensus sequence (Kozak, 1987; Fig. 5.22) although it does contain an adenine residue three bases upstream of the ATG and this is believed to be important for correct initiation. Additionally there is no homology with the "matrix box" which has been found immediately upstream of the initiating ATG in the sheep cortical IFAP genes (Fig. 5.22;

Table 5.3 Comparison of the amino acid and nucleotide content of the coding region for human involucrin (Eckert and Green, 1986) with the homologous portion of the trichohyalin gene (amino acids 124-518).

Amino Acid	Human Involucrin	Trichohyalin (residues 124-518)
	mole percent	*mole percent*
Asp	1.0	1.5
Asn	0.3	0.0
Thr	0.0	0.5
Ser	0.8	2.6
Glu	22.1	29.3
Gln	28.5	18.4
Pro	5.4	1.3
Gly	8.7	1.0
Ala	0.8	3.3
Cys	0.0	0.0
Val	3.1	0.5
Met	0.5	0.0
Ile	0.3	1.0
Leu	16.5	8.9
Tyr	0.3	0.8
Phe	0.3	1.3
Lys	6.2	2.0
His	5.1	0.5
Trp	0.0	1.0
Arg	0.3	26.0
A	27.0	28.0
C	24.1	23.7
G	40.5	41.5
T	8.5	6.8

Figure 5.22 Comparison of the proximal 5' non-coding region of the trichohyalin gene with the "matrix box" sequence of the sheep cortical IFAP genes (see Fig. 1.6) and the consensus sequence of Kozak (1987). The trichohyalin sequence shown little homology with either of the two sequences, although it does contain the purine residue at the -3 position of the Kozak consensus sequence and this is believed to be important for the correct initiation of translation.


-10
-1
Matrix Box:
AACCCAACTCCTGACACCATG
Trichohyalin:
ACTTGAAGAAAGAAAAAAATGT
Kozak:
GCCGCC[A/G]CCATGG

see Section 1.2.3d(ii)). The 3' non-coding region is 1051 bp long and contains a single polyadenylation signal (AATAAA, position 7220). Immediately downstream of the poly(A) addition site, which by comparison with the cDNA clone is after nucleotide 7237, are a GT-rich and a T-rich region (nucleotides 7240-7256) which have been shown to be essential for the efficient and accurate formation of a correct 3' mRNA terminus (Gil and Proudfoot, 1987). The available intron splice sites are highly homologous with the consensus sequences (Mount, 1982) and both introns also contain a consensus DNA branch point motif (5'-PyPyPuAPy-3') which is involved in lariat structure formation for splicing (Ruskin *et al.*, 1984).

5.2.5 Comparison of the Trichohyalin Genomic and cDNA Sequences

a. 3' Non-coding

As both the genomic and cDNA clones contain the complete 3' non-coding region a complete comparison of these sequences was made. The sheep cDNA clone contains a 3' non-coding region of 1025 bp and the corresponding region of the genomic clone consists of 1051 bp. These two regions were aligned by the computer program BESTFIT (Fig. 5.23). The optimal alignment, which contains 25 gaps ranging up to 12 bases in length, has 78% of the aligned nucleotides being identical. The nucleotide differences are spread relatively evenly throughout the 3' non-coding region although it should be noted that within a stretch of 45 bases just upstream of the polyadenylation signal (genomic residues 7140-7184) 43 of the nucleotides are identical.

As the trichohyalin gene has been shown to be present as a single copy within the sheep genome (Section 4.2.5) it is extremely surprising that the 3' non-coding sequences differ so markedly between two sheep sequences, the cDNA and genomic sequences respectively. This was originally attributed to the two different origins of the sequence, namely an Australian Merino (cDNA) and an American breed used by Clontech for the construction of the genomic library.

b. C-terminal coding region

The genomic and cDNA nucleotide and deduced protein sequences which extend from the end of the highly conserved portion of the genomic C-terminal repeat to the stop codon (genomic nucleotides 5878-6183, cDNA nucleotides 1052-1375) are compared in Figure 5.24. Four gaps have been introduced to optimise the alignment and yield an arrangement where 83% of the nucleotides are identical. This is marginally higher than

Figure 5.23 Comparison of the 3' non-coding regions of the trichohyalin cDNA and genomic sequences.

The 3' non-coding regions of the trichohyalin genomic (top) and cDNA (bottom) sequences were aligned using the computer programme BESTFIT (Devereux *et al.*, 1984). Gaps have been introduced into both sequences (.) to maximise alignment, yielding a 78% match between the two sequences. The respective residue numbers are shown.

Continued.....

Figure 5.23 (Cont.)

```
6874 CTTGAAGCAAAAA.TAAGATGTTTCTCAAACAATTTCAGAGCTCATATAT 6922
     | ||||||||||||   |||  |||||||||||||| | ||||| ||| |   |
2024 CATGAAGCAAAAAGCAAGGCGTTTCTCAAACAACTCCAGAGTTCAAACTT 2073

6923 TAAAAC..TATCTCCTGTAGTTTCTAGTGCTGTGTTCTGTTTCCCCCAAG 6970
     ||||||  | |||  ||||||| | |||   |  ||||  |||||||||
2074 TAAAACCATGTCTGCTGTAGTCTGTAGCATTCAGTTC.CTTTCCCCCAGT 2122

6971 TTTTGATGGGTTTGGGAATA....TACCTTTGTGAACTTTGTCTTAGAGG 7016
      || |||| | ||| |||||    | ||||||| || |||  || ||| |
2123 CTTGGATGAGCTTGAGAATATTACTGCCTTTGTTAATTTTAGCTGAGAAG 2172

7017 GAACCTGACCAGGATC..........CTATCTTGCTGTAGGACC..CATG 7054
     | |||||  |||||||          || ||||||| |||  ||  || |
2173 GGACCTGCTCAGGATCCTGTAAACATCTGTCTTGCTCTAGAGCCAACAAG 2222

7055 AGATCACAGAACGTGTTGGGGGATGGAAACGGGGAAAATGTGTGAACAGA 7104
     |||||||||| | | |||||||  || ||  |||||  | |||| |||||
2223 AGATCACAGAGCAT.TTGGGGGTGGGGAAAAGGGAAGTTTTGTGGACAGA 2271

7105 GCGGCAAGTTTCTAGATGCTCTTTAACAAGATTCATCAGCCCACAATCCT 7154
       |||||||  || || |  |||| |||||     |||||||||||||||
2272 AGGGCAAGTCCCTTGAGGACCTTTGACAAGCCCTGTCAGCCCACAATCCT 2321

7155 TTGAGGCCTATTGATACTACCTTGGAGACAGTCTTGTTGAAATAATTGCA 7204
     ||||| |||| |||||||||||||||||||    |||| ||||||||| |
2322 TTGAGCCCTACTGATACTACCTTGGAGACA....TGTTAAAATAATTGGA 2367

7205 CTGTG..AGACATCAATAATAAATGTTTGG 7232
     |||||  | | || |||||||||| ||| |
2368 CTGTGTAAAAAATTAATAATAAATATTTTG 2397
```

Figure 5.24 Comparison of the C-terminal regions of the trichohyalin proteins encoded by the genomic and cDNA clones.

(a) The nucleotide sequence encoding the genomic C-terminal non-repetitive region and the C-flanking domain of the C-terminal repetitive region (top) was aligned with the corresponding sequence of the cDNA clone (bottom) using the computer programme BESTFIT (Devereux *et al.*, 1984). The aligned sequences were found to be 83% identical.

(b) The amino acid sequences encoded by the nucleotide sequences in (a) were aligned using BESTFIT and found to be 82% identical. Note the identity of the last 20 amino acids encoded by both clones.

a.

```
5878 GACCGGCGGTACCGAGCAGAGGAGCAGTTTGCCAGAGAGGAGAAGAGCCG 5927
     |||| || |||||| || |||||||||||||| || |||||||||| ||
1052 GACCAGCAGTACCGGGCGGAGGAGCAGTTTGCTAGGGAGGAGAAGAGGCG 1101

5928 TCGTCTGGAAAGAGAGCTGCGTCAAGAAGAAGAGCAGAGACGCCGCCGGG 5977
     ||||| ||||  |||  || | |   ||||||||||||||||||||| ||
1102 TCGTCAGGAACAAGAATTGAGGC...AAGAAGAGCAGAGACGCCGCCAGG 1148

5978 AACGGGAGAGGAAATTCCGG...GAAGAGCAGCTCCGCCGCCAGCAGGAG 6024
     | ||||||||||||||||||   ||||| |||||||||||||||||| ||
1149 AGCGGGAGAGGAAATTCCGGGAAGAAGAACAGCTCCGCCGCCAGCAGCAG 1198

6025 GAGGAGCAGAGGCGCCGCCAA.........CTGCGGGAGAGACAATTCCG 6065
     |||||||||| ||| ||||||          ||| | |||| |    ||
1199 GAGGAGCAGAAGCGTCGCCAAGAGCGCGACGTGCAGCAGAGCCGCCGCCA 1248

6066 A.........GAAGATCAGAGCCGTCGGCAAGTTCTGGAGCCCGGTACAC 6106
     |         |||||  || |||| ||||| || |||||| | || |  |
1249 AGTGTGGGAGGAAGACAAGGGCCGCCGGCAGGTCCTGGAGGCTGGCAAGC 1298

6107 GTCAGTTTGCTAGGGTCCCGGTGCGCTCCAGCCCTCTTTATGAATACATC 6156
     | |||||||| || | ||| |||||||||| || || || || ||||||
1299 GGCAGTTTGCCAGTGCCCCAGTGCGCTCCAGTCCGCTCTACGAGTACATC 1348

6157 CAAGAGCAGAGATCTCAGTACCGCCCCTAAGA 6188
     |||||||||| |||||||||||| || ||
1349 CAAGAGCAGAGGTCTCAGTACCGCCCTTAGGA 1380
```

b.

```
1306 DRRYRAEEQFAREEKSRRLERELRQEEEQRRRRERERKFR.EEQLRRQQE 1354
     |  |||||||||||| || | |||| |||||| ||||||| ||||||||
 351 DQQYRAEEQFAREEKRRRQEQELRQ.EEQRRRQERERKFREEEQLRRQQQ 399

1355 EEQRRRQLRERQ......FREDQSRRQVLEPGTRQFARVPVRSSPLYEYI 1398
     ||| ||| |  |         ||  |||||| | ||||  ||||||||||
 400 EEQKRRQERDVQQSRRQVWEEDKGRRQVLEAGKRQFASAPVRSSPLYEYI 449

1399 QEQRSQYRP 1407
     |||||||||
 450 QEQRSQYRP 458
```

that seen within the 3' non-coding region. Analysis of the corresponding amino acid alignment has shown that 82% of the amino acids are identical. In particular note that the final 20 amino acids are identical in both protein sequences.

In addition, one of the guinea pig peptide sequences (peptide M) was also derived from the non-repetitive region at the C-terminus of trichohyalin and this was compared with the corresponding cDNA and genomic deduced amino acid sequences (Fig. 5.25). Note that the genomic sequence has no greater similarity to the sheep cDNA sequence than has the guinea pig peptide sequence.

c. C-terminal repeat consensus sequences

As stated in Section 5.2.4d, the C-terminal repetitive sequences of the cDNA and genomic clones differ significantly. This is particularly shown in the consensus sequences with the cDNA repetitive region based on a single 23 amino acid consensus sequence and the genomic repetitive region based on two consensus sequences, one of 24 amino acids (A) and the other of 22 (B). Both the nucleotide and amino acid consensus sequences are compared in Figure 5.26. The cDNA consensus sequence is very similar to that of the genomic repeat A, with most nucleotide changes leaving the encoded amino acid unaltered, but differs markedly from the repeat B sequence, particularly in the central variable region defined above (Section 5.2.4d). These differences in the consensus sequences suggest that the trichohyalin repetitive sequence has either undergone very rapid evolution during the development of the different sheep breeds or that the commercial genomic library with which I was provided was derived not from sheep but from a separate species.

d. Comparison of sequence homology by Zoo Blot

As the identity of the DNA within the genomic library was under doubt it was decided to construct homologous 3' non-coding probes from the sheep cDNA and genomic sequences and probe them to a Zoo Blot containing sheep, human and mouse DNA which had been cut with EcoR I. The 0.47 kb EcoR I fragment was used as the cDNA probe and the 0.8 kb Pst I/EcoR I genomic fragment (see Fig. 5.4) which spans the complete 0.47 kb cDNA EcoR I fragment plus an extra 17 bp upstream and 309 bp downstream was used as the other probe. These probes were hybridised separately to the same Zoo Blot and it was found that the cDNA probe bound specifically to a 6 kb band in the sheep track and the genomic probe bound specifically to a 7 kb band in the human track (Fig. 5.27). A genomic N-terminal coding probe (1.2 kb EcoR I/Pst I fragment)

Figure 5.25 Comparison of the amino acid sequence of the guinea pig trichohyalin proteolytic peptide M with the corresponding regions of the proteins sequences deduced from the cDNA and genomic clones. Note that the cDNA and genomic sequences are from the respective C-terminal non-repetitive sequences. Residues which are present in two or more of the sequences are boxed.

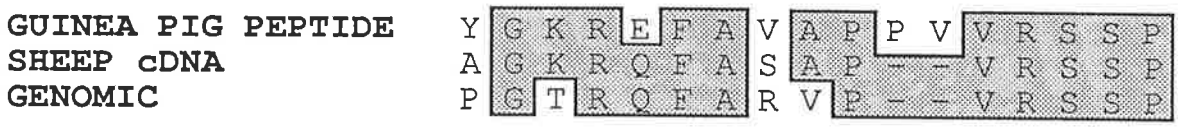

GUINEA PIG PEPTIDE
SHEEP cDNA
GENOMIC
Y G K R E F A V A P P V V R S S P
A G K R Q F A S A P – – V R S S P
P G T R Q F A R V P – – V R S S P

Figure 5.26 Comparison of the consensus sequences for the C-terminal repetitive regions present within the trichohyalin proteins encoded by the genomic and cDNA clones.

(a) The nucleotide sequences of genomic repeats A and B and the cDNA consensus sequence are aligned. Nucleotides within the repeat B and cDNA sequences which are identical to those in repeat A are marked by dots. Note that the majority of the nucleotide changes occur at the third base of the respective codons.

(b) Alignment of the amino acid sequences encoded by the nucleotide sequences of (a). Identical amino acids are indicated as in (a).

a.

```
Genomic(A)    1 GAC AG[G/A] AAA [C/T]TC CGC  GAG GAG GAG CAG CT[C/G] CTT 33
Genomic(B)    1 •[A/C]• ••G ••A C•• ••[C/T] ••• ••• ••• ••• ••G ••G 33
cDNA          1 ••• ••A ••G T•• ••T ••• ••• ••A ••• ••C ••G 33

Genomic(A)   34 CA[G/A] GA[G/A] AGG GAG GAA GAG [A/C]GG CTG CGC CGT CAG GAA CGC 72
Genomic(B)   34 ••G C[A/G]G --- ••• ••G C•• GA• ••C --- ••T ••• ••• ••• 66
cDNA         34 •GC ••A ••• ••A ••• --- CA• ••• ••• ••C ••• ••• ••• 69
```

b.

```
Genomic(A)     D R K [L/F] R E E E Q L L Q E R E E E R L R R Q E R
Genomic(B)     [A/E] • • L • • • • • • • R [Q/R] - • • Q E • - • • • •
cDNA           • • • F • • • • • • • • • • • • - Q • • • • • •
```

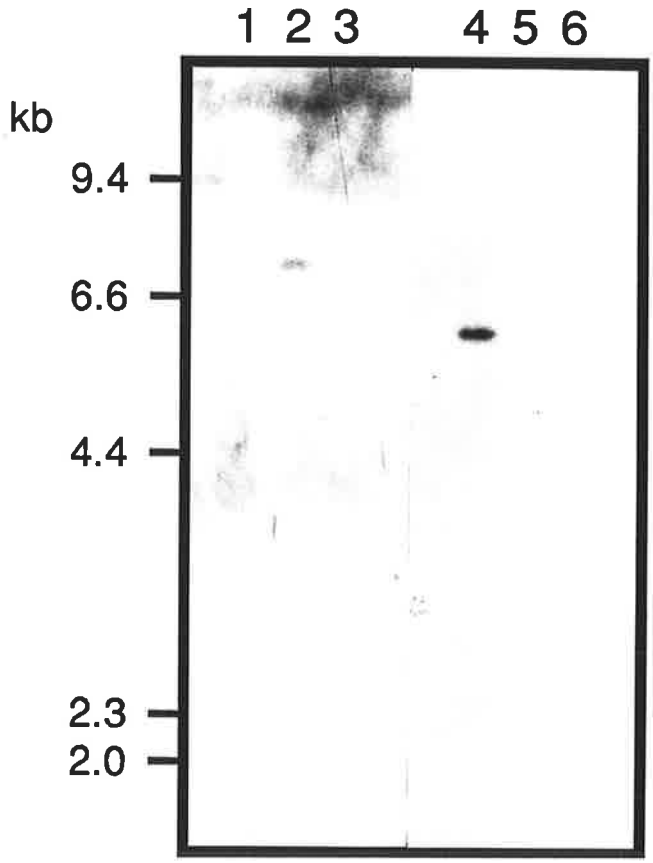
Figure 5.27 Southern analysis of a Zoo blot using 3' non-coding probes from the cDNA and genomic clones.

A Zoo blot, containing sheep (tracks 1 and 4), human (2 and 5) and mouse (3 and 6) genomic DNA which had been digested with EcoR I and separated on a 1% agarose gel, was probed separately with the 0.47 kb EcoR I fragment from the 3' non-coding region of the cDNA clone (tracks 1-3) and the 0.8 kb Pst I/EcoR I (PE0.8, Fig. 5.4) fragment from the 3' non-coding region of the genomic clone (tracks 4-6). The cDNA fragment hybridised specifically to a single band in the sheep digest and the genomic fragment hybridised to a specific band in the human digest. Both hybridisations were washed in 2x SSC, 0.1% SDS at 65°C. Tracks 1-3 were autoradiographed for 18 h and tracks 4-6 for 24 h. Size markers are shown (in kb).

Erratum:

Figure 5.27 Southern analysis of a Zoo blot using 3' non-coding probes from the cDNA and genomic clones.

A Zoo blot, containing sheep (tracks 1 and 4), human (2 and 5) and mouse (3 and 6) genomic DNA which had been digested with EcoR I and separated on a 1% agarose gel, was probed separately with the 0.47 kb EcoR I fragment from the 3' non-coding region of the cDNA clone (tracks 4-6) and the 0.8 kb Pst I/EcoR I (PE0.8, Fig. 5.4) fragment from the 3' non-coding region of the genomic clone (tracks 1-3). The cDNA fragment hybridised specifically to a single band in the sheep digest and the genomic fragment hybridised to a specific band in the human digest. Both hybridisations were washed in 2x SSC, 0.1% SDS at 65°C. Tracks 1-3 were autoradiographed for 24 h and tracks 4-6 for 18 h. Size markers are shown (in kb).


1 2 3 4 5 6
kb
9.4
6.6
4.4
2.3
2.0

also bound to the same human band as the genomic 3' non-coding probe but not to the sheep DNA (data not shown). This conclusively shows that the genomic clone does not derive from sheep DNA. As the genomic probe hybridised to a human fragment which differs in size from the 5.5 kb EcoR I genomic fragment from which the probe was derived, it would appear that the EcoR I site is polymorphic in human DNA or that the library is not human but contains DNA from some other primate species.

5.3 Discussion

The initial aim of the work described in this chapter was the purification, sequencing and analysis of the sheep trichohyalin gene such that the structure and function of trichohyalin could hopefully be determined. Thus a commercial sheep genomic library was ordered and the trichohyalin gene purified and sequenced. Unfortunately, on the basis of evidence presented above and which will be discussed later, it was eventually determined that the provided library originated from a primate species rather than from sheep. Nevertheless, as the structure of trichohyalin appears to be conserved between mammalian species, which is evidenced by the immunocross-reactivity of trichohyalin from numerous mammalian species (Rothnagel and Rogers, 1986), the analysis of the obtained sequence is still valuable in the understanding of the general structure and function of trichohyalin and thus also of sheep trichohyalin. Therefore I will here discuss the analysis of the purified sequence and also compare the primate trichohyalin gene sequence and the sheep cDNA sequence.

5.3.1 Analysis of the Trichohyalin Genomic Sequence

a. Structure of the Trichohyalin Gene

Genomic λ clones, which were detected on the basis of hybridisation to the trichohyalin cDNA clone λsTr1, were purified and the approximate location and the orientation of the trichohyalin gene within the longest clone, λsGT1b, was determined. A region of 7456 bp, which was believed to probably span the trichohyalin gene, was sequenced (Fig. 5.8). Although the 3' end of the gene was easily located by comparison with the cDNA clone sequence the 5' end was not able to be defined experimentally due to both the lack of a cDNA clone covering the 5' end of the mRNA and the sequence differences between the Merino RNA and the purified genomic clone. Fortunately the discovery of two EF hands (calcium-binding domains) within the trichohyalin N-terminal

sequence allowed the beginning of the coding region, and also the structure of the gene, to be predicted.

It therefore appears that the trichohyalin gene consists of three exons (Fig. 5.11): exon 1, which has not been located, is untranslated; exon 2 contains 22 bp of the 5' non-coding region and the first 138 bp of the coding region which encodes the first EF hand; exon 3 contains the remaining 4083 bp of the coding region and the complete 3' non-coding region (1051 bp). The trichohyalin protein is 1407 amino acids long and has a molecular weight of 183,780.

b. Analysis of the Deduced Trichohyalin Protein Sequence

Trichohyalin has an unusual amino acid composition with a high content of hydrophilic amino acids (approx. 80%) and the four amino acids glutamic acid, arginine, glutamine and leucine constituting over 80% of the protein's residues (Table 4.1). The high levels of these four residues may be required to allow α-helix to form within trichohyalin. A number of the glutamine and arginine residues will also act as substrates for follicle transglutaminase and peptidylarginine deiminase. Trichohyalin contains 25 more acidic residues than basic which gives the protein a pI of 5.44. It should be noted that the conversion of arginine residues to citrulline by peptidylarginine deiminase will further acidify the trichohyalin protein. The increased acidic nature of trichohyalin could significantly alter its structure, physical properties and interaction with itself and other proteins. This acidification could be critical in the breakdown of the trichohyalin granules and integration into the filamentous network of the IRS.

The trichohyalin sequence is also characterised by the presence of only two cysteine and two methionine residues, all four of which occur within the 95 amino acid N-terminal hydrophobic sequence. The presence of only a single internal methionine, occurring 56 amino acids from the N-terminus, would explain the very poor cleavage results obtained with cyanogen bromide, i.e., a single correct cleavage product of only 6 kdal would not have been noticed (Section 3.2.2b), whereas the two cysteine residues would be the means by which disulphide cross-linking could occur during protein purification (Section 3.2.1).

Examination of the deduced trichohyalin protein sequence has shown that trichohyalin contains two separate repetitive regions (Fig. 5.12). The large C-terminal repeat, which corresponds to that seen within the cDNA clone extends over most of the

C-terminal 60% of trichohyalin whereas the small N-terminal repetitive region is only 114 residues long.

The C-terminal repetitive region consists of a highly conserved central domain (30 repeats) together with less conserved flanking domains (see Fig. 5.13). Although the average length of the repeats within the central domain is 23 amino acids the sequence is based on two consensus sequences, one of 24 amino acids (repeat A) and one of 22 amino acids (repeat B) (Fig. 5.14b), rather than the single 23 amino acid repeat present within the cDNA clone (Fig. 4.8). The genomic consensus sequences will be discussed and compared with the cDNA consensus sequence later in the discussion.

The N-terminal repetitive region consists of three highly conserved repeats. Unlike the C-terminal repetitive region the N-terminal region does not appear to have any less-conserved flanking regions. The three N-terminal repeats are based on a 40 amino acid consensus sequence that differs from the C-terminal consensus sequences. Although the consensus sequences differ the constituent amino acids are very similar with all but one of the N-terminal consensus amino acids being either glutamic acid, arginine, glutamine or leucine. No lysine residues are present within the N-terminal consensus sequence suggesting that this region is not important in providing lysine residues for cross-linking by transglutaminase.

c. Comparison with Homologous Proteins

Examination of the Genbank and the Swiss Protein Databases for proteins homologous to trichohyalin yielded the epidermal transglutaminase substrate involucrin. The greatest similarity was found when the human involucrin gene sequence (Eckert and Green, 1986) was aligned with the sequence encoding the majority of the N-terminal third of trichohyalin (amino acids 130-521) with 55% of the nucleotides and 23% of the amino acids found to be identical. The nucleotide content within this region was found to be very similar in the two proteins and is characterised by a very low thymidine content and a high level of guanosine (Table 5.3). Also glutamic acid and glutamine together constitute approximately 50% of the amino acids within the overlap region of both proteins. However, the two sequences covering this region were found to differ significantly with regard to the content of a number of other amino acids, especially arginine which constitutes more than a quarter of the trichohyalin residues but less than 1% of those in involucrin. A random arrangement of the amino acids within this portion of the trichohyalin and involucrin sequences would produce over 50% of the observed

amino acid matches due to the high levels of glutamic acid, glutamine and leucine. Further, the involucrin amino acid sequence consists almost completely of tandem 10 amino acid repeats whereas the corresponding trichohyalin sequence is predominantly non-repetitive. These facts suggest that the two sequences have evolved quite separately and that the high degree of similarity is dependent upon the similar biased nucleotide ratios and the high levels of both glutamic acid and glutamine. These similar characteristics may reflect some structural and functional homology between the two proteins which could include their roles as transglutaminase substrates.

The N-terminal 95 amino acids of trichohyalin are quite distinct in amino acid content from the remainder of the protein, containing a higher proportion of hydrophobic residues (Table 5.2). Detailed analysis of this N-terminal hydrophobic region yielded an unexpected discovery; the trichohyalin N-terminus contains two EF hand calcium binding domains. This motif was originally recognised by crystal structure analysis of carp parvalbumin (Kretsinger and Nockolds, 1973) and involves a helix-turn-helix structure with the central turn having a number of oxygen-containing residues positioned at the co-ordination sites of the calcium ion. The arrangement and number of EF hands within trichohyalin is almost identical to that within a family of small calcium-binding proteins (M_r 10k-15k) which are termed the S100-like proteins (Fig. 5.20). The sequence of the first EF hand in these proteins differs from that found in a regular EF hand. It contains two single amino acid insertions, a change in the position of the conserved glycine residue within the turn and an alteration to the positions of the oxygen-containing amino acids (Szebenyi *et al.*, 1981). Although calcium can still bind strongly to the resultant hand it co-ordinates via the carbonyl oxygen atoms in the main chain rather than the oxygen atoms present in the side-chains (Szebenyi and Moffat, 1986). The sequence of the first EF hand of trichohyalin falls between those of the variant and regular EF hands; it contains the two amino acid inserts of the variant EF hand but has the regular positioning of both the glycine and the oxygen-containing residues (Fig. 5.20). If, as expected from the sequence, the first EF hand has the capacity to bind calcium, detailed structural analysis will be required to determine the arrangement of the co-ordinating oxygen atoms.

The vast majority of the proteins which contain EF hands act as modulating proteins. The binding of calcium to the EF hand leads to a structural change which is transmitted to a bound protein, in turn changing its structure or function, e.g., the binding

of calcium to calmodulin causes the activation of the bound enzyme phosphodiesterase (Kakiuchi and Yamazaki, 1970). One exception to this is calpain, a calcium-dependent protease, which contains a domain consisting of four EF hands in a similar arrangement to those in calmodulin that are fused at the C-terminus of a thiol protease (Ohno *et al.*, 1984; Emori *et al.*, 1986). The proteolytic activity of calpain is dependent upon the binding of calcium to the EF hands, although it should be noted that calpain exists as a heterodimer and the binding of calcium to the second protein is also required for activity. Trichohyalin, like calpain, appears to contain an EF hand structure fused to a second protein domain, although the EF hands are present at the N-terminus and trichohyalin has no known enzymatic activity which could be influenced by the binding of calcium. Why then does trichohyalin contain a pair of EF hands? Firstly, trichohyalin may indeed contain some unknown enzymic activity which is only activated in the presence of calcium. Secondly, trichohyalin may be bound to one or both of the calcium-dependent modifying enzymes, i.e., transglutaminase or peptidylarginine deiminase. It is possible that the binding of the enzyme(s) to trichohyalin may lead to the loss of its/their activity and this can only be restored by the binding of calcium to the EF hands of trichohyalin. Conversely, calcium may normally be bound to trichohyalin and the enzymes are only activated when the calcium is released. Alternatively, such a release of calcium could provide the calcium ions required to activate unbound transglutaminase and peptidylarginine deiminase and this could occur if there is a sudden change in the reducing potential of the IRS and medulla cells. For example, if the developing cells have a high reducing potential, as suggested by Rogers (1958a), such that the cysteine residues within the EF hands exist in the sulphydryl form, then a change in the reducing potential may cause the cysteine residues to form disulphide linkages, altering the structure of the EF hands and causing a significant decrease in the binding strength for calcium. At present there is nothing known about the binding capacity of trichohyalin for calcium, the presence of calcium within the granules or the levels of calcium within the cells of the IRS and medulla. Much work needs to be done to examine the purpose of a calcium-binding domain within trichohyalin.

d. Function of Trichohyalin

Although the complete trichohyalin protein sequence now appears to have been obtained, the structural function of trichohyalin is still unclear. Examination of the sequence has shown that trichohyalin does not contain any region with significant

homology to the conserved central α-helical domain of the IF proteins or have any long stretch of heptad repeats which are capable of forming a coiled-coil structure. Nevertheless, almost the complete protein is predicted to form α-helix and this includes five stretches of 80 or more amino acids which do not have any alternative predicted structure. This prediction of α-helix is predominant within the C-terminal and N-terminal repetitive regions. The N-terminal repetitive region contains a stretch including over 90% of the region which has α-helix as the only predicted secondary structure. This, together with the short regions of coiled-coil heptads present within the N-terminal repetitive sequences suggest that these repeats could be involved in some form of coiled-coil α-helical rod. As seen by the predicted structures for the repeat A and B consensus sequences α-helix is also predicted over the complete C-terminal repetitive region (Fig. 5.17). In addition the C-terminal repetitive region also contains short stretches of predicted turn or coil. All of these predicted stretches of turn or coil were found to span a stretch of approximately four residues at the beginning of repeats containing aspartic acid at position 1. This suggests that the C-terminal repetitive region of the trichohyalin protein, i.e., the C-terminal two-thirds, is capable of forming an arrangement containing short regions of α-helix, ranging from one to six repeats in length, which are joined by short stretches of random turn. These short α-helical regions may be able to form intra- and inter-molecular associations with other C-terminal α-helical regions thus producing the granular protein arrangement seen within the developing IRS and medulla cells. Although this secondary structure may explain the granular form of trichohyalin it does not determine whether trichohyalin can form a filamentous structure within the hardened IRS.

One means of further analysing the structure of trichohyalin is to examine the effect that the conversion of arginine residues to citrulline has on the secondary structure of trichohyalin. Although the precise location of the arginine residues which are converted is unknown and the secondary structure analysis program PREDICT only allows the use of the twenty standard amino acids, the effect of deimination could still be examined by replacing arginine residues within the repeat sequences with glutamine, the most related standard amino acid to citrulline. Secondary structure analyses were therefore performed on numerous stretches of sequence within the repetitive regions of trichohyalin in which either individual arginine residues or sets of arginine residues had been replaced with glutamine. Outstandingly, the only changes which altered the

predicted secondary structure occurred within the C-terminal repeats which contained aspartic acid at position 1 and involved the conversion of the two arginine residues immediately adjacent to the aspartic acid residue. This double conversion caused the complete removal of the predicted turn or coil from these repeats (Fig. 5.19) rendering almost the complete C-terminal repetitive region into a single α-helical rod. As these two arginine residues are within the predicted short random turn of the trichohyalin molecule it is highly likely that they would be accessible to peptidylarginine deiminase and thus converted to citrulline. Therefore the conversion of these pairs of arginine residues within the granular trichohyalin protein could allow the complete C-terminal region of trichohyalin to fold out into a large α-helical rod which may then be able to interact with other trichohyalin C-terminal regions to form the filamentous structure seen within the mature IRS. This speculation on the structure of the C-terminal repetitive region does not clearly explain the role of both the repetitive and non-repetitive regions within the N-terminal third of trichohyalin. One possibility is that they act together to provide a matrix between the filaments formed by the C-terminal repeats.

The proposal that trichohyalin forms both the filaments and the matrix of the IRS cells suggests that trichohyalin contains two distinct structural domains, i.e., a C-terminal domain which contains the C-terminal repetitive region and the short non-repetitive sequence at the C-terminus, and a smaller N-terminal domain which contains all of the N-terminal repetitive and non-repetitive sequences. The proposed domain sizes partly correlate with Western analysis of sheep follicle proteins in which a clear set of bands with a molecular weight of approximately 60,000 - 65,000 were bound by the anti-trichohyalin antibody (see Fig. 3.8, track C). These bands are approximately of the size expected for the N-terminal structural domain and could be produced by the cleavage of trichohyalin at a point near the beginning of the C-terminal repetitive region. Further to this, gel analysis of a partially degraded trichohyalin sample, performed by Rothnagel and Rogers (1986), shows two clear sets of bands which appear to be of a size corresponding to both of the predicted C-terminal and N-terminal structural domains. These experimental data cannot prove the proposed model of trichohyalin function but do appear to show that trichohyalin contains two structural domains.

One question that arises from the proposed function of trichohyalin is, "why doesn't the same conversion of granular trichohyalin to filamentous trichohyalin take place within the medulla?". A possible answer to this could be an altered timing of

expression of the enzymes transglutaminase and peptidylarginine deiminase within the IRS and medulla. Within an IRS cell the deimination and formation of filaments could occur before the proteins are cross-linked, whereas in the medulla transglutaminase could be expressed first, such that the trichohyalin molecules are cross-linked and fixed into position prior to the deimination of the arginine residues and the predicted transition to an extended α-helical structure. A second option is that the IRS contains a specific trichohyalin protease, i.e. the 60 kdal trichohyalin breakdown products mentioned above could be caused by specific proteolytic cleavage rather than non-specific degradation. Thus the separation of the two trichohyalin domains may be essential in the formation of the IRS filamentous structure. The lack of expression of such a protease within the medulla would therefore stop the formation of filaments. If trichohyalin is cleaved at the domain junction by a specific protease then there must be some unique sequence which can be recognised at that site. Analysis of the sequence has found that the junction of the proposed N-terminal and C-terminal domains contains two peptide sequences which contain a large number of residues that are not typical within the trichohyalin sequence, i.e., the sequences PGQTWRW (residues 518-524) and HTLYAKPG (residues 535-542). One or both of these sequences could then possibly act as a substrate for the proposed protease thus separating the two structural domains of trichohyalin. Interestingly, the threonine within the second of these sequences is, on the basis of the known recognition sequence (Glass and Smith, 1983; Glass *et al.*, 1986), the only possible phosphorylation site for cAMP- or cGMP-dependent protein kinase within trichohyalin. If this phosphorylation occurs, it could be involved in the cleavage at these sites or it could play some other separate function.

5.3.2 Comparison of the Primate and Sheep Trichohyalin Sequences

a. Sequence Comparisons

As was stated earlier (Section 5.3.1), most of the attempted analysis of the 5' end of the trichohyalin gene was hindered by the sequence differences between the Merino mRNA and the purified genomic clone, whose identity was assumed at the time to be sheep. The initial evidence of the differences between the sheep cDNA sequence and that of the genomic DNA was obtained by Northern analysis which showed that a 1.2 kb fragment from the coding region of the genomic clone hybridised much more weakly to

follicle RNA than a 3' non-coding probe from the cDNA clone (Fig. 5.10). Subsequently, corresponding 3' non-coding probes from both the cDNA and genomic clones were hybridised to a Zoo blot containing sheep, human and mouse genomic DNA. The cDNA probe bound specifically to a fragment within the sheep DNA and the genomic probe bound specifically to a fragment within the human DNA (Fig. 5.27). It therefore appears that the genomic DNA originated from either man or some other primate species. This would explain the poor homology at the 5' end and the inability to experimentally locate the 5' end of the gene with sheep follicle RNA. The discovery of the EF hand motifs has allowed the location of the second and third exons of trichohyalin to be predicted although the proposed first untranslated exon has yet to be located. There are three means by which the first exon can be located and the intron/exon junctions confirmed. Firstly, the source of the genomic clone can be determined either by in situ hybridisation analysis or by Southern analysis of the genomic DNA used in the construction of various Clontech libraries. The corresponding follicle RNA can then be isolated for primer extension analysis, Northern analysis and RNA protection. Secondly, the sheep trichohyalin gene can be purified from a separate library and used for analysis with wool follicle RNA or, thirdly, the isolated gene can be expressed in transgenic mice and the follicle RNA isolated for analysis. The accuracy of the final method would be dependent upon the trichohyalin transcript being processed normally in the mouse follicle.

Although the purification of a primate trichohyalin gene has caused difficulties in the analysis of the 5' end of the trichohyalin gene it has allowed a partial comparison of the trichohyalin gene sequences from two different mammalian species. One of the major points seen within the comparison is the differences between the sheep and primate C-terminal repetitive region consensus sequences. The deduced primate C-terminal repetitive sequence is based on two different consensus sequences, one of 24 amino acids and the other of 22 amino acids, whilst that of the sheep cDNA sequence contains only a single 23 amino acid consensus sequence. Although the three consensus sequences are nearly identical over 15 of their amino acids, i.e. the first 11 and the last 4 of each consensus sequence, they contain a central less-conserved portion ranging between 7 and 9 residues in length (Fig. 5.26). The cDNA and genomic repeat A consensus sequences differ little over the less-conserved portion with the only difference being the substitution of the glutamine within the cDNA sequence at position 17 with the amino acids glutamic acid and arginine. However both of these sequences differ markedly from the genomic

repeat B consensus sequence which contains only three amino acids within the less-conserved portion which align with those in the other two repeat sequences (Fig. 5.26). Although the consensus sequences differ within this region there is no difference in the predicted secondary structure over this portion of the three repeats (Fig. 5.17 and Fig. 4.6(b)). Therefore it is possible that this less-conserved region is involved in the formation of the necessary secondary and tertiary structures of trichohyalin but may not include substrate amino acids for either transglutaminase or peptidylarginine deiminase; possible substrate amino acids for both enzymes are present within the conserved portion of the consensus sequences. It should thus be noted that the glutamine at position 17 within the cDNA consensus sequence which was earlier proposed to act as a transglutaminase substrate (Section 4.3) occurs within the variable region of the consensus sequences and that the surrounding charged environment present within the cDNA sequence does not occur within either of the genomic consensus sequences. Therefore it appears that this glutamine residue probably does not act as a substrate amino acid for transglutaminase. Interestingly the average repeat length within the cDNA and genomic C-terminal repetitive regions is identical due to the genomic region containing equal numbers of repeats A and B. This suggests that an average repeat length of 23 amino acids may be important in obtaining the correct trichohyalin structure.

Comparison of the C-terminal portion of the deduced genomic sequence, i.e., the non-repetitive region at the C-terminus and the less-conserved C-domain repeats, with the corresponding portion of the cDNA clone has shown that 81% of the amino acids are identical. Of particular note the C-terminal 20 amino acids of the sequence in both sheep and primate trichohyalin are identical (Fig. 5.24) suggesting that this sequence may play a critical role in the function of trichohyalin.

Analysis of the two 3' non-coding regions has shown that they have a similar degree of conservation as the C-terminal coding region with 78% of the 3' non-coding nucleotides being identical. Although most of the differences are spread evenly throughout the 3' non-coding region there is a very highly conserved region which is situated approximately 40 bp upstream of the polyadenylation signal (Fig. 5.23). This region spans 45 nucleotides (genomic nucleotides 7140-7184, cDNA nucleotides 2307-2351) of which 43 are identical. It is therefore possible that this region may play some role in either the production or processing of the trichohyalin mRNA although it does not contain the AU rich sequence seen within the 3' non-coding region of transiently

expressed RNA species (Shaw and Kamen, 1986) or the enhancer elements which are present in the 3' non-coding region of the k-fgf mRNA (Curatola and Basilico, 1990). In order to further examine the role of this sequence the trichohyalin gene will need to be expressed in an homologous system and the sequence analysed by subsequent mutation or deletion.

b. Evolution of the Trichohyalin Gene

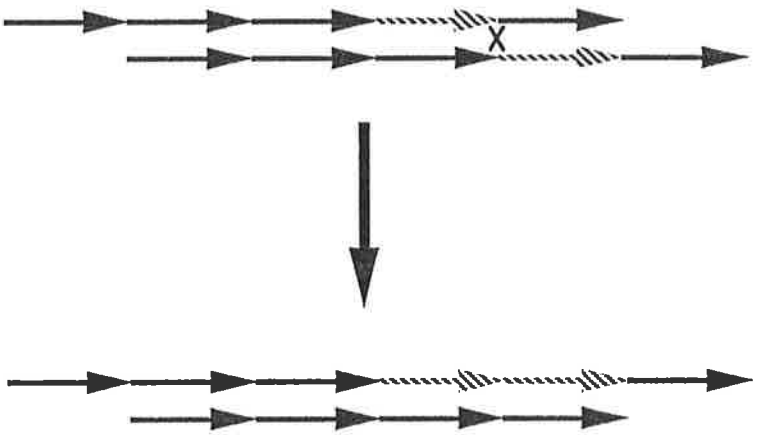
The comparison of the coding regions of sheep and primate trichohyalin also allows the discussion of the evolution of the C-terminal repetitive region. As the sheep and primate C-terminal repetitive regions are based on different consensus sequences it is possible that the repetitive region could have evolved from separate sequences within the ancestral gene, as is the case for the human and lemur involucrin repeats (Tseng and Green, 1988), or that the repetitive regions were produced prior to the species divergence and a subsequent mutation has since spread throughout either one or both of the highly conserved repetitive regions. Analysis of the less-conserved C-domains within the repetitive regions of the sheep (residues 351-428) and primate (residues 1306-1377) trichohyalin has shown that their sequences are based on the C-terminal consensus sequences. However, the alignment of the two sequences (Fig. 5.24) shows that very few gaps are needed to produce an optimal alignment and that the positions of these gaps do not correspond to the differences within the genomic and cDNA consensus sequences. Additionally, both of the less-conserved repetitive domains contain residues which are not normally present within the consensus sequences, as well as unusual arrangements of amino acids. The majority of these residues occur at identical positions in both sequences, e.g., both domains contain the sequence YRAEEQFAREEK. These similarities strongly suggest that the less-conserved C-domains were formed after the production of the repetitive region but prior to the deviation of the species. Therefore the consensus sequence differences must have evolved after the formation of the peptide repeats. This is also suggested by the sequences of the two eight amino acid inserts present in the C-terminal repetitive regions of both the cDNA and genomic sequences. These sequences are almost identical within each of the species but differ significantly from those within the other sequence (Fig. 5.15) which suggests that at least one of these inserts was present in the repetitive region prior to the separation of the sheep and primate species and that the changes to the insert sequence have occurred subsequently.

The spread of a mutation throughout the entire repetitive region may have occurred by the method proposed by Smith (1976) whereby homologous sequences within sister chromatids are able to undergo unequal cross-over events, i.e., the chromatids align out of phase and then undergo a cross-over of DNA strands (Fig. 5.28). Thereby a mutation within a particular repeat can, by a series of such cross-overs, move throughout an entire repetitive region. In the case of the changes to the trichohyalin repetitive sequences, the need to maintain the function of trichohyalin would ensure that only those mutations which do not alter the structure of the repeats would be able to move throughout the repetitive region. It would appear from the differences in the C-terminal consensus sequences (Fig. 5.26) that the region in which the mutations are most likely to occur is the variable stretch of amino acids in the consensus sequences, i.e., residues 12-18 of the genomic repeat A. The selection criteria for a mutation in this region could be the maintenance of the α-helical nature of the repeats and possibly also the preservation of the average length of each repeat. The total length of trichohyalin may also be important and thus the total number of repeats produced by a series of cross-over events may be limited to a given range such that the function of trichohyalin is maintained.

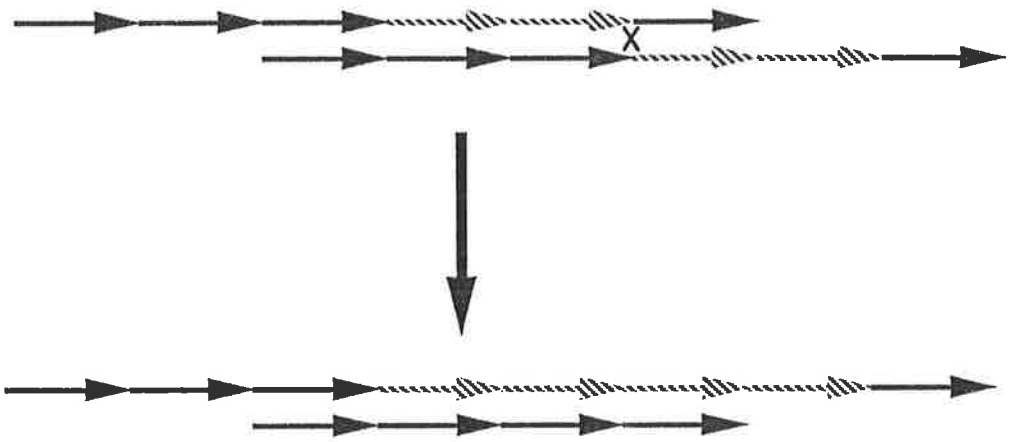
Figure 5.28 Proposed method for the spread of a mutated DNA repeat throughout a complete repetitive region (Smith, 1976).

On the out-of-phase alignment of the repetitive regions within two sister chromatids, a mutated DNA repeat (hatched arrow) can be duplicated on one of the chromatids by a single cross-over event (a). Subsequent cross-over events can enable further duplication of the mutated repeats (b). Note that the normal repeats could then be removed by further unequal cross-overs thus increasing the proportion of mutated repeats and maintaining the original length of the repetitive region.

a.



b.



Chapter 6

Expression Studies on Sheep Trichohyalin

6.1 Introduction

The relationship between the proteins of the trichohyalin granules of the follicle IRS and of the medulla has long been questioned (Rogers, 1962, 1963, 1983; Rogers and Harding, 1976a; Rogers *et al.*, 1977; Rothnagel and Rogers, 1986). They are immunologically, biochemically and histochemically similar (Rogers, 1963, 1964a; Rogers and Harding, 1976a; Rogers *et al.*, 1977; Rothnagel and Rogers, 1986) yet the IRS granules disappear to be replaced by filaments whereas in the medulla they form an amorphous structure. Are the IRS and medulla granular proteins identical or are they just chemically and antigenically similar? The demonstration that trichohyalin is a single copy gene, together with the strong immunoreactivity of the medulla granules to antibodies raised against IRS trichohyalin, suggests that the two proteins are identical. Nevertheless this can only be shown conclusively by determining that the trichohyalin gene is expressed in both tissues. This can be done using the technique of in situ hybridisation which involves the hybridisation of fixed tissue sections with radiolabelled complementary RNA probes and the detection of the hybridised probe with a photographic emulsion. Thus it is possible to determine which follicle tissues express trichohyalin as well as the region within those tissues which contain the trichohyalin mRNA.

Additionally, in situ hybridisation allowed the examination of trichohyalin expression in other epithelial tissues. Hair-like IF proteins, which were once thought to be expressed solely within the hair and other hard-keratinous tissues, have now been shown to also be present within regions of the soft tongue epithelium (Dhouailly *et al.*, 1989). It is therefore possible that trichohyalin also is expressed in the tongue and thus tongue sections, as well as sections containing a number of different epithelia, were probed with a trichohyalin-specific probe. Much of the work presented in this chapter has already been published (Fietz *et al.*, 1990; see Appendix).

6.2 Results

6.2.1 Preparation of cRNA Probes

The two main cRNA probes used for in situ hybridisation analysis were derived from the repetitive coding region and the 3' non-coding region of λsTr1. The repeat-

containing probe was derived from a 124 bp fragment of λsTr1 which extends from the EcoR I site in the 5' linker to the second Pst I site (position 119, see Fig. 4.3). This fragment, which spans approximately one and a half repeat units, was subcloned into pGEM-2 forming the clone pGEM-R (Fig. 6.1). The second probe, which is derived from the 3' non-coding region of λsTr1, was transcribed from pGEM-1.9 (Fig. 6.2). This probe corresponds to the region from the Sac I site (position 1756, see Fig. 4.3) to the EcoR I site at position 1941 (see Fig. 4.3).

A third probe was used in an attempt to determine the identity of the DNA in the genomic clone λsGT1b and was derived from the 3' non-coding region of the gene. The probe spanned the region from the Sac I site (position 6913, see Fig. 5.8) to the Xba I site (position 7117, see Fig. 5.8). It was transcribed from the clone pGXE5.2 (Fig. 6.3), which contains the 5.2 kb EcoR I/Xba I fragment (see Fig. 5.4) subcloned into pGEM-7Zf(+).

6.2.2 In Situ Hybridisation to Wool Follicle Sections

As the cDNA insert of λsTr1 was produced from RNA extracted from Merino follicles the expression of the trichohyalin gene was initially examined in Merino skin sections. As expected, hybridisation of the repeat-containing probe to the Merino follicles produced a strong signal within the IRS (Fig. 6.4). The signal extended from the basal IRS cells positioned near the bottom of the follicle bulb (Fig. 6.4a) through to the cells positioned immediately beneath the zone of hardening (Fig. 6.4b). Within the Henle layer the signal extended to just above the top of the follicle bulb, whereas in the Huxley layer and the IRS cuticle the signal extended much higher up the follicle. No signal was detected with the negative control, i.e., sections probed with the corresponding sense probe (Fig. 6.4c). Hybridisation with the 3' non-coding probe produced a pattern identical to that seen with the repeat-containing probe, although the signal was less intense because the target sequence was not repetitive (Fig. 6.4d).

As the Merino fibres are not medullated, Tukidale skin sections which do produce some medullated fibres were probed to determine whether the trichohyalin gene was expressed within the medulla. Hybridisation of the repeat-containing probe to the Tukidale follicle sections produced a strong signal over both the developing IRS and medulla cells (Fig. 6.5) indicating that the trichohyalin gene is indeed expressed in both tissues. The trichohyalin mRNA within the IRS of Tukidale follicles spans the same

Figure 6.1 Shown is the plasmid pGEM-R which contains a 124 bp fragment (Tr) from the repetitive region of λsTr1 (nucleotides 1-124) cloned into pGEM-2. The anti-sense probe was transcribed with T7 RNA polymerase after linearisation of pGEM-R with EcoR I. The sense probe (negative control) was transcribed with SP6 RNA polymerase after linearisation with Hind III. Both transcriptions were performed in the presence of [^{35}S]UTP.

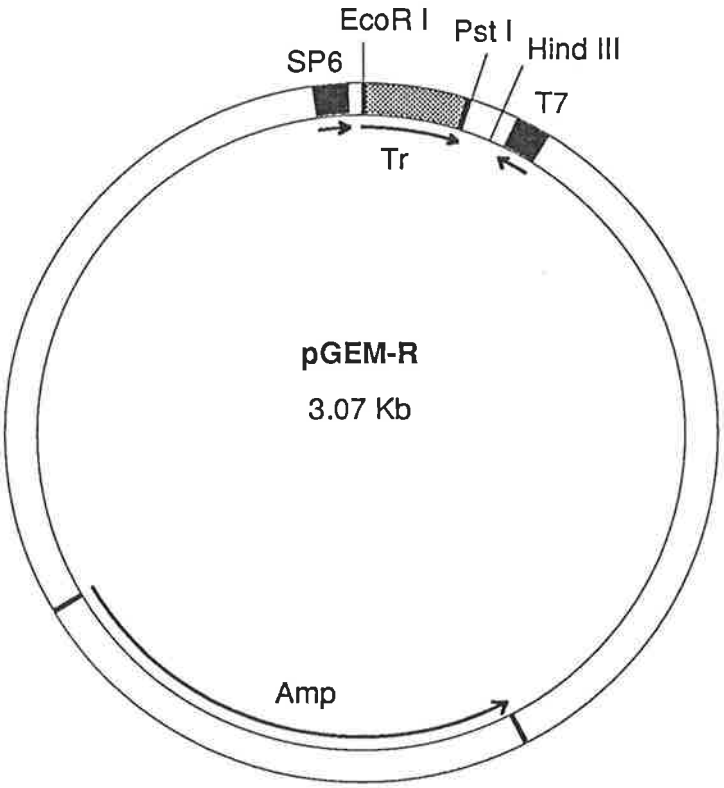

EcoR I
Pst I
Hind III
SP6
T7
Tr
pGEM-R
3.07 Kb
Amp

Figure 6.2 Shown is the plasmid pGEM-1.9 which was formed by the cloning of the 1.9 kb EcoR I fragment of λsTr1 into pGEM-2. An anti-sense probe, derived from the 3' non-coding region of λsTr1, was transcribed with SP6 RNA polymerase after the linearisation of pGEM-1.9 with Sac I. The transcription was performed in the presence of [^{35}S]UTP. Negative control tissue sections were treated with RNase prior to the hybridisation with the anti-sense probe.

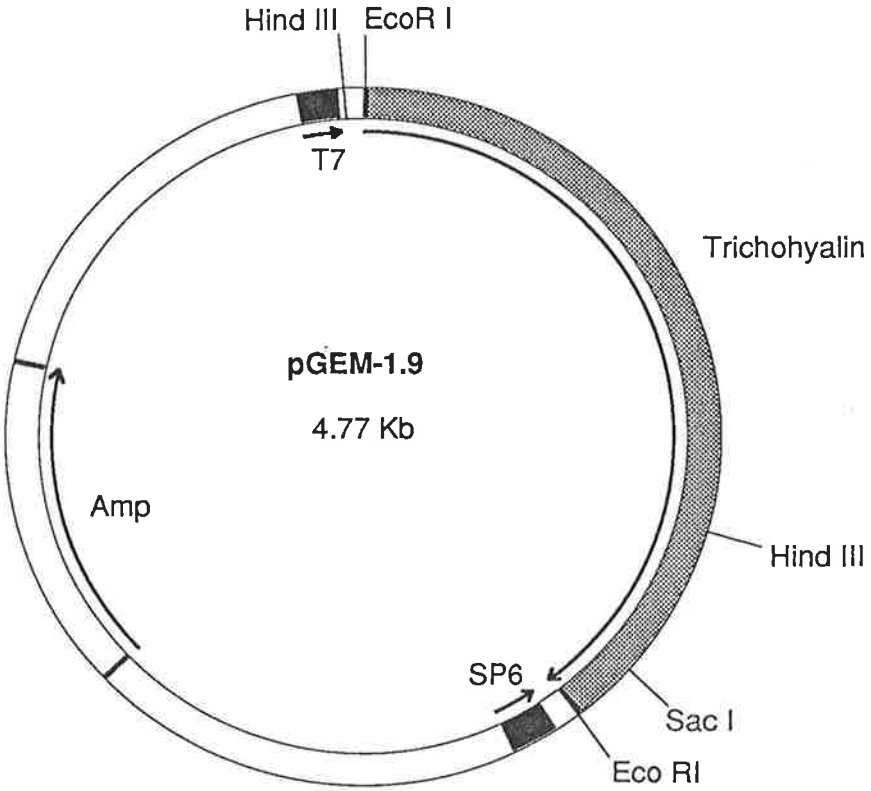

Hind III
EcoR I
T7
Trichohyalin
pGEM-1.9
4.77 Kb
Amp
Hind III
SP6
Sac I
Eco RI

Figure 6.3 The depicted plasmid, pGXE5.2, was used for the transcription of an RNA probe complementary to a portion of the 3' non-coding region of the trichohyalin gene. The plasmid was formed by the cloning of the 5.2 kb Xba I/EcoR I fragment of λsGT1b (XE5.2, Fig. 5.4) into pGEM-7Zf(+). The probe was transcribed by T7 RNA polymerase, in the presence of [^{35}S]UTP, after linearisation of pGXE5.2 with Sac I.

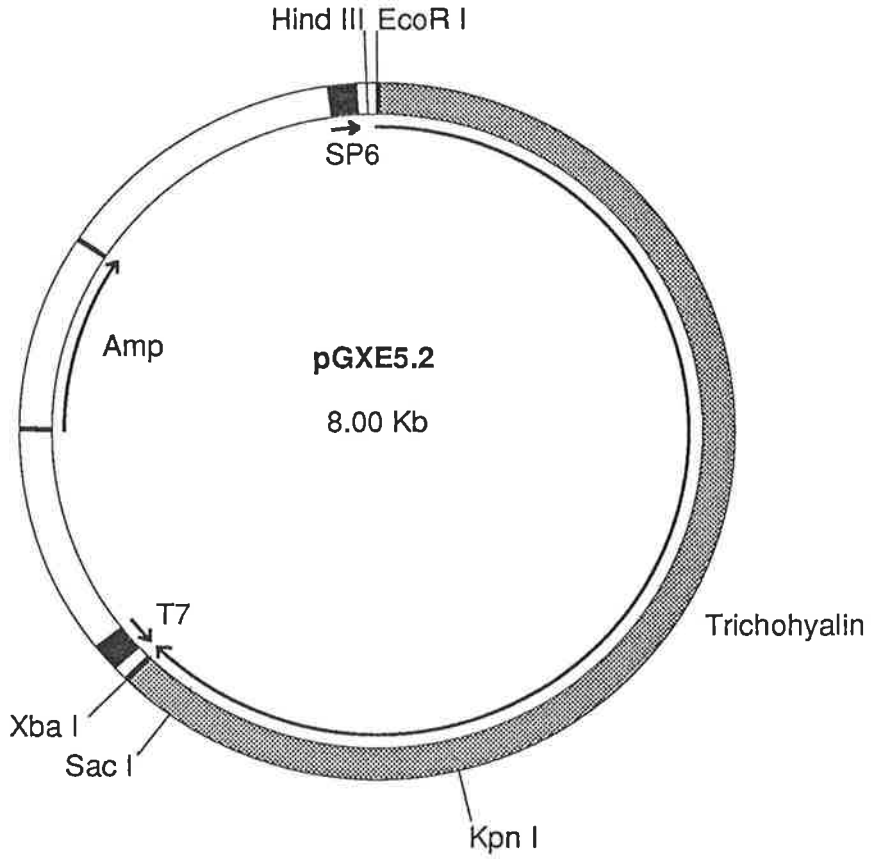

Hind III
EcoR I
SP6
Amp
pGXE5.2
8.00 Kb
T7
Trichohyalin
Xba I
Sac I
Kpn I

Figure 6.4 In situ hybridisation analysis of Merino follicle sections.

(a) Depicted is the lower portion of a Merino follicle which has been hybridised with the repeat-containing cRNA probe. The hybridisation signal (silver grains) appears to commence in the basal cells of the IRS which are positioned near the base of the bulb. The signal then extends up the layer of developing IRS cells. Note that the hybridisation to Henle's layer terminates approximately half way up the follicle section (arrow) and this correlates with the disappearance of the trichohyalin granules. There is no hybridisation to the fibre cortex (C).

(b) Hybridisation of the repeat-containing anti-sense probe to the shaft of two Merino follicles at the level of conversion of the cells within Huxley's layer from the granular to the hardened form. Note the sudden termination of the hybridisation signal which occurs immediately before the hardening of the cells (indicated by transition of the cells to the orange-stained form) and correlates with the disappearance of the trichohyalin granules. Note that on both sides of the two follicles the fibre cuticle and the IRS cuticle have separated during the sectioning procedure.

Continued.....

Figure 6.4 (Cont.)

(c) Hybridisation of a Merino follicle with the repeat-containing sense probe (negative control). There is no hybridisation signal above background.

Continued.....

C.

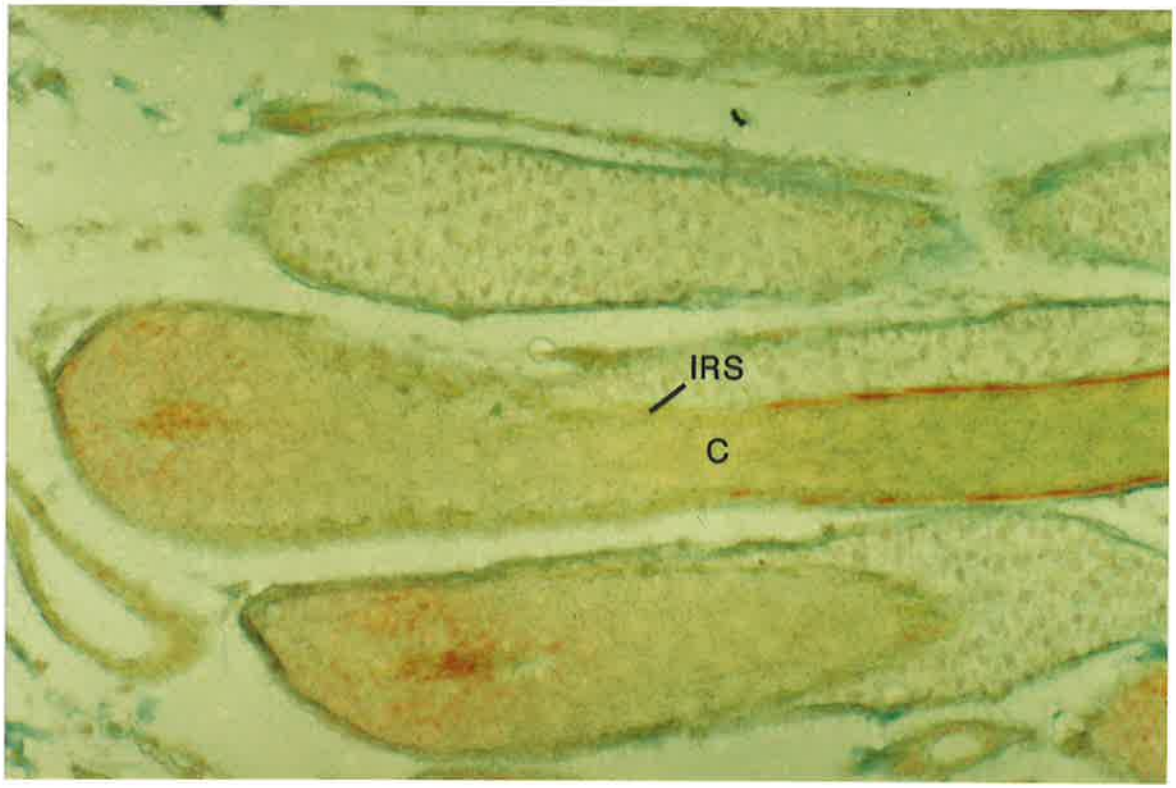


Figure 6.4 (Cont.)

(d) A longitudinal section of a Merino follicle was photographed under dark-field illumination to highlight the low signal strength obtained with the 3' non-coding cRNA probe. The hybridisation signal (white grains) covers the same region seen with the repeat-containing probe (a,b).

C, cortex; IRS, inner root sheath. All sections were stained using the SACPIC procedure of Auber (1950) and are at 200 x magnification.

d.

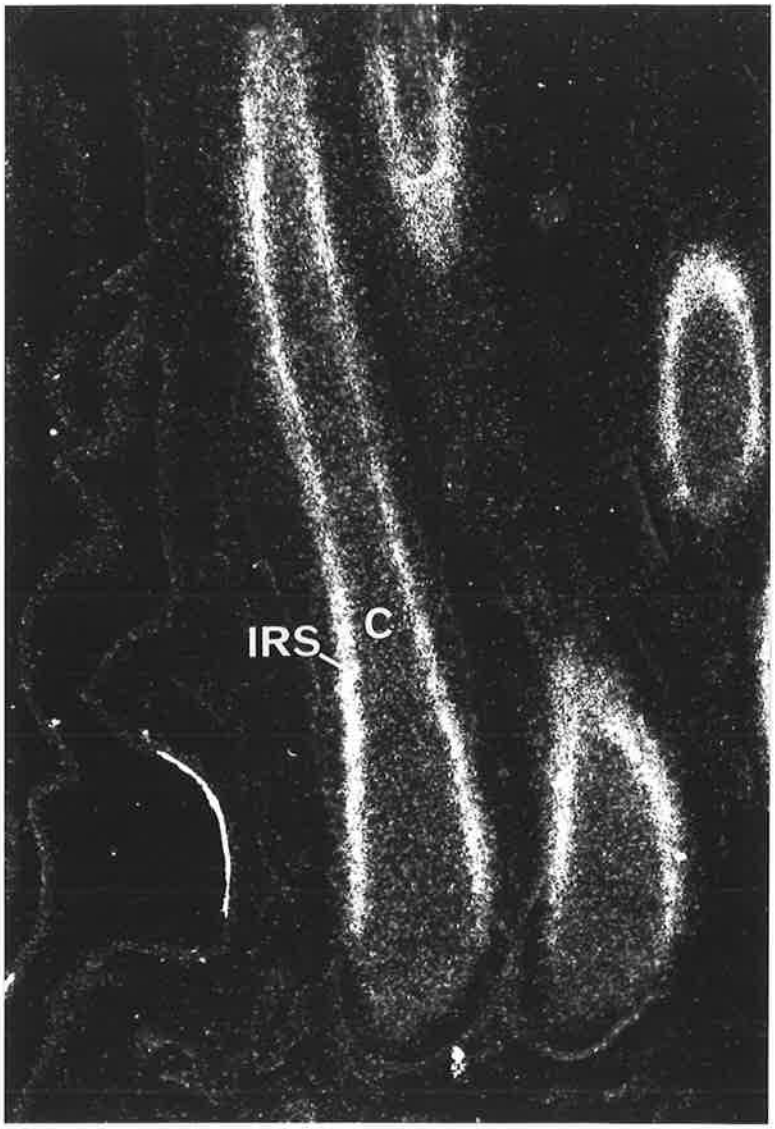


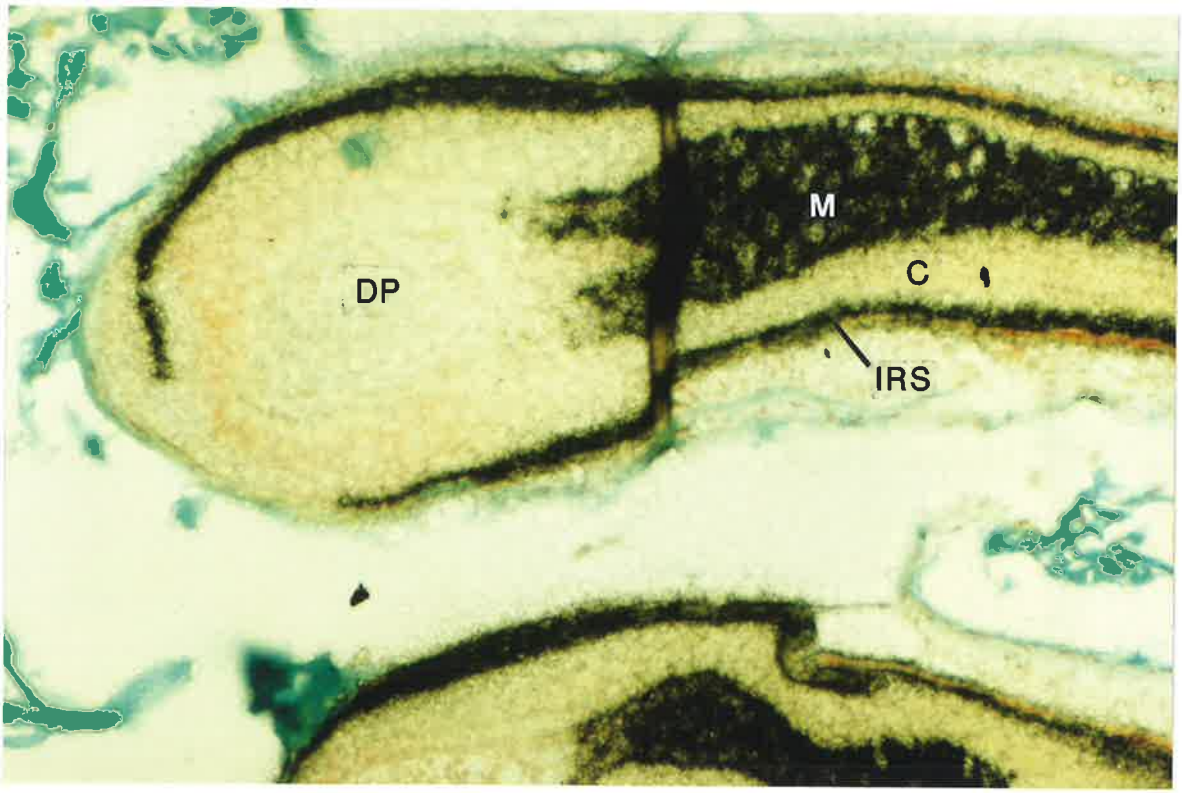
Figure 6.5 In situ hybridisation analysis of Tukidale follicle sections.

(a) Hybridisation of the repeat-containing anti-sense cRNA probe to the lower portion of a medullated Tukidale follicle. The hybridisation signal over the IRS corresponds to that seen with the hybridisation to the Merino follicles (Fig. 6.4a). The hybridisation to the medulla (M) begins with the cells lining the tip of the dermal papilla. The dark line present immediately above the dermal papilla is artefactual and caused by folding of the tissue during sectioning. 200 x

(b) Hybridisation of the repeat-containing cRNA probe to a Tukidale follicle at the level of conversion from the granular to the hardened cells. The hybridisation to the trichohyalin mRNA terminates at the same level as the disappearance of the trichohyalin granules in Huxley's layer of the IRS and the follicle medulla. Note that on both sides of the follicle the fibre cuticle and the IRS cuticle have separated during the sectioning procedure. 200 x

Continued.....

a.



b.

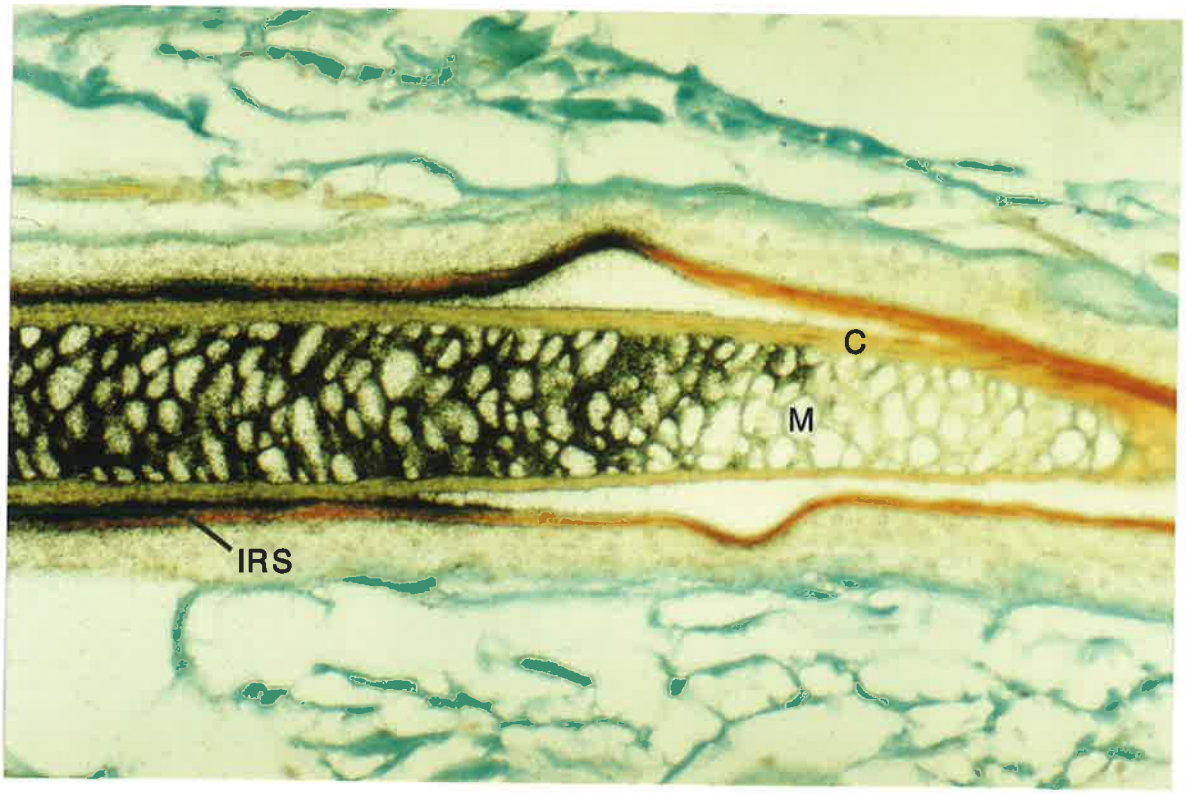


Figure 6.5 (Cont.)

(c) Hybridisation of a Tukidale follicle with the repeat-containing sense cRNA probe, i.e., negative control. There is no signal above background over either the IRS or medulla. Arrowed are trichohyalin granules present in the medulla and IRS cells. 200 x

Continued.....

c.

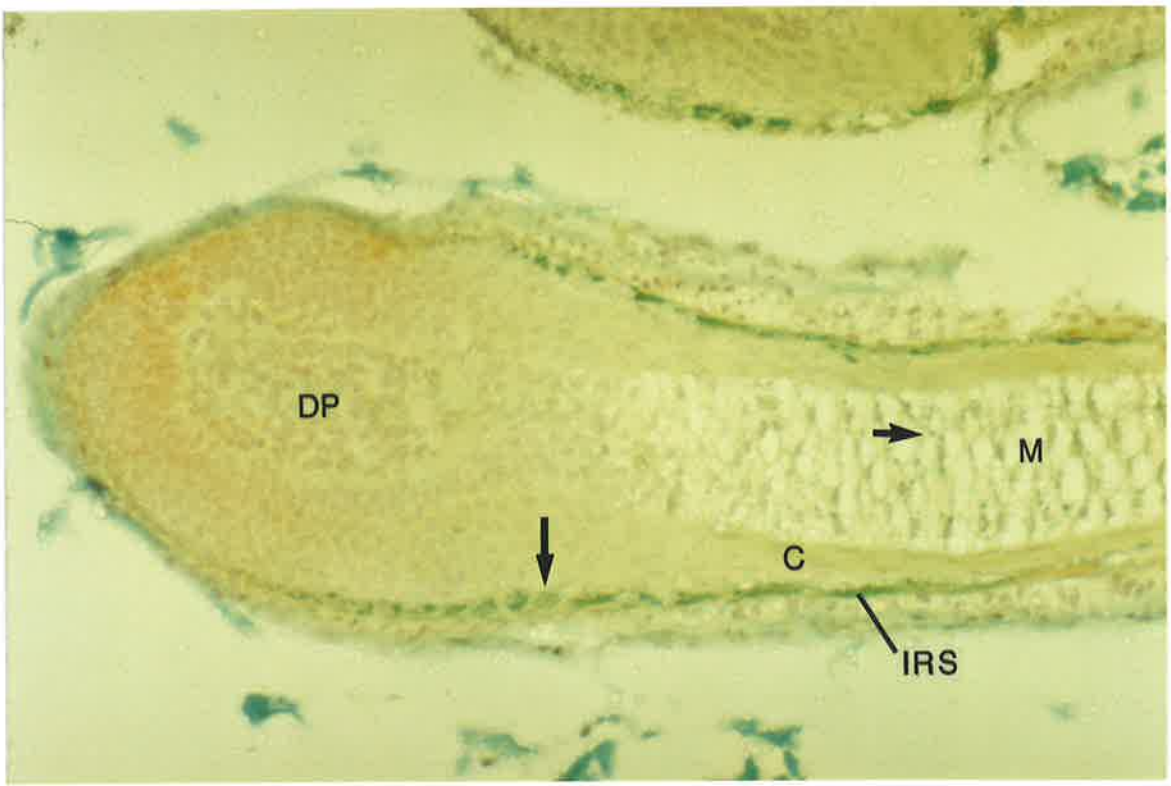
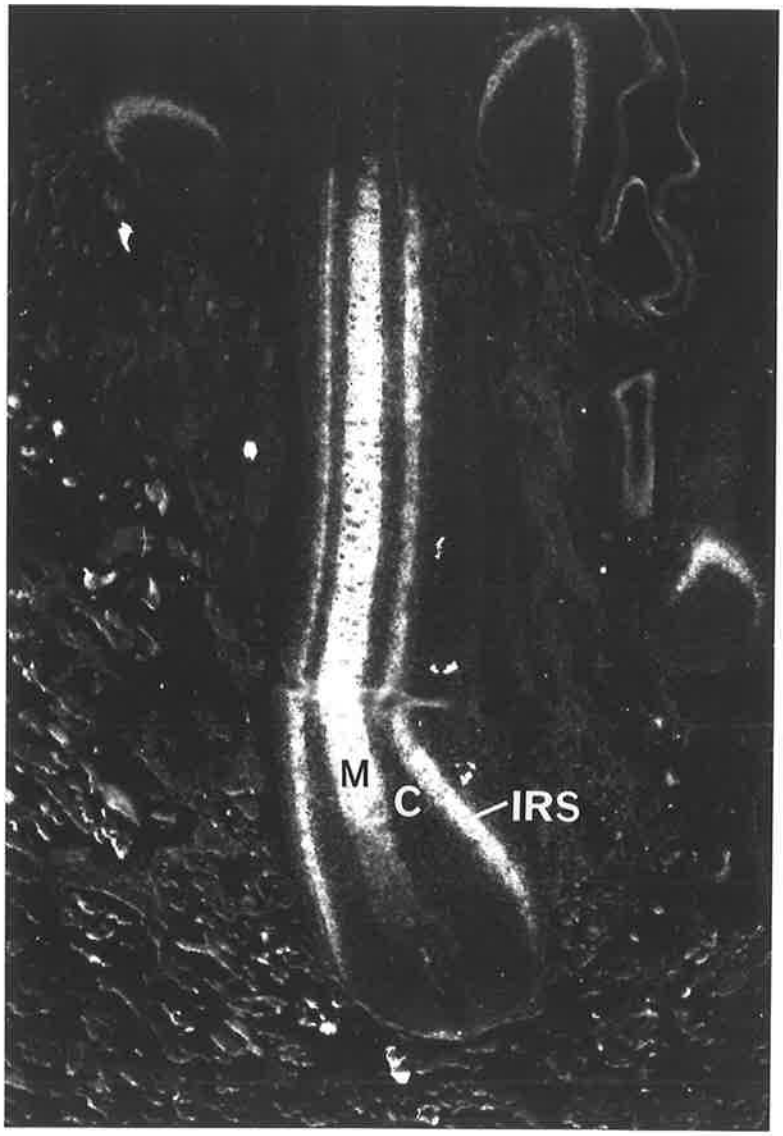


Figure 6.5 (Cont.)

(d) A longitudinal section of a Tukidale follicle was photographed under dark-field illumination after hybridisation with the 3' non-coding cRNA probe derived from the cDNA clone. The 3' non-coding probe hybridised to the same region as the repeat-containing probe (a,b). 100 x

C, cortex; DP, dermal papilla; IRS, inner root sheath; M, medulla.

d.



region as seen in the Merino follicles. In the medulla the trichohyalin gene also appears to be switched on very early in cell development, with the mRNA being detected in the basal cells which line the dome of the dermal papilla. The signal then extends from the basal cells up to the cells positioned immediately below the level of cell hardening. As observed for the Merino sections, the 3' non-coding probe hybridised to exactly the same region as the repeat-containing probe but produced a weaker signal (Fig. 6.5d).

6.2.3 In Situ Hybridisation to Other Keratinised Epithelia

To determine whether trichohyalin is expressed solely in the hair follicle or is also present in other epithelial tissues the repeat-containing cRNA probe was hybridised to sheep epidermis, tongue, oesophagus, rumen and hoof.

Hybridisation to the epidermis could be examined by viewing the previously probed follicle sections. No trichohyalin mRNA was detected within the epidermis (Fig. 6.6).

A positive signal was detected within the tongue epithelium (Fig. 6.7). The tongue epithelium has been divided into four regions on the basis of the expressed IF proteins (Fig. 6.8; Dhouailly *et al.*, 1989); the skin region, the hair region, the oesophageal region and the oesophageal-like region where non-oesophageal proteins are expressed in addition to the oesophageal IF proteins. It is within the developing cells of the oesophageal-like region that the trichohyalin mRNA is detected. The signal begins within the supra-basal cells and extends to the cells which have moved approximately two-thirds of the way to the epithelial surface (Fig. 6.7a). There is a very clear termination of hybridisation suggesting that the mRNA is either rapidly degraded or becomes inaccessible to the cRNA probe. Examination of stained tongue sections has shown that the region containing the trichohyalin mRNA also contains granules with similar staining characteristics to the trichohyalin granules of the follicle IRS and medulla (Fig. 6.7b). The disappearance of the granules correlates with the termination of hybridisation to the trichohyalin mRNA, as is observed within the follicle.

As the trichohyalin mRNA is present within the oesophageal-like portion of the tongue, trichohyalin expression was examined within the oesophagus and also the rumen which is the first segment of the sheep stomach. No trichohyalin mRNA was detected in the oesophagus (Fig. 6.9), although a weak signal was observed in the rumen with the hybridisation occurring just below the cornified layer of the epithelium (Fig. 6.10). The

Figure 6.6 In situ hybridisation analysis of sheep epidermis.

A Merino skin section was hybridised with the repeat-containing anti-sense probe. There is no hybridisation signal present over the epidermis. BL, basal lamina. 400 x

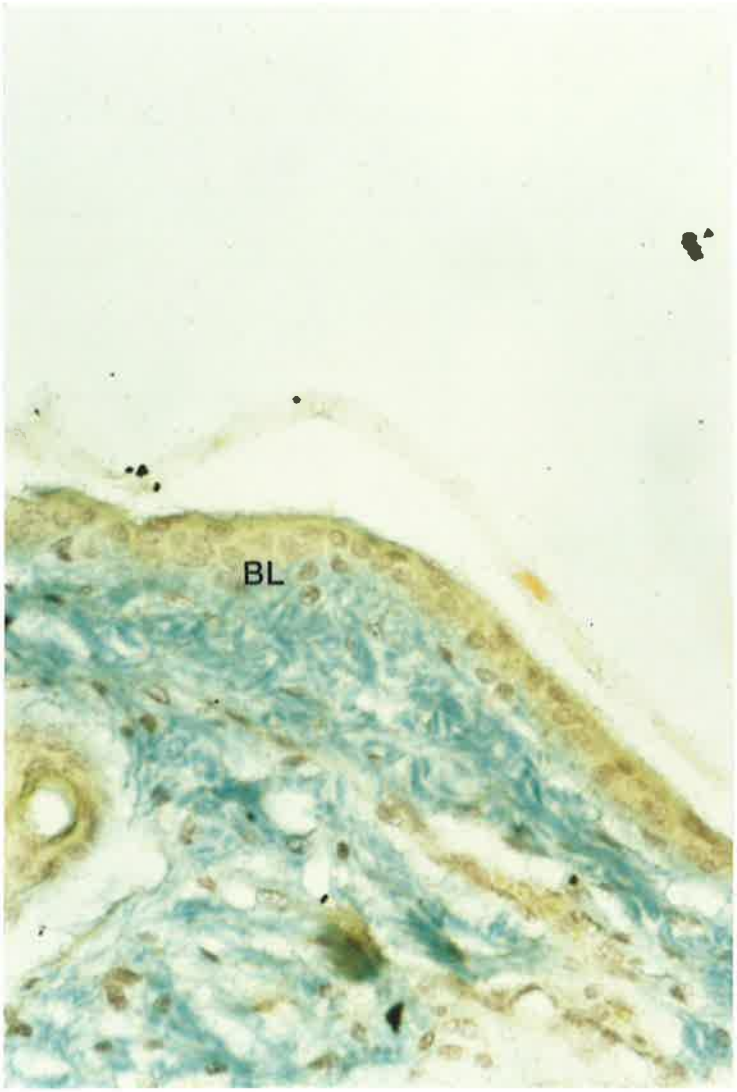
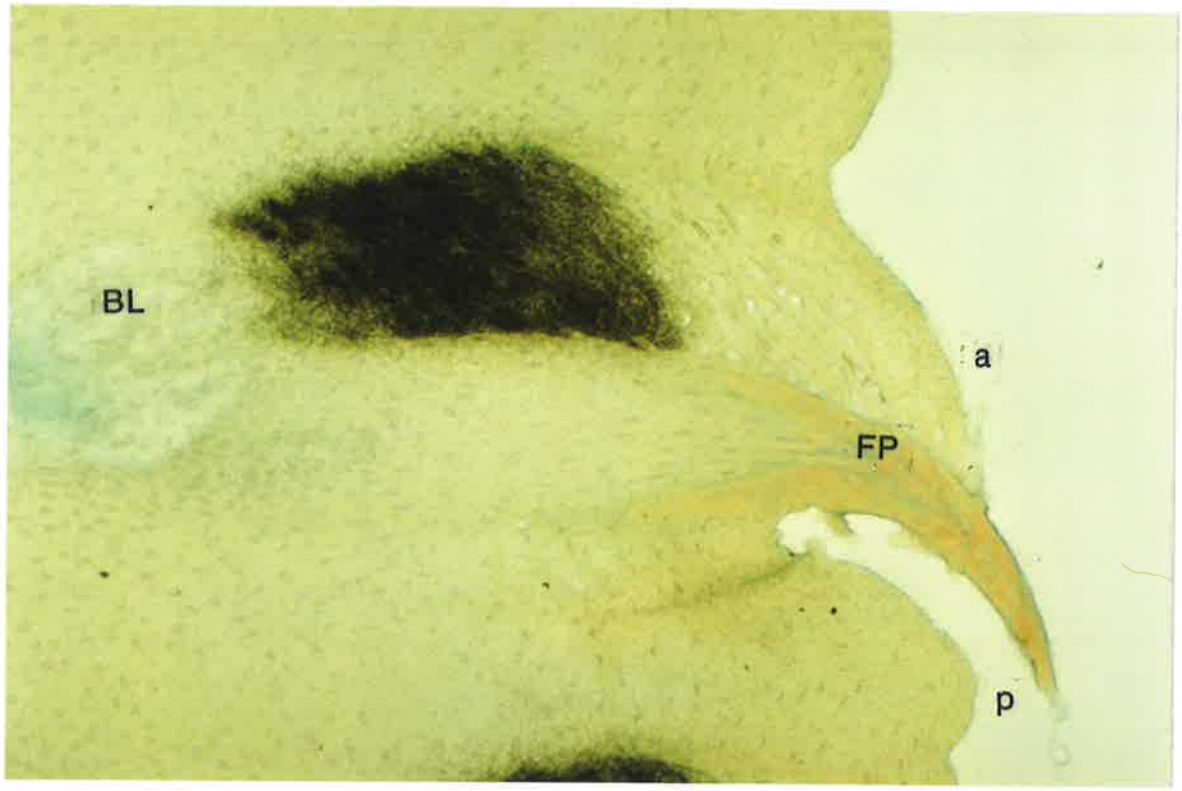

BL

Figure 6.7 In situ hybridisation analysis of sheep tongue.

(a) Hybridisation of a sheep tongue section with the repeat-containing anti-sense cRNA probe. The probe hybridises to a portion of the epithelium immediately anterior of the tongue filiform papilla (FP). Note that the signal commences in the supra-basal epithelial cells. BL, basal lamina; a, anterior; p, posterior. 200 x

(b) Photograph of a sheep tongue section hybridised with the repeat-containing sense cRNA probe. There is no signal present in the tongue epithelium. Note that the region which bound the anti-sense probe (a) contains numerous granules (arrowed) with similar staining characteristics to the trichohyalin granules of the follicle IRS and medulla. FP, filiform papilla. 400 x

a.



b.

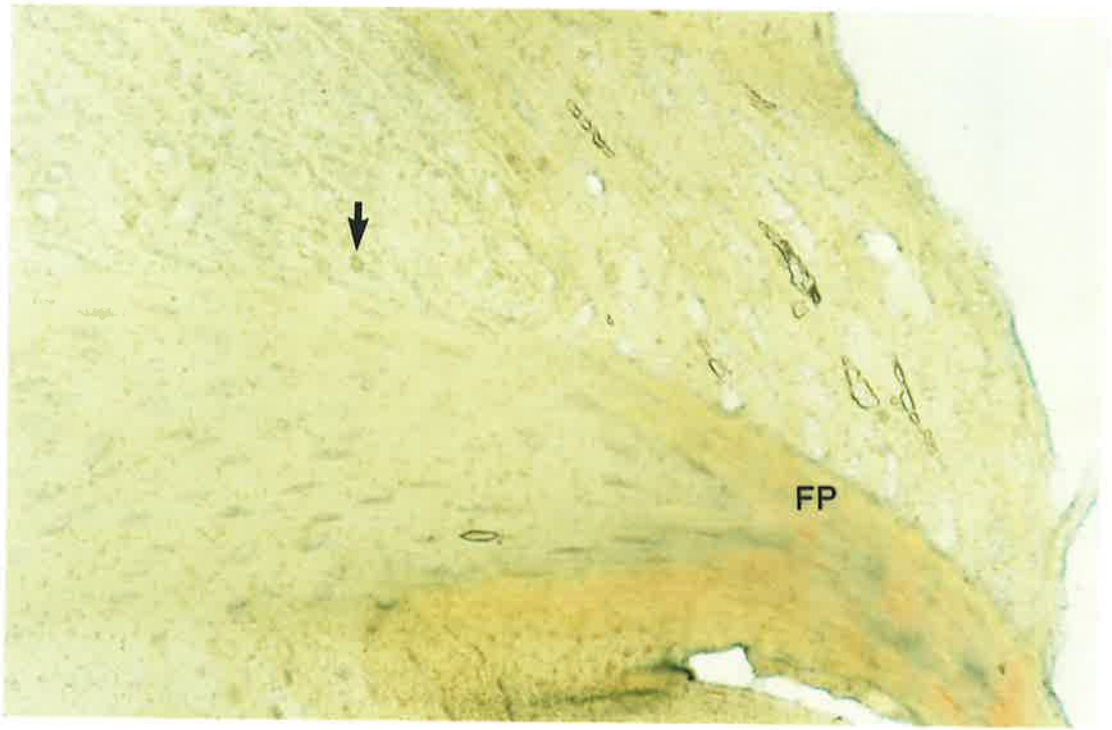


Figure 6.8 Schematic diagram of the compartmentalisation of mouse dorsal tongue epithelium.

The dorsal epithelium of the mouse tongue has been subdivided on the basis of the expressed IF proteins (Dhouailly *et al.*, 1989). Regions E, H, and S express IF proteins which are also expressed in the oesophagus, hair and skin respectively. Examination of the sheep tongue epithelium suggests that the arrangement of the various compartments is similar to that present in the mouse epithelium. The region containing the trichohyalin mRNA (Fig. 6.7a) appears to correspond with the E' region which has been shown to contain non-oesophageal proteins in addition to the oesophageal IF proteins. a, anterior; p, posterior.

(Reproduced from Dhouailly *et al.*, 1989.)

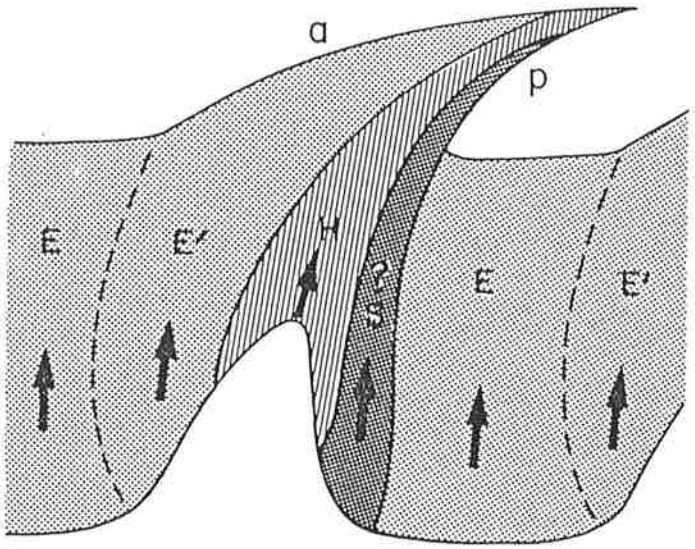

a
p
E
E'
H
S?
E
E'

Figure 6.9 In situ hybridisation analysis of sheep oesophagus.

Hybridisation of the repeat-containing anti-sense cRNA probe to a sheep oesophageal section. No hybridisation signal was detected over the oesophageal epidermis. BL, basal lamina. 200 x

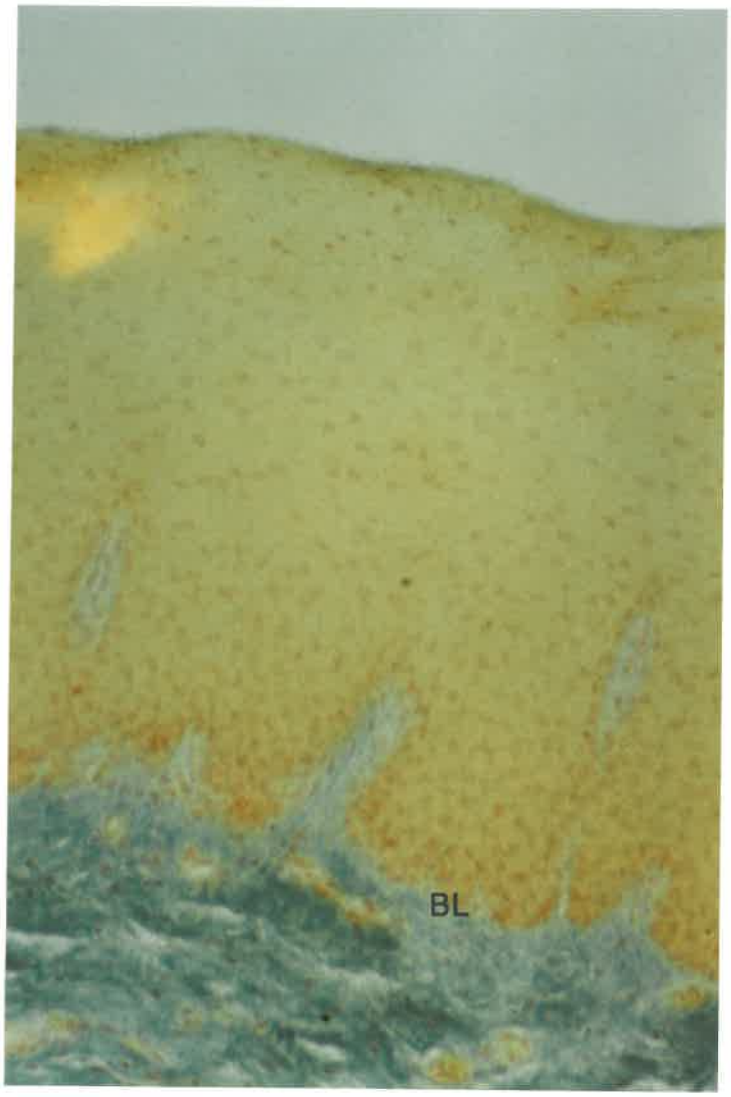
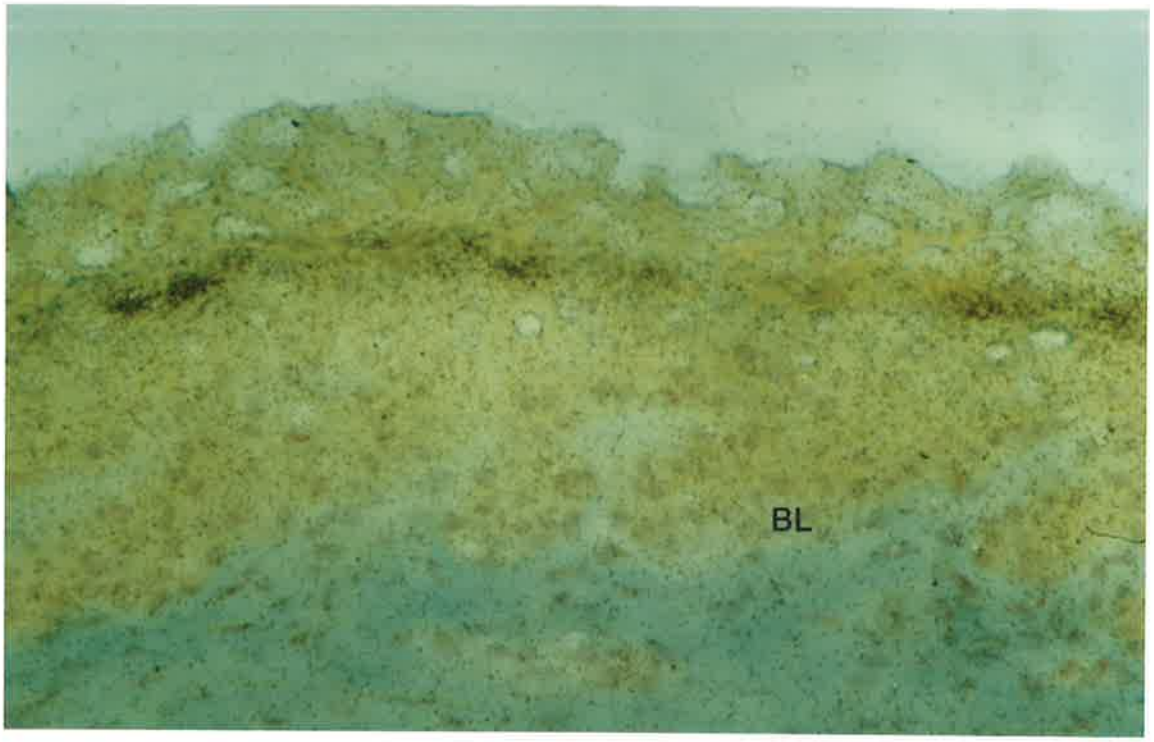

BL

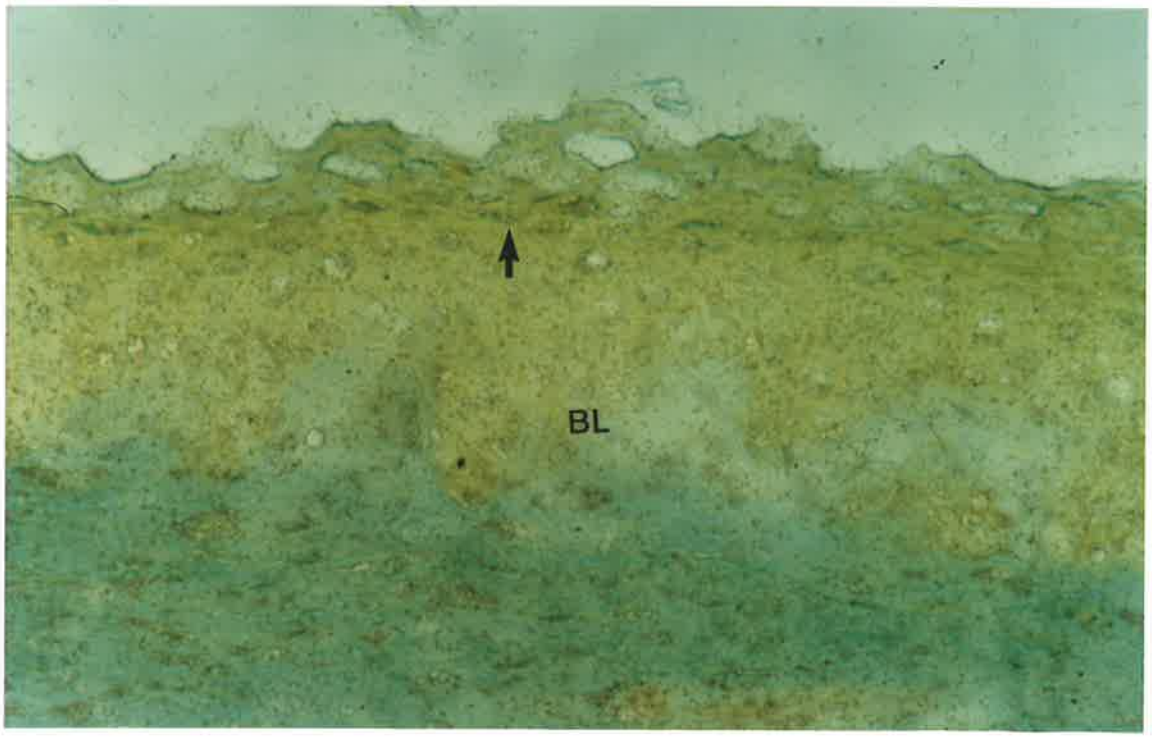
Figure 6.10 In situ hybridisation analysis of sheep rumen.

Hybridisation of a sheep ruminal section with the repeat-containing anti-sense probe (a) yielded a weak signal just below the stratum corneum of of the epithelium. No signal was detected with the corresponding sense probe (b). Note that a few small trichohyalin-like granules are present at the same stage of development as the trichohyalin mRNA (b, arrowed). BL, basal lamina. 400 x

a.



b.



signal within the rumen is patchy with regions of the ruminal epithelium containing little or no signal. On analysis of stained ruminal sections a small number of granules resembling those of the IRS and medulla were found at the same level as the hybridisation signal (Fig. 6.10b).

Trichohyalin mRNA was also detected within the epidermis of the sheep hoof (Fig. 6.11). The signal was mainly found in the ventral and anterior epithelia (Fig. 6.11a). It is notable that the expression is not detected within the basal epithelial cells but is detected soon after the cells have begun moving toward the hoof surface. The signal then terminates before the cells reach the outer surface of the hoof. Trichohyalin-like granules are also visible in histological sections of the hoof epidermis (Fig. 6.11c). These granules are not only visible in the region shown to contain the trichohyalin mRNA but remain present within the cells after the hybridisation signal disappears.

6.2.4 Inter-species In Situ Hybridisation Analysis

The similarities between the sheep, human and mouse trichohyalin repetitive regions have been examined on a gross scale by Zoo blot analysis (Section 5.2.5d). To extend this analysis the sheep cRNA probe containing the repeat sequence was hybridised to both human and mouse skin sections. In neither case was a significant signal detected above background in either the IRS or medulla (Fig. 6.12). Note that sheep skin sections were used as a positive control and the expected follicle hybridisation signal was detected (data not shown).

To investigate further the similarity of the genomic clone and human genomic sequences, a probe derived from the 3' non-coding region of the trichohyalin gene (see Section 6.2.1) was hybridised to human skin sections. No signal was detected above background (Fig. 6.13) suggesting that the genomic clone is not derived from human DNA.

6.3 Discussion

The related nature of the trichohyalin granules within the hair follicle IRS and medulla raises two important questions regarding the two tissues. Firstly, are the proteins within the IRS and medulla granules identical? Secondly, if they are identical what causes the two tissues to mature into quite distinct hardened forms? Although the purification and sequencing of the trichohyalin gene has not given a conclusive answer to

Figure 6.11 In situ hybridisation analysis of a foetal sheep hoof.

(a) Depicted is a montage showing hybridisation of the repeat-containing anti-sense cRNA probe to a foetal sheep hoof section. The hybridisation signal is present in the anterior (left) and ventral (bottom) regions of the developing hoof epithelium. The signal extends along the epithelium between the points marked by the two arrows. 20 x

Continued.....

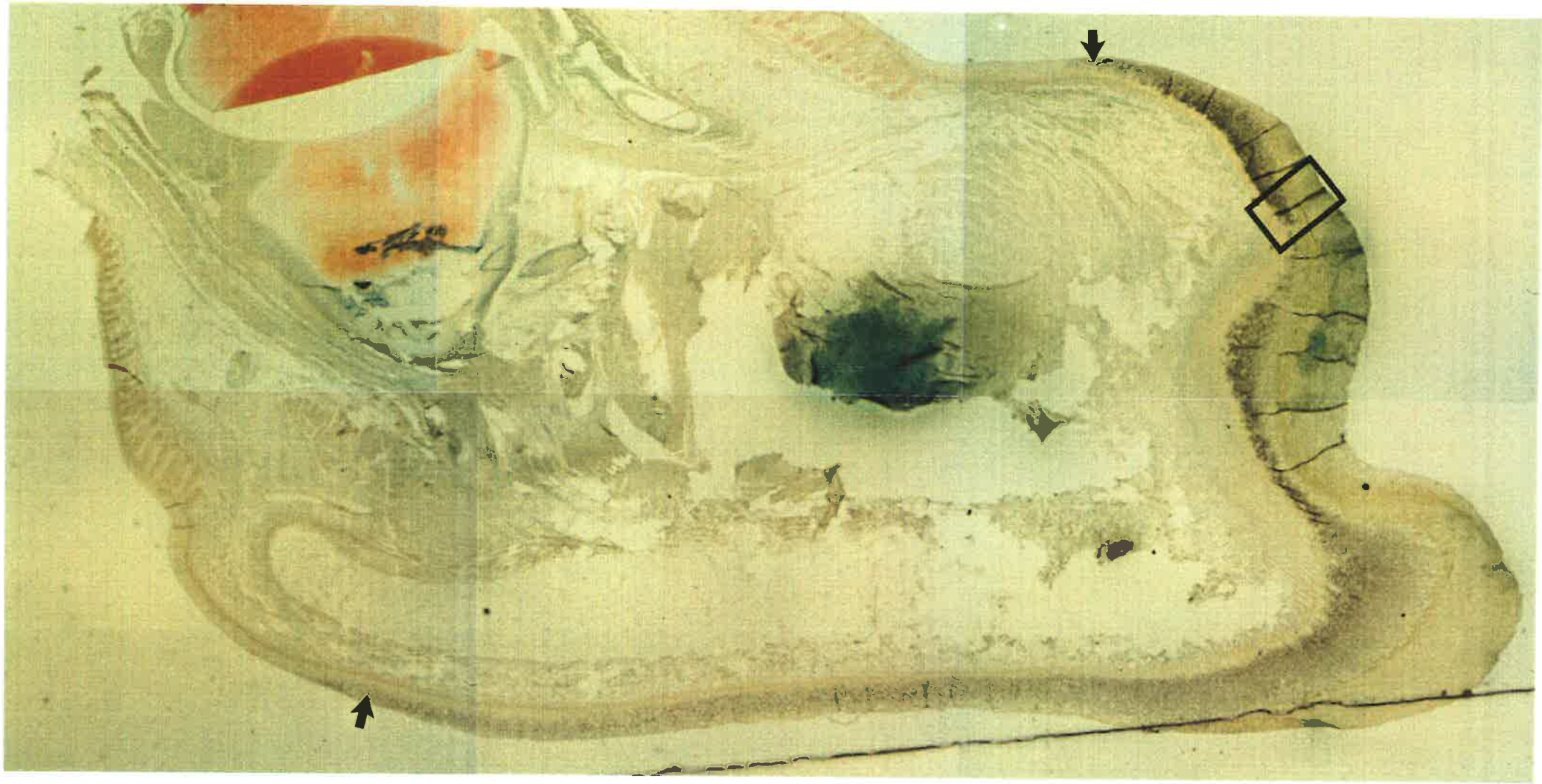


Figure 6.11 (Cont.)

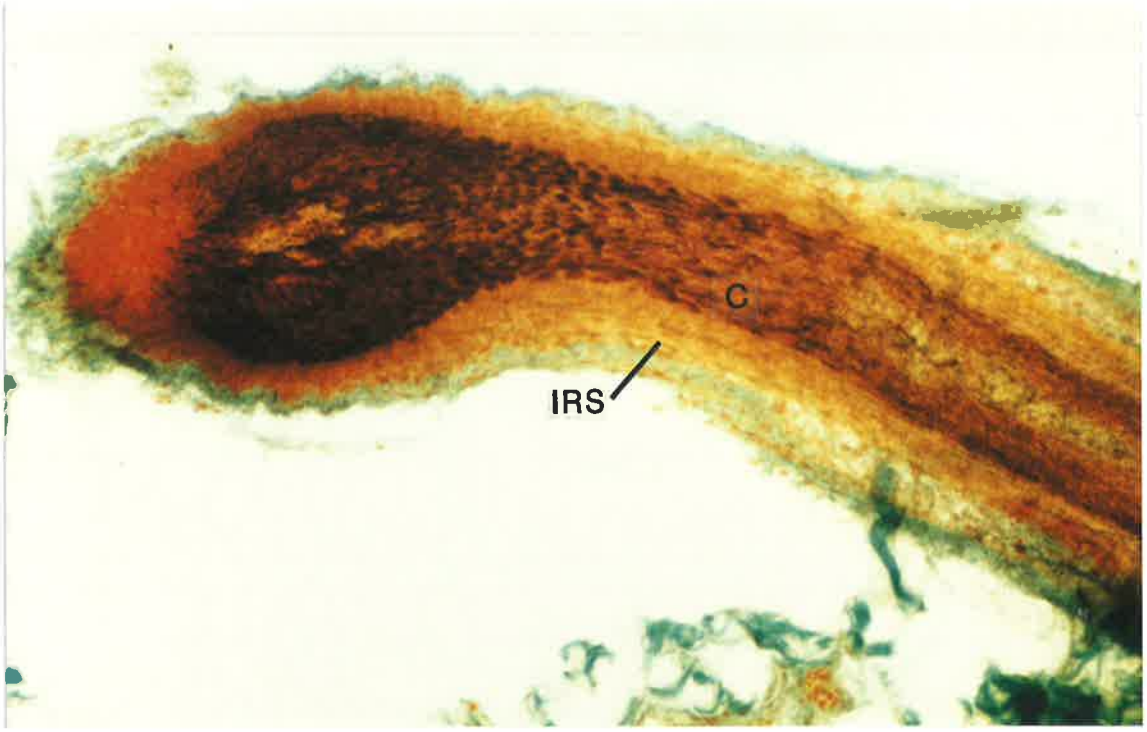
(b) Enlargement of the boxed region in (a). The hybridisation signal begins within the supra-basal cells of the hoof epithelium. Note that the presence of trichohyalin-like granules (arrowed) extends beyond the region of hybridisation. BL, basal lamina. 400 x

(c) Hybridisation of the corresponding sense probe to a foetal hoof section. There is no hybridisation signal present over the hoof epithelium. Trichohyalin-like granules (arrowed) are present in the region containing the trichohyalin mRNA (b). BL, basal lamina. 400 x

Figure 6.12 In situ hybridisation of human and mouse follicle sections with a sheep trichohyalin probe.

Depicted are longitudinal sections of human (a) and mouse (b) follicles which had been hybridised with the sheep repeat-containing anti-sense cRNA probe. No hybridisation signal is present over the IRS of either tissue. Note that the follicles of both species contain melanin within the cortical tissue. C, cortex; IRS, inner root sheath. 200 x

a.



b.

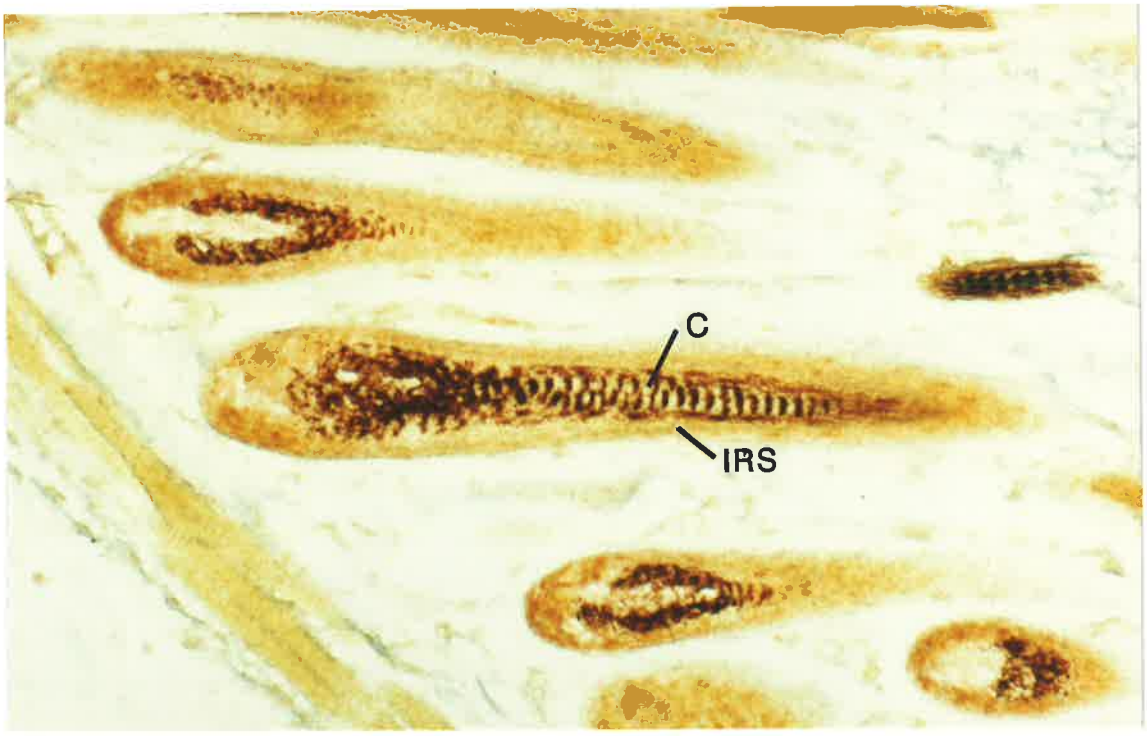
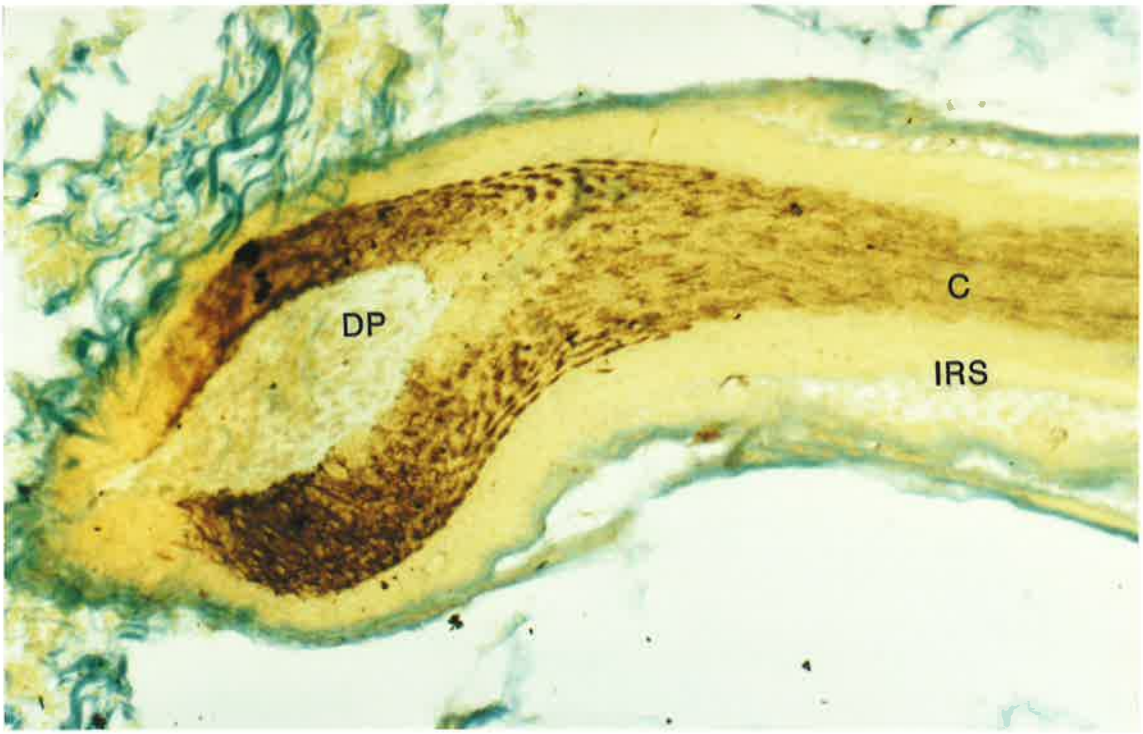


Figure 6.13 A longitudinal section of a human follicle was hybridised with the genomic 3' non-coding anti-sense cRNA probe. There is no hybridisation signal present over the IRS. Note that melanin is present within the cortical tissue. C, cortex; DP, dermal papilla; IRS, inner root sheath.


DP
C
IRS

the second question (Section 5.3.1d) it has demonstrated that the majority of the trichohyalin molecule within the IRS and medulla granules is identical. Hybridisation of Tukidale follicle sections with cRNA probes derived from both the trichohyalin repetitive region and the 3' non-coding region has shown the presence of trichohyalin mRNA in both the developing IRS and medulla (Figs. 6.5). Therefore trichohyalin is present in both tissues. It still remains possible however that differential splicing at the 5' end of the trichohyalin gene may produce differing proteins within the two tissues, e.g., there may be two separate promoters to the trichohyalin gene which may produce trichohyalin molecules with different N-termini. Examination of this will require either the isolation of the sheep trichohyalin gene or the determination of the origin of the DNA within the genomic clone.

The in situ hybridisation experiments have revealed that in both the medulla and the IRS, trichohyalin mRNA is expressed within the basal cells and remains present in the cells until just prior to cell hardening (Figs. 6.4, 6.5). This extended presence could be due to either the continuous expression of the trichohyalin gene or a prolonged half-life of the trichohyalin mRNA. The presence of trichohyalin mRNA during almost all stages of cell development emphasises the large amount of trichohyalin produced and thus its importance in growth and differentiation of the follicle. It is clear that within the IRS and medulla trichohyalin is a very early differentiation marker, produced much earlier than the IF proteins or the IF-associated proteins present in the cortical cells of the hair fibre (Kopan and Fuchs, 1989; MacKinnon *et al.*, 1990; Powell and Rogers, 1990b; Powell, B.C., manuscript in preparation) and apparently is expressed very soon after the initiation of differentiation. The factors involved in the initiation of differentiation are at present unknown and their expression may be dependent upon the relative positioning of the basal cells to the dermal papilla. Studies of the factors involved in the control of the trichohyalin gene may prove to be of great importance in understanding the initiation and control of differentiation in the hair follicle.

Analysis of the structure and sequence of the trichohyalin gene has not allowed the function of trichohyalin within the IRS to be absolutely determined (Section 5.3.1d). If trichohyalin acts as an IFAP then IF proteins must also be synthesised in the developing cells of the IRS. Immunological evidence suggests that IF proteins are synthesised within the IRS. Heid *et al.* (1988) showed that monoclonal antibodies to the hyperproliferative epidermal IF proteins K6 and K16 bind to the IRS in the human hair

follicle. Additionally, Ito *et al.* (1986a,b) and Lane *et al.* (1985) have shown that certain IF antibodies also bind to the IRS. Contrary to this, workers in our laboratory using the in situ hybridisation technique have been unable to show the presence of IF mRNA within the IRS. It was found that a gene-specific probe derived from a sheep cDNA clone encoding a K6-like protein and also a probe derived from the conserved α-helical region of a follicle IF gene did not hybridise to the IRS (L. Whitbread, B. Powell, personal communication). Thus, at present, it is not clear whether IF proteins are expressed in the IRS. This problem could be further examined by performing in situ hybridisation analysis using conserved epithelial IF probes or by attempting to examine the IF content of isolated IRS cells.

The examination of numerous keratinised epithelia by in situ hybridisation has shown that trichohyalin, like at least one hair-like IF protein (Dhouailly *et al.*, 1989), is not expressed exclusively within the hair follicle. Trichohyalin mRNA was detected in the epithelia of ovine tongue (Fig. 6.7), hoof (Fig. 6.11) and rumen (Fig. 6.10) but not in epidermis (Fig. 6.6) or oesophagus (Fig. 6.9). The hybridisation to the tongue and hoof agrees with the results of Manabe *et al.* (1989) who showed that antibodies to human trichohyalin bound to the human tongue and nail bed. Within both the tongue and hoof the hybridisation signal is strong suggesting that large amounts of trichohyalin are produced and that it is a major intracellular component in the mature structures. The hybridisation to the rumen epithelium is very weak (Fig. 6.10) indicating that only a small amount of trichohyalin is produced. Trichohyalin would therefore be expected to play only a minor role in the structure of the rumen epithelial cells. IF proteins are also known to be produced in each of the tissues expressing the trichohyalin mRNA, i.e., hair-like IF proteins are present in the hoof cells (Marshall and Gillespie, 1977; Baden and Kubilus, 1983), K6 and an unidentified type I IF protein are expressed in the ruminal epithelium (L. Whitbread, personal communication), and the epithelial IF proteins K4 and K13 are produced in the portion of the tongue which hybridises to the trichohyalin probe (Dhouailly *et al.*, 1989). It is possible that trichohyalin acts as an IFAP for the IF proteins present in each of these tissues.

To examine the similarities in trichohyalin development in the each of the tissues expressing trichohyalin, the hoof, rumen and the oesophageal-like portion of the tongue could be examined for the presence of ε-(γ-glutamyl)lysine cross-links and peptidylcitrulline. If trichohyalin plays a different role in the non-follicular tissues, it is

possible that one or both of the post-translational modifications occurring within the follicle may not be required. One apparent difference between the follicular and the non-follicular tissues is that the expression of trichohyalin within the non-follicular tissues only begins in the supra-basal epithelial cells (Figs. 6.7, 6.10, 6.11). Thus the trichohyalin gene appears to be switched on at a later stage in the differentiation of the hoof, rumen and tongue epithelial cells than is the case within the IRS and medulla.

Immunological studies performed on follicle proteins extracted from numerous mammalian epithelia have shown that the trichohyalin proteins present within these species are strongly immunocross-reactive (Rothnagel and Rogers, 1986). It was therefore surprising that a DNA probe derived from the repetitive region of the sheep cDNA clone, when washed at low stringency, hybridised with only moderate strength to the human gene and very weakly to the mouse gene (Section 5.2.5d). Furthermore, hybridisation with the cRNA probe, which was derived from the repetitive region of the sheep cDNA, gave no signal above background when the probe was hybridised to human and mouse follicle sections (Fig. 6.12). This lack of hybridisation may be due to nucleotide differences occurring between the C-terminal repeats of sheep trichohyalin and those of mouse and human trichohyalin. Although these differences do not stop hybridisation to the genomic blot (Fig. 4.10), the increased stringency of the in situ hybridisation procedure, which is caused by treatment with RNase prior to the washing of the hybridised sections, probably brings about the removal of any hybridised probe. The trichohyalin molecules from numerous mammalian species have been shown to be strongly immunocross-reactive (Rothnagel and Rogers, 1986) which indicates that the nucleotide differences occurring between the sheep, human and mouse sequences do not appear to effect the antigenic determinants of trichohyalin. Therefore the nucleotide differences must either occur at the third base of a codon, i.e., the encoded amino acid is not changed, or the differences must be limited to a portion of the repeat sequence such that the remainder produces the constant antigenic sites. Both of these possibilities are seen in the differences between the cDNA and genomic C-terminal repeats (Fig. 5.26).

The source of the DNA within the genomic trichohyalin clone λsGT1b, which was purified from what was purchased as a sheep genomic library, has been strongly questioned. A 3' non-coding probe, derived from λsGT1b, hybridised to human genomic DNA but not to sheep genomic DNA (Fig. 5.27). The species-specificity, shown by hybridisation with the sheep trichohyalin cRNA probes, indicated that the

identity of λsGt1b could be further examined by in situ hybridisation analysis. A 3' non-coding probe was transcribed from the trichohyalin gene and hybridised to human follicle sections. No signal was detected in either the IRS (Fig. 6.13) indicating that the purified gene is not of human origin. As the 3' non-coding probe from the genomic clone did hybridise to human DNA on the genomic blot (Fig. 5.27) it would therefore appear that the clone is derived from a related primate species. As a monkey genomic library was the only other primate library available from Clontech in λEMBL3 it would appear therefore that the purified trichohyalin gene is from monkey.

Chapter 7

General Discussion

The role of trichohyalin in the development of the IRS has long been enigmatic. It is synthesised during the development of the IRS cells and is stored as a precursor in non-membrane bound trichohyalin granules. These granules eventually disappear and the trichohyalin is incorporated into the organised array of parallel filaments (8-10 nm diameter) which fills the mature IRS cells. The trichohyalin is then present in an altered form; it is cross-linked by ε-(γ-glutamyl)lysine iso-peptide bonds and many of its arginine residues have been converted to citrulline. What function does the altered trichohyalin play within this filamentous arrangement? There are three possibilities diagrammatically outlined in Figure 1.13; trichohyalin may act as an IF protein which is cross-linked by separate IFAPs, or as an IFAP, cross-linking separately synthesised IF proteins, or as both an IF and an IFAP. Previous data on the role of trichohyalin have been equivocal. A comparison of the trichohyalin amino acid composition with that of the IRS filaments (Steinert, 1978) has shown that the two are similar (Rothnagel, 1985). Additionally, early ultrastructural studies seem to show that the IRS filaments stream out from the trichohyalin granules (Rogers, 1958b, 1964a,b). Together these data appear to suggest that trichohyalin acts as an IF protein. Contrary to this, immunological studies support the idea that trichohyalin is an IFAP (Rothnagel and Rogers, 1986) and that the filaments are formed by separate IF proteins. It was therefore the central aim of the work described in this thesis to further examine the role of trichohyalin within the hardened IRS and this was to be achieved by obtaining the complete trichohyalin amino acid sequence.

As the trichohyalin protein is too large to sequence directly, the trichohyalin gene was isolated and sequenced. Unfortunately, the experimental mapping of the 5' end of the gene was precluded by the sequence differences between the gene and the available Merino mRNA. These differences were found to be due to the DNA within the purchased library probably originating from monkey rather than sheep (Sections 5.2.5d, 6.3). Fortuitously, the majority of the trichohyalin gene was contained within a 4083 bp ORF and the predicted amino acid sequence at the N-terminus of the ORF was found to be highly homologous with sequences within the calcium-binding domains present in the S100-like calcium-binding proteins. This homology allowed the location of the initiation codon and the structure of the gene to be confidently predicted (Section 5.2.4a). The encoded trichohyalin protein is 1407 amino acids long, has a predicted molecular weight of 183,780 and is very hydrophilic in nature. The predominant feature within the

trichohyalin sequence is the C-terminal repetitive region which consists of tandem repeats of an approximately 23 amino acid sequence and spans 60% of the protein.

Unfortunately analysis of the deduced trichohyalin amino acid sequence has not conclusively determined whether or not trichohyalin can form the filaments of the IRS. Although trichohyalin does not contain any region with significant homology to the conserved α-helical core region of the IF proteins, so precluding it from forming a normal intermediate filament, the large C-terminal repetitive region of trichohyalin has the capacity to form a long α-helical rod (Section 5.3.1). A number of these rods might therefore be able to interact and form a stable α-helical aggregate thus producing a filament with the dimensions of the IFs. It should be emphasised that such α-helical aggregates could not be of the two-chain form found in IFs because trichohyalin does not contain the required heptad repeats (Section 1.5.2). If the C-terminal repetitive region forms the filaments of the IRS then it is possible that trichohyalin constitutes the complete filamentous structure of the IRS cells, with the remaining N-terminal region of trichohyalin acting as the required IFAP.

An alternative interpretation is that the α-helical rods, which have been proposed to be formed from the C-terminal repeats, may not have the capacity to aggregate and form a filamentous structure. Therefore the complete trichohyalin molecule would act as an IFAP providing a bridge spanning the gaps which occur between the filaments of the IRS. The synthesis of separate IF proteins in the IRS cells is suggested by a comparison of the amino acid contents of the deduced trichohyalin sequence and of filaments isolated from guinea pig IRS cells (Steinert, 1978) which shows that trichohyalin contains a significantly higher level of arginine and a lower level aspartic acid/asparagine than the filaments of the IRS (Table 7.1). In addition, the IRS filaments contain almost no citrulline or ε-(γ-glutamyl)lysine cross-links (Steinert, 1978). Further to this, recent immunological evidence has strongly suggested that derivatives of the K1 and K10 IF proteins are present within the IRS (Stark *et al.*, 1990) implying that these proteins form the filaments of the IRS. Trichohyalin would therefore act as an IFAP cross-linking the filaments formed from the K1-like and K10-like proteins. To examine the possible aggregation of K1 and K10 filaments with trichohyalin, <u>in vitro</u> analysis of the type used to determine the role of filaggrin (Dale *et al.*, 1978) could be performed using purified K1, K10 and trichohyalin. This analysis may also require the presence of follicle transglutaminase and/or peptidylarginine deiminase. One or both of these enzymes may

Table 7.1 Comparison of the amino acid content of the inner root sheath filaments (Steinert, 1978) with that of the complete trichohyalin protein (see Table 5.1).

SCM, S-carboxymethylated.

Amino Acid	Inner Root Sheath Filaments	Trichohyalin Gene Sequence
	mole percent	*mole percent*
SCM-cys	0.7	0.0
Asp/Asn	10.0	2.6/0.1
Thr	3.2	0.4
Ser	6.5	1.7
Glu/Gln	24.0	28.6/14.9
Pro	4.1	0.9
Gly	7.7	1.1
Ala	6.6	2.4
Cys	0.0	0.1
Val	4.8	0.7
Met	2.0	0.1
Ile	3.4	0.9
Leu	9.8	12.7
Tyr	1.9	0.9
Phe	1.7	1.9
Lys	4.6	3.8
His	1.4	0.4
Trp	0.2	0.4
Arg	3.8	25.2
Citrulline	3.9	0.0

be necessary for trichohyalin to either interact with the filaments or produce stable cross-links. Thus it may be necessary to purify, or at least partially purify, both of these enzymes prior to these experiments. If peptidylarginine deiminase is required for the aggregation of the filaments this analysis may also enable the role of citrulline in the IRS cells to be examined.

The presence of trichohyalin in its native granular form may also be required for the attempted aggregation of the K1/K10 filaments. Trichohyalin is currently extracted from follicle tissue using either of the strong chaotropic agents urea or guanidine hydrochloride. Present attempts have been unable to return the extracted trichohyalin to its native form (data not shown). The cloning of the trichohyalin gene behind an efficient promoter may allow trichohyalin to be highly expressed in transformed eukaryotic cells. If the synthesised trichohyalin forms granules of the type seen in the IRS and medulla cells these granules could then be purified using separation on the basis of factors such as density and sedimentation coefficient thus providing the pure granular form of trichohyalin.

The determination of the function of trichohyalin within the IRS could also be assisted using transgenic mice technology. Histological changes brought about in the IRS cells of adult mice by either the expression of the complete or truncated trichohyalin mRNAs or the ablation of the trichohyalin gene by homologous recombination in mouse embryonic stem cells may provide information regarding the interaction of trichohyalin with both itself and the IF proteins.

The role of trichohyalin in the follicle medulla and the cause for the production of different hardened forms in the IRS and medulla has also long been unclear. Although the production of IF proteins solely in the IRS would explain the difference in hardened structures, the work of Stark *et al.* (1990) appears to show similar amounts of K1- and K10-derivatives in both the IRS and medulla. If the trichohyalin in the IRS acts as an IFAP, and IF proteins are present in the medulla, what stops the medulla cells from forming the filaments seen within the IRS cells. One possibility, raised by electrophoretic analysis of partially degraded trichohyalin (Section 5.3.1d), is that the IRS contains a protease which is able to cleave trichohyalin into two separate domains. Two short "non-typical" sequences are present immediately before the C-terminal repetitive region and one or both of these could be cleaved by such a protease. Further analysis of this proposal requires a search for an IRS-specific protease and, if found, an examination of its

cleavage of trichohyalin. Another possible cause for the structural differences in the IRS and medulla cells could be an altered timing of expression of the follicle transglutaminase and peptidylarginine deiminase. In the IRS the deimination of trichohyalin may occur first, allowing the filaments to be formed prior to cross-linking by transglutaminase whereas in the medulla, the initial expression may be of transglutaminase which could promote cross-linking prior to the organisation of the filaments. This could be examined using the in vitro aggregation analysis mentioned above, i.e., the timing of addition of transglutaminase and peptidylarginine deiminase could be altered and the resultant structures examined and compared. A further possibility could be that alternative splicing occurs at the 5' end of the trichohyalin gene such that the trichohyalin molecules within the medulla and the IRS have different N-termini. Once the sheep trichohyalin gene has been purified or the identity of λsGT1b determined, this possibility could be examined by in situ hybridisation analysis of medullated follicles using probes derived from the 5' end of the trichohyalin gene.

Although it is not known whether trichohyalin acts as an IF protein or an IFAP it has always been assumed that its role is purely structural. This assumption can now be questioned by the discovery of two EF hand calcium-binding domains in the trichohyalin sequence. The EF hand calcium-binding proteins are believed to act as modulating proteins, i.e., the binding of calcium alters the structure of the binding protein and this change is transmitted to a bound protein thus altering its activity. An example of this is the activation of phosphodiesterase by the binding of calcium to calmodulin (Kakiuchi and Yamazaki, 1970). The presence of the EF hands would therefore suggest that a separate protein is bound to trichohyalin and that the structures of both trichohyalin and the bound protein are altered by the binding of calcium to the EF hands. As the only apparent changes to trichohyalin during the development of the IRS and medulla cells occurs at the point of conversion from the granular to the hardened forms, it is likely that calcium-binding may be important in the signalling required for this conversion. The signalling could occur by either the binding or the release of calcium ions from trichohyalin. One possibility is that an influx of calcium ions into the IRS and medulla cells and the subsequent binding of calcium to the EF hands is the signal involved in the breakdown of the trichohyalin granules. This raises the interesting question, "what direct effect could be brought about by the binding of calcium to trichohyalin?". Another possibility is that at least one of the enzymes transglutaminase or peptidylarginine

deiminase, both of which require calcium for their activity, is bound in an inactive state to the granular form of trichohyalin. On the binding of calcium to trichohyalin the bound enzyme could be either activated or released. Trichohyalin would therefore be activating an enzyme which then uses trichohyalin as its major substrate. This binding and control may be important in that a high concentration of the inactive enzyme could be located in the vicinity of trichohyalin such that upon the introduction of the appropriate signal the enzyme could be activated and trichohyalin rapidly altered and immediately incorporated into the mature cellular structure. In order to examine this possibility it will be important to purify both enzymes. The purification of the enzymes will allow the raising of antibodies to the enzymes and the subsequent examination of their location in the IRS and medulla cells.

It should be noted that the importance of calcium is also shown by the presence of parathyroid hormone related protein (PTHrP) in the developing IRS cells (Hayman *et al.*, 1989). The actions of PTHrP are very similar to those of parathyroid hormone (PTH) acting via the PTH receptor to produce the resorption of calcium from bone, increased uptake of calcium from the intestine and decreased excretion of calcium from the kidneys. Although the role of PTHrP in the IRS is unknown it is highly probable that it acts as a paracrine agent, i.e., after its release it binds to a receptor on the adjacent IRS cells producing an alteration in the intracellular calcium levels. At present the timing of expression of PTHrP within the IRS is unknown and it is impossible to correlate its presence and the breakdown of the trichohyalin granules or to determine the effect of PTHrP on the intracellular concentration of calcium. It would be interesting to compare PTHrP expression in all of the tissues where trichohyalin expression has been demonstrated, to see if there is a consistent relationship which may provide clues to their functional roles.

The use of <u>in situ</u> hybridisation analysis has shown the presence of trichohyalin mRNA in the epithelia of sheep tongue, hoof and rumen (Section 6.2.3) which correlates with the presence of trichohyalin granules (Section 6.2.3) and in part with the immunological detection of trichohyalin in human tongue and nail bed (Manabe *et al.*, 1989). The role of trichohyalin in these tissues is unknown although the mature nail epithelial cells contain a cellular envelope which differs in amino acid content from the epidermal cellular envelope (Baden and Fewkes, 1983) and is believed to contain ε-(γ-glutamyl)lysine cross-links (Baden and Fewkes, 1983; Shono and Toda, 1983). This

suggests that trichohyalin may be involved in the formation of the hoof cellular envelope. As IF proteins are expressed in the oesophageal-like cells of the tongue epithelium and the ruminal epithelium, namely K4 and K13 in the tongue (Dhouailly *et al.*, 1989) and K6 and an unidentified type I protein in the rumen (L. Whitbread, personal communication), it is possible that trichohyalin is acting as an IFAP within these two tissues. If trichohyalin is an IFAP it may therefore also have the capacity to cross-link a number of different pairs of IF proteins. Further examination of the tongue, hoof and rumen for the presence of citrulline residues and ε-(γ-glutamyl)lysine cross-links, together with detailed ultrastructural examinations of the respective tissues in both normal and transgenic mice, may give some indication of the role of trichohyalin in each of these tissues. Such an examination may give further clues which may eventually help to determine both the function of trichohyalin and its role in the development of the hair follicle IRS and medulla.

Bibliography

Aebi, U., Cohn, J., Buhle, L., and Gerace, L. (1986). *Nature (Lond.).* **323**:560-564.

Aebi, U., Engel, A., and Eichner, R. (1985). *J. Ultrastruct. Res.* **90**:323-335.

Auber, L. (1950). *Trans. Roy. Soc. Edinburgh.* **62**:191-254.

Baden, H.P., and Fewkes, J. (1983). *In* Biochemistry and Physiology of the Skin. Vol. 1. Goldsmith, L.A., ed. Oxford University Press, New York. pp. 553-566.

Baden, H.P., and Kubilus, J. (1983). *J. Invest. Dermatol.* **81**:220-224.

Benton, W.D., and Davies, R.W. (1977). *Science (Wash. DC).* **196**:180-182.

Birbeck, M.S.C., and Mercer, E.H. (1957). *J. Biophys. Biochem. Cytol.* **3**:223-230.

Birnboim, H.C., and Doly, J. (1979). *Nucleic Acids Res.* **7**:1513-1523.

Blumenthal, K.M., Moon, K., and Smith, E.L. (1975). *J. Biol. Chem.* **250**:3644-3654.

Borck, K., Beggs, J.D., Brammar, W.J., Hopkins, A.S., and Murray, N.E. (1976). *Mol. Gen. Genet.* **146**:199-208.

Brown, R.K. (1967). *Methods Enzymol.* **11**:917-927.

Bullock, W.O., Fernandez, J.M., and Short, J.M. (1987). *BioTechniques* **5**:376.

Burgess, A.W., Ponnuswamy, P.K., and Scheraga, H.A. (1974). *Israel J. Chem.* **12**:239-286.

Cabral, F., Gottesman, M.M., Zimmerman, S.B., and Steinert, P.M. (1981). *J. Biol. Chem.* **256**:1428-1431.

Chen, R., and Doolittle, R.F. (1971). *Biochemistry.* **10**:4486-4491.

Chomczynski, P., and Sacchi, N. (1987). *Anal. Biochem.* **162**:156-159.

Chou, P.Y., and Fasman, G.D. (1974). *Biochemistry.* **13**:222-245.

Chung, S.I., and Folk, J.E. (1972). *Proc. Natl. Acad. Sci. USA.* **69**:303-307.

Ciment, G., Resler, A., Letourneau, P.C., and Weston, J.A. (1986). *J. Cell Biol.* **102**:246-251.

Cohen, J. (1961). *J. Embryol. Exp. Morphol.* **9**:117-127.

Conway, J.F., and Parry, D.A.D. (1988). *Int. J. Biol. Macromol.* **10**:79-98.

Cotsarelis, G., Sun, T.-T., and Lavker, R.M. (1990). *Cell.* **61**:1329-1337.

Cox, K.H., DeLeon, D.V., Angerer, L.M., and Angerer, R.C. (1984). *Dev. Biol.* **101**:485-502.

Crewther, W.G. (1976). *In* Proc. 5th Int. Wool Text. Res. Conf., Aachen, 1975. Vol. I. pp. 1-103.

Crewther, W.G., Dowling, L.M., Steinert, P.M., and Parry, D.A.D. (1983). *Int. J. Biol. Macromol.* **5**:267-274.

Crick, F.H.C. (1953). *Acta Cryst.* **6**:689-697.

Curatola, A.M., and Basilico, C. (1990). *Mol. Cell. Biol.* **10**:2475-2484.

Dale, B.A., Holbrook, K.A., and Steinert, P.M. (1978). *Nature (Lond.).* **276**:729-731.

Dale, B.A., Resing, K.A., Haydock, P.V., Fleckman, P., Fisher, C., and Holbrook, K.A. (1989). *In* The Biology of Wool and Hair. Rogers, G.E., Reis, P.J., Ward, K.A., and Marshall, R.C., eds. Chapman and Hall Ltd., London. pp. 97-115.

Darwish, H.M., Krisinger, J., Strom, M., and DeLuca, H.F. (1987). *Proc. Natl. Acad. Sci. USA.* **84**:6108-6111.

Desplan, C., Heidmann, O., Lillie, J.W., Auffray, C., and Thomasset, M. (1983). *J. Biol. Chem.* **258**:13502-13505.

Devereux, J., Haeberli, P., and Smithies, O. (1984). *Nucleic Acids Res.* **12**:387-395.

Dhouailly, D., Xu, C., Manabe, M., Schermer, A., and Sun, T.-T. (1989). *Exp. Cell Res.* **181**:141-158.

Dufton, M.J., and Hider, R.C. (1977). *J. Mol. Biol.* **115**:177-193.

Dunn, R., Landry, C., O'Hanlon, D., Dunn, J., Allore, R., Brown, I., and Marks, A. (1987). *J. Biol. Chem.* **262**:3562-3566.

Ebling, F.J.G. (1988). *Clinics Dermatol.* **6(4)**:67-73.

Eckert, R.L., and Green, H. (1986). *Cell.* **46**:583-589.

Eichner, R., Rew, P., Engel, A., and Aebi, U. (1985). *Ann. N. Y. Acad. Sci.* **455**:381-402.

Eliopoulos, E., Geddes, A.J., Brett, M., Pappin, D.J.C., and Findlay, J.B.C. (1982). *Int. J. Biol. Macromol.* **4**:263-268.

Elöd, E., and Zahn, H. (1944). *Melliand Textilber.* **25**:361.

Emori, Y., Ohno, S., Tobita, M., and Suzuki, K., (1986). *FEBS Lett.* **194**:249-252.

Feinberg, A.D., and Vogelstein, B. (1983). *Anal. Biochem.* **132**:6-13.

Fietz, M.J., Presland, R.B., and Rogers, G.E. (1990). *J. Cell Biol.* **110**:427-436.

Folk, J.E. (1980). *Ann. Rev. Biochem.* **49**:517-531.

Folk, J.E., and Finlayson, J.S. (1977). *Adv. Protein Chem.* **31**:1-133.

Forster, A.C., McInnes, J.L., Skingle, D.C., and Symons, R.H. (1985). *Nucleic Acids Res.* **13**:745-761.

Fraser, R.D.B., and MacRae, T.P. (1973). *Polymer.* **14**:61-67.

Fraser, R.D.B., Jones, L.N., MacRae, T.P., Suzuki, E., and Tulloch, P.A. (1980). *In* Proc. 6th Int. Wool Text. Res. Conf., Pretoria, 1980. pp. 1-33.

Fraser, R.D.B., MacRae, T.P., and Rogers, G.E. (1972). Keratins: Their Composition, Structure and Biosynthesis. Charles C. Thomas, Publisher, Springfield. 304 pp.

Fraser, R.D.B., MacRae, T.P., and Suzuki, E. (1976). *J. Mol. Biol.* **108**:435-452.

Frenkel, M.J., Powell, B.C., Ward, K.A., Sleigh, M.J., and Rogers, G.E. (1989). *Genomics.* **4**:182-191.

Fujisaki, M., and Sugawara, K. (1981). *J. Biochem.* **89**:257-263.

Fullmer, C.S., and Wasserman, R.H. (1981). *J. Biol. Chem.* **256**:5669-5674.

Gahlmann, R., Wade, R., Gunning, P., and Kedes, L. (1988). *J. Mol. Biol.* **201**:379-391.

Garnier, J., Ostguthorpe, D.J., and Robson, B. (1978). *J. Mol. Biol.* **120**:97-120.

Geisler, N., and Weber, K. (1981). *J. Mol. Biol.* **151**:565-571.

Geisler, N., and Weber, K. (1982). *EMBO J.* **1**:1649-1656.

Georgatos, S.D., and Blobel, G. (1987). *J. Cell Biol.* **105**:117-125.

Georgatos, S.D., and Marchesi, V.T. (1985). *J. Cell Biol.* **100**:1955-1961.

Georgatos, S.D., Weaver, D.C., and Marchesi, V.T. (1985). *J. Cell Biol.* **100**:1962-1967.

Georgatos, S.D., Weber, K., Geisler, N., and Blobel, G. (1987). *Proc. Natl. Acad. Sci. USA.* **84**:6780-6784.

Gil, A., and Proudfoot, N.J. (1987). *Cell.* **49**:399-406.

Gillespie, J.M. (1983). *In* Biochemistry and Physiology of the Skin. Vol. 1. Goldsmith, L.A., ed. Oxford University Press, New York. pp. 475-510.

Gillespie, J.M., and Frenkel, M.J. (1974a). *Comp. Biochem. Physiol.* **47B**:339-346.

Gillespie, J.M., and Frenkel, M.J. (1974b). *Aust. J. Biol. Sci.* **27**:617-627.

Glass, D.B., and Smith, S.B. (1983). *J. Biol. Chem.* **258**:14797-14803.

Glass, D.B., El-Maghrabi, M.R., and Pilkis, S.J. (1986). *J. Biol. Chem.* **261**:2987-2993.

Goldsmith, L.A., Baden, H.P., Roth, S.I., Colman, R., Lee, L., and Fleming, B. (1974). *Biochim. Biophys. Acta.* **351**:113-125.

Granger, B.L., and Lazarides, E. (1980). *Cell.* **22**:727-738.

Granger, B.L., and Lazarides, E. (1984). *Mol. Cell. Biol.* **4**:1943-1950.

Gross, E., and Witkop, B. (1962). *J. Biol. Chem.* **237**:1856-1860.

Gruen, L.C., and Woods, E.F. (1983). *Biochem. J.* **209**:587-598.

Hanahan, D., and Meselson, M. (1980). *Gene (Amst.).* **10**:63-67.

Hansen, A.J., Elferink, L.A., and May, B.K. (1989). *DNA.* **8**:179-191.

Harding, C.R., and Scott, I.R. (1983). *J. Mol. Biol.* **170**:651-673.

Harding, H.W.J., and Rogers, G.E. (1971). *Biochemistry.* **10**:624-630.

Harding, H.W.J., and Rogers, G.E. (1972a). *Biochim. Biophys. Acta.* **257**:37-39.

Harding, H.W.J., and Rogers, G.E. (1972b). *Biochemistry.* **11**:2858-2863.

Hardy, M.H. (1952). *Am. J. Anat.* **90**:285-337.

Hashimoto, K. (1970). *Br. J. Dermatol.* **83**:167-176.

Hatzfeld, M., and Franke, W.W. (1985). *J. Cell Biol.* **101**:1826-1841.

Haydock, P.V., and Dale, B.A. (1990). *DNA Cell Biol.* **9**:251-261.

Hayman, J.A., Danks, J.A., Ebeling, P.R., Moseley, J.M., Kemp, B.E., and Martin, T.J. (1989). *J. Pathol.* **158**:293-296.

Heid, H.W., Moll, I., and Franke, W.W. (1988). *Differentiation.* **37**:137-157.

Herrmann, H., and Wiche, G. (1987). *J. Biol. Chem.* **262**:1320-1325.

Hirokawa, N., Glicksman, M.A., and Willard, M.B. (1984). *J. Cell Biol.* **98**:1523-1536.

Horii, I., Kawasaki, K., Koyama, J., Nakayama, Y., Nakajima, K., Okazaki, K., and Seiji, M. (1983). *In* Normal and Abnormal Epidermal Differentiation. Seiji, M., and Bernstein, J.A., eds. University of Tokyo Press, Tokyo. pp. 301-315.

Hunkapiller, M.W., Hewick, R.M., Dreyer, W.J., and Hood, L.E. (1983). *Methods Enzymol.* **91**:399-413.

Huynh, T.V., Young, R.A., and Davis, R.W. (1985). *In* DNA Cloning: A Practical Approach. Vol. 1. Glover, D.M., ed. IRL Press, Oxford. pp. 49-78.

Ish-Horowicz, D., and Burke, J.F. (1981). *Nucleic Acids Res.* **9**:2989-2998.

Ito, M., Tazawa, T., Ito, K., Shimizu, N., Katsuumi, K., and Sato, Y. (1986a). *J. Histochem. Cytochem.* **34**:269-275.

Ito, M., Tazawa, T., Shimizu, N., Ito, K., Katsuumi, K., Sato, Y., and Hashimoto, K. (1986b). *J. Invest. Dermatol.* **86**:563-569.

Jensen, R., Marshak, D.R., Anderson, C., Lukas, T.J., and Watterson, D.M. (1985). *J. Neurochem.* **45**:700-705.

Jones, J.C.R., and Goldman, R.D. (1985). *J. Cell Biol.* **101**:506-517.

Jones, J.C.R., Yokoo, K.M., and Goldman, R.D. (1986). *Cell Motility Cytoskel.* **6**:560-569.

Jones, L.N. (1975). *Biochim. Biophys. Acta.* **412**:91-98.

Jones, L.N. (1976). *Biochim. Biophys. Acta.* **446**:515-524.

Kabat, E.A. (1973). *Proc. Natl. Acad. Sci. USA.* **70**:1473-1477.

Kakiuchi, S., and Yamazaki, R. (1970). *Biochem. Biophys. Res, Commun.* **41**:1104-1110.

Kao, F.T., Hartz, J.A., Law, M.L., and Davidson, J.N. (1982). *Proc. Natl. Acad. Sci. USA.* **79**:865-869.

Kartenbeck, J., Schweicheimer, K., Moll, R., and Franke, W.W. (1984). *J. Cell Biol.* **98**:1072-1081.

Kopan, R., and Fuchs, E. (1989). *Genes & Dev.* **3**:1-15.

Kozak, M. (1987). *Nucleic Acids Res.* **15**:8125-8148.

Kretsinger, R.H. (1980). *Crit. Rev. Biochem.* **8**:119-174.

Kretsinger, R.H., and Nockolds, C.E. (1973). *J. Biol. Chem.* **248**:3313-3326.

Krieg, P.A., and Melton, D.A. (1987). *Methods Enzymol.* **155**:397-415.

Krisinger, J., Darwish, H., Maeda, N., and DeLuca, H.F. (1988). *Proc. Natl. Acad. Sci. USA.* **85**:8988-8992.

Kubilus, J., and Baden, H.P. (1983). *Biochim. Biophys. Acta.* **745**:285-291.

Kubilus, J., Waitkus, R.F., and Baden, H.P. (1980). *Biochim. Biophys. Acta.* **615**:246-251.

Kuczek, E.S., and Rogers, G.E. (1985). *Eur. J. Biochem.* **146**:89-93.

Kuczek, E.S., and Rogers, G.E. (1987). *Eur. J. Biochem.* **166**:79-85.

Kuwano, R., Usui, H., Maeda, T., Fukui, T., Yamanari, N., Ohtsuka, E., Ikehara, M., and Takahashi, Y. (1984). *Nucleic Acids Res.* **12**:7455-7465.

Laemmli, U.K. (1970). *Nature (Lond.).* **227**:680-684.

Lagasse, E., and Clerc, R.G. (1988). *Mol. Cell. Biol.* **8**:2402-2410.

Lane, E.B., Bartek, J., Purkis, P.E., and Leigh, I.M. (1985). *Ann. NY Acad. Sci.* **455**:241-258.

Lawson, D. (1983). *J. Cell Biol.* **97**:1891-1905.

Lendahl, U., Zimmerman, L.B., and McKay, R.D.G. (1990). *Cell.* **60**:585-595.

Liem, R.K.H., and Hutchinson, S.B. (1982). *Biochemistry.* **21**:3221-3226.

Lieska, N., Yang, H.-Y., and Goldman, R.D. (1985). *J. Cell Biol.* **101**:802-813.

Lim, V.I. (1974a). *J. Mol. Biol.* **88**:857-872.

Lim, V.I. (1974b). *J. Mol. Biol.* **88**:873-894.

Lin, J.J.-C., and Feramisco, J.R. (1981). *Cell.* **24**:185-193.

Lipman, D.J., and Pearson, W.R. (1985). *Science (Wash. DC).* **227**:1435-1441.

Lonsdale-Eccles, J.D., Teller, D.C., and Dale, B.A. (1982). *Biochemistry.* **21**:5940-5948.

Lynley, A.M., and Dale, B.A. (1983). *Biochim. Biophys. Acta.* **744**:28-35.

MacKinnon, P.J. (1989). Molecular Analysis of the Ultra-High-Sulphur Keratin Proteins. Ph.D. thesis. University of Adelaide, Adelaide, Australia. 94 pp.

MacKinnon, P.J., Powell, B.C., and Rogers, G.E. (1990). *J. Cell Biol.* **112**: in press.

Manabe, M., Oguin, W.M., Loomis, C., Eckert, F., Sanchez, M., Ackerman, A.B., Freedberg, I.M., and Sun, T.-T. (1989). *J. Invest. Dermatol.* **92**:475(A).

Mangeat, P.H., and Burridge, K. (1984). *J. Cell Biol.* **98**:1363-1377.

Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982). Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. 545 pp.

Marshall, R.C., and Gillespie, J.M. (1977). *Aust. J. Biol. Sci.* **30**:389-400.

McNab, A.R., Wood, L., Theriault, N., Gierman, T., and Vogeli, G. (1989). *J. Invest. Dermatol.* **92**:263-266.

Melton, D.A., Krieg, P.A., Rebagliati, M.R., Maniatis, T., Zinn, K., and Green M.R. (1984). *Nucleic Acids Res.* **12**:7035-7056.

Mercer, E.H. (1961). Keratin and Keratinization. An Essay in Molecular Biology. Pergamon, Oxford.

Merril, C.R., Goldman, D., and Van Keuren, M.L. (1984). *Methods Enzymol.* **104**:441-447.

Messing, J. (1979). *Recomb. DNA Tech. Bull.* **2(2)**:43-48.

Messing, J., and Vieira, J. (1982). *Gene (Amst.).* **19**:269-276.

Messing, J., Crea, R., and Seeburg, P.H. (1981). *Nucleic Acids Res.* **9**:309-321.

Miller, J.H. (1972). Experiments in Molecular Genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. 466 pp.

Mount, S.M. (1982). *Nucleic Acids Res.* **10**:459-472.

Murray, N.E., Brammar, W.J., and Murray, K. (1977). *Mol. Gen. Genet.* **150**:53-62.

Nagano, K. (1973). *J. Mol. Biol.* **75**:401-420.

Napolitano, E.W., Pachter, J.S., Chin, S.S.M., and Liem, R.K.H. (1984). *J. Cell Biol.* **101**:1323-1331.

Norrander, J., Kempe, T., and Messing, J. (1983). *Gene (Amst.).* **26**:101-106.

Ogawa, H., and Goldsmith, L.A. (1977). *J. Invest. Dermatol.* **68**:32-35.

Ohno, S., Emori, Y., Imajoh, S., Kawasaki, H., Kisaragi, M., and Suzuki K. (1984). *Nature (Lond.).* **312**:566-570.

Oliver, R.F. (1970). *J. Embryol. Exp. Morphol.* **23**:219-236.

Parakkal, P.F. (1969). *In* Advances in Biology of Skin. Vol. 9. Montagna, W., and Dobson, R.L., eds. Pergamon Press, Oxford. pp. 441-469.

Parakkal, P.F., and Matoltsy, A.G. (1964). *J. Invest. Dermatol.* **43**:23-34.

Parmacek, M.S., and Leiden, J.M. (1989). *J. Biol. Chem.* **264**:13217-13225.

Parry, D.A.D., and Fraser, R.D.B. (1985). *Int. J. Biol. Macromol.* **7**:203-213.

Parry, D.A.D., Crewther, W.G., Fraser, R.D.B., and MacRae, T.P. (1977). *J. Mol. Biol.* **3**:1146-1156.

Parry, D.A.D., Steven, A.C., and Steinert, P.M. (1985). *Biochem. Biophys. Res. Commun.* **127**:1012-1018.

Peterson, L.L., and Buxman, M.M. (1981). *Biochim. Biophys. Acta.* **657**:268-276.

Powell, B.C., and Rogers, G.E. (1986). *In* Biology of the Integument. Vol. 2. Vertebrates. Bereiter-Hahn, J., Matoltsy, A.G., and Richards, K.S., eds. Springer-Verlag, Berlin. pp. 695-721.

Powell, B.C., and Rogers, G.E. (1990a). *In* Cellular and Molecular Biology of Intermediate Filaments. Goldman, R.D., and Steinert, P.M., eds. Plenum Press, New York. pp. 267-300.

Powell, B.C., and Rogers, G.E. (1990b). *EMBO J.* **9**:1485-1493.

Powell, B.C., Sleigh, M.J., Ward, K.A., and Rogers, G.E. (1983). *Nucleic Acids Res.* **11**:5327-5346.

Price, M.G., and Lazarides, E. (1983). *J. Cell Biol.* **97**:1860-1874.

Pruss, R.M., Mirsky, R., Raff, M.C., Thorpe, R., Dowding, A.J., and Anderton, B.J. (1981). *Cell.* **27**:419-428.

Quinlan, R.A., and Franke, W.W. (1982). *Proc. Natl. Acad. Sci. USA.* **79**:3452-3456.

Radding, C.M., and Shreffler, D.C. (1966). *J. Mol. Biol.* **18**:251-261.

Radloff, R., Baver, W., and Vinograd, J. (1967). *Proc. Natl. Acad. Sci. USA.* **57**:1514-1521.

Reed, K.C., and Mann, D.A. (1985). *Nucleic Acids Res.* **13**:7207-7221.

Renner, W., Franke, W.W., Schmid, E., Geisler, N., Weber, K., and Mandelkow, E. (1981). *J. Mol. Biol.* **149**:285-306.

Rice, R.H., and Green, H. (1977). *Cell.* **11**:417-422.

Rice, R.H., and Green, H. (1979). *Cell.* **18**:681-694.

Rigby, P.W.J., Dieckmann, M., Rhodes, C., and Berg, P. (1977). *J. Mol. Biol.* **133**:237-251.

Robins, E.J., and Breathnach, A.S. (1969). *J. Anat.* **104**:553-569.

Rogers, G.E. (1958a). *Biochim. Biophys. Acta.* **29**:33-43.

Rogers, G.E. (1958b). *Exp. Cell Res.* **14**:378-387.

Rogers, G.E. (1962). *Nature (Lond.).* **194**:1149-1151.

Rogers, G.E. (1963). *J. Histochem. Cytochem.* **11**:700-705.

Rogers, G.E. (1964a). *In* The Epidermis. Montagna, W., and Lobitz, W., eds. Academic Press Inc., New York. pp. 179-236.

Rogers, G.E. (1964b). *Exp. Cell Res.* **33**:264-276.

Rogers, G.E. (1983). *In* Biochemistry and Physiology of the Skin. Vol. 1. Goldsmith, L.A., ed. Oxford University Press, New York. pp. 511-521.

Rogers, G.E., and Harding, H.W.J. (1976a). *In* Biology and Disease of the Hair. Kobori, T., and Montagna, W., eds. University of Tokyo Press, Tokyo. pp. 411-435.

Rogers, G.E., and Harding, H.W.J. (1976b). *In* Proc. 5th Int. Wool Text. Res. Conf., Aachen, 1975. Vol. II. pp. 212-222.

Rogers, G.E., and Rothnagel, J.A. (1983). *In* Normal and Abnormal Epidermal Differentiation. Seiji, M., and Bernstein, J.A., eds. University of Tokyo Press, Tokyo. pp. 171-184.

Rogers, G.E., and Simmonds, D.H. (1958). *Nature (Lond.).* **182**:186-187.

Rogers, G.E., Harding, H.W.J., and Llewellyn-Smith, I.J. (1977). *Biochim. Biophys. Acta.* **495**:159-175.

Rogers, G.E., Kuczek, E.S., MacKinnon, P.J., Presland, R.B., and Fietz, M.J. (1989). *In* The Biology of Wool and Hair. Rogers, G.E., Reis, P.J., Ward, K.A., and Marshall, R.C., eds. Chapman and Hall Ltd., London. pp. 69-85.

Rothnagel, J.A. (1985). Biochemical Studies on Trichohyalin. Ph. D. Thesis. University of Adelaide, Adelaide, Australia. 147 pp.

Rothnagel, J.A., and Rogers, G.E. (1986). *J. Cell Biol.* **102**:1419-1429.

Rothnagel, J.A., Mehril, T., Idler, W.W., Roop, D.R., and Steinert, P.M. (1987). *J. Biol. Chem.* **262**:15643-15648.

Ruskin, B., Krainer, A.R., Maniatis, T., and Green, M.R. (1984). *Cell.* **38**:317-331.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual, Second Edition. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H., and Roe, B.A. (1980). *J. Mol. Biol.* **143**:161-178.

Scott, I.R., and Harding, C.R. (1981). *Biochim. Biophys. Acta.* **669**:65-78.

Scott, I.R., Harding, C.R., and Barrett, J.G. (1982). *Biochim. Biophys. Acta.* **719**:110-117.

Shaw, G., and Kamen, R. (1986). *Cell.* **46**:659-667.

Shono, S., and Toda, K. (1983). *In* Normal and Abnormal Epidermal Differentiation. Seiji, M., and Bernstein, J.A., eds. University of Tokyo Press, Tokyo. pp. 317-326.

Simon, M., and Green, H. (1985). *Cell.* **40**:677-683.

Smith, G.P. (1976). *Science (Wash. DC).* **191**:528-535.

Southern, E.M. (1977). *Methods Enzymol.* **68**:152-176.

Staden, R. (1982). *Nucleic Acids Res.* **10**:2951-2961.

Staden, R. (1984). *Nucleic Acids Res.* **12**:521-538.

Stark, H.-J., Breitkreutz, D., Limat, A., Ryle, C.M., Roop, D., Leigh, I., and Fusenig, N. (1990). *Eur. J. Cell Biol.* **52**:359-372.

Steinert, P.M. (1978). *Biochemistry.* **17**:5045-5052.

Steinert, P.M. (1981). *In* Electron Microscopy of Proteins. Vol. 1. Harris, J.R., ed. Academic Press Inc., London. pp. 123-166.

Steinert, P.M. (1983). *In* Stratum Corneum. Proc. Int. Symp., 1981. Marks, R., and Plewig, G., eds. Springer-Verlag, Berlin. pp. 25-38.

Steinert, P.M., and Roop, D.R. (1988). *Ann. Rev. Biochem.* **57**:593-625.

Steinert, P.M., Cantieri, J.S., Teller, D.C., Lonsdale-Eccles, J.D., and Dale, B.A. (1981a). *Proc. Natl. Acad. Sci. USA.* **78**:4097-4101.

Steinert, P.M., Dyer, P.Y., and Rogers, G.E. (1971). *J. Invest. Dermatol.* **56**:49-54.

Steinert, P.M., Harding, H.W.J., and Rogers, G.E. (1969). *Biochim. Biophys. Acta.* **175**:1-9.

Steinert, P.M., Idler, W.W., and Goldman, R.D. (1980). *Proc. Natl. Acad. Sci. USA.* **77**:4534-4538.

Steinert, P.M., Idler, W.W., and Zimmerman, S.B. (1976). *J. Mol. Biol.* **108**:547-567.

Steinert, P.M., Idler, W.W., Aynardi-Whitman, M., Zackroff, R.V., and Goldman, R.D. (1982). *Cold Spring Harbor Symp. Quant. Biol.* **46**:465-474.

Steinert, P.M., Idler, W.W., Cabral, F., Gottesman, M.M., and Goldman, R.D. (1981b). *Proc. Natl. Acad. Sci. USA.* **78**:3692-3696.

Steinert, P.M., Idler, W.W., Zhou, X.-M., Johnson, L.D., Parry, D.A.D., Steven, A.C., and Roop, D.R. (1985a). *Ann. N. Y. Acad. Sci.* **455**:451-461.

Steinert, P.M., Parry, D.A.D., Idler, W.W., Johnson, L.D., Steven, A.C., and Roop, D.R. (1985b). *J. Biol. Chem.* **260**:7142-7149.

Steinert, P.M., Parry, D.A.D., Racoosin, E.L., Idler, W.W., Steven, A.C., Trus, B.L., and Roop, D.R. (1984). *Proc. Natl. Acad. Sci. USA.* **81**:5709-5713.

Steinert, P.M., Rice, R.H., Roop, D.R., Trus, B.L., and Steven, A.C. (1983). *Nature (Lond.).* **302**:794-800.

Steven, A.C., Trus, B.L., Hainfeld, J.T., Wall, J.S., and Steinert, P.M. (1985). *Ann. N. Y. Acad. Sci.* **455**:371-380.

Steven, A.C., Wall, J.S., Hainfeld, J.F., and Steinert, P.M. (1982). *Proc. Natl. Acad. Sci. USA.* **79**:3101-3105.

Straile, W.E. (1965). *In* Biology of the Skin and Hair Growth. Lyne, A.G., and Short, B.F., eds. Angus and Robertson Ltd., Sydney. pp. 35-57.

Strynadka, N.C.J., and James, M.N.G. (1989). *Ann. Rev. Biochem.* **58**:951-998.

Sugawara, K. (1977). *In* Biochemistry of Cutaneous Epidermal Differentiation. Seiji, M., and Bernstein, I.A., eds. University of Tokyo Press, Tokyo. pp. 387-397.

Sugawara, K. (1979). *Agric. Biol. Chem.* **43**:2543-2548.

Sugawara, K., Oikawa, Y., and Ouchi, T. (1982). *J. Biochem.* **91**:1065-1071.

Sun, T.-T., and Green, H. (1976). *Cell.* **9**:511-521.

Svoboda, M., Meuris, S., Robyn, C., and Christophe, J. (1985). *Anal. Biochem.* **151**:16-23.

Swank, R.T., and Munkres, K.D. (1971). *Anal. Biochem.* **39**:462-477.

Swart, L.S., Jourbert, F.J., and Parris, D. (1976). *In* Proc. 5th Int. Wool Text. Res. Conf., Aachen, 1975. Vol. II. pp. 254-263.

Szebenyi, D.M.E., and Moffat, K. (1986). *J. Biol. Chem.* **261**:8761-8777.

Szebenyi, D.M.E., Obendorf, S.K., and Moffat, K. (1981). *Nature (Lond.).* **294**:327-332.

Takahara, H., Oikawa, Y., and Sugawara, K. (1983). *J. Biochem.* **94**:1945-1953.

Takahara, H., Tsuchida, M., Kusubata, M., Akutsu, K., Tagami, S., and Sugawara, K. (1989). *J. Biol. Chem.* **264**:13361-13368.

Thacher, S.M., and Rice, R.H. (1985). *Cell.* **40**:685-695.

Thomas, P.S. (1983). *Methods Enzymol.* **100**:255-256.

Thompson, E.O.P., and O'Donnell, I.J. (1965). *Aust. J. Biol. Sci.* **18**:1207-1225.

Tseng, H., and Green, H. (1988). *Cell.* **54**:491-496.

Van der Bliek, A.M., Meyers, M.B., Biedler, J.L., Hes, E., and Borst, P. (1986). *EMBO J.* **5**:3201-3208.

Vanaman, T.C., Sharief, F., and Watterson, D.M. (1977). *In* Calcium-Binding Proteins and Calcium Function. Wasserman, R.H., Corradino, R., Carafoli, E., Kretsinger, R.H., MacLennan, D., and Siegel, F., eds. North-Holland, New York. p.107.

Voet, D., and Voet, J.G. (1990). Biochemistry. John Wiley & Sons, New York. 1223 pp.

Vogelstein, B., and Gillespie, D. (1979). *Proc. Natl. Acad. Sci. USA.* **76**:615-619.

Vörner, H. (1903). *Dermatol. Z. (Berlin).* **10**:357-376.

Walker, J.M., and Mayes, E.L.V. (1983). *In* Techniques in Molecular Biology. Walker, J.M., and Gaastra, W., eds. Croom Helm Ltd., Kent. pp. 63-86.

Wang, E., Asai, D.J., and Lazarides, E. (1980). *Proc. Natl. Acad. Sci. USA.* **77**:1541-1545.

Wang, E., Cairncross, J.G., Yung, W.K.A., Garber, E.A., and Liem, R.K.H. (1983). *J. Cell Biol.* **97**:1507-1514.

Wang, E., Gomer, R.H., and Lazarides, E. (1981). *Proc. Natl. Acad. Sci. USA.* **78**:3531-3535.

Watanabe, K., Akiyama, K., Hikichi, K., Ohtsuka, R., Okuyama, A., and Senshu, T. (1988). *Biochim. Biophys. Acta.* **966**:375-383.

Wiche, G., Krepler, R., Artlieb, U., Pytela, R., and Denk, H. (1983). *J. Cell Biol.* **97**:887-901.

Wilbur, W.J., and Lipman, D.J. (1983). *Proc. Natl. Acad. Sci. USA.* **80**:726-730.

Williams, R.C., and Aamodt, E.J. (1985). *Ann. N. Y. Acad. Sci.* **455**:509-524.

Wilson, B.W., Edwards, K.J., Sleigh, M.J., Byrne, C.R., and Ward, K.A. (1988). *Gene (Amst.).* **73**:21-31.

Winter, G., and Fields, S. (1980). *Nucleic Acids Res.* **8**:1965-1974.

Woods, E.F., and Inglis, A.S. (1984). *Int. J. Biol. Macromol.* **6**:277-283.

Yang, H.-Y., Lieska, N., Goldman, A.E., and Goldman, R.D. (1985). *J. Cell Biol.* **100**:620-631.

Yanisch-Perron, C., Vieira, J., and Messing, J. (1985). *Gene.* **33**:103-119.

Zackroff, R.V., Idler, W.W., Steinert, P.M., and Goldman, R.D. (1982). *Proc. Natl. Acad. Sci. USA.* **79**:754-757.

Zeugin, J.A., and Hartley, J.L. (1985). *Focus* **7(4)**:1-2.

Appendix

Appendix

Publications

Rogers, G.E., Kuczek, E.S., MacKinnon, P.J., Presland, R.B., and Fietz, M.J. (1989). Special biochemical features of the hair follicle. *In* The Biology of Wool and Hair. Rogers, G.E., Reis, P.J., Ward, K.A., and Marshall, R.C., eds. Chapman and Hall Ltd., London. pp. 69-85.

Fietz, M.J., Presland, R.B., and Rogers, G.E. (1990). The cDNA-deduced amino acid sequence for trichohyalin, a differentiation marker in the hair follicle, contains a 23 amino acid repeat. *J. Cell Biol.* **110**:427-436.

Rogers, G.E., Fietz, M.J., and Fratini, A. (1991). Trichohyalin and matrix proteins. *Ann. N. Y. Acad. Sci.* in press.

The cDNA-deduced Amino Acid Sequence for Trichohyalin, A Differentiation Marker in the Hair Follicle, Contains a 23 Amino Acid Repeat

Michael J. Fietz, Richard B. Presland, and George E. Rogers

Commonwealth Centre for Gene Technology, Department of Biochemistry, University of Adelaide, Adelaide, South Australia, 5001 Australia

Abstract. Trichohyalin is a highly expressed protein within the inner root sheath of hair follicles and is similar, or identical, to a protein present in the hair medulla. In situ hybridization studies have shown that trichohyalin is a very early differentiation marker in both tissues and that in each case the trichohyalin mRNA is expressed from the same single copy gene. A partial cDNA clone for sheep trichohyalin has been isolated and represents ~40% of the full-length trichohyalin mRNA. The carboxy-terminal 458 amino acids of trichohyalin are encoded, and the first 429 amino acids consist of full- or partial-length tandem repeats of a 23 amino acid sequence. These repeats are characterized by a high proportion of charged amino acids. Secondary structure analyses predict that the majority of the encoded protein could form α-helical structures that might form filamentous aggregates of intermediate filament dimensions, even though the heptad motif obligatory for the intermediate filament structure itself is absent. The alternative structural role of trichohyalin could be as an intermediate filament–associated protein, as proposed from other evidence (Rothnagel, J. A., and G. E. Rogers. 1986. *J. Cell Biol.* 102: 1419–1429).

THE mammalian hair follicle is a derivative of the epidermis that develops within the dermal layer of the skin. In the follicle bulb, the epidermal cells surround a dermal papilla that is essential for the initial development of the follicle and the growth of the fiber (Oliver and Jahoda, 1989). The positioning of the epidermal cells relative to the papilla determines the follicle layer that they will form, namely the different hair components, the cuticle, cortex, and medulla, and the three layers of the inner root sheath (IRS),[1] Henle, Huxley, and IRS cuticle, which accompany the developing fiber in its outward growth.

The IRS is a cylinder of cells surrounding the developing fiber and is believed to fulfill a structural role within the follicle, supporting and directing the fiber cells (Straile, 1965). The basal IRS cells are positioned near the periphery of the follicle bulb, and the daughter cells mature as they move up the follicle hardening into a rigid sheath that eventually degenerates by an unknown process before reaching the surface of the skin. Electron-dense nonmembrane-bound trichohyalin granules appear at a very early stage within all of the developing IRS cells and contain the protein trichohyalin which, between species, has been found to vary in size from 190 to 220 kD (Rothnagel and Rogers, 1986). As the cells move up the follicle, 8–10-nm-diameter filaments appear in close association with the trichohyalin granules (Rogers, 1964). Finally, the granules disappear, the IRS cells become completely filled with intermediate-like filaments aligned parallel with the direction of hair growth (Rogers, 1964), and these filaments harden into the insoluble contents of the mature IRS cell (Birbeck and Mercer, 1957).

The insoluble proteins within the mature cells of the IRS are readily distinguishable from those of the hardened α-keratin in the hair fiber (Rogers, 1964). The IRS proteins are cross-linked by γ-(ϵ-glutamyl)lysine bonds, formed between glutamine and lysine residues, instead of the disulphide bonds typical of keratin in the hair cortex (Harding and Rogers, 1971). In addition, the IRS proteins also contain the amino acid citrulline, which is posttranslationally formed from arginine by peptidylarginine deiminase (Rogers et al., 1977).

The medulla, when present in hair, is a central core of cells within the fiber that develops from cells covering the dome of the dermal papilla. The maturation process of these cells is similar to that of the IRS cells with one important distinction. Nonmembrane-bound granules, which are immunologically cross-reactive with the IRS trichohyalin (Rothnagel and Rogers, 1986) and thus termed trichohyalin granules, begin to appear within the developing medulla cells soon after they move up from the papilla. The trichohyalin granules finally coalesce to form the interior of the hardened medulla

R. B. Presland's current address is Department of Periodontics, Oral Biology and Medicine/Dermatology, University of Washington, Seattle, WA 98195.

1. *Abbreviations used in this paper*: IF, intermediate filaments; IRS, inner root sheath.

cells (Parakkal and Matoltsy, 1964). Protein-bound citrulline and γ-(ε-glutamyl)lysine cross-links, typical of the IRS cells, are also present in the hardened medulla cells (Steinert et al., 1969; Harding and Rogers, 1971, 1972). The essential difference between the IRS and medulla is that the hardened medulla cells are filled with an amorphous protein mass and not the oriented filamentous structures of the IRS.

It is presently unknown what causes the visible differences in the structure of the medulla and the IRS. If trichohyalin is the major precursor in both cell types, it is possible that either the two tissues contain very similar but nevertheless differing forms of trichohyalin or that trichohyalin contains a region capable of folding into an intermediate filament (IF) structure but which is only able to do so within the developing IRS. Alternatively, IF proteins could be expressed within the developing IRS and are cross-linked to the trichohyalin to form the hardened tissue.

In the present paper, we report the isolation and characterization of a cDNA clone encoding a portion of sheep trichohyalin. To determine the probability of trichohyalin forming IFs, the resultant protein sequence was examined for predicted secondary structure and for similarity to known conserved IF sequences. We have also used in situ hybridization analysis to localize trichohyalin expression within the follicle and to examine the relationship of mRNAs coding for medulla and IRS trichohyalin.

Materials and Methods

Peptide Isolation and Sequencing

Sheep and guinea pig trichohyalin were prepared from follicle tissue as described by Rothnagel and Rogers (1986) with a number of modifications. Follicle extracts were concentrated using Centriflo CF25 ultrafiltration cones (Amicon Corp., Danvers, MA) and chromatographed at 0.13 ml/min on a CL-4B gel filtration column (85 × 1.5 cm; Pharmacia Fine Chemicals, Uppsala, Sweden) that had been equilibrated with 7 M guanidine-HCl, 50 mM Tris-HCl, pH 7.5, 100 mM NaCl, 1 mM EDTA, 1 mM DTT. Trichohyalin-containing fractions were concentrated, and the buffer was changed to 7.5 M urea, 0.5 M guanidine-HCl, 50 mM Tris-HCl, pH 7.5, 1 mM EDTA, 1% β-mercaptoethanol by serial filtration in Centriflo CF25 cones.

A sample (500 μg) of pure guinea pig trichohyalin was digested with endoproteinase lysine C (Boehringer Mannheim GmbH, Mannheim, West Germany) for 18 h at 37°C. The resulting peptides were separated on a Superose 12 gel filtration column (30 × 1 cm; Pharmacia Fine Chemicals) that had been equilibrated with 2 M urea, 50 mM Tris-HCl, pH 7.5, 100 mM NaCl, 1 mM EDTA. A sample containing 130 μg of the proteolytic digest was chromatographed at a flow rate of 0.2 ml/min, and 0.5-ml fractions were collected. Protein-containing fractions were loaded onto an Aquapore Butyl reverse-phase cartridge (30 × 2.1 mm; Brownlee Labs, Santa Clara, CA) equilibrated with 0.1% trifluoroacetic acid, and a number of linear gradients of 10–35% acetonitrile were applied over 25–30 min at a flow rate of 0.2 ml/min. The absorbance at 215 nm was recorded, and fractions were collected corresponding to absorbance peaks.

Purified peptides were then sequenced using a gas-phase protein sequencer (470A; Applied Biosystems, Inc., Foster City, CA).

Amino acid analysis of sheep trichohyalin was kindly performed by Dr. R. C. Marshall (Division of Biotechnology, Commonwealth Scientific and Industrial Research Organization, Melbourne, Australia).

Isolation of RNA

Total cellular RNA was isolated from the wool follicles of Merino-Dorset Horn crossed with Border Leicester sheep as described by Powell et al. (1983). High molecular weight genomic DNA was removed from the nucleic acid preparation either by LiCl precipitation (Diaz-Ruiz and Kaper, 1978) or by treatment with RNase-free DNase I (Bresatec, Adelaide, South Australia). Poly A^+ RNA was then isolated from the total RNA preparation by oligo(dT)-cellulose chromatography (Boehringer Mannheim GmbH) (Bantle et al., 1976).

Construction and Screening of the Wool Follicle cDNA Library

A cDNA library was constructed from sheep wool follicle poly A^+ RNA in the expression vector λgt11 (Huynh et al., 1985). Briefly, double stranded cDNA was prepared essentially as described by Gubler and Hoffman (1983) except that murine Moloney leukemia virus reverse transcriptase (Bethesda Research Laboratories, Gaithersburg, MD) was used to synthesize the first strand. The cDNA was ligated to Eco RI linkers (Bresatec) (Huynh et al., 1985) and size selected on 10–40% sucrose gradients to remove cDNAs <1 kb. The cDNA was ligated into λgt11, packaged in vitro (Gigapack Plus; Stratagene, La Jolla, CA), and plated on *Escherichia coli* strain Y1090 (Huynh et al., 1985).

The library was screened essentially as detailed elsewhere (Huynh et al., 1985) using a polyclonal antibody that had been raised in rabbits against sheep trichohyalin by the procedure described by Rothnagel and Rogers (1986). Antibody-bound plaques were detected by incubation with a goat anti–rabbit IgG/alkaline phosphatase conjugate (Sigma Chemical Co., St. Louis, MO) followed by colorimetric staining (Forster et al., 1985).

Recombinant phage DNA was prepared by the liquid culture method of Kao et al. (1982).

DNA Subcloning and Sequencing

The two Eco RI fragments from the longest immunopositive clone, λsTr1, were subcloned into pUC19, and a detailed 6-base restriction enzyme map was derived (see Results). DNA fragments, prepared by restriction enzyme digestion or deletion with Bal31 endonuclease (New England Biolabs, Beverly, MA) (Maniatis et al., 1982), were subcloned into appropriate M13 vectors (Messing and Vieira, 1982; Norrander et al., 1983). These clones were sequenced by the dideoxy chain termination method (Messing et al., 1981; Sanger et al., 1980) using the Bresatec dideoxy sequencing kit.

DNA sequence data was compiled and analyzed on a VAX 11-785 computer (Digital Equipment Corp., Marlboro, MA) using the programs ANALYSEQ (Staden, 1984) and DIAGON (Staden, 1982). Secondary structure analyses were obtained using the PREDICT program, which is a modified version of the joint prediction program of Eliopoulos et al. (1982).

Searches of the Genbank and National Biomedical Research Foundation databases were performed using the programs MATCH, MATCH TRANSLATE, and MATCH FAST (Wilbur and Lipman, 1983; Lipman and Pearson, 1985).

DNA and RNA Hybridizations

Sheep genomic DNA (prepared by Dr. G. R. Cam, Department of Biochemistry, University of Adelaide, Adelaide, Australia) was digested with the appropriate restriction enzymes (Toyobo, Osaka, Japan), electrophoresed on a 1% agarose gel (Sigma Chemical Co.), and transferred to Zeta Probe membrane (Bio-Rad Laboratories, Richmond, CA) by alkaline DNA blotting. After transfer, the membranes were hybridized according to the instructions for Gene Screen (New England Nuclear, Boston, MA) and washed in 0.3 M NaCl, 30 mM tri-sodium citrate, and 0.1% SDS at 65°C.

For Northern blot analysis, total wool follicle RNA was denatured with glyoxal and then fractionated on a 1% agarose gel containing 10 mM sodium phosphate, pH 7.0 (Thomas, 1983). The RNA samples were then blotted onto Zeta Probe membrane, and the filters were hybridized as above and washed in 15 mM NaCl, 1.5 mM tri-sodium citrate, and 0.1% SDS at 20°C.

In Situ Hybridizations to Wool Follicle Sections

High specific activity cRNA probes for in situ hybridization analysis were produced by cloning DNA fragments from λsTr1 into pGEM-2 and transcribing these clones, after appropriate linearization, in the presence of [α-^{35}S]UTP. To obtain the repeat-containing probe, the fragment extending from the Eco RI site in the 5′ linker to the second Pst I site (position 119) was cloned into pGEM-2 and transcribed with T7 RNA polymerase (Promega Biotec, Madison, WI). The 3′ noncoding probe was produced by cloning the 1.9-kb Eco RI fragment and transcribing the Sac I cut clone (λsTr1 position 1,756) with SP6 RNA polymerase (Bresatec).

The in situ hybridization procedure was based on the method of Cox et al. (1984) with the modifications of Powell, B. C., and G. E. Rogers (manuscript in preparation).

Results

Isolation of Peptide Sequences

To confirm the identity of purified cDNA clones a partial amino acid sequence of trichohyalin was required. Due to low yields and poor proteolysis obtained with sheep trichohyalin, guinea pig trichohyalin, which cross reacts immunologically with sheep trichohyalin, was used. Purified guinea pig trichohyalin was proteolysed with endoproteinase lysine C, and after chromatographic purification three of the resultant peptides were subjected to automated Edman degradation. The three peptide sequences showed considerable cross-homology (Fig. 1). All three peptides contain the sequence glutamine-leucine, which is surrounded by a region of charged amino acids. In addition, two peptides also begin with phenylalanine-arginine. The homology observed suggests that at least a portion of the trichohyalin protein consists of repeats. Note that protease-specific cleavage indicates that each peptide should be preceded by a lysine residue in the total protein sequence (see Fig. 1).

Isolation of Sheep Trichohyalin cDNA Clones and Northern Blot Analysis

To increase the probability of isolating full-length trichohyalin cDNA clones, estimated from the size of sheep trichohyalin (190 kD) to be 5–6 kb in length, cDNA that had been prepared by oligo-dT priming of sheep follicle poly A^+ RNA was size selected to remove cDNAs <1 kb before cloning into λgt11. The resultant clones were screened with a polyclonal anti-trichohyalin antibody. This screening yielded three positive cDNA clones, and the longest of these, λsTr1, was 2.4 kb long, which is less than half the expected size for a full-length trichohyalin cDNA clone.

The nucleotide sequence of λsTr1, together with the deduced protein sequence, is shown in Fig. 2 *A*. λsTr1 is 2,408 bp long, with an open reading frame spanning the first 1,375 bp. This is flanked by a 3′ noncoding region of 1,030 bp of which the final 5 bases possibly belong to the poly(A) tail. A putative polyadenylation signal is present at position 2,385 (Fig. 2 *A*).

Northern blot analysis of sheep follicle RNA demonstrated that λsTr1 hybridizes to a 6-kb mRNA species (Fig. 3) which, as stated above, is sufficient to encode the estimated 1,600 amino acids of sheep trichohyalin.

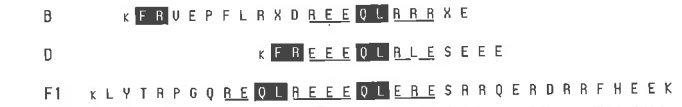


Figure 1. The amino-terminal sequences of three guinea pig trichohyalin peptides (*B*, *D*, and *F1*) are shown. They were purified from endoproteinase lysine C digests of pure trichohyalin and sequenced by the gas-phase method. Unassigned residues are indicated by the letter X. A lysine residue (*small letters*) has been placed at the amino end of each peptide because of the site-specific cleavage of endoproteinase lysine C. The peptides have been aligned to emphasize their similarity. Each peptide contains at least one QL combination (*boxed*) which is surrounded by a highly charged region (*underlined*). Two peptides also begin with FR (*boxed*), indicating an association of these residues with lysine, a substrate amino acid for transglutaminase. The sequence similarity suggests the presence of a repeat structure within trichohyalin.

Predicted Amino Acid Sequence of Trichohyalin

The 1,375-bp open reading frame of λsTr1 encodes a protein of 458 amino acids with a predicted molecular mass of 60 kD. Therefore, the cDNA clone encodes ~30% of the native 190-kD trichohyalin. The deduced protein is hydrophilic, with 59% of the amino acids being charged (Table I). Analysis of the amino acid composition of the deduced protein indicates that glutamic acid/glutamine, arginine, leucine, and lysine are, in decreasing order, the most abundant amino acids, and this feature corresponds with an analysis of total sheep trichohyalin (Table I). The mole percents of glutamic acid/glutamine and arginine are much higher than in the total protein, suggesting that they are enriched within this coded segment. Interestingly, there are no sulphur-containing amino acids in the deduced sequence, which correlates with their low levels in total trichohyalin and also indicates that no disulphide cross-links, which are typical of hair IF proteins, are present within this region of the trichohyalin molecule. Portions of the deduced protein sequence show strong homology with all three guinea pig peptide sequences (Fig. 2 *A*), confirming that the cDNA clone does indeed encode trichohyalin.

Secondary structure predictions, performed with the program PREDICT, indicate that the majority of the protein could adopt an α-helical structure (Fig. 4 *A*). Comparison of the trichohyalin and available IF amino acid sequences (Conway and Parry, 1988) showed no significant homology between the two, indicating that this portion of trichohyalin is unrelated to epithelial, epidermal, or hair IF proteins.

A dot matrix plot of the λsTr1 amino acid sequence that had been compared with itself revealed numerous diagonals spaced in the main by ~25 amino acids (Fig. 5). These diagonals, which indicate internal repeats, cover a region that extends from the beginning of the protein sequence to ~50 amino acids from the carboxy terminus. More detailed analysis of the sequence reveals that there is a 23 amino acid repeat, and a consensus sequence, as determined from the initial 14 repeats, is shown at the top of Fig. 6. The deduced protein sequence was aligned with the consensus sequence and contains 25 full or partial length repeats (Fig. 6). Secondary structure analysis of the consensus sequence predicts that the complete 23 amino acid sequence could form an α-helical rod, although the region from aspartate (at position 1) to phenylalanine (at position 4) could also form random coil (Fig. 4 *B*).

Detection of Trichohyalin Sequences in Sheep Genomic DNA

Sheep genomic DNA samples were digested with Bam HI, Eco RI, and Hind III, blotted, and hybridized with two probes, namely the 1.9-kb (predominantly coding) and 0.47-kb (3′ noncoding) Eco RI fragments of λsTr1 (see Fig. 2 *B*). The coding probe detected a single band within all three tracks (Fig. 7), indicating that the sheep genome contains only a single gene encoding the trichohyalin repeat. Additionally, the hybridization to one, rather than two, Hind III fragments indicates that the genomic sequence appears to differ from that of the cDNA clone since the Hind III site within the 1.9-kb Eco RI fragment of λsTr1 is not detected in the genomic DNA. Upon hybridization with the 0.47-kb probe to the Hind III genomic digest, three bands were de-

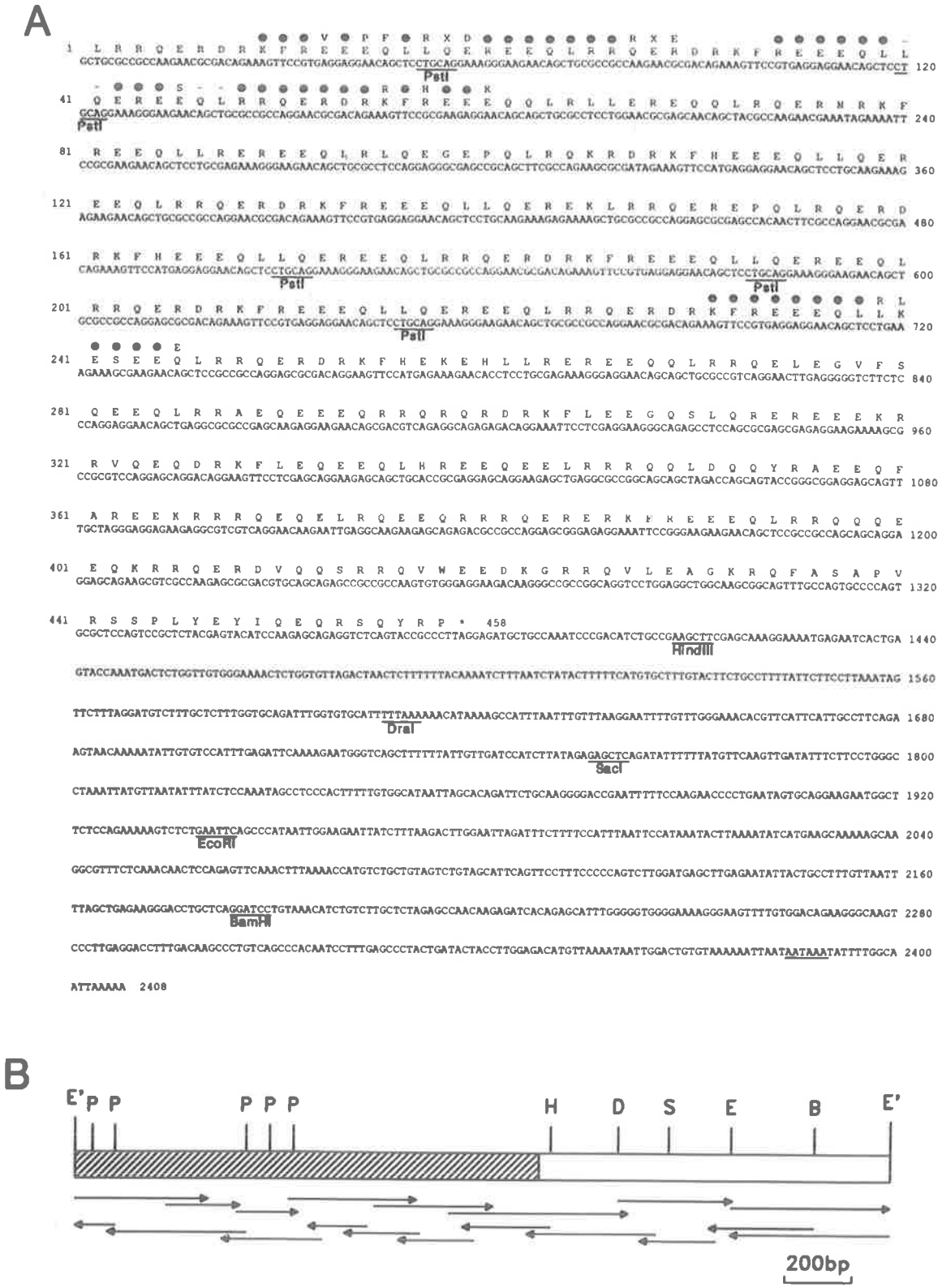


Figure 2. (*A*) The nucleotide and predicted amino acid sequence of λsTr1 is shown. The restriction endonuclease sites depicted in *B* and the likely polyadenylation signal within the 3′ noncoding region are underlined. The guinea pig peptide sequences (see Fig. 1) are positioned in the order B, F1, and D immediately above the corresponding deduced amino acid sequence. Identical residues are indicated by a dot and deletions by a dash. (*B*) The partial restriction map of λsTr1 is shown and includes all the restriction sites used for subcloning, sequencing, and transcription. The sequencing strategy is also depicted, with arrows indicating the extent and direction of sequencing reactions. Hatched and open bars, show the coding and 3′ noncoding regions, respectively. *B*, Bam HI; *D*, Dra I; *E*, Eco RI; *H*, Hind III; *P*, Pst I; *S*, Sac I. E′ indicates terminal Eco RI linkers added during cDNA clone formation.

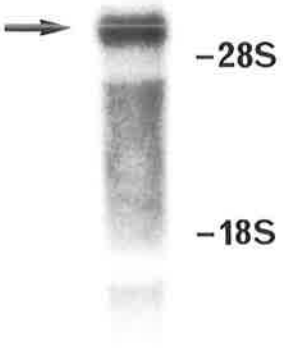


Figure 3. Northern blot analysis of sheep follicle RNA. A sample of total sheep RNA (5 μg) was denatured with glyoxal and fractionated on a 1% agarose gel. The resultant filter was probed with the 1.9-kb Eco RI fragment of λsTrl. Full-length trichohyalin mRNA (6 kb) is detected (*arrow*) together with partially degraded message. Ribosomal RNA marker positions are indicated.

tected (Fig. 7, lane *6*), namely the 13-kb band bound by the coding probe (not seen in Fig. 7 but visible after longer exposure) and also two additional fragments that are both >0.47 kb (4 and 1.5 kb). This suggests that an intron is present within the 3′ noncoding region of the trichohyalin gene.

Localization of Trichohyalin mRNA in Sheep Wool Follicles

The expression of trichohyalin within the follicle was examined by in situ hybridizations performed on skin sections from Merino and Tukidale sheep. Merino wool follicles, which are equivalent to those from which the trichohyalin cDNA was derived, produce fibers that are nonmedullated, whereas a percentage of the Tukidale wool follicles produce medullated fibers. Hybridizations to the Tukidale follicles therefore enabled a comparison of the trichohyalin mRNA expression in the IRS with that in the medulla. cRNA probes were made from both the coding repeat and the 3′ noncoding region by cloning fragments from λsTrl into pGEM-2 and synthesizing transcripts with either SP6 or T7 RNA polymerase. Hybridization of the repeat-containing probe to the Tukidale follicle sections produced a strong signal over both the IRS and the medulla cells (Fig. 8, *A* and *B*). Within both layers, the signal extended from the basal cells within the follicle bulb (Fig. 8 *A*) to the cells positioned immediately beneath the zone of hardening (Fig. 8 *B*). The signal intensity over the medulla cells was similar to that covering the IRS cells. No signal was detected over any layer of the Tukidale epidermis (data not shown). Hybridization with the 3′ noncoding probe produced a pattern identical to that seen with the repeat-containing probe, although the signal was less intense due to the presence of a nonrepeated target sequence (Fig. 8 *C*). With the Merino wool follicle sections, both probes bound to the IRS and produced signals of similar positioning and intensity to that seen in the IRS of the Tukidale follicles (data not shown).

Table I. Amino Acid Composition of the Deduced Trichohyalin Protein and Native Wool Follicle Trichohyalin

Amino acid	Deduced trichohyalin sequence	Wool follicle trichohyalin*
	mole percent	*mole percent*
Asp/Asn‡	3.3/0.2	6.4
Thr	0.0	2.9
Ser	1.7	5.4
Glu/Gln§	26.0/18.1	28.0
Pro	1.1	3.3
Gly	1.1	5.3
Ala	1.3	4.7
½-Cys	0.0	0.6
Val	1.3	4.0
Met	0.0	0.1
Ile	0.2	2.5
Leu	10.9	10.0
Tyr	0.9	2.1
Phe	3.7	2.4
Lys	5.0	6.7
His	1.1	1.7
Arg	23.8	13.7
Trp	0.2	0.2

* Amino acid analysis of purified wool follicle trichohyalin was performed by Dr. R. C. Marshall (Division of Biotechnology, Commonwealth Scientific Industrial Research Organization, Melbourne, Australia).
‡ Aspartic acid and asparagine are denoted separately for the deduced sequence but combined for the native trichohyalin analysis.
§ Glutamic acid and glutamine are denoted separately for the deduced sequence but combined for the native trichohyalin analysis.

Discussion

We report here the isolation and characterization of a cDNA clone, λsTrl, that encodes 2.408 kb of the ~6-kb mRNA of sheep wool follicle trichohyalin. This clone contains the partial coding sequence (1,375 bp) and probably the complete 3′ noncoding region that is >1 kb in length. The coding region encodes 458 amino acids of trichohyalin, extending up to the carboxy terminus, and would form a partial protein with a predicted mass of 60 kD. The amino acid composition of the encoded protein is similar to that of intact sheep trichohyalin (Table I), and regions within the deduced protein sequence are nearly identical to the sequence of purified peptides isolated from guinea pig trichohyalin (Fig. 2 *A*).

Proteolytic peptide sequences (Fig. 1) and two-dimensional gel analysis of trichohyalin breakdown products (Rothnagel and Rogers, 1986) predicted that peptide repeats are present within trichohyalin. Analysis of λsTrl indicates that trichohyalin contains a 23 amino acid repeat and that 95% of the deduced protein sequence consists of full- or partial-length repeats (Fig. 6). The 23 amino acid consensus sequence is highly charged, containing 19 charged or polar residues. Although 15 of these residues are charged, the repeat has an net charge of only −1. Arginine, glutamine, and lysine residues, which are the substrate amino acids for the follicle enzymes peptidylarginine deiminase and transglutaminase, are present within the repeat unit, and some or all of these residues may act as substrates for the enzymes. Since only a single lysine residue is present within the repeat, it is possible that the adjacent highly conserved phenylalanine residue and also the surrounding charged environment are necessary for transglutaminase function. In addition, the glutamine at position 17 within the repeat (Fig. 6) is in a charged environment (EE*Q*LRR) which is similar to that of

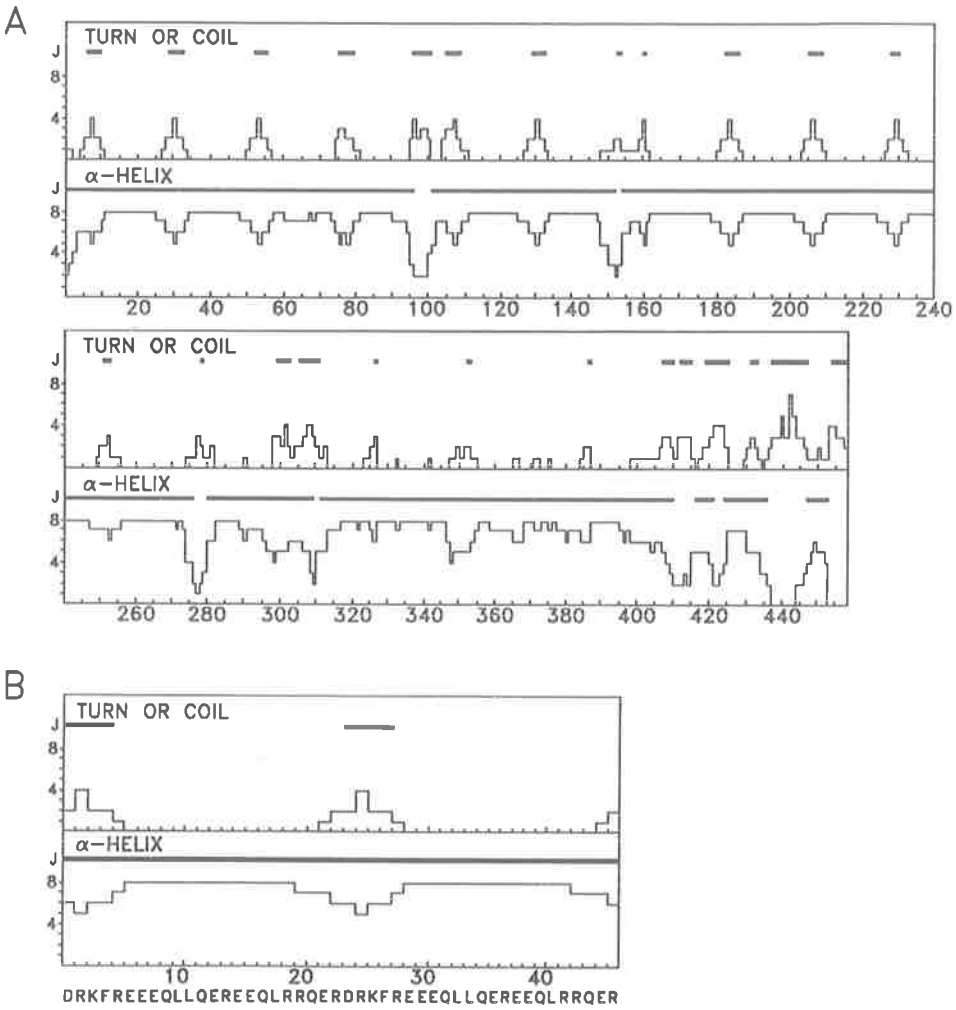


Figure 4. The secondary structure of the deduced amino acid sequence was analyzed by the program PREDICT (see Materials and Methods), and the number of predictions for α-helix and for turn or coil are graphed. Predicted regions of the given secondary structure are indicated by the solid bars at the top of each section (*J*). (*A*) Secondary structure predictions for the total deduced protein sequence. Note that α-helix is predicted for the vast majority of the sequence and that small regions of coil, which predominantly overlap regions of α-helix, are also predicted. (*B*) The predicted structure for two consecutive repeats is enlarged (e.g., from residue 183 to 208). Note that the region from asp(1) to phe(4) (positions 1–4 and 24–27) could be either α-helix or random coil.

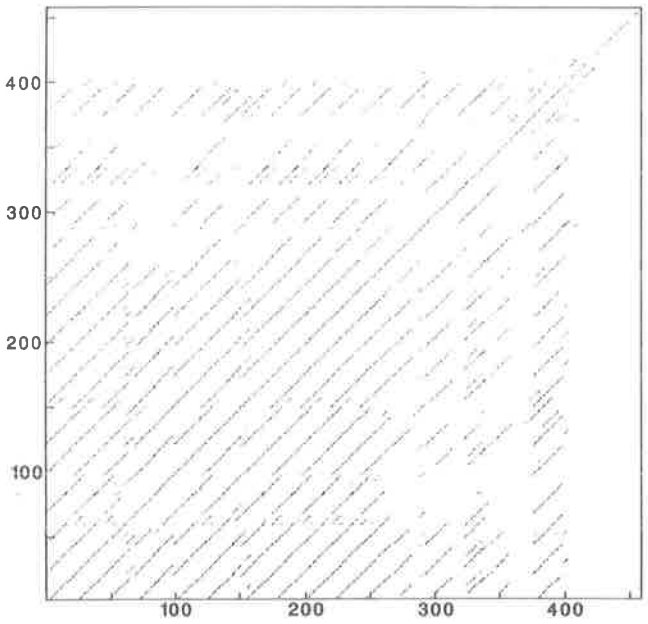


Figure 5. The predicted amino acid sequence of λsTr1 was analyzed for similar internal sequences using the computer program DIAGON (Staden, 1982) with a window length of 23. The output of this comparison was then plotted. Internal similarities are represented by the lines parallel to the central diagonal. The axes are labeled in residue numbers.

the cross-linked glutamine residue within fibrin (EG*Q*QHH), another transglutaminase substrate (Chen and Doolittle, 1971). Thus, gln (at position 17) may therefore be cross-linked by the follicle transglutaminase. Interestingly, the two proposed sites for transglutaminase activity are nearly identical in sequence to the two common sequences present in the guinea pig peptides (Fig. 1). Enzymatic studies on synthetic oligopeptides or cDNA-derived proteins will be necessary to determine the exact sites of transglutaminase and peptidylarginine deiminase action.

The repeat region of the protein can be divided on the basis of the length of each repeat and the level of amino acid conservation into amino- and carboxy-terminal segments. The 14 repeats within the amino-terminal segment are mainly full length and are highly conserved, with 75–100% of the amino acids within each repeat identical to those of the consensus sequence (Fig. 6). Of the 11 repeats within the carboxy-terminal segment, all but one are of partial length and together they show a lower degree of conservation with the consensus sequence (40–89%), although each does retain an eight amino acid stretch (residues 15–22; see Fig. 6). This overall structure suggests that within the amino-terminal segment of the deduced sequence the full 23 amino acid repeat is functionally required, whereas at the carboxy terminus only a smaller region is necessary. The differences in the respective levels of arginine and glutamic acid/glutamine in the deduced sequence and in the native sheep trichohyalin

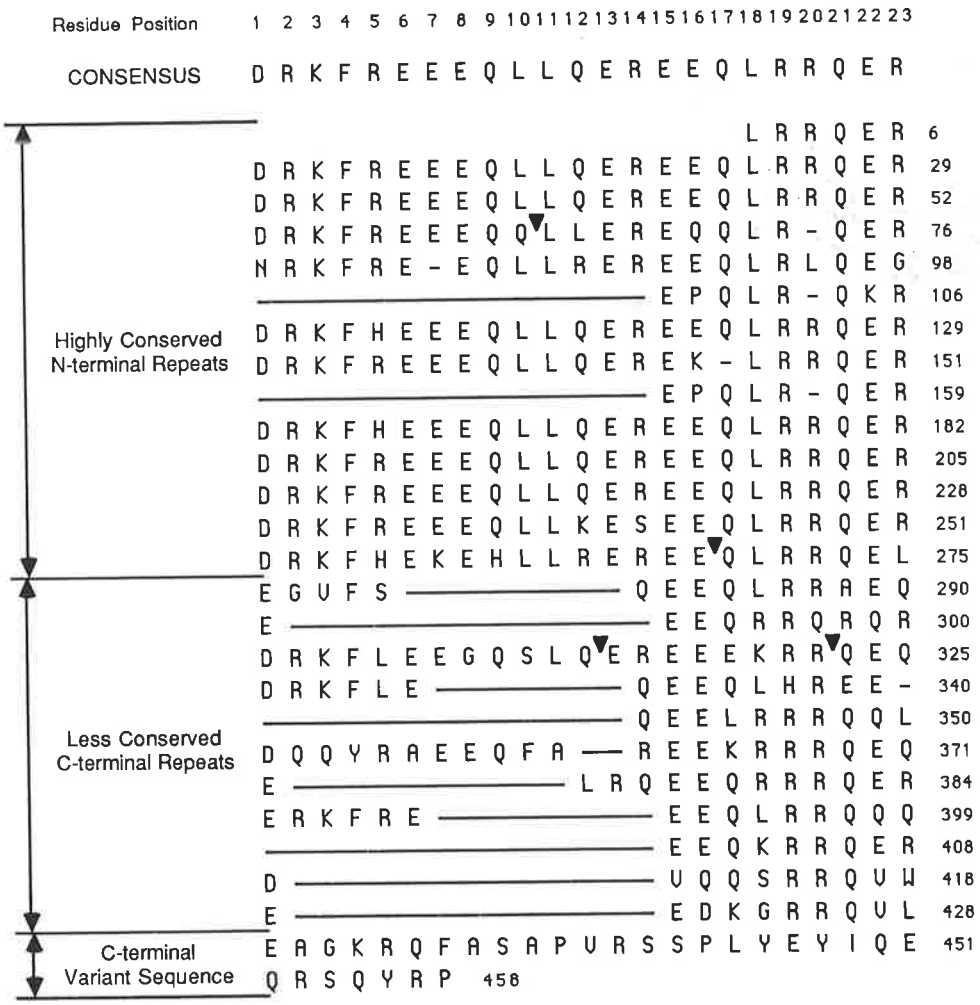


Figure 6. The predicted trichohyalin amino acid sequence is aligned with respect to the 23 amino acid consensus sequence (*top*) which was derived from the amino-terminal segment of highly conserved repeats. Dashes and arrowheads indicate space insertions or sequence deletions that have been introduced for optimal alignment. Each arrowhead indicates the removal of only one or two amino acids.

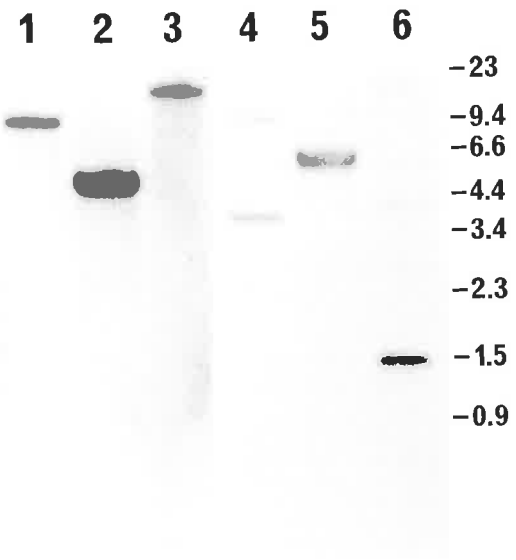


Figure 7. Southern blot analysis of sheep genomic DNA. Total sheep genomic DNA (4 μg/lane) was digested with Bam HI (lanes *1* and *4*), Eco RI (*2* and *5*), and Hind III (*3* and *6*). Lanes *1–3* were probed with the 1.9-kb Eco RI fragment of λsTr1 (coding) and lanes *4–6* were probed with the 0.47-kb Eco RI fragment (3′ noncoding). After prolonged exposure of lane *6*, an additional band is seen and is the same size as the band present in lane *3* (13 kb). The autoradiograph of lanes *1–3* was exposed for 20 h and that of lanes *4–6* for 6 d. Size markers are shown (in kilobase pairs).

(Table I) suggest that the 23 amino acid repeat is not present throughout the whole of the trichohyalin molecule and that sequences distinct from it may be present.

The repeat structure of trichohyalin is similar in certain respects to that of involucrin, an epidermal transglutaminase substrate. Human involucrin contains a highly conserved central region consisting of 39 tandem repeats of a 10 amino acid sequence and this region is flanked by segments that have little homology with the 10 amino acid repeat (Eckert and Green, 1986). The involucrin repeat unit contains a large proportion of hydrophilic residues (70%). Of these, glutamic acid and glutamine residues constitute 50% of the involucrin repeat and this is very similar to the trichohyalin repeat where they account for 48% of the consensus sequence. Although there is no homology between the two protein repeats, glutamic acid residues are positioned within close proximity to the glutamine residues in both repeat sequences, suggesting that neighboring glutamic acid residues may be required for glutamine to be cross-linked by transglutaminase. For example, the involucrin repeat contains the sequence EG*Q*LKH, in which the glutamic acid positioning and charged environment are similar to those within the trichohyalin and fibrin sequences mentioned above. In addition to the similar amino acid composition, there are also similarities in the nucleotide compositions, with both highly conserved repeat segments having a very low T content, being 9% in trichohyalin and 8% in involucrin (Eckert and Green, 1986).

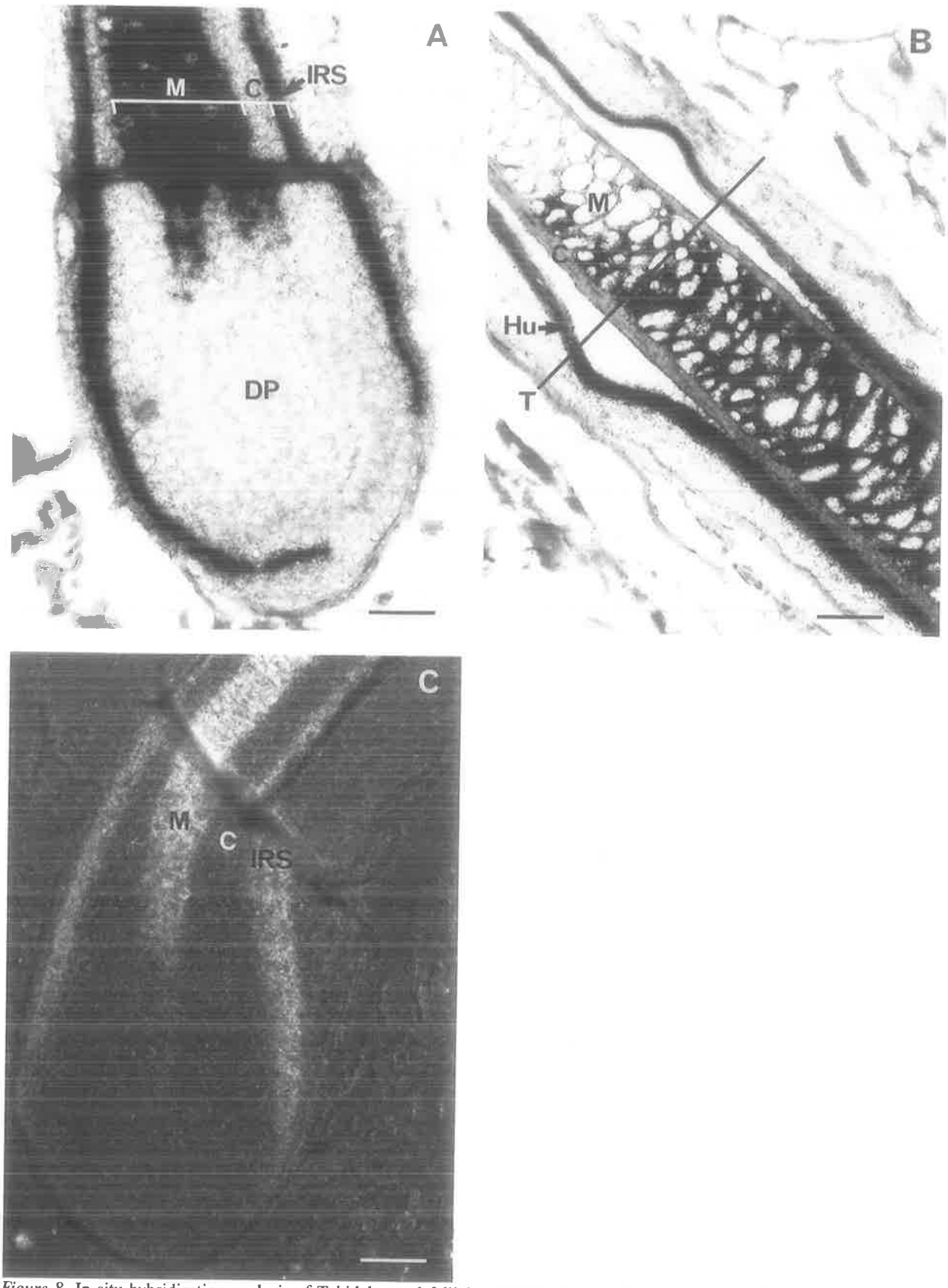


Figure 8. In situ hybridization analysis of Tukidale wool follicles. (*A*) The bulb region of a medullated Tukidale follicle is shown after hybridization with the repeat-containing cRNA probe. Note that the hybridization to the medulla begins with the cells lining the tip of the dermal papilla (*DP*). The hybridization to the IRS extends from the edge of the bulb on one side and the base of the bulb on the other. The latter appearance is caused by the obliqueness of the follicle section. There is no hybridization to the fiber cortex. The dark line present immediately above the dermal papilla is artefactual and caused by folding of the tissue during sectioning. (*B*) Hybridization of the repeat-

The similarity of the proteins within the trichohyalin granules of the IRS and the medulla and the functions of these proteins has long been questioned (Rogers, 1962, 1963, 1983; Rogers and Harding, 1976; Rogers et al., 1977; Rothnagel and Rogers, 1986). Genomic southern blot analysis has shown that the repeat region of λsTrl hybridizes to only a single fragment when sheep DNA is digested with either Bam HI, Eco RI, or Hind III (Fig. 7), indicating that only a single gene encodes the repeat sequence found in λsTrl. In situ hybridization analysis has shown that both repeat and 3′ noncoding probes derived from λsTrl hybridize to mRNAs in both the medulla and IRS of Tukidale wool follicles (Fig. 8). Therefore, the trichohyalin protein within the medulla and IRS is encoded by the same single copy gene. It remains possible that differential splicing of an intron positioned in the 5′ end of the trichohyalin transcript may lead to variant mRNA species in the IRS and the medulla and that the structure and function of the resultant proteins may be somewhat different. Once the complete gene has been purified, examination of the gene structure will establish whether introns are present in the coding region. If they are, then in situ hybridization with fragments corresponding to the 5′ end of the mRNA will be required to determine whether differential splicing occurs.

The trichohyalin that is synthesized in both the IRS and the medulla appears to be the same. The difference between these two hardened tissues could be due to either the additional synthesis of IF proteins in the IRS but not the medulla, such that in the IRS trichohyalin acts as an IF-associated protein (Rothnagel and Rogers, 1986), or a different chemical environment occurring in the two tissues so that trichohyalin itself forms filaments only in the IRS. Evidence for the presence of IF proteins and the independent synthesis of IF in IRS cells has come from Heid et al. (1988) who showed that monoclonal antibodies to the hyperproliferative epidermal IF proteins K6 and K16 bind to the IRS in the human hair follicle. Further, Ito et al. (1986*a*,*b*) and Lane et al. (1986) have shown that certain IF antibodies also bind to the IRS. However, contrary evidence to the presence of IF proteins has come from two in situ hybridization studies using, respectively, a sheep cDNA probe encoding a K6-like protein (Whitbread, L., personal communication) and a probe derived from the conserved α-helical region of a follicle IF gene (Powell, B., personal communication). In both instances, mRNAs were not detected in the IRS cells.

With the primary structure for ~33% of the trichohyalin protein available from the present study, the secondary structure was examined for its theoretical capacity to form α-helical coiled coils with the dimensions specific for the formation of IFs. We have found that although the majority of the trichohyalin sequence and the complete consensus sequence could adopt an α-helical formation (Fig. 4) the predicted trichohyalin sequence has no significant homology with the primary structure of the IF α-helical region. This region is characterized by tandem heptads (a-b-c-d-e-f-g) where the presence of hydrophobic residues at a and d allows the formation of α-helical coiled coils (Parry et al., 1977; McLachlan and Karn, 1982). The α-helical trichohyalin repeats contain a different heptad where only the first residue (phe 4, leu 11, leu 18) is hydrophobic (Fig. 6). This suggests that the trichohyalin repeats, although not involved in coiled coil formation, could nevertheless form a linear cluster of α-helical rods producing filaments with an 8–10 nm diameter, characteristic of genuine IFs. Alternatively, the secondary structure predictions indicate that random coil is also possible between asp(1) and phe(4) (Fig. 4 *B*), producing short α-helical units that could, by random aggregation, form the nonfilamentous structure seen in the medulla cells. Since only 30% of the complete sequence is available, the structural predictions are not yet conclusive and it remains possible that trichohyalin could act as either an IF or an IF-associated protein in the IRS. Despite this equivocal situation, we believe that the follicle enzymes transglutaminase and peptidylarginine deiminase play a central role in the divergent development of the IRS and medulla (Rogers et al., 1989). Differences in the timing of enzyme expression and the level of enzyme activity within the two tissues could allow trichohyalin to adopt different conformations in the two hardened structures.

It is important to note that the secondary structure predictions are based on the precursor protein and that conversion of certain arginine residues to citrulline may alter the overall secondary structure. Further studies will be required to determine which arginine residues are converted and to assess the effect that these changes have on the secondary structure. In addition, complete sequence analysis may determine whether regions within the remaining 70% of the trichohyalin protein are capable of forming IFs.

The in situ hybridization studies have revealed that in both the medulla and the IRS follicle layers, trichohyalin mRNA is expressed within the basal cells. It remains within the developing cells until immediately before tissue hardening (Fig. 8) and this could be due to either the continuous expression or a prolonged half-life of the trichohyalin mRNA. Therefore, trichohyalin mRNA is present during almost all levels of cell development emphasizing the large amount of trichohyalin produced and its importance in growth and differentiation within the follicle. The absence of trichohyalin mRNA in the epidermis suggests that trichohyalin is a follicle-specific protein, although further in situ analysis using other keratinised tissues is required to confirm this. It is clear that in the follicle trichohyalin is a very early

containing cRNA probe to a medullated follicle shaft at the level of conversion from granular to hardened cells. The developing cells of the medulla and the Huxley layer of the IRS (*Hu*) reach, in their movement up the follicle, a transition level (marked approximately by the line *T*) at which the trichohyalin granules disappear and are replaced by the hardened cellular material. The hybridization to the trichohyalin mRNA terminates in both tissues at the same level as the disappearance of the trichohyalin granules. Note that on both sides of the follicle the fiber cuticle and the IRS cuticle have separated during the sectioning procedure. (*C*) The longitudinal section of a medullated Tukidale follicle was photographed under dark-field illumination to highlight the low signal strength obtained with the 3′ noncoding probe (*white grains*). The probe hybridizes to both the IRS and the medulla and spans the same regions as the repeat-containing probe (*A* and *B*). The line present above the bulb was caused by folding of the tissue during sectioning. *C*, cortex; *IRS*, inner root sheath; *M*, medulla. Bars, 50 μm.

differentiation marker, produced much earlier than the IF proteins or the IF-associated proteins present in the cortical cells of the hair fiber (Kopan et al., 1989; MacKinnon, 1989; Powell, B. C., J. L. Arthur, L. A. Crocker, M. J. Fietz, A. Fratini, R. A. Keough, P. J. MacKinnon, A. Nesci, M. T. O'Donnell, and G. E. Rogers, manuscript in preparation) and is apparently expressed immediately after the initiation of differentiation. The factors involved in the initiation of differentiation are at present unknown and their expression may be dependent upon the relative positioning of the basal cells to the dermal papilla. We anticipate that the further investigation of factors controlling the expression of trichohyalin will prove to be of great importance in understanding the initiation and control of differentiation in the hair follicle.

We are very grateful to Dr. B. Powell for help with the in situ hybridizations and for critical analysis of the manuscript. We thank Dr. R. C. Marshall for the amino acid analysis and Mrs. Guo Xiao Hui for assistance with the cDNA sequencing.

This work was supported by a Research Associateship awarded by the Australian Research Council (R. B. Presland), a Commonwealth Postgraduate Research Award (M. J. Fietz), and funds from a Commonwealth Special Centre Grant and the Australian Wool Corporation (G. E. Rogers).

Received for publication 14 August 1989 and in revised form 9 October 1989.

References

Bantle, J. A., I. H. Maxwell, and W. E. Hahn. 1976. Specificity of oligo(dT)-cellulose chromatography in the isolation of polyadenylated RNA. *Anal. Biochem.* 72:413–427.

Birbeck, M. S. C., and E. H. Mercer. 1957. The electron microscopy of the human hair follicle. III. The inner root sheath and trichohyalin. *J. Biophys. Biochem. Cytol.* 3:223–230.

Chen, R., and R. F. Doolittle. 1971. γ-γ cross-linking sites in human and bovine fibrin. *Biochemistry.* 10:4486–4491.

Conway, J. F., and D. A. D. Parry. 1988. Intermediate filament structure. III. Analysis of sequence homologies. *Int. J. Biol. Macromol.* 10:79–98.

Cox, K. H., D. V. DeLeon, L. M. Angerer, and R. C. Angerer. 1984. Detection of mRNAs in sea urchin embryos by *in situ* hybridization using asymmetric RNA probes. *Dev. Biol.* 101:485–502.

Diaz-Ruiz, J. R., and J. M. Kaper. 1978. Isolation of viral double-stranded RNAs using a LiCl fractionation procedure. *Prep. Biochem.* 8:1–17.

Eckert, R. L., and H. Green. 1986. Structure and evolution of the human involucrin gene. *Cell.* 46:583–589.

Eliopoulos, E., A. J. Geddes, M. Brett, D. J. C. Pappin, and J. B. C. Findlay. 1982. A structural model for the chromophore-binding domain of ovine rhodopsin. *Int. J. Biol. Macromol.* 4:263–268.

Forster, A. C., J. L. McInnes, D. C. Skingle, and R. H. Symons. 1985. Non-radioactive hybridization probes prepared by the chemical labelling of DNA and RNA with a novel reagent, photobiotin. *Nucleic Acids Res.* 13:745–761.

Gubler, U., and B. J. Hoffman. 1983. A simple and very efficient method for generating cDNA libraries. *Gene (Amst.).* 25:263–269.

Harding, H. W. J., and G. E. Rogers. 1971. ε-(γ-glutamyl)lysine cross-linkage in citrulline-containing protein fractions from hair. *Biochemistry.* 10:624–630.

Harding, H. W. J., and G. E. Rogers. 1972. The occurrence of the ε-(γ-glutamyl)lysine cross-link in the medulla of hair and quill. *Biochim. Biophys. Acta.* 257:37–39.

Heid, H. W., I. Moll, and W. W. Franke. 1988. Patterns of expression of trichocytic and epithelial cytokeratins in mammalian tissues. I. Human and bovine hair follicles. *Differentiation.* 37:137–157.

Huynh, T. V., R. A. Young, and R. W. Davis. 1985. Constructing and screening cDNA libraries in λgt10 and λgt11. *In* DNA Cloning: A Practical Approach. Vol. 1. D. M. Glover, editor. IRL Press, Oxford. 49–78.

Ito, M., T. Tazawa, K. Ito, N. Shimizu, K. Katsuumi, and Y. Sato. 1986*a*. Immunological characteristics and histological distribution of human hair fibrous proteins studied with anti-hair keratin monoclonal antibodies HKN-2, HKN-4 and HKN-6. *J. Histochem. Cytochem.* 34:269–275.

Ito, M., T. Tazawa, N. Shimizu, K. Ito, K. Katsuumi, Y. Sato, and K. Hashimoto. 1986*b*. Cell differentiation in human anagen hair and hair follicles studied with anti-hair keratin monoclonal antibodies. *J. Invest. Dermatol.* 86:563–569.

Kao, F. T., J. A. Hartz, M. L. Law, and J. N. Davidson. 1982. Isolation and chromosomal localization of unique DNA sequences from a human genomic library. *Proc. Natl. Acad. Sci. USA.* 79:865–869.

Kopan, R., and E. Fuchs. 1989. A new look into an old problem: keratins as tools to investigate determination, morphogenesis, and differentiation in skin. *Genes & Dev.* 3:1–15.

Lane, E. B., J. Bartek, P. E. Purkis, and I. M. Leigh. 1985. Keratin antigens in differentiating skin. *Ann. NY Acad. Sci.* 455:241–258.

Lipman, D. J., and W. R. Pearson. 1985. Rapid and sensitive protein similarity searches. *Science (Wash. DC).* 227:1435–1441.

MacKinnon, P. J. 1989. Molecular Analysis of the Ultra-High-Sulphur Keratin Proteins. Ph.D. thesis. University of Adelaide, Adelaide, Australia. 94 pp.

Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. 545 pp.

McLachlan, A. D., and J. Karn. 1982. Periodic charge distributions in the myosin rod amino acid sequence match cross-bridge spacing in muscle. *Nature (Lond.).* 299:226–233.

Messing, J., and J. Vieira. 1982. A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. *Gene (Amst.).* 19:269–276.

Messing, J., R. Crea, and P. H. Seeburg. 1981. A system for shotgun DNA sequencing. *Nucleic Acids Res.* 9:309–321.

Norrander, J., T. Kempe, and J. Messing. 1983. Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene (Amst.).* 26:101–106.

Oliver, R. F., and C. A. B. Jahoda. 1989. The dermal papilla and maintenance of hair growth. *In* The Biology of Wool and Hair. G. E. Rogers, P. J. Reis, K. A. Ward, and R. C. Marshall, editors. Chapman and Hall Ltd., London. 51–67.

Parakkal, P. K., and A. G. Matoltsy. 1964. A study of the differentiation products of the hair follicle cells with the electron microscope. *J. Invest. Dermatol.* 43:23–34.

Parry, D. A. D., W. G. Crewther, R. D. B. Fraser, and T. P. MacRae. 1977. Structure of α-keratin: structural implications of the amino acid sequences of the type I and type II chain segments. *J. Mol. Biol.* 113:449–454.

Powell, B. C., M. J. Sleigh, K. A. Ward, and G. E. Rogers. 1983. Mammalian keratin gene families: organisation of genes coding for the B2 high-sulphur proteins of sheep wool. *Nucleic Acids Res.* 11:5327–5346.

Rogers, G. E. 1962. Occurrence of citrulline in proteins. *Nature (Lond.).* 194:1149–1151.

Rogers, G. E. 1963. The localization and significance of arginine and citrulline in proteins of the hair follicle. *J. Histochem. Cytochem.* 11:700–705.

Rogers, G. E. 1964. Structural and biochemical features of the hair follicle. *In* The Epidermis. W. Montagna and W. Lobitz, editors. Academic Press Inc., New York. 179–236.

Rogers, G. E. 1983. The occurrence of citrulline in structural proteins of the hair follicle. *In* Biochemistry and Physiology of the Skin. Vol. 1. L. A. Goldsmith, editor. Oxford University Press, New York. 511–521.

Rogers, G. E., and H. W. J. Harding. 1976. Molecular mechanisms in the formation of hair. *In* Biology and Disease of the Hair. T. Kobori and W. Montagna, editors. University of Tokyo Press, Tokyo. 411–435.

Rogers, G. E., H. W. J. Harding, and I. J. Llewellyn-Smith. 1977. The origin of citrulline-containing proteins in the hair follicle and chemical nature of trichohyalin, an intracellular precursor. *Biochim. Biophys. Acta.* 495:159–175.

Rogers, G. E., E. S. Kuczek, P. J. MacKinnon, R. B. Presland, and M. J. Fietz. 1989. Special biochemical features of the hair follicle. *In* The Biology of Wool and Hair. G. E. Rogers, P. J. Reis, K. A. Ward, and R. C. Marshall, editors. Chapman and Hall Ltd., London. 69–85.

Rothnagel, J. A., and G. E. Rogers. 1986. Trichohyalin, an intermediate filament-associated protein of the hair follicle. *J. Cell Biol.* 102:1419–1429.

Sanger, F., A. R. Coulson, B. G. Barrell, A. J. H. Smith, and B. A. Roe. 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol.* 143:161–178.

Staden, R. 1982. An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucleic Acids Res.* 10:2951–2961.

Staden, R. 1984. Graphic methods to determine the function of nucleic acid sequences. *Nucleic Acids Res.* 12:521–538.

Steinert, P. M., H. W. J. Harding, and G. E. Rogers. 1969. The characterization of protein-bound citrulline. *Biochim. Biophys. Acta.* 175:1–9.

Straile, W. E. 1965. Root sheath-dermal papilla relationships and the control of hair growth. *In* Biology of the Skin and Hair Growth. A. G. Lyne and B. F. Short, editors. Angus and Robertson Ltd., Sydney. 35–57.

Thomas, P. S. 1983. Hybridization of denatured RNA transferred or dotted to nitrocellulose paper. *Methods Enzymol.* 100:255–256.

Wilbur, W. J., and D. J. Lipman. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA.* 80:726–730.