Alignment of Time Course Microarray Data with Hidden Markov Models

Sean Robinson

Primary Supervisor: Associate Professor Gary Glonek

Secondary Supervisor: Associate Professor Inge Koch

Thesis submitted towards Master of Philosophy program, December 2012.

SCHOOL OF MATHEMATICAL SCIENCES
DISCIPLINE OF STATISTICS

THE UNIVERSITY of ADELAIDE

Declaration

I, Sean Robinson certify that this work contains no material that has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed by student:

Date:

Signed by supervisor:

Date:

Acknowledgements

I would like to extend a huge thank you to Gary and Inge for their supervision during this project. Particular thanks are also due to Brian Danilko for help with software, Chris Davies for help with the data, Jess Kasza for help with math, and Trent Mattner and Matt Roughan for their help as postgraduate coordinators.

Abstract

Time course microarray experiments allow for insight into biological processes by quantifying changes in gene expression over a time period of interest. This project is motivated by time course microarray data from an experiment conducted on grapevines over the development cycle of the grape berries at a number of different vineyards in South Australia. Although the underlying biological process is the same at each vineyard, there are differences in the timing of the development cycle at different vineyards due to local conditions.

The aim of this project is to construct a methodology to align the data from different vineyards in order to obtain a common representation of the gene expression over the development cycle of the grape berries for each gene. Hidden Markov models (HMMs) have been used to model time series data in a number of domains and have also been used to model time course microarray data. We review these applications in addition to the use of HMMs for particular alignment problems in genomic sequence data. We present an extension of HMMs and propose a novel alignment methodology based on this extension. We evaluate the proposed alignment methodology by applying it to simulated data prior to using it to align the grapevine data.

Contents

1 Introduction **1**

1.1 Motivation 3

1.2 Aims 6

1.3 Overview of Thesis 6

Part I

2 Notation, Preliminary Definitions and Results **9**

2.1 Random Variables and Independence 9

2.2 Sequences of Random Variables 12

2.3 Graphical Models 13

3 Hidden Markov Models **16**

3.1 Definition 16

3.2 Parameterisation 18

4 Associated Problems for HMMs **20**

4.1 Evaluation of the Marginal Emission Density 21

4.2 Most Likely Corresponding Realised State Sequence 26

4.3 Maximum Likelihood Estimator of the Parameters 29

5 Alignment with HMMs **34**

5.1 Pair HMMs for Alignment of Genomic Sequences 34

5.2 Extensions of Pair HMMs 37

5.2.1 Binary Markov Dynamics with Continuous Emissions 37

5.2.2 Additional Information Incorporated into Model 39

5.3 Simple Alignment with an HMM 41

Part II

6 $L(t)$-fold HMMs **45**

6.1 Definition 45

6.2 Parameterisation 48

7 Associated Problems for $L(t)$-fold HMMs **49**

7.1 Evaluation of the Marginal Emission Density 50

7.2 Most Likely Corresponding Realised State Sequence 52

7.3 Maximum Likelihood Estimator of the Parameters 54

8 Alignment with $L(t)$-fold HMMs **58**

8.1 An Example Model 58

8.2 Model for the Grapevine Data 61

9 Model Fitting Methodology **64**

9.1 Model Fitting Algorithm 64

9.2 Simulation Example 65

10 Alignment of the Grapevine Data **70**

10.1 Fitting the Model 70

10.2 Common Representations 75

11 Alignment Diagnostics **78**

11.1 Ordering Based on Log-likelihood 79

11.2 Fitting an HMM and the Estimated HMM Parameters 81

11.2.1 Comparison of Estimated Parameters 82

11.2.2 Difference in Log-likelihood 83

12 Conclusion **87**

12.1 Summary of Thesis 87

12.2 Discussion 88

Appendices

A The Grapevine Data **91**

B The EM Algorithm **110**

C Proof of Theorem 4.6 **113**

Bibliography **121**