

Mathematical Methods for Comparative *Ab Initio* Gene Prediction

Irmtraud Margret Meyer

Trinity College

Cambridge

August 2002

A dissertation submitted for
the degree of Doctor of Philosophy
at the University of Cambridge

Preface

The work presented in this dissertation was carried out at the Sanger Institute in Cambridge between January **2000** and August 2002. This dissertation is the result of my own work and includes nothing which is the outcome **of** work done in collaboration except where specifically indicated in the text. **No** part of this dissertation nor anything substantially the same **has** been or is being submitted for any qualification at any other university.

Summary

This dissertation introduces two novel methods for the comparative prediction of protein coding genes in eukaryotic genomes. The first method, implemented in a program called **DOUBLESCAN**, is an ab initio method which simultaneously predicts the gene structures and the alignment of two evolutionarily related input **DNA** sequences from the sequence of their **A, C, G, T** bases only. The second method, implemented in a program called **PROJECTOR**, is a homology based method which predicts gene structures in one **DNA** sequence according to the known gene structures of a related **DNA** sequence and which simultaneously aligns the two **DNA** sequences. Both methods employ a probabilistic pair Hidden Markov model and are capable of predicting partial, complete and multiple genes as well as pairs of genes which are related by events of exon-fusion or exon-splitting. Predictions are generated using two different algorithms: the Hirschberg algorithm whose predictions are generated in linear memory and quadratic time and a new algorithm, called the Stepping Stone algorithm, whose memory and time requirements scale both linearly with the length of the input sequence. This work describes the theoretical concepts underlying the two novel methods and their implementation into computer programs and demonstrates the validity and generality of the approach by evaluating the performance of the gene prediction on a test set of mouse (*Mus musculus*) and human (*Homo sapiens*) as well as *Caenorhabditis elegans* and *Caenorhabditis briggsae* **DNA** sequence pairs.

Acknowledgements

First of all, I would like to thank my supervisor, Richard Durbin, for his advice, support and encouragement. I thank Kevin Howe, Matthew Pocock, Raphael Leplae, Aaron Levine, Marc Sohrmann, Ashwin Hajarnavis, Lachlan Coin, Thomas Down and all the other members of the Wellcome Trust Genome Campus who have made both science and life on the campus interesting and enjoyable.

I am grateful to Trinity College, Cambridge, for an External Research Studentship and to the Wellcome Trust for a Prize Studentship.

Contents

Summary	ii
1 Introduction	1
1.1 Motivation	1
1.2 Biological background	6
1.3 Existing non-comparative methods	9
1.3.1 Types of evidence	9
1.3.2 Methods	9
1.3.3 Summary	15
1.4 Existing comparative methods	15
1.4.1 Conservation detection methods	16
1.4.2 Methods for comparative functional prediction	17
1.4.3 Summary	21
1.5 Theoretical background	23
1.5.1 Pair hidden Markov models	24
1.5.2 Alignment algorithms	24
2 The pair HMM of DOUBLESCAN and PROJECTOR	30
2.1 Introduction and motivation	30
2.2 States and transitions of the pair HMM	32
2.3 Determination of the pair HMM's parameters	37
2.4 The Stepping Stone algorithm	42
3 <i>Ab initio</i> prediction of mouse and human genes	46
3.1 Introduction and motivation	46

3.2	Results	47
3.3	Prediction of conserved subsequences	51
3.4	Validation of the Stepping Stone algorithm	51
3.5	Summary and discussion	52
4	Prediction of mouse and human genes	55
4.1	Introduction and motivation	55
4.1.1	Implementation	56
4.2	Results	57
4.3	Summary and discussion	59
5	Prediction of <i>C. elegans</i> and <i>C. briggsae</i> genes	67
5.1	Introduction and motivation	67
5.2	Training of the pair HMM's parameters	68
5.3	Results	69
5.3.1	Performance on test set 1	72
5.3.2	Performance on test set 2	74
5.3.3	Comparison of the performance of DOUBLESCAN and FGENESH	83
5.4	Summary and discussion	86
6	DOUBLEBUILD	88
6.1	Introduction and motivation	88
6.2	Special transitions within DOUBLEBUILD	89
6.3	Special emissions within DOUBLEBUILD	93
6.4	The main classes	95
6.4.1	The <code>Pairhmm</code> class	95
6.4.2	The <code>Pairhmm.State</code> class	95
6.4.3	The <code>Sequence</code> class	99
6.5	Alignment algorithms	102
6.5.1	The Viterbi algorithm	102
6.5.2	The Hirschberg algorithm	102
6.5.3	The Stepping Stone algorithm	102
7	Conclusion	104

A Mouse human training and test sets	107
A.1 The training set	107
A.2 The test set	108
A.3 Post-processing of the predicted mouse and human genes	110
B Mouse human parameter tables	111
C <i>C. elegans</i> <i>C. briggsae</i> training and test sets	121
C.1 The training set.	121
C.2 Test set 1	122
C.3 Test set 2	122
D <i>C. elegans</i> <i>C. briggsae</i> parameter tables	126
E The DOUBLESCAN web-server	132
Bibliography	142