

Studies in Probabilistic  
Sequence Alignment and Evolution

Thesis by

Ian Holmes

In Partial Fulfilment of the Requirements  
for the Degree of  
Doctor of Philosophy

University of Cambridge

Queens' College

and

The Sanger Centre  
Wellcome Trust Genome Campus  
Hinxton Hall, Hinxton  
Cambridge

(Submitted on December 18, 1998)

## Acknowledgements

During the past three years I have been very fortunate in having had the opportunity to discuss computational biology and Bayesian methods with Alex Bateman, Ewan Birney, Bill Bruno, Richard Durbin, Sean Eddy, David Hausler, Des Higgins, Tim Hubbard, Anders Krogh, Chip Lawrence, David MacKay, Graeme Mitchison, Richard Mott, Kimmen Sjölander, Guy St.Clair Slater, Victor Solovyev, Erik Sonnhammer and many other workers at the Wellcome Trust Genome Campus in Hinxton.

I especially thank my supervisor, Richard Durbin, for his thoughtful encouragement, guidance and advice, for being a fair and consistent critic and for his tireless patience and support.

I am very grateful for the assistance I have received from the Sanger Centre. I would particularly like to thank Sharon Thornton and Sheila Skingley, the Informatics group secretaries during the course of my PhD, and the Systems Support team for their ever-rapid response.

I would also like to thank my proofreaders, whose comments were vital: Alex Bateman, Ewan Birney, Tony Cox, Richard Durbin and Guy Slater.

Thank you to the people who have supported me most during the last year: Andrew Holmes; Jennifer Holmes; and Claire Hibbitt.

This work was supported by a grant from the Medical Research Council.

## Summary

The complete sequencing of whole genomes presents opportunities for detailed study of molecular evolution. This thesis combines theoretical developments of Bayesian approaches in bioinformatics with analysis of duplications in the recently completed *C.elegans* genome.

Developments in the Bayesian probabilistic framework for sequence analysis using hidden Markov models (HMMs) are described. The principal HMM algorithms are reviewed including alignment, training and model comparison. Theory is developed for prediction of alignment accuracy and tested using simulations. Software to provide accuracy measures for multiple alignments, based on the popular HMMER suite of profile-based alignment algorithms, is presented and evaluated with reference to the Pfam database of multiple alignments.

Several of these statistical techniques are applied to an analysis of genomic duplications in the *C.elegans* genome. The completion of this - the first animal genome - offers an opportunity to study the random duplication that are believed to be the first step in the evolution of a new gene. The construction of a database of non-coding duplications is described and measurements of molecular evolutionary parameters in *C.elegans* are calculated from the data and reported. A method of dating gene duplications using alignments between conserved introns is presented and compared to existing methods using Bayesian techniques developed earlier in the dissertation. Amongst the principal agents involved in creating genomic duplications are transposons; one of the simplest families of transposon is the Tc1-*mariner* family, of which two distinct active subfamilies are well-known in *C.elegans*. Using HMM profiles, six new subfamilies of *mariner*-like transposon have been identified in the *C.elegans* genome. Several of the new subfamilies display interesting homologies to one another, suggestive of common mechanisms of transpositional catalysis.

Finally, the software tools developed during this project are described and made available for public retrieval from the Sanger Centre web site.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Preamble . . . . .	2
1.2	Sequence analysis . . . . .	4
1.2.1	Bayesian methods . . . . .	5
1.2.2	Summary of Part I of the thesis . . . . .	9
1.3	Gene duplications . . . . .	11
1.3.1	Agents of change . . . . .	12
1.3.2	Summary of Part II of the thesis . . . . .	15
1.4	Statement of originality . . . . .	17
<b>I</b>	<b>Studies in Probabilistic Sequence Alignment</b>	<b>18</b>
<b>2</b>	<b>Bayesian Methods for Hidden Markov Models in Biological Sequence Analysis</b>	<b>19</b>
2.1	Introduction . . . . .	20
2.2	Notation . . . . .	20
2.2.1	Other formulations of HMMs . . . . .	22
2.3	Aligning sequences to HMMs . . . . .	23
2.3.1	Maximising the alignment likelihood: the Viterbi algorithm	23
2.3.2	Summing alignment likelihoods: the Forward algorithm .	24

2.3.3	Posterior probabilities of alignments: the Forward-Backward algorithm . . . . .	25
2.3.4	Comparing alignments . . . . .	26
2.4	Hidden Markov models in molecular evolution . . . . .	28
2.4.1	Time-dependent substitution matrices . . . . .	28
2.4.2	Time-dependent gap probabilities . . . . .	30
2.5	Likelihood derivatives and Fisher scores . . . . .	32
2.6	Fitting parameters to HMMs . . . . .	34
2.6.1	Maximising the likelihood in parameter space: training . . . . .	36
2.6.2	Integrating the likelihood over parameter space: model comparison . . . . .	37
2.6.3	Incremental Baum-Welch and sparse envelopes . . . . .	38
2.7	Score and length distributions of an HMM . . . . .	40
2.8	Generalised HMMs . . . . .	42
<b>3</b>	<b>Dynamic Programming Alignment Accuracy</b>	<b>46</b>
3.1	Introduction . . . . .	47
3.2	Definitions and notation . . . . .	48
3.2.1	Definition of the alignment fidelity . . . . .	48
3.2.2	Choice of scoring parameters . . . . .	48
3.2.3	Probabilistic interpretation . . . . .	50
3.2.4	A simple point substitution model . . . . .	51
3.2.5	Relationship between probabilistic model and alignment algorithm . . . . .	51
3.3	Results . . . . .	53
3.3.1	Simulation 1: Optimisation of the alignment fidelity with respect to the scoring scheme . . . . .	53
3.3.2	Simulation 2: Measurement of the alignment fidelity . . . . .	53
3.3.3	The probabilistic prediction $\hat{\lambda}$ is supported experimentally . . . . .	53
3.3.4	The fidelity decreases as $p_G$ and $p_S$ are increased . . . . .	54

3.3.5	An analytic approximation to the alignment fidelity . . .	57
3.3.6	Calculation of the edge wander . . . . .	58
3.3.7	Estimating the fidelity of a particular alignment . . . . .	65
3.3.8	An optimal accuracy alignment algorithm . . . . .	65
3.3.9	Simulation 3: Evaluation of the optimal accuracy algorithm	66
3.4	Discussion . . . . .	67
<b>4</b>	<b>postal: Software for Checking Multiple Alignment Accuracy</b>	<b>70</b>
4.1	Introduction . . . . .	71
4.1.1	Mathematical overview . . . . .	72
4.2	The <code>postal</code> software . . . . .	74
4.2.1	Usage . . . . .	76
4.2.2	A note on interpretation . . . . .	76
4.2.3	The optimal accuracy algorithm and <code>postal</code> . . . . .	77
4.2.4	More complex models . . . . .	77
4.3	Evaluation: using <code>postal</code> as a semi-automated quality check for Pfam . . . . .	80
4.4	Discussion . . . . .	83
4.4.1	Availability . . . . .	83
<b>II</b>	<b>Studies in Evolution</b>	<b>84</b>
<b>5</b>	<b>Wormdup: a Database of DNA Duplications in <i>Caenorhabditis</i> <i>elegans</i></b>	<b>85</b>
5.1	Chapter introduction . . . . .	86
5.2	Methods . . . . .	88
5.2.1	Overview of methods . . . . .	88
5.2.2	Filtering low-complexity regions . . . . .	91
5.2.3	Preliminary scan for repetitive elements . . . . .	91
5.2.4	Finding duplicated blocks . . . . .	93

5.2.5	Excluding genes and repetitive elements . . . . .	94
5.2.6	Gene duplications . . . . .	95
5.3	Statistics of duplications in Wormdup . . . . .	96
5.3.1	Age distribution of duplications: the duplication fixation rate . . . . .	97
5.3.2	Length distribution of duplications: indel rates . . . . .	100
5.4	Repetitive element-mediated duplications . . . . .	105
5.5	Comparison of non-coding duplications and coding duplications .	105
5.6	Discussion . . . . .	109
5.6.1	Availability . . . . .	110
<b>6</b>	<b>Intron Clocks: Time-Dependent Models of Intron Evolution</b>	<b>111</b>
6.1	Introduction . . . . .	112
6.2	General patterns of intron evolution . . . . .	113
6.2.1	Conserved signals in introns . . . . .	114
6.2.2	Sizes of indels in introns . . . . .	115
6.2.3	Intron mobility . . . . .	117
6.3	Fitting time-dependent models to pairs of introns . . . . .	119
6.3.1	Down-weighting uninformative pairs . . . . .	120
6.3.2	Testing intron clocks . . . . .	121
6.4	Discussion . . . . .	123
6.4.1	Availability . . . . .	124
<b>7</b>	<b>Classification of DNA Transposons in <i>Caenorhabditis elegans</i></b>	<b>125</b>
7.1	Abstract . . . . .	126
7.2	Introduction . . . . .	126
7.3	Methods . . . . .	129
7.3.1	Construction of the transposon family data set . . . . .	129
7.3.2	Analysis of the transposon family data set . . . . .	132
7.4	Results . . . . .	132

7.4.1	Previously characterised transposon families . . . . .	132
7.4.2	Previously uncharacterised transposon families . . . . .	134
7.4.3	Variation between transposon families . . . . .	135
7.4.4	Variation within transposon families . . . . .	137
7.4.5	Location of transposons within the <i>C.elegans</i> genome . . .	138
7.5	Discussion . . . . .	141
<b>8</b>	<b>Conclusion</b>	<b>144</b>
	<b>Appendices</b>	<b>149</b>
<b>A</b>	<b>Software</b>	<b>149</b>
A.1	Introduction . . . . .	150
A.2	LogSpace: C++ classes for working with HMMs . . . . .	150
A.2.1	Posterior probabilities for profile HMMs . . . . .	152
A.2.2	Availability . . . . .	153
A.3	BayesPerl: Perl modules and scripts for working with tables of log-likelihood data . . . . .	153
A.3.1	Availability . . . . .	155
A.4	GFFTools: Perl scripts for processing GFF files . . . . .	155
A.4.1	Availability . . . . .	160
A.5	bigdp: A program for assembling BLAST hits by dynamic pro- gramming . . . . .	160
A.5.1	Availability . . . . .	163
	<b>Bibliography</b>	<b>164</b>



# List of Figures

1.1	Multiple alignment of rhodopsin-like protein sequences. . . . .	6
1.2	The dynamic programming matrix. . . . .	8
1.3	Unequal crossing-over during recombination. . . . .	14
2.1	Hidden Markov model for global pairwise alignment. . . . .	31
2.2	Looping models that tend towards flat length distributions. . . .	44
2.3	The Bayes block aligner. . . . .	45
3.1	Coupled Markov model of sequence evolution. . . . .	50
3.2	Contours of the predicted optimal effective gap penalty $\hat{\lambda}$ in parameter space. . . . .	54
3.3	Variation of alignment fidelity with effective gap penalty $\lambda$ . . . .	55
3.4	Optimal values of the effective gap penalty $\lambda$ . . . . .	56
3.5	Variation of alignment fidelity with mutation parameters. . . . .	57
3.6	Illustration of edge wander. . . . .	58
3.7	Alignment score fluctuations near a gap. . . . .	60
3.8	Edge wander predictions for the alignment fidelity. . . . .	64
3.9	Evaluation of the optimal accuracy algorithm. . . . .	67
4.1	The Forward-Backward algorithm. . . . .	73
4.2	post-al accuracy levels for the Pfam rhodopsin-like domain. . . .	75
4.3	Alignment ambiguity vs. alignment size. . . . .	78
4.4	Alignment ambiguity vs. alignment quality. . . . .	79

4.5	Comparative sequence rankings of <code>postal</code> and HMMER. . . . .	81
5.1	Schematic view of the construction of Wormdup. . . . .	89
5.2	Frequency distribution of the number of times a base is involved in a high-scoring duplication. . . . .	95
5.3	Age distribution of high-scoring duplications. . . . .	98
5.4	Mean separation of same-chromosome duplications plotted by age. . . . .	99
5.5	Variation of observed duplication size with age. . . . .	101
5.6	Size distribution of recent duplications. . . . .	102
5.7	Distribution of indel sizes. . . . .	104
6.1	Distribution of differences in intron lengths. . . . .	116
6.2	Failure of the null model for <i>C.elegans</i> introns. . . . .	118
7.1	Putative domain structure of the Tc1 transposase. . . . .	128
7.2	Alignment of D35E motifs. . . . .	131
7.3	Alignment of <i>mariner</i> -like transposases in <i>C.elegans</i> . . . . .	136
7.4	Phylogenetic tree of <i>mariner</i> -like transposases in <i>C.elegans</i> . . . . .	137
7.5	Phylogenetic tree for Tc1 and Tc3 transposases. . . . .	138
7.6	Phylogenetic tree for Tc11-Tc13 transposases. . . . .	139
A.1	<code>gffdp.pl</code> model file for finding repeat-mediated duplications. . . . .	161

# List of Tables

3.1	Edge wander for various common amino acid substitution matrices.	63
4.1	The 20 most suspicious alignments in Pfam, October 1998. . . .	82
5.1	Copy numbers of CeRep elements in <i>C.elegans</i> . . . . .	92
5.2	The 30 largest duplicated gene families in <i>C.elegans</i> . . . . .	107
6.1	Unexplained intron homologies in <i>C.elegans</i> . . . . .	119
6.2	Log-odds-ratios of “intron clock” hypotheses. . . . .	122
7.1	Previously characterised DNA transposons in <i>C.elegans</i> . . . . .	133
7.2	Previously uncharacterised DNA transposons in <i>C.elegans</i> . . . . .	135
7.3	Proximity of transposons to coding sequence. . . . .	140
7.4	Propensities for <i>C.elegans</i> repeat types to be found within 1kb of each other. . . . .	142