LIBRARY
Te Tumu Herenga
THE UNIVERSITY OF AUCKLAND

RESEARCHSPACE@AUCKLAND

http://researchspace.auckland.ac.nz

ResearchSpace@Auckland

Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.
http://researchspace.auckland.ac.nz/feedback

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.

Note : Masters Theses

The digital copy of a masters thesis is as submitted for examination and contains no corrections. The print copy, usually available in the University Library, may contain alterations requested by the supervisor.

Computational Statistical Inference for Molecular Evolution and Population Genetics

A thesis submitted in partial fulfilment of the
requirements for the Degree of
Doctorate of Philosophy in Biological Sciences,
The University of Auckland,
New Zealand
June 2002

© Alexei James Drummond

0.1 Abstract

This research aims to develop new methods and software for evolutionary inference. The focus will be on two challenges that analysis of molecular data in the genomic age provides: (i) *measurably evolving populations* and (ii) evolution of RNA secondary structure. Molecular sequence data is increasing in length, and also acquiring a depth in the time dimension (for example, HIV-1, human influenza A, and ancient mtDNA). This has provided an innovative research direction, for which explicit evolutionary inference methods are required. The first aim of this research is to provide new statistical methods and new bioinformatic tools (software packages) to assist in tackling this new problem in evolutionary biology. Both maximum likelihood and Bayesian inference methods are developed for the purpose of estimating substitution rates and concerted changes in the substitution rate. In addition, with the rapid succession of newly sequenced full genomes, researchers can no longer use simple molecular sequence similarity to infer homology. Knowledge of molecular structure needs to be incorporated into evolutionary inference methods. The evolutionary relationship between sequence and structure is still poorly understood and the new wealth of data provides an exciting opportunity to guide theoretical developments. The second major objective of this research is to use the wealth of sequence data available to explore the role and impact of RNA secondary structure on evolution. To this end, empirical studies and simulations are undertaken to explore the role of RNA secondary structure in the evolution of 16S-like rRNA-encoding genes. Finally the inference of spatially resolved populations from gene sequences is briefly investigated.

This research project has both computational and conceptual objectives. In both cases, the concrete result of these objectives will be new statistical models and computer software for evolutionary inference and a better understanding of the action of molecular and population processes during evolution.

0.2 Acknowledgements

Firstly I thank my supervisors, Dr Allen Rodrigo, Dr David Saul and Associate Professor Peter Wills. Without their guidance I would have doubtless got lost on some fruitless path, and without their support of my curiosity I would have strayed too little from the grip of Dogma.

For erudite exposition on the powers and pitfalls of sample-based Bayesian inference and Metropolis-Hastings Markov chain Monte Carlo, and for forthright criticism I am indebted to Dr Geoff Nicholls.

For letting me poke my nose into the world of ancient DNA, and learning something about scientific collaboration, I thank Professor David Lambert, Dr Craig Millar and Dr Peter Ritchie.

For providing the brown bear sequence data and valuable commentary I thank Dr Alan Cooper.

For lively discussion and being great people I thank Roald and Mette.

For hanging out in retro clubs and comedy bars in between writing papers I thank Roman Biek.

I thank all my fellow students at the Santa Fe Institute Complex Systems Summer School 2000 for a great voyage into the world of complexity theory.

For making me feel welcome during my visit to Oxford and making me feel like one of the guys, I thank Andrew Rambaut, Oliver Pybus, Korbinian Strimmer and Mike Charleston.

For helping me relax with games of hacky during my writeup, I thank Josh Guilbert, Charlotte Hardy and David Clarke and for Starcraft sessions I thank Revel Drummond, Peter Mclaren and Athena Ferreira.

For just being my friends and supporting me, I thank all the people I hang out with whenever I am not writing a thesis :-) you know who you are.

For helping debug my thesis at the going away party I thank Caroline Holliss, Ellen Campbell, Jeremy, Joline, Jim Mooney, Jo Simons, Lee Moss and Phil Sneddon.

For tireless effort during final proofing I thank Caroline Holliss, Revel Drummond and Athena Ferreira.

For loving me unconditionally I thank my Mum and Dad, and my grandparents Yo & Grant Drummond and Jean & James Babson. For pointing out to me that section 9.2.2 is wrong, but may still be useful, I thank Dad. For excellent editorial comments on Chapter 1 and 7, I thank Y.F.D.

For help and support, and the cutest smile, (and for reminding me to blink, breath and focus) I thank my sweetheart, Athena Ferreira.

0.3 Table of Contents

0.1 ABSTRACT II
0.2 ACKNOWLEDGEMENTS III
0.3 TABLE OF CONTENTS V
0.4 LIST OF ABBREVIATIONS IX
0.5 LIST OF SYMBOLS AND FUNCTIONS X
0.6 LIST OF TABLES XI
0.7 LIST OF FIGURES XII

1 INTRODUCTION 1

1.1 OVERVIEW 2
1.2 THEORETICAL POPULATION GENETICS 2
1.2.1 Fisher-Wright-Haldane population genetics 3
1.2.2 Kimura and the neutral theory 4
1.2.3 Kingman and the coalescent 5
1.2.4 Computational population genetics 6
1.2.5 Complexity theory and population genetics 8
1.3 PHYLOGENETIC SYSTEMATICS 9
1.3.1 Models of molecular evolution 10
1.3.2 Distance-based estimation of evolutionary trees 14
1.3.3 Maximum parsimony estimation of evolutionary trees 15
1.3.4 Maximum likelihood estimation of evolutionary trees 16
1.3.5 Uncertainty in phylogenetic reconstruction 19
1.3.6 Bayesian inference in phylogenetics 20
1.4 THE INFORMATION AGE OF BIOLOGY 20
1.4.1 Sequence databases 21
1.4.2 Parallelisation of computer power 21
1.4.3 Open source development 22
1.5 STATISTICAL INFERENCE IN EVOLUTIONARY BIOLOGY 22
1.5.1 Maximum likelihood 23
1.5.2 Bayesian statistical inference 23
1.5.3 Computational Techniques 25
1.6 CONCLUSION 26

2 MEASURABLY EVOLVING POPULATIONS 29

2.1 INTRODUCTION 30
2.1.1 Rapidly evolving pathogens 30
2.1.2 Fossil and Pre-fossil DNA 30
2.2 CONCEPTS 32
2.2.1 Non-independence of temporally spaced sequences 32
2.2.2 Estimating mutation rates using temporally spaced sequences 33
2.2.3 Estimating generation length using temporally spaced sequences 35
2.2.4 Estimating population size dynamics using temporally spaced sequences 35
2.2.5 Estimating divergence times using temporally spaced sequences 38
2.2.6 Phylogenetic inference using temporally spaced sequence data 39
2.2.7 MEPs and the neutral theory of evolution 39
2.2.8 Hypothesis testing and experimental design 40
2.3 CONCLUSION 40

3 RECONSTRUCTING GENEALOGIES OF SERIAL SAMPLES USING SERIAL-SAMPLE UPGMA (SUPGMA). 43

3.1 Overview 44
3.2 Introduction 44
3.3 Serial sample UPGMA 45
3.3.1 Estimation of δs. 47
3.3.2 Correction of pairwise distances 49
3.3.3 Cluster using UPGMA. 49
3.3.4 Trim back branches. 49
3.4 Estimating Population Parameters and Mutation Rate 50
3.5 Efficiency of Tree Reconstruction 51
3.6 Efficiency of Parameter Estimation 54
3.7 An Example Dataset 57
3.8 Discussion 59
3.9 Acknowledgements 61

4 INFERENCE OF STEP-WISE CHANGES IN HIV-1 *ENV* MUTATION RATE USING MAXIMUM LIKELIHOOD 62

4.1 Overview 63
4.2 Introduction 63
4.3 Likelihood model 65
4.4 Least-squares model 68
4.5 Example 69
4.6 Discussion 74
4.7 Acknowledgements 77

5 BAYESIAN EVOLUTIONARY INFERENCE OF MEASURABLY EVOLVING POPULATIONS 78

5.1 Overview 79
5.2 Introduction 79
5.2.1 Kingman coalescent with temporally offset leaves 80
5.2.2 DNA Substitution Model 82
5.2.3 Bayesian Inference for scale parameters 84
5.3 Markov chain Monte Carlo for evolutionary parameters 85
5.3.1 Proposal mechanisms 88
5.3.2 Implementation, convergence checking and debugging 91
5.4 Exponential growth and relative rates of substitution 93
5.5 Examples 94
5.5.1 HIV-1 *env* data 94
5.5.2 Simulated sequence data 102
5.6 Discussion 104

6 EXTENDING BAYESIAN EVOLUTIONARY INFERENCE BY EXAMPLE 107

6.1 Overview 108
6.2 Ancient DNA of Beringian Brown Bears: A case study 108
6.2.1 Data 108
6.2.2 Results 109
6.2.3 Discussion 116
6.3 Molecular dating of undated and old sub-fossil bones 119
6.3.1 Methods and materials 120
6.3.2 Priors 120
6.3.3 Results 121

6.3.4 Discussion 122
6.4 Hepatitis C in Egypt: A non-MEP analysis 123
6.4.1 The piecewise logistic demographic model 123
6.4.2 Codon position rate heterogeneity 125
6.4.3 Data 126
6.4.4 Results 127
6.4.5 Conclusion 129
6.5 Discussion 130
7 EVOLUTION OF RNA SECONDARY STRUCTURE 132
7.1 Introduction 133
7.2 Secondary structure prediction 133
7.2.1 Thermodynamic structure prediction 135
7.2.2 Kinetic structure prediction 137
7.2.3 Phylogenetic (or comparative) structure prediction 138
7.2.4 Hybrid techniques 139
7.3 Properties of RNA sequence-structure maps 139
7.3.1 Structural Discontinuity and Punctuated Equilibrium 141
7.4 Prokaryotic systematics 142
7.5 Discussion 143
8 RNA-BASED EVOLUTIONARY INFERENCE 145
8.1 Introduction 146
8.2 Structurally-specific RNA substitution models 146
8.2.1 Methods 147
8.2.2 Models for structural environment assignment 148
8.2.3 Empirically-derived substitution models constructed from a large sequence alignment 149
8.2.4 Simulations of RNA substitution models 153
8.2.5 A maximum-likelihood comparison on a small dataset 155
8.2.6 Conclusions 156
8.3 Multiple sequence alignment of RNA 158
8.3.1 Pair-wise alignments 158
8.3.2 Multiple sequence alignments 159
8.3.3 Statistical alignment 159
8.3.4 Sequence alignment of RNA sequences 160
8.3.5 Combining RNA secondary structure and sequence information 161
8.3.6 Conclusions 166
8.4 Bayesian inference of substitution parameters and phylogeny from sequence-structure data 167
8.4.1 Bayesian inference of combined sequence-structure characters 167
8.4.2 Data 168
8.4.3 Results 168
8.5 Discussion 170
9 A TANGENT: SPATIAL POPULATION GENETICS 172
9.1 Overview 173
9.2 A simple model of movement for inference 173
9.2.1 Bayesian estimation of diffusion in *n*-dimensions on a tree 173
9.2.2 Ecological data and spherical geometry 176
9.3 Simple models of movement for simulation 178
9.3.1 Lattice models 178

9.3.2 The box neighbourhood 178
9.3.3 The Gaussian neighbourhood 181
9.3.4 Restrictions of "constant-organisation" models 182
9.4 DISCUSSION 182

10 MEPI 183

10.1 INTRODUCTION 184
10.2 PROGRAMS 184
10.2.1 mepi 185
10.2.2 mepix 185
10.2.3 tracer 187
10.2.4 treesummary 188
10.3 THE MEPIX FILE FORMAT 190
10.3.1 Example file: hiv1.mepix 191
10.3.2 mepix elements 192
10.3.3 The alignment element 193
10.3.4 The data element 194
10.3.5 The demographicmodel element 195
10.3.6 The evolutionmodel element 196
10.3.7 The frequencies element 197
10.3.8 The mcmc element 197
10.3.9 The mutationratemodel element 198
10.3.10 The node element 199
10.3.11 The operator element 200
10.3.12 The operators element 201
10.3.13 The parameter element 202
10.3.14 The prior element 204
10.3.15 The ratematrix element 205
10.3.16 The sequence element 206
10.3.17 The simulatedata element 207
10.3.18 The simulatetree element 208
10.3.19 The sitemodel element 209
10.3.20 The time element 210
10.3.21 The timedata element 211
10.3.22 The tree element 212
10.4 CONCLUSION 212

11 PERSONAL CONCLUSION 213

11.1 OPEN PROBLEMS 213
11.1.1 The rate invariance problem 214
11.2 FUTURE DIRECTIONS 216
11.3 CONCLUSION 217

12 REFERENCES 218

13 APPENDIX: ADDITIONAL PUBLISHED PAPERS 229

0.4 List of Abbreviations

3D	three-dimensional
AIDS	acquired immunodeficiency syndrome
bp	base pairs
BP	before present
BSC	biological species concept
CPU	central processing unit
DNA	deoxyribonucleic acid
EM	expectation-maximization algorithm
ESS	effective sample size
F81	Felsenstein 1981 (model of substitution)
GPL	GNU public licence
GTR	general time-reversible (model of substitution)
HCV	Hepatitis C virus
HIMDU	hairpin, internal bulge, multi-stem loop, downstream-paired and upstream-paired (model of structure)
HIV-1	human immunodeficiency virus, subtype 1
HKY	Hasegawa, Kishino and Yano (model of substitution)
HOM	homogeneous model of structure (i.e. no structure)
HPD	highest posterior density
HVR1	hyper-variable region 1 (of the mitochondrial control region)
IACT	integrated autocorrelation time
indels	insertion-deletion events
JC	Jukes-Cantor (model of substitution)
LGPL	lesser GPL
LS	least-squares
MCMC	Markov chain Monte Carlo
MEP	measurably evolving population
ML	maximum likelihood
MRCA	most recent common ancestor
MRDT	multiple rate dated-tips (model of mutation rate)
mtDNA	mitochondrial DNA
NIH	National Institutes of Health
NJ	neighbour joining (method of phylogenetic reconstruction)
NNI	nearest neighbour interchange (method of branch swapping)
PAL	Phylogenetics Analysis Library
RNA	ribonucleic acid
rRNA	ribosomal RNA
SDI	symmetric difference index
SPR	subtree-prune and reattachment (method of branch swapping)
SR	single rate (model of mutation rate)
SRDT	single rate dated tips (model of mutation rate)
sUPGMA	serial-sample UPGMA
tRNA	transfer RNA
UP	unpaired/paired model of structure
UPGMA	unweighted pair-group method using arithmetic means
WPGMA	weighted pair-group method using arithmetic means

0.5 List of symbols and functions

Unless defined otherwise in the text, these symbols are defined as in the table below.

~	distributed as
$\wedge$	logical 'and' operation
$\vee$	logical 'or' operation
α	the shape parameter of the gamma distribution of rate heterogeneity among sites.
δ	divergence (measured in substitutions/site)
E_g	a set of edges defining a bifurcating tree
Exp(x)	exponentially distribution with a mean of x
g	a genealogy = (E_g, t_Y)
κ	kappa, the ratio between the instantaneous rate of a particular transition and the instaneous rate of a particular transversion.
μ	mutation rate
N_A	ancestral effective population size
N_C	current effective population size
N_e	effective population size
π	equilibrium base frequencies
Q	the instantaneous substitution rate matrix
Θ	intra-specific diversity: $2N_e\mu$ for haploid, $4N_e\mu$ for diploid
r	exponential growth rate
R	relative rate matrix
t_I	the set of times/ages of the leaves of a genealogy.
t_{MRCA}	time to the most recent common ancestor
t_{root}	synonym for t_{MRCA}
t_Y	the set of times/ages of the ancestral nodes of a genealogy.
Unif(x, y)	uniform distribution with a lower limit of x and an upper limit of y
ω	mutation rate (in mutations per site per calendar unit)
Z	normalizing constant

0.6 List of Tables

Table 2.1 Mutation rates of various genetic regions in different organisms. 32
Table 2.2 Historical and current uses of MEP methods. 33
Table 3.1 Sampling stategies under which phylogenetic reconstruction was tested. 52
Table 3.2 Threshold values of total divergence over which sUPGMA outperforms UPGMA. 53
Table 3.3 Parameter estimates under the δ-*parameterisation* for 1000 simulated datasets of four samples of five sequences. 55
Table 3.4 Parameter estimates under the ω-*parameterisation* for 1000 simulated datasets of four samples of five sequences. 55
Table 3.5 Estimated parameters for example dataset. 58
Table 4.1 ML and LS estimates of substitution rates under the SR, SRDT and MRDT models. 71
Table 5.1 Parameter estimates for ten independent analyses of the pre-treatment dataset with a simple model. 98
Table 5.2 Parameter estimates for ten independent analyses of the pre-treatment dataset with a complex model. 100
Table 5.3 Simulation studies to assess the performance of Bayesian inference. 104
Table 6.1 Experiment 1: Analysis of 30 ancient sequences. 110
Table 6.2 Experiment 2: Analysis of 30 ancient sequences and 17 modern sequences. 113
Table 6.3 Experiment 3: Analysis of 30 ancient sequences and 17 modern sequences with gamma-distributed rate heterogeneity among sites. 115
Table 6.4 The maximum time frame over which different rate categories can provide phylogenetic and molecular rate information. 116
Table 6.5 Estimates of bone ages based on sequence comparison to known ages using MCMC. 121
Table 6.6 Parameter estimates for two datasets of partial E1 genes. 129
Table 8.1 Log-likehood values of HOM and UP models of RNA evolution. 156

0.7 List of Figures

Figure 2.1 The genealogies of two unlinked genes. 36
Figure 2.2 Genealogy-based population genetics. 38
Figure 3.1 An outline of the sUPGMA procedure. 46
Figure 3.2 Phylogenetic reconstruction performances of sUPGMA and UPGMA. 53
Figure 3.3 Θ estimates for four samples of five sequences for 1000 simulated trees. 54
Figure 3.4 δ_1, δ_2 and δ_3 estimates for four samples of five sequences. 56
Figure 3.5 Estimated mutation rate, ω, from 1000 simulations of four samples of five sequences. 57
Figure 3.6 Two sUPGMA trees constructed from an example dataset. 58
Figure 4.1 Three different models of substitution rates through time. 65
Figure 4.2 Maximum-likelihood solutions for the full example dataset. 72
Figure 4.3 The likelihood surface of μ parameters. 73
Figure 5.1 Diagrams of two proposal mechanisms used to modify tree topology during an MCMC analysis. 86
Figure 5.2 Marginal posterior densities for the pre-treatment dataset. 97
Figure 5.3 Marginal posterior density for the pre-treatment dataset assuming exponential growth rate and a GTR model of substitution. 99
Figure 5.4 Marginal posterior densities for the post-treatment dataset. 101
Figure 5.5 Marginal posterior density for the post-treatment mutation rate assuming a upper limit on t_{root}. 102
Figure 6.1 Sample tree from Experiment 1. 112
Figure 6.2 Sample tree from Experiment 2. 118
Figure 6.3 Two proposal mechanisms for estimating the relative contributions of codon positions 1, 2, and 3 to the overall rate of substitution. 125
Figure 6.4 Demography of the Egypt HCV epidemic. 128
Figure 7.1 A transfer RNA (tRNA) secondary structure. 135
Figure 7.2 The concept of common secondary structures. 141
Figure 8.1 Annotation of rRNA-encoding sequences with structure category information. 149
Figure 8.2 The collection of substitution statistics from a pair of sequences. 150
Figure 8.3 Empirical base frequency histograms of the UP and HIMDU models. 152
Figure 8.4 A comparison of the structural stability of the HOM and UP models. 154
Figure 8.5 Comparison of different *evaluation schemes* where only sequence information was considered. 162
Figure 8.6 Comparison of different *evaluation schemes* where both sequence and structure information was considered. 163
Figure 8.7 Comparison of different alignment *generation schemes.* 165
Figure 8.8 Comparison of sequence alignments versus sequence-structure alignments. 166
Figure 8.9 Estimated relative rate matrix (R_{comb}) for combined characters. 169
Figure 8.10 Sample tree of life established using sequence and structure information. 170
Figure 9.1 The neighbourhood model. 179
Figure 9.2 Three different populations evolving spatially under the neighbourhood model. 180
Figure 9.3 The effect of mutation rate on spatial patterns of a single nucleotide site under the neighbourhood model. 181
Figure 9.4 The Gaussian neighbour model on a lattice. 182

Figure 10.1 “Enter times” dialog. 186
Figure 10.2 “Parameters” tab of the “MCMC Analysis settings” dialog. 187
Figure 10.3 A screenshot of tracer program. 188
Figure 10.4 A screenshot of treesummary program. 189
Figure 10.5 An example tree file. 189