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

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

0.4 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.

$\sim$	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
$\mathrm{Exp}(x)$	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.
$\mathrm{Unif}(x, y)$	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

0.7 List of Figures

1 Introduction

"*Nearly every great advance in science arises from a crisis in the old theory, through an endeavor to find a way out of the difficulties created.*"

(Albert Einstein, 1938)

1.1 Overview

What processes have produced the wonder of life we see around us? Since Darwin's *The Origin of Species* (DARWIN 1859), evolution has been central to understanding the diversity of life; it is a fundamental aspect of biology, from humans to bacteria to viruses. However, the detailed mechanisms of the evolutionary process are still poorly understood. In this thesis I will describe the development of novel statistical models and techniques for understanding some of the details of molecular evolution and population dynamics. The work I have undertaken is computationally based and brings together both theoretical and empirical aspects of population genetics and evolutionary inference. To understand the direction taken in this research, it is important first to consider the historical development of population genetics and evolutionary inference. What follows is a short history of these fields, written to establish the context and the relevance of the research described in the subsequent chapters.

1.2 Theoretical population genetics

The field of theoretical population genetics was born from the reconciliation of Mendelian genetics and Darwinian evolution. The particulate nature of Mendelian genetics had, in some eyes, threatened the action of Darwin's *natural selection* as a mechanism of evolution. However, a seminal paper written by Ronald A. Fisher in 1918 showed that Mendelian genetics and Darwinian evolution were actually quite complementary (FISHER 1918). It had previously been thought that the discrete nature of Mendelian characters and their assortment by sexual reproduction precluded the gradual change that dominated Darwinian evolutionary theory. The reconciliation of these ideas provided the foundation for the discipline of population genetics, which under the influence of Fisher, Wright and Haldane quickly took shape in the following decades. Each of these researchers had a distinctive focus, but the research of all three stemmed from an interest in the quantitative consequences of stochasticity and natural selection in populations with Mendelian inheritance.

It is interesting to ponder the fact that their ideas were developed at a time when almost nothing was known about the empirical reality of genetics in natural populations. Therefore most of this initial work was carried out with very little concrete knowledge of (i) selection strengths and differentials, (ii) genetic variation or (iii) gene flows, population structure and population subdivision. Surprisingly, much of the debate in theoretical population genetics remains because of the continued lack of discriminating empirical evidence. In this computational and data-rich age, many of the shortfalls can be resolved, but first let us consider the founders of population genetics.

1.2.1 Fisher-Wright-Haldane population genetics

"*The investigation of natural selection may be compared to the analytic treatment of the Theory of Gases, in which it is possible to make the most varied assumptions as to the accidental circumstances, and even the essential nature of the individual molecules, and yet to develop the natural laws as to the behaviour of gases, leaving but a few fundamental constants to be determined by the experiment.*"

(Ronald A. Fisher, 1922)

Fisher's views, collected in his book, *The Genetical Theory of Natural Selection* (FISHER 1930) were dominated by the notion that deterministic selection in large populations is the most important factor in evolutionary change. The primacy that Fisher gave to selection seems to have come almost entirely from his assumption that natural populations had large effective populations and selection effects were small, finely graded, independent and additive between genes. This assumption appears to have come, at least in part, from his view that the analytical treatment of natural populations could be compared to the thermodynamic Theory of Gases: also a statistical theory describing the properties of large populations (FISHER 1922a). In this setting of large panmictic populations, even very small selection pressures would overcome stochastic effects.

In contrast, Sewall Wright's *Shifting Balance Theory* (WRIGHT 1931), was based on the assumption that natural populations were fragmented rather than panmictic, having sub-divided populations with limited gene flow and complex genetic interactions that were not necessarily additive. In this situation, genetic drift and selection both play an important role in evolutionary change. Wright's assumptions of local populations required that non-adaptive genetic drift played a significant role in evolution, quite contrary to Fisher's view. This difference of opinion resulted in a scientific debate that played itself out over three decades. While the validity of Wright's specific model is still debated, in hindsight it seems that his view of population genetic reality was more accurate.

Along with Haldane, these two theoreticians are known as the founders of population genetics. These three researchers largely shaped the view of evolutionary theory in the 1920's and 1930's and laid the foundations of the *modern synthesis* of evolution. Despite their various differences, each researcher worked with very similar mathematical frameworks and the statistical methods they developed are still used today.

1.2.2 Kimura and the neutral theory

"*...we believe that definitely advantageous mutant substitutions are a minority when compared with a relatively large number of "non-Darwinian" type mutant substitutions, that is, fixations of mutant alleles in the population through the process of random drift of gene frequency.*"

(Motoo Kimura, 1974, p2851)

Motoo Kimura, a student of Hitoshi Kihara and Jim Crow, first introduced the *neutral theory of molecular evolution* in 1968 (KIMURA 1968). The idea behind the neutral theory is simply that the vast majority of substitutions that drive molecular evolution are not selectively advantageous. The neutral theory hypothesizes that genetic drift, not positive Darwinian selection, is the primary force behind molecular evolution and variation. The neutral theory does not preclude natural selection, but rather states that positive natural selection plays a minor role compared to genetic drift in producing evolutionary changes in molecules (KIMURA 1983; KIMURA and OHTA 1971; KIMURA and OHTA 1974). This assertion is at odds with the predominantly *selectionist* evolutionary theory of the early population geneticists, Fisher, Wright and Haldane, for which positive natural selection was the central force of change.

Kimura was developing his theory at a time when the molecular structure of DNA had already been discovered (WATSON and CRICK 1953) and molecular diversity data was becoming available (for example, LEWONTIN and HUBBY 1966). Although there was still only limited empirical data, it was becoming clear to some that *selectionist* models could not account for the similarities in the molecular heterozygosity of diverse species, and for the large amount of molecular variation found within individual populations. Although subsequent refinements of the neutral theory such as the *nearly neutral theory* (OHTA and KIMURA 1971) focused attention on negative (or purifying) selection to account for the conservation of protein structure and function, positive Darwinian selection remained absent. Testable hypotheses can easily be formulated for the neutral theory and its variants. Unfortunately many expectations of *neutralist* and *selectionist* theories overlap, especially in scenarios with fluctuating environment-mediated selection (GILLESPIE 1989). Another difficulty in testing the neutral theory (or any other theory) is that it cannot be separated from the underlying mathematical models used to express it. The assumptions of the model are chosen at least partially for mathematical tractability, not realism, and are invariably only approximations of the researcher's actual beliefs/assumptions about the process of evolution.

Despite some difficulties, the neutral theory has gained strong support. This is not only because of its attractive simplicity, but also because of its exceptional power to describe many of the observed patterns (for example, divergence, polymorphism and clock-like evolution) of molecular change without any recourse to positive Darwinian selection. This raises serious questions about the understanding of the role of natural selection in molecular evolution.

1.2.3 Kingman and the coalescent

One of the most significant recent developments in population genetics modelling was the introduction of *coalescent* or genealogical methods (KINGMAN 1982a; KINGMAN 1982b). The coalescent is a stochastic process that provides good approximations to the distribution of ancestral histories that arise from classical forward-time models such as the Fisher-Wright (FISHER 1930; WRIGHT 1931) and Moran population models (MORAN 1958). In this way, the coalescent links genealogies with the effective population size (N_e)[1]. The coalescent inherits the assumption of neutral evolution from the population models it is based on. The coalescent was an important step in its explicit use of genealogies[2] to estimate population parameters. This allowed the non-independence of ancestral relationships between genetic samples to be taken into account. Like the founders of population genetics before him, Kingman dealt predominantly with closed-form mathematical models that could be solved with pencil and paper. Simple population models such as Fisher-Wright-Moran can easily be simulated exactly in a computer, however an exact mathematical solution would be tedious: hence the coalescent approximation[3]. Slightly more complex population models, while still trivial to simulate, can be very difficult to solve analytically. In part, this thesis is an argument for a move to computational methods that allow more realism by sacrificing exact pencil and paper solutions.

[1] The effective population size (N_e) is the number of individuals that will produce offspring and thus contribute genes to the next generation (Wright, 1931). It is equal to the census population size (N) in an idealized randomly mating population. In real populations, N_e will differ from N because of factors such as overlapping generations, population structure, fluctuating population sizes, non-random mating and unequal sex ratios. In a fluctuating population, N_e is dominated by the smallest census sizes and N_e will be equal to the harmonic mean of N, all else being equal.

[2] 'Genealogy' and 'tree' are used interchangeably throughout. A haploid genealogy is a collection of edges, nodes and node times that together completely specify an acyclic rooted history of evolutionary relationships.

[3] But for the Moran model it should be noted that the exact distribution of coalescence times is almost as accessible as the coalescent.

1.2.4 Computational population genetics

This section will outline some of the recent developments in computational population genetics. The integration of likelihood-based phylogenetic methods and population genetics through the coalescent has provided fertile ground for new developments. Many coalescent-based estimation methods focus on a single genealogy (FELSENSTEIN 1992b; FU 1994; NEE *et al.* 1995; PYBUS *et al.* 2000) that is typically obtained using standard phylogenetic reconstruction methods. For example a maximum likelihood tree (under clock constraints) can be obtained and then used to obtain a maximum likelihood estimate of $\Theta = 2N_e\mu$ (where μ is mutation rate) using coalescent theory[4]. However, there is often considerable uncertainty in the reconstructed genealogy. In order to allow for this uncertainty it is necessary to compute the average likelihood of the population parameters of interest. The calculation involves integrating over genealogies distributed according to the coalescent (FELSENSTEIN 1988; FELSENSTEIN 1992a; GRIFFITHS 1989; GRIFFITHS and TAVARE 1994; KUHNER *et al.* 1995). Integration for some models of interest can be carried out using Monte Carlo methods. Importance-sampling algorithms have been developed to estimate the population parameter Θ (GRIFFITHS and TAVARE 1994; STEPHENS and DONNELLY 2000), migration rates (BAHLO and GRIFFITHS 2000) and recombination (FEARNHEAD and DONNELLY 2001; GRIFFITHS and MARJORAM 1996). Metropolis-Hastings Markov chain Monte Carlo (MCMC) (HASTINGS 1970; METROPOLIS *et al.* 1953) has been used to obtain sample-based estimates of Θ (KUHNER *et al.* 1995), exponential growth rate (KUHNER *et al.* 1998), migration rates (BEERLI and FELSENSTEIN 1999; BEERLI and FELSENSTEIN 2001; KUHNER *et al.* 1998) and recombination rate (KUHNER *et al.* 2000).

In addition to developments in coalescent-based population genetic inference, sequence data sampled at different times are now available from both rapidly evolving viruses such as HIV-1 (HOLMES *et al.* 1992; RODRIGO *et al.* 1999; SHANKARAPPA *et al.* 1999; WOLINSKY *et al.* 1996), and from ancient DNA sources (BARNES *et al.* 2002; HANNI *et al.* 1994; LAMBERT *et al.* 2002; LEONARD *et al.* 2000; LOREILLE *et al.* 2001). Temporally spaced data provides the potential to observe the accumulation of mutations over time,

[4] Although Θ is defined here simply as two times the product of N_e and μ, the factor of 2 is used only in haploid (asexual) populations, such as viruses. In a diploid population such as humans, a factor of 4 is used instead. The reason that N_e and μ must be estimated as a product is that divergence times in molecular genealogies are expressed in mutations rather than generations. In most cases the mutation rate per generation must be independently obtained to yield coalescent estimates of N_e directly from molecular genealogies. However, in Chapters 2, 3, 4 and 5 situations in which molecular genealogies can be used to obtain joint estimates of N_e and μ are discussed. Joint estimation of N_e and μ is a major topic of this thesis.

thus allowing the estimation of mutation rate (DRUMMOND and RODRIGO 2000; RAMBAUT 2000). In fact it is even possible to estimate variation in the mutation rate over time (DRUMMOND *et al.* 2001). This leads naturally to the more general problem of simultaneous estimation of population parameters and mutation parameters from temporally spaced sequence data (DRUMMOND *et al.* 2001; DRUMMOND and RODRIGO 2000; RODRIGO and FELSENSTEIN 1999; RODRIGO *et al.* 1999).

Chapter 5 is concerned with sample-based Bayesian inference of population and mutation parameters, dates of divergence and tree topology from sequence data. The important novelties in this kind of inference are the data type (for example, temporally sampled sequences and secondary structure information), the relatively large number of unknown model parameters, and the MCMC sampling procedures. The coalescent gives the approximate expected frequency with which any particular genealogy arises under the Fisher-Wright-Moran population model. The coalescent may then be treated, either as part of the observation process defining the likelihood of the population parameters, or as the prior distribution for the unknown true genealogy. In either case the likelihood must be integrated over the state space of the coalescent. Both Bayesian and purely likelihood-based population genetic inference use the same reasoning, and algorithms, up to the point where prior distributions are given for the parameters of the coalescent and mutation processes. Bayesian methods permit priors while likelihood analyses do not.

Bayesian reasoning has recently been applied to both phylogenetic inference (HUELSENBECK *et al.* 2000; MAU *et al.* 1999; THORNE *et al.* 1998; YANG and RANNALA 1997) and population genetic inference (WILSON and BALDING 1998).

The current, empirically-derived beliefs of most evolutionary biologists are not accurately represented by theoretical population genetic models. This is largely because of the simplifying assumptions required to make the models mathematically tractable. Thus the hypotheses generated from a rigorous mathematical model (that we know to be a simplification) only allow us to say that either the evolutionary theory, or the simplifications of the model, should be rejected. The increasingly complex hypotheses of evolutionary biologists can’t be adequately captured by the simplifying mathematical assumptions of most current population genetic models.

One example of this shortcoming is that the formulation of the neutral theory as a mathematical model involved a notable assumption. The neutral models of Kimura (KIMURA 1983; KIMURA and OHTA 1971) recognize only two types of mutation - those that are neutral and those that are definitely deleterious. Neutral (or nearly neutral) mutations can lead to substitutions in the model and deleterious mutations are removed by negative selection. Thus although it is clear that Kimura (KIMURA 1983; KIMURA and OHTA 1974) believed that positive Darwinian evolution must occur sometimes, his models completely disallow the possibility of advantageous mutants being fixed. Perhaps more importantly, the choice of mathematical model used to formalize population genetics theories can have pronounced quantitative and qualitative effects on predictions. This was well illustrated by John Gillespie who showed that the predictions of *shift models* and *house-of-cards models* of the nearly neutral theory were qualitatively different (GILLESPIE 1995). In light of this, it seems extremely important that models are chosen for realism rather than mathematical tractability!

Recent work on lineage-specific mutation rate models (HUELSENBECK *et al.* 2000; THORNE *et al.* 1998) suggests that the future of computational population genetics will hold increasingly complex statistical models of molecular evolution that have no tractable closed-form solutions. Thus computational population genetics will, to a large extent, break from the long-standing tradition of pencil and paper analysis. The large array of complex mechanisms involved in molecular evolutionary processes and the increasing gap between empirical knowledge and simplifying mathematical assumptions will speed this transition. Chapter 5 develops a general framework for Bayesian inference of molecular evolution and population dynamics from molecular sequences that attempts to answer some of these difficulties.

1.2.5 Complexity theory and population genetics

Complexity theory is the study of systems of interacting agents that give rise to self-organising phenomena. A frequently used concept to describe complex systems is that of *emergent phenomena*; “the whole is more than the sum of its parts”. The formalization of complexity theory is still in its infancy. There are a large number of seemingly disjoint phenomena that have been claimed to have primacy in the mechanistic underpinning of complexity. Nevertheless, from a physicist’s point of view complexity most often comes from non-linear dynamics in a many-agent system. With this view in mind, it seems likely that a large class of models in evolution and ecology should display emergent phenomena characteristic of complex systems. For example, reaction-diffusion systems are the prototype for a host of spatially distributed systems that occur in nature, and form the basis of many spatial models of ecology and evolution.

1.2.5.1 Ecological and spatial population genetics

Since the predator-prey Lotka-Volterra equation was solved independently by Lotka (1925) and Volterra (1926), theoretical models of ecology with complex dynamics have been widely demonstrated. A classic demonstration of this was Robert May's investigation of the logistic growth equation (MAY 1976). In conjunction with spatial diffusion, many ecological models are examples of reaction-diffusion equations. Models of this kind are in general non-linear and often exhibit self-organising properties (for example, HASSELL *et al.* 1994). Hassell, Comins & May showed by simulation that parasite-host dynamics could generate stable spatial patterns, qualitatively similar to niche-adaptation, in completely homogeneous environments. Results of this kind challenge the strictly adaptive interpretations of biological distributions in nature.

In addition, non-linear population models have been used to uncover chaotic behaviour in experimental systems. For example, a system of three stochastic difference equations were used to model flour beetle (*Tribolium* sp.) population dynamics, demonstrating extensive chaotic behaviour (CONSTANTINO *et al.* 1997).

Recently, even some very simple models that population geneticists have regarded as 'solved' have been shown to behave in surprising ways when spatial properties are considered. These results suggest that the genetic distribution of spatially structured populations is not simply interpreted. Some preliminary work in this direction is described in Chapter 9.

1.3 Phylogenetic systematics

Phylogenetic systematics is the study of the evolutionary relationships between different populations, species and higher-level taxonomic groups. Modern systematics evolved from the synthesis of taxonomic classification and evolutionary theory. Building on the development of theoretical population genetics, many evolutionary biologists became involved in a movement now known as the *modern synthesis* of evolution, or alternatively, Neo-Darwinism. One of the chief figures in this synthesis was Ernst Mayr, who in 1942 published *Systematics and the Origin of Species*. In it, Mayr developed the *biological species concept* (BSC). The modern synthesis began a conceptual overhaul of the field of systematics (taxonomy) that is still occurring today. Systematics has almost completed the transformation from an essentialist philosophy of *classification* (following the tradition of Plato through to Linnaeus) to an enterprise predominantly involved in understanding and inferring the *evolutionary* relationships between living and extinct populations, species and higher taxonomic groups. This reflects the transition from a static rationalist view of biological diversity based on a notion of *kinds* or *archetypes* to a dynamic empiricist view.

With this revolution came a surge of methodological developments in the inference of evolutionary trees. Some of these methods arose from taxonomic and classification problems. In an attempt to automate and systematise the practice of classification the field of numerical taxonomy was born (Sneath and Sokal 1973; Sokal 1961; Sokal and Michener 1958). The methods were empirical and distance-based and were most often applied to morphological characters. As the availability of molecular sequence data has increased it has arguably become the data type of choice because of its ubiquitous nature and relative ease of modelling. As early as 1967, a least-squares technique for phylogenetic reconstruction from protein sequences was described (Fitch and Margoliash 1967). In this section I give a brief overview of three classes of phylogenetic reconstruction techniques: distance-based, maximum parsimony and maximum likelihood. I am primarily interested in molecular sequences and will largely restrict myself to molecular sequence data for the following discussion. A precursor to understanding these methods is some knowledge of explicit models of molecular evolution.

1.3.1 Models of molecular evolution

The simplest measure of distance between a pair of molecular sequences is the number of sites at which they differ. This is known as the *Hamming distance* (*H*). This raw score can be normalised for the length of a sequence (*L*) to get the proportion of sites that differ between the two sequences, $p = H / L$. Consider two hypothetical nucleotide fragments of length $L = 20$:

```
1                  20
ACGTCGTAAGCGTACTCAGC
ACGTAGCTAGCTTACTCAGC
    * **   *
```

In these sequences $H = 4$ and $p = 4/20 = 0.2$. The proportion of sites that are different, p, is an estimate of the evolutionary distance between these two sequences. A single nucleotide site can, given enough time, undergo multiple substitution events. Because the alphabet of nucleotide sequences is small, multiple substitutions can rapidly be hidden by reversals and parallelisms. If this is the case, some substitutions will not be observed. Therefore the estimate of 0.2 substitutions/site in this example could be an underestimate. This is easily recognised if one considers two hypothetical sequences separated by a very large evolutionary distance – for example 10 substitutions per site. Even though the two sequences will be essentially random with respect to each other they will still, by chance alone, have matches at about 25% of the sites. This would give them an uncorrected distance, p, of 0.75 substitutions/site, despite being actually separated by 10 substitutions/site.

To compensate for this tendency to underestimate large evolutionary distances, a technique called *distance correction* is used. Distance correction requires an explicit model of molecular evolution. The simplest of these models is the Jukes-Cantor (JC) model (JUKES and CANTOR 1969). Under the JC model the evolutionary distance between two nucleotide sequences is:

$$d = -\frac{3}{4}\ln(1-\frac{4}{3}p)$$

This model assumes that all substitutions are equally likely and that the frequencies of all nucleotides are equal and at equilibrium. All the models considered in this section are simple time-reversible Poisson jump processes, independent and identical across sites. The nucleotide character at site s mutates in forward time according to a Poisson jump process with 4×4 instantaneous rate matrix Q. Here, $Q_{i,j}$ is the instantaneous rate for the transition from state i to state j, and $A \leftarrow 1, C \leftarrow 2, G \leftarrow 3, T \leftarrow 4$. Let $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ be a 1×4 vector of base frequencies, corresponding to the stationary distribution of the mutation process $\pi Q = (0,0,0,0)$. These are termed the equilibrium base frequencies.

The time units of the rate $Q_{i,j}$ can be chosen so that the mean number of mutations per unit time occurring at a site is equal to one. The JC model is then:

$$Q = \begin{bmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{bmatrix}, \quad \pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$$

1.3.1.1 Unequal base frequencies

The JC model is quite unrealistic once empirical data is considered. One obvious departure from this model is the observation of unequal base frequencies. Felsenstein (1981) suggested a model (F81) that allows unequal base frequencies:

$$Q \propto \begin{bmatrix} -\pi_C - \pi_G - \pi_T & \pi_C & \pi_G & \pi_T \\ \pi_A & -\pi_A - \pi_G - \pi_T & \pi_G & \pi_T \\ \pi_A & \pi_C & -\pi_A - \pi_C - \pi_T & \pi_T \\ \pi_A & \pi_C & \pi_G & -\pi_A - \pi_C - \pi_G \end{bmatrix}$$

Where $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$. All subsequent models considered will assume unequal base frequencies.

1.3.1.2 Transition/transversion bias

In addition to unequal base frequencies, there is often a bias in transitions (A↔G, C↔T) over transversions (A↔C, A↔T, C↔G, G↔T). Hasegawa, Kishino and Yano (1985) suggested a model (HKY) that allowed for unequal base frequencies and a transition/transversion bias κ:

$$Q \propto \begin{bmatrix} -\pi_C - \kappa\pi_G - \pi_T & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & -\pi_A - \pi_G - \kappa\pi_T & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & -\kappa\pi_A - \pi_C - \pi_T & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & -\pi_A - \kappa\pi_C - \pi_G \end{bmatrix}$$

1.3.1.3 General time-reversible model

The general time-reversible model (RODRIGUEZ *et al.* 1990) allows both unequal base frequencies and individual rates for each pair of nucleotides. Pairs (A↔C, A↔G, A↔T, C↔G, C↔T) have rates (a, b, c, d, e) relative to G↔T. This model is the most general time-reversible model.

$$Q \propto \begin{bmatrix} -a\pi_C - b\pi_G - c\pi_T & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -a\pi_A - d\pi_G - e\pi_T & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -b\pi_A - d\pi_C - \pi_T & \pi_T \\ c\pi_A & e\pi_C & \pi_G & -c\pi_A - e\pi_C - \pi_G \end{bmatrix}$$

1.3.1.4 Rate heterogeneity among sites

In the simplest case, the models of molecular evolution discussed above are assumed to act independently and identically at all sites in a molecule. In this situation each site can be regarded as an independent sample of the substitution process. However rate heterogeneity among sites has become an unavoidable conclusion drawn from empirical data. In reality, rates of evolution often vary across sites, for example due to structural constraints (negative/purifying selection), sequence-specific mechanisms of replication, and/or positive selection. The unjustified assumption of identical rates across sites can lead to errors in phylogenetic reconstruction (YANG 1993). One of the most widely used models developed to account for rate heterogeneity among sites is the gamma distribution model of rate heterogeneity (YANG 1993). The gamma distribution shape parameter (α) can be used to define a continuum of rate distributions from a single uniform rate across sites ($\alpha \to \infty$) to a distribution in which most sites have low rates but a few sites have extremely high rates ($\alpha \to 0$). As α increases the gamma distribution becomes more symmetrical and peaked around the central rate.

1.3.1.5 The molecular clock and rate heterogeneity among lineages

The molecular clock hypothesis states that the rate of molecular evolution is uniform across different lineages and that therefore a pair of taxa sharing a common ancestor should have equal molecular divergences from that ancestor. Evidence for the existence of a molecular clock is one of the main arguments in support of the neutral theory of evolution (KIMURA 1968).

While it is widely accepted, and in fact obvious, that there must be a correlation between time and molecular divergence, there is argument over the degree and applicability of this correlation. In rapidly evolving organisms, such as some RNA viruses, there is good evidence for the existence of a molecular clock in at least some cases (GOJOBORI *et al.* 1990; LEITNER and ALBERT 1999). However, even among viruses there is some departure from the strictest interpretation of the molecular clock hypothesis (JENKINS *et al.* 2002). Despite these problems, the molecular clock hypothesis has been very successful as a null hypothesis of molecular evolution. It has also led to a healthy scientific industry of inferring dates of evolutionary events from molecular phylogenies (for example, KORBER *et al.* 2000).

Recently, modifications of the molecular clock hypothesis, known informally as 'relaxed molecular clocks' have been described (HUELSENBECK *et al.* 2000; THORNE *et al.* 1998). These modifications to the molecular clock attempt to model lineage-specific substitution rates by assuming that the rate of evolution itself is subject to evolution over time. There are a number of reasons why one might expect lineage-specific substitution rates, even under the neutral theory. One of the primary reasons may be that body temperature and body mass, by affecting metabolic rate, affect the underlying mutation rate. According to the neutral theory the substitution rate is equal to the mutation rate, and thus large changes in body temperature and body mass would be expected to result in changes in the substitution rate (GILLOOLY *et al.* 2001).

For large divergences, across diverse taxa, the strict molecular clock hypothesis is often rejected in favour of models that allow for lineage-specific rates. This suggests that there is a growing need for explicit statistical models of rate heterogeneity among lineages, such as the 'relaxed molecular clock' models, for molecular phylogenies to continue to be useful for dating evolutionary events of the past.

The molecular clock hypothesis is an explicit assumption of the methods developed in Chapters 3, 4, 5 and 6. The molecular clock hypothesis is a useful starting point, and a fairly accurate assumption for the organisms considered in those chapters, but in light of the above discussion, it should also be regarded as a limitation of the methods presented.

1.3.2 Distance-based estimation of evolutionary trees

Phylogenetic reconstruction methods can be grouped by the data type that they use and by whether they use an optimality criterion. Distance-based methods are a class of heuristic (or so-called 'algorithmic') phylogenetic estimation methods that use pair-wise distances between sequences as input. Most of these methods do not involve an optimality criterion (however least-squares methods do). These methods discard sequence character information, such as the distribution of character states across taxa at a particular nucleotide site in the sequence alignment. Their benefit lies entirely in being computationally efficient.

It is possible to construct an estimate of the phylogenetic tree of a group of sequences using a simple clustering algorithm on pair-wise distances. The 'Unweighted Pair-Group

Method using Arithmetic averaging' (UPGMA) is one of the earliest such pair-grouping algorithms (SOKAL and MICHENER 1958). UPGMA is best described as an *ad hoc* heuristic procedure. Roughly speaking, the algorithm proceeds by selecting the closest matching pair at each step and amalgamating them into a single cluster. When all sequences/clusters have been amalgamated into a single cluster the algorithm halts. UPGMA has the property that two clusters amalgamated during this procedure will be sister clades in the resulting tree. UPGMA can be shown to have a statistical justification, under a molecular clock. It is consistent in that it will reconstruct the correct tree at the limit of infinite sequence length.

The related and widely used method Neighbour Joining (NJ: SAITOU and NEI 1987) uses essentially the same algorithm, but has an additional re-weighting step after each amalgamation that accounts for rate heterogeneity among lineages. This method is much more widely used than UPGMA because it doesn't assume a strict molecular clock. Recently, two improvements have been described for the NJ method. One improvement involves a fast method to calculate multiple low-cost tree topologies (PEARSON *et al.* 1999). The second method takes advantage of statistical models of molecular evolution to agglomerate by a minimum variance technique, thereby improving topological reconstruction accuracy (BIONJ: GASCUEL 1997). Although some of these sophisticated agglomeration methods are impressively accurate for the amount of computer time used, they suffer from a lack of power.

1.3.3 Maximum parsimony estimation of evolutionary trees

The maximum parsimony criterion is an optimality criterion for evolutionary trees, and has been widely used for evolutionary inference and classification. Under maximum parsimony, a candidate tree is evaluated by considering the minimum number of evolutionary events it requires to explain the observed data. The most parsimonious tree is the tree that requires the minimum number of evolutionary events to explain the observed data. In the case of molecular sequence data, the maximum parsimony criterion selects the tree that requires the least substitutions to have occurred to explain the observed sequences. If the number of characters in the dataset is small or the number of taxa is large, there may be many equally parsimonious trees. Typically, more data must be collected to distinguish between equally parsimonious trees.

Maximum parsimony has a long, and at times acrimonious, history in phylogenetics. The field of cladistics serves as the philosophical underpinning of the maximum parsimony criterion. Proponents of cladistic philosophy have argued that one of the advantages of parsimony is its lack of an explicit evolutionary model, because they say that any particular model is bound to be wrong (FARRIS 1973). Another argument in support of maximum parsimony is that simpler hypotheses are better. Superficially this second argument may hold some appeal to the scientifically minded. However, it has been shown (FELSENSTEIN 1978) that maximum parsimony is inconsistent (i.e. increasing data does not increase accuracy of estimation) for almost all reasonable models of evolution. This becomes apparent when one realises that maximum parsimony makes the implicit assumption that evolutionary events are extremely rare. This is simply not true over long periods of evolutionary time, and in the case of rapidly evolving pathogens such as human immunodeficiency virus type 1 (HIV-1) it is not even true over short periods of time (for example a few years). Neither of these arguments in support of parsimony will be credited with any more discussion.

One of the possible advantages of maximum parsimony is that as an optimality criterion it can be calculated rapidly. For example, using branch and bound techniques, exact results can be obtained for quite large trees. The branch and bound algorithm eliminates regions of 'tree space' during a depth-first search, without evaluating them, based on knowledge of the best solution so far. It was first applied to phylogenetic trees by Hendy and Penny (1982).

It has also been shown that various modifications of maximum parsimony, such as weighting rare evolutionary events more than frequent evolutionary events, can make parsimony more consistent and more accurate at phylogenetic estimation (HILLIS *et al.* 1994). However, the focus of this thesis is on explicit statistical modelling of evolution, which being philosophically remote to the maximum parsimony criterion will preclude its further discussion.

1.3.4 Maximum likelihood estimation of evolutionary trees

If a technique to calculate the likelihood $P(D|T)$ of a tree is available, then the maximum likelihood (ML) optimality criterion can be used to choose a tree that makes the data most probable. The earliest systematic attempt to use maximum likelihood in estimation of evolutionary trees was by Edwards & Cavalli-Sforza (CAVALLI-SFORZA and EDWARDS 1967; EDWARDS and CAVALLI-SFORZA 1964). These papers were seminal in attempting to apply standard statistical inference methods to the problem of estimating the branching history of genetically related populations. The difficulties that these authors found with maximum likelihood phylogenetic reconstruction were due in part to the ambitious nature of their models. In their treatment, gene frequencies were transformed into a Euclidean space and then treated as if undergoing Brownian motion on a branching tree. The Yule process (YULE 1924) was used as a prior on tree topologies, and both the time of each branching event and its position in the Euclidean genetic space were objects of inference. In this framework, the authors found that the likelihood surface had singularities and short of describing the entire likelihood surface, a pure maximum likelihood treatment seemed doomed. In light of the current work of this thesis it is noted with some satisfaction that the authors suggested the use of Monte Carlo methods to alleviate these problems. However, in this intellectually stimulating early work they restricted themselves to a hybrid method of maximum likelihood.

In 1973 Joseph Felsenstein published two papers describing general techniques for calculating the likelihood $P(D|T)$ of a tree for discrete (FELSENSTEIN 1973a) and continuous characters (FELSENSTEIN 1973b). He avoided the problems of earlier workers by ignoring the population dynamic aspects of the evolutionary process and by collapsing mutation rate and divergence times together. In 1981 he followed this work up with a paper describing a technique for calculating the maximum likelihood tree from molecular sequence data under the assumption of unequal base frequencies (FELSENSTEIN 1981). One of the key strengths of ML is the use of explicit statistical models (such as Poisson process models of mutation at a site) to describe evolutionary processes. This means that is it relatively simple to modify the underlying statistical model and assumptions. Ironically, as mentioned previously, early detractors saw this as a disadvantage (FARRIS 1973). Farris argued that any particular model was bound to be wrong in the vast majority of cases and advocated the principle of maximum parsimony as an alternative. I will not investigate this line of reasoning further for reasons stated in section 1.3.3.

The flexibility that ML provides can be seen in the complexity of models studied within the ML framework. Models of secondary structure for proteins (GOLDMAN *et al.* 1998; THORNE *et al.* 1996), models of secondary structure for RNA molecules (SAVILL *et al.* 2001; TILLIER and COLLINS 1995; TILLIER and COLLINS 1998), empirical models of protein evolution (for example, WHELAN and GOLDMAN 2001) and models of codon evolution (GOLDMAN and YANG 1994; YANG *et al.* 2000) have all been developed and investigated in a likelihood framework. Models with gamma-distributed rate variation among sites (YANG 1993; YANG 1994) and models with correlated rates among sites (FELSENSTEIN and CHURCHILL 1996) have been described. Recently models that include selection in the form of unequal rates of synonymous and non-synonymous substitutions have also been developed (NIELSEN and YANG 1998; YANG 1998).

1.3.4.1 Tree searching algorithms

Maximum likelihood is far more computationally intensive than the distance-based methods described in the previous section. Exhaustive evaluation of the likelihood for all trees of a given size is not feasible for large numbers of taxa (>>10). This has led to the development of a number of heuristic techniques. In his 1981 paper Felsenstein suggested a greedy algorithm of stepwise addition to generate a starting tree, followed by a number of branch-swapping operations in an attempt to find a near-maximum likelihood tree without searching all of tree space. Heuristic algorithms of this sort still dominate today in ML phylogenetic construction, despite the lack of a guarantee that the ML tree will in fact have been found by algorithms of this sort. Most of the techniques employed use some kind of *tree operation* to iteratively refine a starting estimate of the phylogenetic tree. A *tree operation* defines the set of neighbours, in 'tree space', of a given tree. For the purpose of traversal, repeated stochastic applications of a tree operator should eventually lead to all possible trees. The nearest-neighbour interchange (NNI) and subtree-pruning and regrafting (SPR) are two examples of tree operators implemented in PAUP* (SWOFFORD 1999) to carry out heuristic searches of 'tree space'.

1.3.4.2 Hypothesis testing

One of the strengths of the maximum likelihood method of phylogenetic reconstruction is the ability to objectively select between two models using the likelihood ratio test. The simplest form of the likelihood ratio test occurs when the models are nested. In this case, the test statistic is twice the difference in the log likelihood of the two models. This test statistic is asymptotically χ^2 distributed with degrees of freedom equal to the difference in number of parameters of the two models [but see (OTA *et al.* 2000) for an example where the nested model represents a boundary value and the resulting test statistic is not χ^2 distributed]. Thus, while a complex model will always have a likelihood score greater than or equal to a simpler nested model, the difference in likelihood may not be great enough to justify the extra parameters. The *p*-values obtained in this kind of test can be interpreted in the standard way with regard to type 1 and type 2 errors.

In the case of selecting between alternative tree topologies there is no obvious nesting of the two models. This precludes standard use of the likelihood ratio test. Both parametric and non-parametric statistical tests have been employed for selecting between tree topologies. One of the earliest tests of this variety was the Kishino-Hasegawa (KH) test (KISHINO and HASEGAWA 1989). The KH test was widely used to compare the maximum likelihood tree to a tree representing some *a priori* hypothesis. However it has recently been pointed out that this usage is incorrect as both tree topologies must be *a priori* specified (GOLDMAN *et al.* 2000). More recent tests for topology include the Shimodaira-Hasegawa (SH) test (SHIMODAIRA and HASEGAWA 1999), which takes into account multiple test scenarios. This area of phylogenetic inference still seems to be under scrutiny, as subsequent tests such as the SH test have been shown to contain biases and are still unsatisfactory (SHIMODAIRA and HASEGAWA 2001; STRIMMER and RAMBAUT 2002).

Despite some difficulties and computational burden, parameter estimation and model comparison in a likelihood framework have a number of attractive statistical properties that are not shared by other methods such as heuristic distance-based methods and maximum parsimony.

1.3.5 Uncertainty in phylogenetic reconstruction

Phylogenetic reconstruction is a statistical enterprise, subject to all of the problems inherent in statistical estimation in general, such as sampling error. In most cases, the amount of data analysed will be insufficient for a single estimate of the evolutionary relationships to be an honest description of our knowledge. One way of assessing the strength of support for an estimated phylogeny is non-parametric bootstrapping (FELSENSTEIN 1985). Bootstrapping involves the generation of pseudoreplicates of the original data matrix (the sequence alignment) by sampling with replacement. In this way, each pseudoreplicate will resemble the original dataset in the sites it contains and differ in the frequency. Each site in the original dataset may appear 0, 1 or many times in a pseudoreplicate. Each pseudoreplicate is used to estimate a phylogeny and the set of phylogenies obtained from repeated generation of pseudoreplicates will provide information about the sampling error associated with the original dataset. Strongly informative datasets will show little variation in estimates of phylogenies between pseudoreplicates. The result of bootstrapping is a set of trees. The "bootstrap support" for a clade is the proportion of trees in a bootstrap analysis that contain that clade. A single consensus tree that contains the clades with the highest bootstrap support is usually used to summarize the bootstrap trees. Bootstrapping can be used with many phylogenetic reconstruction methods, including distance-based, maximum parsimony and maximum likelihood methods.

A related method is jackknifing. Jackknifing is an alternative method for generating pseudoreplicates that subsamples the original data matrix without replacement. A jackknife pseudoreplicate is thus a copy of the original dataset with some sites deleted. Each site in the original dataset may appear 0 or 1 times. Jackknifing was suggested in a phylogenetic setting (PENNY and HENDY 1985) about the same time as bootstrapping, however it has received far less attention despite being potentially more efficient.

1.3.6 Bayesian inference in phylogenetics

Bayesian inference is a methodology that naturally lends itself to quantifying uncertainty. Bayesian phylogenetic inference is a recent phenomenon and coincides with a growth of Bayesian statistical methods in all areas of science. The aim of a Bayesian phylogenetic inference is not necessarily to establish the 'true' tree, but instead to provide information about all plausible trees given a data set. Bayesian inference allows analyses to proceed without exact knowledge of the tree topology. In many instances evolutionary hypotheses can be explored in a Bayesian setting by weighting a set of plausible trees by their relative probabilities. The tree topology is then effectively treated as a nuisance parameter that is taken account of but then discarded. This avoids the potential bias of assuming a particular tree topology, or even worse assuming no evolutionary correlations in the observed data. The concept of likelihood is still fundamental to Bayesian inference and the relationship between these two inferential frameworks will be elaborated in section 1.5 and again in Chapter 5.

Bayesian inference in phylogenetics is a rapidly growing area (HUELSENBECK *et al.* 2001; LEWIS 2001) and will become increasingly important in evolutionary inference. Although it has many advantages, they do not come without a price. Demonstrating that a particular Bayesian analysis is correct is notoriously difficult and often dependent on the input data. A large part of this thesis will deal with some novel applications of this methodology to answer fundamental questions in evolutionary biology.

1.4 The Information Age of Biology

The fields of computational biology and bioinformatics have in recent years become almost as central to biology as evolution. The recent elucidation of the complete human genomic sequence by two independent parties (LANDER *et al.* 2001; VENTER *et al.* 2001) was heralded as one of humankind's greatest achievements in biology. This endeavour was fundamentally one of computational biology and bioinformatics. The human genomic sequence constructed was in both cases pure information. They are idealized virtual genomes consisting of 23 sentences written in a 4-letter alphabet (A, C, G, T) and have no physical reality. The power of this 'discovery' is in its information content – that is, in the patterns that computers will find in it. Without computers the human genome sequence would be almost useless.

In the case of the private effort (VENTER *et al.* 2001) the assembly of the entire sequence was achieved by an automated computational method. Modern molecular biology is increasingly focused on sequence analysis of whole genomes. This world is fundamentally a computational one.

1.4.1 Sequence databases

GenBank is the National Institutes of Health (NIH) genetic sequence database and is an annotated collection of all publicly available DNA sequence. As of the 30th of May 2002, GenBank accommodated 2.971 x 10^{10} nucleotides and more than doubles in size every year (ROOS 2001). Along with the databases of the European Molecular Biology Laboratory (EMBL) and the National Institute of Genetics, Japan (DNA Data Bank of Japan: DDBJ), which both exchange entries with it, is the largest public sequence database. In the face of this vast sea of sequence data, specialist databases are becoming important filters allowing researchers to focus on areas of interest.

Molecular sequence databases are growing faster than computational power is increasing. Therefore the number of CPU cycles per nucleotide is actually decreasing at a worrying rate. This situation suggests that more focus on development of sequence analysis techniques is necessary to manage the available information. In addition, parallelisation of computer power will also become increasingly important.

1.4.2 Parallelisation of computer power

In the face of large increases in molecular sequence data and increasing complexity of computational analyses, parallelisation of computer power is becoming essential. Parallelisation allows an analysis that might otherwise be impossible on a single desktop computer to be conducted relatively cheaply on 'farms' of desktop computers. Consequently algorithms that lend themselves to parallelisation will be attractive to researchers in the field of bioinformatics. Some results presented in this thesis would not have been possible without access to medium-sized arrays (25-50) of desktop computers.

1.4.3 Open source development

In a research environment in which software development is becoming an increasingly important aspect of basic research, the software source code itself becomes a valuable asset of the scientific community. Open source licences such as the GNU Public Licence (GPL) and the Lesser GNU Public Licence (LGPL) are important benchmarks in scientific programming. The use of these licence structures or similar ones such as Apache-style (http://www.apache.org/LICENSE.txt) increase the rapidity of verification and development that are essential to science. A large part of the software development during this thesis has been made available as open source and the remaining has been made freely available as Java class libraries and executables.

1.5 Statistical Inference in Evolutionary Biology

How do we begin to understand the patterns and complexity of molecular sequence data? There are a number of computational inference techniques that can be employed to investigate statistical models of evolution. The kinds of algorithms that are employed are often determined by the kind of statistical inference and models that are being used. Broadly speaking statistics can be separated into "frequentist" and Bayesian schools of thought[5]. These two approaches are philosophically separated by their interpretations of truth and probability. This separation translates into a bias in the way investigators in each group describe evolutionary processes. This section describes the similarities and differences in these two approaches.

Before investigating these two inferential strategies it is constructive to define some terms. Both maximum likelihood and Bayesian inference provide inference through the concepts of a (statistical) *model* and a (statistical) *hypothesis*. The *model* is the part of the description of the observed data that is taken as known: for example, a tree topology, an alignment, or the Markovian nature of a substitution model. The *hypothesis* is the focus of uncertainty: for example, the branch lengths of the tree or the relative rates of different substitution rates in the Markov jump process. From this description it should be obvious that the boundary between model and hypothesis is blurry at best. In one situation the tree topology or alignment may be given, and therefore part of the model, and in another situation one or the other may be an object of inference and therefore part of the hypothesis.

[5] Some may regard this as an abuse of terminology. 'Likelihoodists' are really a separate philosophical branch of statistical theorists in their own right, apart from both traditional frequentists and 'subjective' Bayesians.

Estimation finds the 'best' parameter values of a given statistical model for some observed data. The difference between maximum likelihood and Bayesian inference is in the criterion by which we select the 'best' hypothesis (i.e. the 'best' parameter values of the model) given the data. Analogously, *model comparison* allows us to choose between alternative statistical models (model comparison is sometimes referred to, somewhat confusingly, as hypothesis testing), and again the criterion we use will depend on the inferential framework.

1.5.1 Maximum likelihood

The maximum likelihood criterion was introduced to statistics as an optimality criterion that gives evidentiary support precedence (EDWARDS 1972; FISHER 1922b; FISHER 1925). Likelihood is a unifying principle of statistical inference that focuses on making statements about relative support rather than absolute belief. This is reflected in the focus on log-likelihood ratios, known as *support* (EDWARDS 1972). The likelihood axiom states that for a given statistical model, all the information about the relative merit of two hypotheses is contained in the likelihood ratio of the hypotheses on the data. The likelihood ratio is the degree of support for one hypothesis over the other. The likelihood axiom is only concerned with weighing the two hypotheses and decision theory is not within its scope. The decision to *select* one hypothesis over another (for instance selecting one tree topology over another as discussed in section 1.3.4 above) must involve additional concepts, such as repeated sampling theory or information theory.

Maximum likelihood has been widely accepted in phylogenetics as a method of phylogenetic reconstruction and in population genetics as a method of estimating population parameters using coalescent theory. In relatively simple settings (such as branch length estimation) maximum likelihood lends itself to numerical optimisation techniques. One of the most attractive properties of maximum likelihood is the invariance of likelihood ratios under a transformation of variables (EDWARDS 1972; FISHER 1921).

1.5.2 Bayesian statistical inference

Bayesian inference is a methodology that focuses on quantifying uncertainty. The focus of a Bayesian inference is not to establish an estimate of a parameter of interest, but instead to provide information about the uncertainty of the parameter. An example of the interpretation that can be made of a Bayesian analysis is the relative probability of the parameter lying in one of two intervals.

Technically, Bayesian statistical inference is closely related to likelihood-based methods, in that the likelihood function plays a central role. Bayesian statistical inference generally involves the same statistical models and in some cases, even the same sampling machinery. However, philosophically Bayesian inference is almost the antithesis of maximum likelihood. Bayesian inference treats parameters of interest as random variables, requiring all aspects of the hypothesis to have prior probability densities – in some sense this denies that a parameter has a true value. It also requires that the investigator describe probability densities for all parameters of interest, even when no repeatable experiment exists. However, in practise this paradox can be difficult to grasp, especially when Bayesian inference is used with so-called non-informative priors. Probability calculus tells us that *Posterior probability* $\propto$ *Likelihood* $\times$ *Prior probability*. Thus the mathematical relationship between the two methods is striking, especially when the prior probability is uniform. In this situation it is often the case that Bayesian inference is simply sampling the likelihood surface, and the difference between the methods boils down to how the sampled distribution is summarized. Nevertheless, there is an active debate about the relative merits of the two methodologies.

Detractors of the Bayesian approach argue that Bayesian inference is subjective because a necessary step in a Bayesian analysis, prior elicitation, is subjective. For example, even when using so-called 'non-informative' priors, the posterior density may still be sensitive to boundary conditions applied to the prior. In general, the posterior density can be arbitrarily transformed by selection of an appropriate prior. Also, transformation of a parameter can render uninformative priors informative. Uninformative-ness is not invariant under transformations.

On the other hand, proponents of Bayesian inference would argue that prior elicitation allows a researcher to investigate different systems of assumptions. In this sense, Bayesian inference is a powerful exploratory tool and offers the researcher tools that are completely lacking in a maximum likelihood analysis. While it is true that the prior can modify posterior beliefs, it is also true that the prior should be scrutinized as a visible part of any Bayesian analysis.

One of the most flexible computational techniques for inference in a Bayesian framework is the sample-based inference technique described by Metropolis *et al* (1953) and Hastings (1970). Chapter 5 describes a Metropolis Hastings Markov chain Monte Carlo sampler developed for population genetic inference.

1.5.3 Computational Techniques

The following section describes computation methods for optimisation and integration.

1.5.3.1 Numerical optimisation

The term *numerical optimisation* describes a host of computational methods that aim to find maxima of an arbitrary function of interest, h:

$$\max_{\theta \in \Theta} h(\theta)$$

The function h may have no closed form but must be able to be evaluated for a given set of parameter values, θ. Traditionally the optimisation problem has been approached by using deterministic methods that make use of analytical properties of the target function, such as convexity and smoothness. However, these properties are not always easy to guarantee for functions of interest in evolutionary inference. It has recently been shown that these analytical properties are generally less important in Monte Carlo optimisation techniques where h has a probabilistic interpretation.

The initial description of maximum likelihood analysis of molecular sequences in the context of evolutionary trees (FELSENSTEIN 1981) was accompanied with an application of the expectation-maximisation (EM) algorithm (DEMPSTER *et al.* 1977) for branch length estimation. In its original form the EM algorithm is not a stochastic algorithm, and deterministic algorithms of this type have been frequently employed in maximum likelihood settings. Each branch length is taken in turn and optimised, conditional on the current values of the other branches. This reduces the problem to a series of univariate maximization problems that are individually trivial under the above assumptions of convexity and smoothness. The EM algorithm causes an increase of the likelihood at each step that ensures convergence to the maximum likelihood estimator for a unimodal likelihood function. However, in a multi-modal function, initial conditions will determine which local maximum is reached. In an unconstrained model, such as an unrooted tree with no molecular clock, the likelihood surface will often be uni-modal. It was assumed by some that the branch length likelihood surface was uni-modal for all phylogenetic trees of interest. However, Rogers and Swofford recently showed that data simulated on realistic trees could contain multiple local maxima in tree space (ROGERS and SWOFFORD 1999). Although computationally intensive, this can be overcome by multiple starting points. Additionally it has been shown that multiple *global* maxima, especially in the form of ridges in the likelihood surface, are also possible in branch length space (CHOR *et al.*

2000; STEEL 1994). Both of these sets of results are of practical importance because most software programs that implement maximum likelihood phylogenetic inference use simple hill-climbing techniques from a single starting point. As a result, maximum likelihood implementations can fail to return the correct maximum likelihood value during branch length estimation, and tree topology searches, and will typically do so without the user realising it.

1.5.3.2 Monte Carlo optimisation

Monte Carlo methods may be attractive when the function of interest is not uni-modal, as has been shown in a number of phylogenetic settings (CHOR *et al.* 2000; ROGERS and SWOFFORD 1999; STEEL 1994). Monte Carlo methods are stochastic and therefore rarely guarantee an exact solution. However, some trade-offs between deterministic and stochastic search strategies may be of interest in this context. The first and most obvious is hill-climbing from multiple starting points. Another possibility is the use of *simulated annealing*, in which the search begins in a stochastic mode, jumping randomly in space, and slowly becomes more deterministic in its hill-climbing through the gradual decrease in a 'temperature' parameter. Both of these methods will still suffer shortcomings if the number of maxima is large or the basins of attraction are small.

1.5.3.3 Metropolis-Hastings Markov chain Monte Carlo

Markov chain Monte Carlo is a general method of generating samples from a density of interest when the normalizing constant is unknown (HASTINGS 1970; METROPOLIS *et al.* 1953). It is a sampling method rather than a method of optimisation, and the Markov chain generated is expected to visit all volumes of parameter space in proportional to their probability density in the function of interest. This method is heavily utilised in Chapters 5, 6 and 8 to solve phylogenetic inference problems. Chapters 5 and 6 describe the estimation of mutation rate from temporally spaced sequences. This problem has been shown to have multiple local likelihood maxima for real data sets (Rodrigo *et al*, in preparation).

1.6 Conclusion

This chapter has briefly explored the developments of the last century of research in theoretical and empirical evolutionary biology, with specific attention to population genetics and phylogenetics. Recent advances in molecular sequencing and computer power has ushered a new information age into biology. This has important implications for the study of evolution, and this introduction has covered the major areas of impact.

The following chapters detail progress I have made in this interdisciplinary area of research.

This thesis concentrates on two emerging problems in evolutionary inference: (i) *measurably evolving populations* and (ii) the role and impact of molecular structure in molecular evolution.

Chapter 2 introduces the concept of *measurably evolving populations*, especially in light of recent technological advances, such as rapid molecular sequencing and ancient DNA sequencing, and methodological advances, many of which are a direct result of work described in this thesis.

Chapter 3 describes the first method I developed to analyse *measurably evolving populations*. This method is based on pair-wise distance data and sacrifices sophistication for speed. This chapter has been published in the international peer-reviewed journal, *Molecular Biology and Evolution*.

Chapter 4 describes a new method of maximum-likelihood estimation of mutation rate from *measurably evolving populations* that allows for multiple rates. This method has implications for the analysis of rapidly evolving pathogens interacting with the immune system, and drug therapy. This chapter has been published in the international peer-reviewed journal, *Molecular Biology and Evolution*.

Chapter 5 describes a new Bayesian inference method for the joint estimation of mutation rate, population history and genealogical relationships from *measurably evolving populations*. This chapter has been accepted for publication in the international peer-reviewed journal, *Genetics*.

Chapter 6 explores some extensions to the Bayesian framework outlined in Chapter 5. To illustrate the flexibility and power of computational intensive statistically explicit modelling strategies, three case studies are undertaken. The effect of rate heterogeneity among sites is analysed, and estimation of the age of fossils from genetic material is undertaken.

Chapter 7 introduces the subject of RNA evolution and Chapter 8 investigates a number of problems motivated by the development in Chapter 7. The focus is on RNA secondary structure and its role in the molecular evolution of rRNA-encoding genes. The estimation of structure-specific substitution models is undertaken, and the inclusion of structural information in both sequence alignment and phylogenetic reconstruction is evaluated.

Chapter 9 outlines a brief excursion into the inference of spatial patterns from genetic data and Chapter 10 outlines the use of the MEPI software and the XML language MEPIX used to describe MCMC analyses for phylogenetics and population genetics.

Chapter 11 discusses the open problems in evolutionary inference and discusses possible directions that future research in computational evolutionary inference might take, in the hope that this lineage of scientific endeavour survives a while longer.

It will become apparent to the reader that the research program followed herein is philosophically aligned with those workers that strive for explicit and realistic statistical models to describe the process of evolution.

2 Measurably Evolving Populations

"*Our genetic reasoning will be almost entirely confined to the analysis of gene frequencies among present-day populations, though it is clearly possible to extend it to other cases. In particular once methods have been set up for estimating the course of evolution from present-day data, they can be extended without difficulty to include data from the past.*"

(Cavalli-Sforza & Edwards, 1967)

This chapter is based on a leading-author manuscript in preparation for submission entitled "Measurably Evolving Populations" by A.J. Drummond and A.G. Rodrigo.

2.1 Introduction

Population genetic and phylogenetic studies that utilize molecular sequences typically rely on samples that have been obtained contemporaneously. However, there has been increasing interest in the analysis of samples that are temporally spaced[6]. If (i) the mutation rate is fast, (ii) the time frame is sufficiently long and/or (iii) the sequences are sufficiently long, then temporally spaced sequence data can provide the opportunity to measure evolutionary rates. Populations for which we can collect datasets of this type are called *measurably evolving populations* (MEPs). Measurably evolving populations allow us to separate intra-specific diversity into contributions from population size and contributions from mutation rate. Temporally spaced sequences have begun to revolutionize the study of viral evolution and ancient DNA based population genetics. They provide the opportunity for detailed analysis of temporal patterns of population size and mutation rates, and a new means of estimating divergence times and mutation rates that is independent of external calibrations. The concept of measurably evolving populations promises to further illuminate our understanding of evolutionary processes from viruses to vertebrates.

2.1.1 Rapidly evolving pathogens

One source of temporally spaced sequence data and measurably evolving populations is obtained from intra- and inter-host samples of rapidly evolving viruses such as HIV-1 and human influenza A. These viruses evolve at such high rates (~1% per year in the case of the HIV-1 envelope (*env*) gene) that measurable differences in a virus population are often readily apparent from one year to the next (LEITNER and ALBERT 1999; LI *et al.* 1988; SHANKARAPPA *et al.* 1999).

2.1.2 Fossil and Pre-fossil DNA

Another source of temporally spaced sequence data is ancient DNA. With increasing regularity short DNA sequence fragments (100-1000bp) are being recovered from pre-fossil material of considerable age (BARNES *et al.* 2002; HANNI *et al.* 1994; LAMBERT *et al.* 2002; LOREILLE *et al.* 2001). Especially successful has been the recovery of mitochondrial DNA (mtDNA) (which has a high copy number) from permafrost remains at least 60,000 years old (BARNES *et al.* 2002). Even in the absence of relatively cold and dry

[6] "Temporally spaced" will be the canonical term used to refer to a set of sequences that are not all of the same age. Synonyms for "temporally spaced" that are used are "serially sampled" and "longitudinally sampled" in the case of many virus evolution studies and alternatively "time-stamped" or "dated-tips".

conditions some researchers have claimed recovery of mtDNA hyper-variable region 1 (HVR1) sequences from remains as old as 60,000 years (ADCOCK *et al.* 2001). With the widespread success of mtDNA recovery from recent (<100,000 years) fossil material, researchers are starting to embark on the recovery of DNA sequence from nuclear loci such as microsatellites (David M. Lambert, *personal communication*).

This growing wealth of ancient DNA sequences brings with it the opportunity to treat organisms with much lower mutation rates than viruses as measurably evolving populations. Recently a study of 5669 genes in mammals found an average substitution rate of 2.2 x 10^{-9} (KUMAR and SUBRAMANIAN 2002). This is 5 million times slower than HIV-1 *env* substitution rates of up to 1% per year. In contrast early phylogenetic estimates of mtDNA in birds suggested an overall rate of 2 mutations per 100 million years (SHIELDS and WILSON 1987), ten times faster than that of nuclear genes in mammals. The HVR1 region, located in the *control region* of the mitochondrial genome, exhibits the fastest rates of evolution in vertebrates, and the use of temporally spaced sequence data has already had an impact here. Previous estimates of HVR-1 evolution in birds were about 0.2 mutations per site per million years. However, a recent analysis of ancient DNA in Adelie penguins using MCMC methods to compare the temporally spaced ancient sequences with modern sequences suggested that the mutation rate in this region is probably 2-7 times faster (0.4-1.4 mutations per site per million years) (LAMBERT *et al.* 2002). Table 2.1 provides a comparison of rate estimates from various taxa and methods.

Table 2.1 Mutation rates of various genetic regions in different organisms.

Both MEP and non-MEP estimation methods are shown.

Genetic region	Rate (substitutions per site per year)	Method	Reference
Mammalian nuclear genes	2.2×10^{-9}		(KUMAR and SUBRAMANIAN 2002)
Avian whole mitochondrial genome	2×10^{-8}	Phylogenetic method	(SHIELDS and WILSON 1987)
Adelie penguin hypervariable region 1 (326 bp)	$4 - 14 \times 10^{-7}$	MEP Bayesian method (ancient DNA)	(LAMBERT *et al.* 2002)
human mitochondrial control region (673 bp)	$^*7.7 \times 10^{-7}$	Pedigree/familial direct method	(HEYER *et al.* 2001)
Inter-host influenza A non-structural gene	2.3×10^{-3}	MEP linear regression	(FITCH *et al.* 1991)
Inter-host influenza A hemagglutinin gene	$5.7 - 6.7 \times 10^{-3}$	MEP linear regression	(FITCH *et al.* 1997; FITCH *et al.* 1991)
Inter-host HIV-1 env gene	6.2×10^{-3}	Triplet based distance method	(LI *et al.* 1988)
Inter-host HIV-1 env gene	6.7×10^{-3}	MEP linear regression	(LEITNER and ALBERT 1999)
Intra-host HIV-1 env gene	9.2×10^{-3}	MEP linear regression	(SHANKARAPPA *et al.* 1999)

* Assuming a generation length in humans of 20 years.

2.2 Concepts

The concept of a measurably evolving population (MEP) becomes useful when describing populations where the mutation rate, time frame or sequence length is substantial enough that the accumulation of substitutions can be detected over the sampling period. Intuitively, a population is measurably evolving if one can feasibly construct a sampling scheme that shows a (statistically) significant accumulation of substitutions in a gene or gene fragment when two or more temporally distinct samples are obtained from the population. The relevance of this concept is demonstrated by the large number of studies that have explicitly, or implicitly, used it (see Table 2.2).

2.2.1 Non-independence of temporally spaced sequences

Many evolutionary methods assume that substitutions follow a strict molecular clock (i.e. all sequences that are of the same age have the same expected evolutionary distance from the root of the tree). This assumption is made in ultrametric tree-building methods such as UPGMA (SOKAL and MICHENER 1958), tests of the molecular clock based on trees (FELSENSTEIN 1981) and most coalescent methods in population genetics (HUDSON 1990; KINGMAN 1982a; KUHNER *et al.* 1995). However, when sequences are temporally spaced it may be necessary to explicitly model this time structure to avoid biases in an analysis. If a population sample consists of data of different ages and a substantial number of substitutions occur over the sampling time frame, then the use of models that do not incorporate this information will produce unpredictable biases in inference and hypothesis-testing procedures. One plausible solution would be to treat each sample as an independent replicate, and derive estimates (or make inferences) from sequences of each sampling occasion separately. However, this approach is flawed, since the genealogies of the samples may overlap extensively. For example, the correlation across samples will bias the variance of the derived estimates. Treating serially sampled sequences as independent replicates is analogous to treating moving averages as independent. The non-independence caused by shared ancestry is a familiar problem in evolutionary inference. However, in this setting this potential problem can be exploited in a number of novel ways.

Table 2.2 Historical and current uses of MEP methods.

Problem	MEP-based solution	References
Estimate mutation rate per unit time	"Simple counting"	(HAYASHIDA *et al.* 1985; KRYSTAL *et al.* 1983; MARTINEZ *et al.* 1983)
	Distance method of Li Tanimura & Sharp (LI *et al.* 1988)	(GOJOBORI *et al.* 1990; LI *et al.* 1988; LUKASHOV *et al.* 1995)
	Least-squares regression (on pairwise distances or on a single reconstructed tree)	(BUONAGURIO *et al.* 1986; DRUMMOND and RODRIGO 2000; FITCH *et al.* 1997; FITCH *et al.* 1991; GOJOBORI *et al.* 1990; LEITNER and ALBERT 1999; PAGEL 1999; SHANKARAPPA *et al.* 1999)
	ML methods	(DRUMMOND *et al.* 2001; RAMBAUT 2000; SEO *et al.* 2002b)
	Pseudo-likelihood coalescent method	(SEO *et al.* 2002a)
	Bayesian coalescent method	(DRUMMOND *et al.* 2002)
	RNA virus meta-study	(JENKINS *et al.* 2002)
	Ancient mtDNA study	(LAMBERT *et al.* 2002)
Estimate generation length	Coalescent-based method	(RODRIGO *et al.* 1999)
	Distanced-based method	(FU 2001)
Estimate population size dynamics	Coalescent theory	(RODRIGO and FELSENSTEIN 1999)
	Pseudo-likelihood coalescent method	(SEO *et al.* 2002a)
	Bayesian coalescent method	(DRUMMOND *et al.* 2002)
Estimate divergence times	ML method	(RAMBAUT 2000)
	Bayesian coalescent method	(DRUMMOND *et al.* 2002)
	Ancient mtDNA study	(LAMBERT *et al.* 2002)
Inferring phylogeny	Distance-based method	(DRUMMOND and RODRIGO 2000)
	Bayesian coalescent method	(DRUMMOND *et al.* 2002)
Experimental design	Likelihood method	(SEO *et al.* 2002b)

2.2.2 Estimating mutation rates using temporally spaced sequences

Some of the first attempts to estimate mutation rates using temporally spaced sequences were made by comparing sequences of human influenza A strains isolated at different times (KRYSTAL *et al.* 1983; MARTINEZ *et al.* 1983). This early work was later expanded by the analysis of more genes and the comparison of synonymous and non-synonymous evolutionary rates (HAYASHIDA *et al.* 1985). All of the early work done in human influenza A involved comparing the genetic distance with the time interval between pairs of sequences. This kind of analysis is based on the assumption that the population diversity at any one time is negligible, and thus two sequences isolated at different times differ only by the accumulation of substitutions over the time interval. It is now known that this simplification is not valid, and that there is significant genetic diversity in human influenza A and many other measurably evolving populations. Methods to account for substantial population polymorphism were developed for studying both human influenza A (BUONAGURIO *et al.* 1986; SAITOU and NEI 1986) and human immunodeficiency virus type 1 (HIV-1) (LI *et al.* 1988). Least-squares linear regression on a reconstructed tree was used by two of these methods (BUONAGURIO *et al.* 1986; SAITOU and NEI 1986). The method of Li, Tanimura & Sharp (LI *et al.* 1988) took a different approach and compared each sister pair of isolates of different ages, with a closely related outgroup to account for any divergence pre-dating the time interval between the sister pair.

More recently Leitner and Albert (1999) estimated the rate of molecular evolution in HIV-1 using a linear regression technique from a known transmission history. In the same year Shankarappa *et al* (1999) used a similar least-squares regression method to study the long-term intra-host rate of HIV-1 evolution in 9 infected patients. All of these studies suggest very high mutation rates in both HIV-1 and human influenza A viruses (see Table 2.1).

Since these initial attempts, a number of researchers have developed and validated methods that accommodate the time structure of temporally spaced data both in phylogenetic and population genetic frameworks. With these developments, accurate estimation of mutation rate (DRUMMOND *et al.* 2002; RAMBAUT 2000) and its variation over time (DRUMMOND *et al.* 2001; DRUMMOND and RODRIGO 2000) has become possible using temporally spaced data. These advances have included Bayesian inference frameworks that allow for uncertainty in the tree topology (DRUMMOND *et al.* 2002). This development has made it possible to re-assess mutation rates of mtDNA in birds, using ancient DNA where the tree topology is unknown (LAMBERT *et al.* 2002).

In some cases, the evolutionary rate *per se* may not be the primary focus of interest, but may instead be used as an indicator of other biological processes. For example, in intra-patient studies of evolution in HIV-1 patients undergoing drug therapy, the accumulation of substitutions per unit time can be used as a proxy for the rate of viral replication. A measure of viral replication rates could be useful in assessing the efficiency of a drug regime (DRUMMOND *et al.* 2001).

2.2.3 Estimating generation length using temporally spaced sequences

If the ages of sequences are known in calendar units (for example, days or years) then it is possible to estimate the mutation rate per site per calendar unit. However, population genetic theory tells us that in a haploid population the expected genetic diversity, Θ, is two times the product of population size and mutation rate *per generation*. Hence in order to estimate population size we need to know the conversion factor ρ, the number of calendar units per generation (i.e. the generation length).

This problem can be turned on its head if the mutation rate is already known from some external source. In this case, one can estimate the generation length from serially sampled genetic data, given the mutation rate. A number of methods have been described to do this (FU 2001; RODRIGO *et al.* 1999; SEO *et al.* 2002a) and all agree closely with methods based on viral load dynamics. This congruence between genetic methods and viral load dynamics is encouraging because they arrive at the same result, despite completely different sources of data and methodology. The most recent of these methods, a pseudo-likelihood method (SEO *et al.* 2002a), was used to estimate the generation length of 9 intra-patient data sets. Assuming a single underlying mutation rate, they estimated that generation length in HIV-1 varied from 0.73 to 2.43 days among the 9 patients, again showing close congruence with early work.

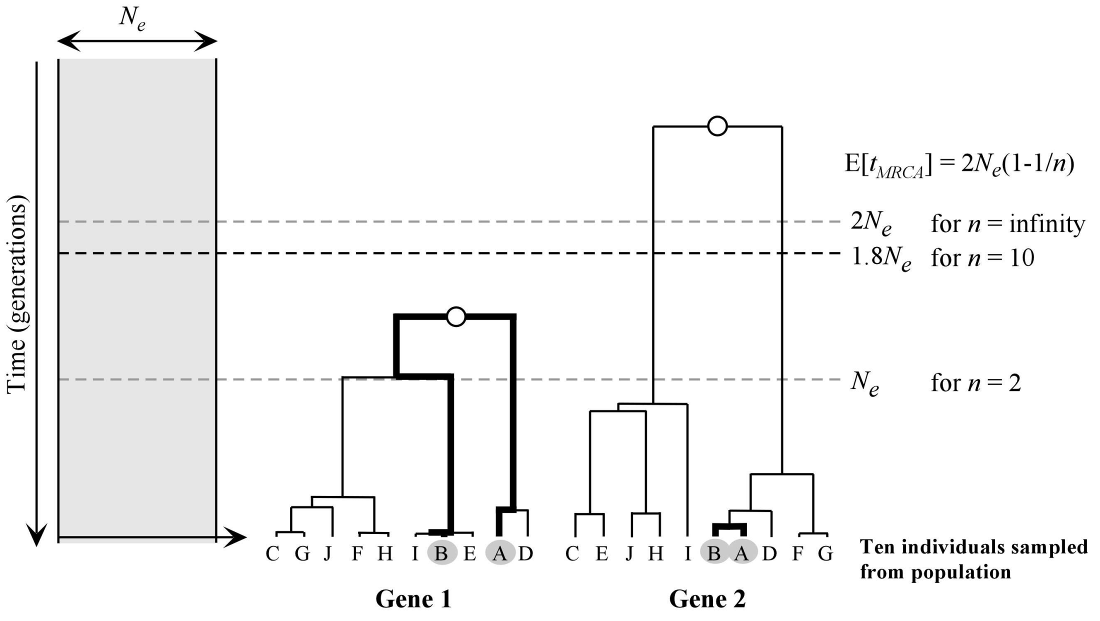
2.2.4 Estimating population size dynamics using temporally spaced sequences

A pair of contemporary homologous sequences drawn randomly from a haploid population with effective population size N_e, are expected to have a common ancestor on average N_e generations in the past (WRIGHT 1931). If the mutation rate, μ, is constant, the pair of sequences will accumulate an average of $N_e\mu$ mutations each, so that between them one expects to see $2N_e\mu$ differences. These statements about common ancestry and mutation rates provide us with a way of working backwards from sequence data to derive estimates of effective population size, rates of growth or decline, migration, and selection. These expectations are illustrated in Figure 2.1 and Figure 2.2.

If the sequences are not contemporary but each has a separate age, then the expected number of mutations that separate the two is no longer a function of N_e alone – rather the total number of mutations separating them will become a function of both N_e and the time interval separating them. For example, consider two ancient sequences and two modern sequences from the same population. The expected number of mutations between the pair of ancient sequences will not be the same as that expected between an ancient and a modern sequence. In fact, the expected difference between an "ancient-modern" pair and an "ancient-ancient" pair will be equal to the product of the mutation rate and the time interval separating the ancient and modern samples (Figure 2.2B), as has been demonstrated previously (DRUMMOND and RODRIGO 2000; FU 2001).

Figure 2.1 The genealogies of two unlinked genes.

Gene 1 and Gene 2 are two unlinked genes from the same ten individuals sampled from an idealized haploid population of size N_e. This diagram shows that two random individuals (such as A and B) are expected to share a common ancestor, on average, N_e generations ago. Although this is the expectation, a particular pair of sequences may be more closely or more distantly related. In general, a sample of n individuals is expected to share a common ancestor, on average, $2N_e(1-1/n)$ generations ago.



2.2.4.1 The coalescent and temporally spaced sequences

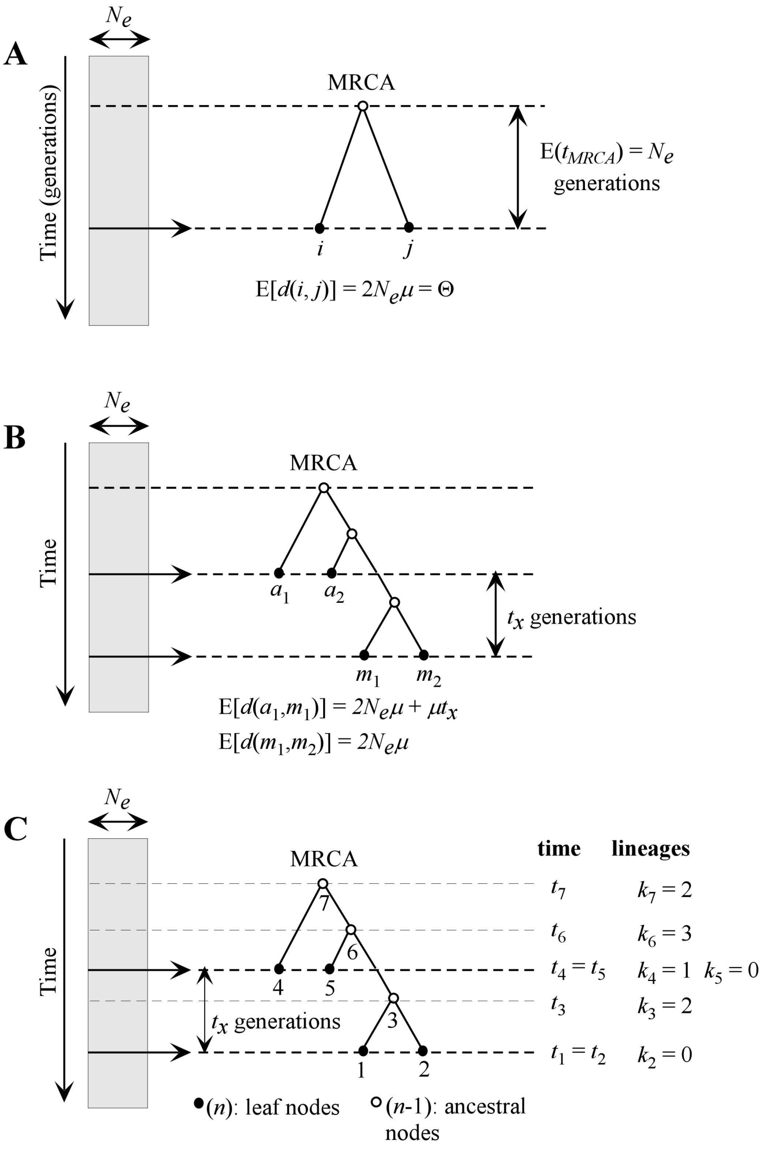
The historical population processes that shape the genetic diversity of a population can be illuminated by genealogical methods such as the *coalescent* (KINGMAN 1982a). The *coalescent* is the most appropriate framework for studying the evolutionary genetics of a large population from which a sample of sequences is drawn. A description of the coalescent for serially sampled sequences has recently been given (RODRIGO and FELSENSTEIN 1999). Consider a genealogy with n terminal nodes (individuals or sequences) and $n-1$ ancestral nodes. For node i, let t_i denote the age of the node in generations from the present. The class of all "legal" trees is the set of all rooted binary trees with n leaves fixed at ages t_i and $n-1$ ancestral nodes assigned ages t_j in such a way that all ancestral nodes are at least as old as their children. The coalescent is a probability density for a tree, $f(G \mid N_e)$, and is computed as follows. First, order the nodes of the tree by increasing age. Then, for i=1,2…2n-1, $t_1 \leq t_2 \leq \ldots t_i \leq \ldots \leq t_{2n-1}$. Let k_i denote the number of lineages present in the interval of time between the node $i-1$ and the node i. The coalescent density is:

$$f(G \mid N_e) = \frac{1}{(N_e)^{n-1}} \prod_{i=2}^{2n-1} \exp\left(-k_i(k_i - 1)\frac{t_i - t_{i-1}}{2N_e} \right) \qquad (2.1)$$

This formulation of the coalescent has been used to develop methods that estimate population sizes from serially sampled sequences (DRUMMOND *et al.* 2002; SEO *et al.* 2002a). We have used this theory to describe a method for jointly estimating population size, population growth rate and mutation rates while taking into account the uncertainty of the tree topology and mutation model using MCMC (DRUMMOND *et al.* 2002). Others have used the coalescent to described a pseudo-maximum likelihood method when the tree is known (SEO *et al.* 2002a). Together, these methods will assist in analyses aimed at elucidating population size dynamics in measurably evolving populations.

Figure 2.2 Genealogy-based population genetics.

(A) Two random individuals selected from a population are expected to share a common ancestor N_e generations in the past. (B) The expected difference between an "ancient-modern" pair [for example $d(a_1, m_1)$] and a "modern-modern" [for example, $d(m_1, m_2)$] pair will be equal to the product of the mutation rate and the time interval (t_x) separating the ancient and modern samples. (C) The coalescent for serial sampled sequences. k_i denotes the number of lineages present in the interval of time between the node i-1 and the node i. Nodes are numbered so as to increase with age.



2.2.5 Estimating divergence times using temporally spaced sequences

One important use of measurably evolving populations is the independent estimation of divergence times in phylogenetic and population genetic studies. Traditionally, independent information (such as fossil evidence) has been used to determine the divergence time of an anchor node and then, by assuming a molecular clock, to estimate the ages of other divergences in the tree (SHIELDS and WILSON 1987). Often a particular class of substitutions (for example, transversions in *Adh* gene) are chosen as a *molecular clock*. This selection will depend on the time-scale and sequence fragment in question. The molecular clock calibration method suffers when there is rate heterogeneity across lineages and when substitution rates over long time-scales are used to calibrate divergences over short time-scales. Recent advances have addressed rate heterogeneity across lineages by explicitly incorporating it into the analysis (HUELSENBECK *et al.* 2000; THORNE *et al.* 1998).

MEPs provide the unique opportunity to estimate substitution rates over relatively short time scales (~5,000 years in mtDNA) and consequently to estimate divergence times in individual populations where ancient DNA is available (LAMBERT *et al.* 2002). Effectively, the ages of the sequences themselves are used as calibration points (DRUMMOND *et al.* 2002). This is a powerful alternative in situations where ancient DNA is available, as it allows for an independent assessment of molecular evolutionary rates at the population level. In the future this method should also allow species differences in the rate of evolution of homologous DNA sequences to be directly compared. As a result, this will provide a test of methods that incorporate lineage specific rate heterogeneity.
The use of methods based on temporally spaced samples to estimate divergence times, has led to the suggestion that molecular evolutionary rates appear to be faster over short time frames than over longer time frames (LAMBERT *et al.* 2002). This raises questions about our understanding of molecular evolution, especially in relation to Motoo Kimura's neutral theory. The neutral theory implies invariance of rates at different time-scales, which is contradicted by recent findings.

2.2.6 Phylogenetic inference using temporally spaced sequence data

When mutation rates are fast or time-scales are long, ignoring the correlation between genetic distance and isolation time will result in biases in tree reconstruction methods that assume contemporaneous sequences. We have been involved in the development of two methods that allow for phylogenetic reconstruction in the face of temporally spaced samples. The first is a variant of UPGMA (DRUMMOND and RODRIGO 2000) and the second is a Bayesian inference method that assumes a coalescent prior on trees (DRUMMOND *et al.* 2002). These methods have not yet been widely used, however in later chapters I will show the utility of our Bayesian coalescent method. We are not aware of any other methods available to perform phylogenetic reconstruction for measurably evolving populations.

2.2.7 MEPs and the neutral theory of evolution

The clock-like nature of many rapidly evolving viruses has been used to support both the molecular clock hypothesis (LEITNER and ALBERT 1999) and Kimura's neutral theory of evolution (GOJOBORI *et al.* 1990). Although there is now fairly strong evidence of positive selection in HIV-1 (NIELSEN and YANG 1998) it still appears to be a relatively minor contribution to the evolution of the HIV-1 genome as a whole.

Recent preliminary evidence of a negative correlation between population size and mutation rate suggests that negative selection imposed by functional constraints is more important and ubiquitous in HIV-1 evolution than positive selection (SEO *et al.* 2002a). This observation provides support for the *nearly neutral theory* (discussed in GILLESPIE 1995; OHTA and KIMURA 1971).

2.2.8 Hypothesis testing and experimental design

Recent maximum likelihood and Bayesian methods of analysis have filled an important gap in the study of measurably evolving populations. These methods both provide a wealth of options for hypothesis testing and model comparison. Of first and foremost concern is the extent to which the molecular clock hypothesis survives careful scrutiny. For example, only 7 out of 50 RNA viruses fit a strict molecular clock when tested in one recent comprehensive study (JENKINS *et al.* 2002). However, the researchers went on to show by simulation that even for the viruses that did not obey a strict molecular clock, the substitution rates estimated could still be regarded as an accurate reflection of the average substitution rate.

Most previous tests of the constancy of evolutionary rate only tested the uniformity of rates across lineages. A concerted change in evolutionary rate over time would not be detectable using only contemporaneous sequence data. However, with temporally spaced sequence data it is possible to both estimate (DRUMMOND *et al.* 2001; DRUMMOND and RODRIGO 2000) and test for (SEO *et al.* 2002b) concerted changes in evolutionary rate.

2.3 Conclusion

Measurably evolving populations provide an opportunity to ask questions about population dynamics and molecular evolution that are not possible with slow-evolving organisms and/or contemporaneous sequence data.

All populations accumulate mutations over time, but whether or not we treat a population as a MEP will depend on the amount of information obtainable over time. This, in turn, depends on the sampling scheme (i.e. the number and length of sequences obtained and the length of time over which they are sampled), and on the underlying biological system (i.e. the mutation and population dynamics of the genetic element being studied). Thus the concept of measurable evolution is an empirical one, related to the logistical constraints of data collection for the population of interest.

For a particular set of biological parameters, the concept of a MEP allows us to partition sampling schemes into those that will show measurable amounts of evolution and those that will not. Hence, given our current knowledge of these biological parameters, we can use the MEP concept to guide us in designing sampling strategies that fall into one of these two categories depending on the purpose of the analysis. Furthermore, we can assess whether the MEP concept is of importance for a particular population and if not, what changes in technology or understanding might make it so. Some sampling strategies will clearly be non-accessible for many populations due to technological and financial limitations. This may, for instance, be true of many extinct populations. It may also be true of slowly evolving species that have little or no reservoirs of non-degraded ancient DNA. On the other hand, fast-evolving populations such as RNA viruses can readily be sampled to show measurable amounts of evolution.

Finally, at the limit of infinite sequence length it has been demonstrated that evenly spacing sample times among all sequences provides more precise estimates of population size than sampling multiple sequences at the same time (SEO *et al.* 2002b).

A solid theoretical basis for developments in this area has been put in place in recent years based on coalescent theory and likelihood models of molecular evolution. However, the methods available are still limited by the simplifying assumptions of the models used. Substantial population subdivision, recombination or selection may adversely affect many current methods of analysis for serially sampled sequences. Most assume single panmictic populations, free of recombination and selection. Methods that take into account migration between subpopulations, substantial recombination and selection effects are needed.

Most of these effects fall squarely within the purview of population genetics and are already well understood in the context of contemporaneous samples of sequences. We expect that in the near future methods that allow incorporation of all of these effects will exist for analysis of measurably evolving populations. In fact, very early on it was predicted that temporally spaced data would provide the opportunity to shed new light on these forces:

"*To sum up, selective trends will be detectable only if data from the past are available.*"

(Cavalli-Sforza & Edwards, 1967)

The use of the methods outlined in this chapter, and their derivatives, will assist in answering fundamental questions about the tempo and mode of molecular evolution from viruses to vertebrates.

3 Reconstructing genealogies of serial samples using serial-sample UPGMA (sUPGMA).

This chapter is based on a leading-author paper published in *Molecular Biology and Evolution* entitled "Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA" by A.J. Drummond & A.G. Rodrigo (2000).

3.1 Overview

Reconstruction of evolutionary relationships from non-contemporaneous molecular samples provides a new challenge for phylogenetic reconstruction methods. With recent biotechnological advances there has been an increase in molecular sequencing throughput, and the potential to obtain serial samples of sequences from populations, including rapidly evolving pathogens, is fast being realized. A new method called serial sample UPGMA (sUPGMA) is presented that reconstructs a genealogy or phylogeny of sequences sampled serially in time, using a matrix of pair-wise distances. The resulting tree depicts the terminal lineages of each sample ending at a different level consistent with the samples' temporal order. Since sUPGMA is a variant of UPGMA, it will perform best when sequences have evolved at a constant rate (i.e., according to a molecular clock). On simulated data, this new method performs better than standard cluster analysis under a variety of longitudinal sampling strategies. Serial sample UPGMA is particularly useful for analysis of longitudinal samples of viruses and bacteria as well as ancient DNA samples, with the minimal requirement that samples of sequences are ordered in time.

3.2 Introduction

It is well known that some of the more pernicious human viral pathogens evolve rapidly. Indeed, it is their evolution that stymies attempts to battle infection with antiviral drugs – resistance evolves too quickly. With HIV-1, for instance, 10^{-5} to 10^{-4} substitutions accumulate at each site each generation, and there are an estimated 140 – 300 generations per year (PERELSON *et al.* 1996; RODRIGO *et al.* 1999). Parts of the HIV genome have been shown to accumulate substitutions at a rate of 0.92% per year (SHANKARAPPA *et al.* 1999). There is some thought in the research community that understanding how these viruses evolve is the key to understanding how one may control disease. Recent results give us cause to think that this may be true: a study by Shankarappa *et al* (1999) found that in nine individuals infected with HIV, the pattern of viral evolution within each patient was strikingly similar, with certain features that appeared predictive of progression to AIDS. If such commonality of pattern is universal, then generalizations can be made about the process of evolution that such patterns suggest, and this, in turn, may lead to a strategy to control progression.

The study by Shankarappa *et al* (1999) involved repeated sampling of the viral population from each individual over several years, but such sampling schemes are not uncommon for such rapidly evolving pathogens (HOLMES *et al.* 1992; RODRIGO *et al.* 1999; WOLINSKY *et al.* 1996). A starting point for many evolutionary and population genetic methods is a reconstructed phylogeny of sampled sequences (FELSENSTEIN 1992b; FU 1994; NEE *et al.* 1995; PYBUS *et al.* 2000), often under the assumption of a molecular clock but, until now, there has been no method to reconstruct evolutionary trees of serially sampled sequences under this assumption. Here, we present such a method. Serial sample UPGMA (sUPGMA) is a fast, flexible phylogenetic reconstruction method that can be used whenever samples have been obtained at different times. These samples may be of sequences from a rapidly evolving viral population obtained from within a patient over the course of infection, or from cohorts of individuals sampled over time. We demonstrate the efficiency of sUPGMA at recovering the true topology and describe accessory analyses that allow the estimation of population parameters and mutation rate. Finally, we discuss various extensions of sUPGMA and its associated analyses.

3.3 Serial sample UPGMA

Consider the following sampling scheme: a population is sampled several times over the course of a study period, and at each sampling time a number of sequences are obtained. If these sequences have evolved so that all lineages accumulate substitutions at the same rate over the same period of time (i.e., according to a molecular clock), then the best representation or model of their phylogeny will look something like that shown in Figure 3.1E. Here, six sequences have been sampled, two at each of three time points. One would expect, if clock-like evolution were occurring, that sequences from the same time point would terminate at identical times. One method for reconstructing phylogenies of sequences according to a molecular clock is UPGMA [Unweighted Paired-Group Method with Arithmetic Means (SOKAL and MICHENER 1958)]. However, with UPGMA all tips on the tree terminate at the same time (i.e., the tree is ultrametric). What is required to reconstruct the phylogeny shown in Figure 3.1E is a method that will allow the tips to terminate at different times, but constrains tips sampled at the same time to terminate at identical distances from the root. Serial sample UPGMA allows for this. The method consists of four sequential steps.

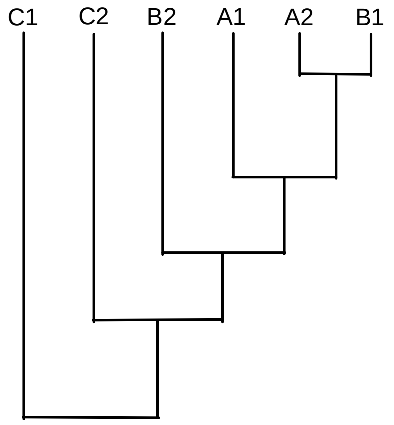
Figure 3.1 An outline of the sUPGMA procedure.

(A) First a distance matrix of the sequences sampled must be collected. (B) A matrix is constructed that relates each observed distance to the parameters to be estimated. Each row in B corresponds to an instance of Equation 3.2, and the binary values in the columns correspond to the X's in Equation 3.3. For convenience, only a single Θ is estimated in this example. Once this matrix is constructed the least squares solution (Equation 3.4) can be used to estimate the parameters. (C) The estimated values of δ are then used to correct the original distance matrix (Equation 3.6). (D) A standard UPGMA tree is constructed from these corrected distances. (E) The branches in the UPGMA tree are then trimmed using the estimated deltas to produce the serially sampled genealogy.

A

	A1	A2	B1	B2	C1	C2
A1		*a*	*b*	*c*	*d*	*e*
A2	0.0271		*f*	*g*	*h*	*i*
B1	0.0317	0.005		*j*	*k*	*l*
B2	0.0547	0.0153	0.0582		*m*	*n*
C1	0.0693	0.0875	0.0089	0.0736		*o*
C2	0.0327	0.0584	0.0383	0.0352	0.0512	

D

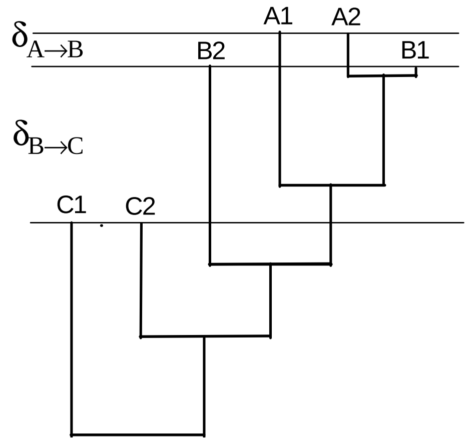


B

	$d(m_i, n_j)$	Θ	$\delta_{A\to B}$	$\delta_{B\to C}$
			(δ_1)	(δ_2)
d(A1, A2)	**0.0271**	**1**	**0**	**0**
d(B1, B2)	**0.0582**	**1**	**0**	**0**
d(C1, C2)	**0.0512**	**1**	**0**	**0**
d(A1, B1)	**0.0317**	**1**	**1**	**0**
d(A1, B2)	**0.0547**	**1**	**1**	**0**
d(A2, B1)	**0.005**	**1**	**1**	**0**
d(A2, B2)	**0.0153**	**1**	**1**	**0**
d(A1, C1)	**0.0693**	**1**	**1**	**1**
d(A1, C2)	**0.0327**	**1**	**1**	**1**
d(A2, C1)	**0.0875**	**1**	**1**	**1**
d(A2, C2)	**0.0584**	**1**	**1**	**1**
d(B1, C1)	**0.0089**	**1**	**0**	**1**
d(B1, C2)	**0.0383**	**1**	**0**	**1**
d(B2, C1)	**0.0736**	**1**	**0**	**1**
d(B2, C2)	**0.0352**	**1**	**0**	**1**

(estimated values: Θ = 0.0326, δ_1 = 0.00368, δ_2 = 0.016)

E



C

	A1	A2	B1	B2	C1	C2
A1		*a*	$b+\delta_1$	$c+\delta_1$	$d+\delta_1+\delta_2$	$e+\delta_1+\delta_2$
A2	0.0271		$f+\delta_1$	$g+\delta_1$	$h+\delta_1+\delta_2$	$i+\delta_1+\delta_2$
B1	0.0354	0.0087		$j+2\delta_1$	$k+2\delta_1+\delta_2$	$l+2\delta_1+\delta_2$
B2	0.0584	0.0190	0.0656		$m+2\delta_1+\delta_2$	$n+2\delta_1+\delta_2$
C1	0.0890	0.1072	0.0323	0.0970		$o+2\delta_1+2\delta_2$
C2	0.0524	0.0781	0.0617	0.0586	0.0906	

3.3.1 Estimation of δs.

Simply, this step involves estimating the expected number of substitutions per site that accumulates between sampling times. It has been shown how this may be done for pairs of samples (FU 2001). The expected distance between a pair of sequences, one from a later time point and the other from an earlier time point is:

$$E[dist(S_{early}, S_{late})] = E[dist(S_{early}^{(1)}, S_{early}^{(2)})] + \delta_{early \to late} \qquad (3.1)$$

The first term on the right hand side is simply the expected average distance between any two sequences from the earlier time point. To obtain an estimate of δ, we substitute the average pairwise distance between early and late sequences calculated from our sample for the term on the left, and the average pairwise distance between pairs of early sequences for the first term on the right, and solve. The problem becomes tricky when there are more than two time points, because then it becomes possible to calculate δs for every possible pair of sampling times. The problem with this approach is that it may be that for any three time points *A, B, C* (where *C* is earlier than *B*, which is earlier than *A)*, $\hat{\delta}_{CA} \neq \hat{\delta}_{CB} + \hat{\delta}_{BA}$ (where $\hat{\delta}$ is the estimated value) when, in fact, under any reasonable model the equivalent equality must be true. To overcome this problem, we have adopted a general regression approach to estimate δ, as follows. Consider a dataset of p samples, with sample i obtained more recently than sample i+1 $(i \in 1,\ldots,p)$. Let $d(m_i, n_j)$ be the evolutionary distance between the ith sequence of the mth sample and the jth sequence of the nth sample; by convention we will assume that $m \geq n$, i.e., we will only consider elements in the diagonal and lower triangular matrix of pairwise distances.

We can model each $d(m_i, n_j)$ by its expectation $E[d(m_i, n_j)]$, and from Equation 3.1, obtain

$$E[d(m_i, n_j)] = E[d(m_i^{(1)}, m_i^{(2)})] + \delta_{m \to n} \qquad (3.2)$$

For reasons that will become obvious below, we will designate $E[d(m_i^{(1)}, m_i^{(2)})] = \Theta_m$. In addition, $\delta_{m\to n}$ can be written as the sum of $\delta_{m\to m-1}, \delta_{m-1\to m-2}, \ldots, \delta_{n+1\to n}$. Thus, we can write the linear equation relating the distances to the parameters as:

$$d(m_i, n_j) = \sum_{k=1}^{p} \Theta_k X_k + X_{(2\to 1)m,j}\delta_{2\to 1} + X_{(3\to 2)m,j}\delta_{3\to 2} + \cdots + X_{(p\to p-1)m,j}\delta_{p\to p-1} + \varepsilon_{m_i,n_j} \tag{3.3a}$$

where $\delta_{k\to k-1}$ is the expected number of substitutions that have accumulated between the k^{th} and $(k-1)^{\text{th}}$ sample;

$$X_k = \begin{cases} 1 & \text{if } k = m \\ 0 & \text{otherwise} \end{cases} \tag{3.3b}$$

$$X_{(k\to k-1)m,n} = \begin{cases} 1 & \text{if } m \geq k \text{ and } n \leq k - 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.3c}$$

and ε_{m_i,n_j} is the error due to natural variation, measurement and sampling.

The vector of estimated parameters $\hat{\mathbf{a}} = \{\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_p, \hat{\delta}_{2\to 1}, \ldots, \hat{\delta}_{p-1\to p}\}$ is obtained by the standard least squares solution:

$$\hat{\mathbf{a}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{d} \tag{3.4}$$

where **d** is a vector of pairwise distances. With this approach, the estimate of the δs satisfies the condition that $\hat{\delta}_{CA} = \hat{\delta}_{CB} + \hat{\delta}_{BA}$. One additional constraint that we make to the δs is to set any value of δ that has been estimated as a negative value to zero.

For the estimation approach above, it is not essential to know the actual sampling times, only the order in which the samples were drawn. If the actual sampling times are known, then an alternative approach to estimating δ is to estimate a single constant, ω (effectively the number of substitutions per unit time), and multiply this by the time interval between two sampling occasions, i.e. $\omega(t_1 - t_2)$.

Once again, we estimate ω using a regression procedure. In this case,

$$d(m_i, n_j) = \sum_{k=1}^{p} \Theta_k X_k + \omega(t_m - t_n) + \varepsilon_{m_i, n_j} \qquad (3.5)$$

where t_k is the time at which the k^{th} sample was obtained. Note that ω is not the mutation rate per generation, unless time is expressed in generation units. However, ω can be converted to the mutation rate (i.e., number of substitutions per site per generation) if the generation time is known.

3.3.2 Correction of pairwise distances

Each pairwise distance d_{ij} in the distance matrix is now transformed to a corrected distance, $c(m_i, n_j)$ as follows:

$$c(m_i, n_j) = d(m_i, n_j) + \delta_{m \to 1} + \delta_{n \to 1} \qquad (3.6$$

where $\delta_{m \to 1}$ and $\delta_{n \to 1}$ are the δs associated with the divergence between samples m and n and the most recent sampling occasion (labeled "1"). What this does, in effect, is extend the distances of sequences sampled earlier to a value that approximates the expected divergences of sequences obtained most recently.

3.3.3 Cluster using UPGMA.

UPGMA or WPGMA (Sneath and Sokal 1973) is applied to the corrected distance matrix, giving a tree in which all branches terminate at the same time.

3.3.4 Trim back branches.

Once the UPGMA tree has been constructed, for any terminal lineage which extends to sequences in sample m, $\delta_{t(i) \to 0}$ is subtracted from the branch length. The sUPGMA tree has the topology recovered by UPGMA (on corrected distances) with tips terminating in the appropriate order of sampling.

3.4 Estimating Population Parameters and Mutation Rate

As described above, a vector of parameters is estimated as part of the tree-building algorithm. This vector takes the form $\hat{\mathbf{a}} = \{\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_p, \hat{\delta}_{2\to1}, \ldots, \hat{\delta}_{p-1\to p}\}$ when the order of samples is known and $\hat{\mathbf{a}} = \{\hat{\Theta}_1, \ldots, \hat{\Theta}_p, \hat{\omega}\}$ when exact times are known. Of course, within this framework, there is no requirement to specify a model with different values of Θ; instead, we could estimate a single parameter, Θ_0 such that $\hat{\mathbf{a}} = \{\hat{\Theta}_0, \hat{\omega}\}$. In this case, the average pairwise diversity at each time point is effectively a random variable with expectation Θ_0. Setting Θ_0 as a constant is equivalent to assuming a population model with constant effective size; under such a model $\Theta_0 = 2N_e\mu$ where N_e is the effective population size, and μ is the mutation rate per site per generation (TAJIMA 1983).

Although the interpretation of a single Θ_0 is easily accommodated within a simple constant-sized population model, this is not the case when multiple Θs are estimated. Multiple Θs should not be taken as (independent) estimates of different $2N_e\mu$ values because the overlap in genealogies from one sample to the next affects the pair-wise distances of the sequences in a complex way. The simple assignment of different Θs in our model does not incorporate these complexities.

However we choose to define our model, the variance of the estimates cannot be easily calculated analytically. However, at least for a constant-sized population model a parametric bootstrap method for obtaining the variance of these estimates can be implemented. For a given set of parameter estimates, a large number (typically > 1000) of serially sampled genealogies can be simulated using the estimated parameters (and assuming a constant population size) to generate pseudo-replicate datasets. For each generated pseudo-replicate the sUPGMA procedure is then repeated, resulting in a range of estimates for Θ(s) and δs or ω. In the case of a 95% confidence interval and 1000 replicates, the 26th and 975th estimates (when ranked) are taken as the upper and lower 95% confidence limits of the original estimate.

3.5 Efficiency of Tree Reconstruction

To test the efficiency of sUPGMA, simulated datasets were created for which the real phylogenetic tree was known. Rodrigo and Felsenstein (1999) described how Kingman's n-coalescent (KINGMAN 1982a), essentially a diffusion approximation of the times of $n-1$ coalescent events on an n-taxon tree, could be extended to coalescent trees with non-contemporaneous tips. One of the novel properties of coalescent trees of serial samples is that sampling a direct descendent of a sequence sampled in an earlier time point becomes possible (although unlikely when the effective population size, N_e, is very large). The probability of a single lineage from a later time point having a direct ancestor in a earlier sample is equal to the fraction of the total population size sampled at the earlier time ($n_{earlier}$ / N_e) (EPPERSON 1999). This possibility was also permitted in the simulations performed, representing an extension of the original description of the serial sample coalescent (FELSENSTEIN *et al.* 1999; RODRIGO and FELSENSTEIN 1999). It should be noted that this inclusion results in the possibility of multiple coalescent events occurring at the same time point when more than one direct ancestor is sampled at one time point. However, this happens at an appreciable rate only when the assumption of a very large population size is broken (i.e. when $n^2 \geq N_e$). At this point the diffusion approximation of coalescent intervals itself is no longer valid. To avoid this problem, values of N_e were selected so that n^2 was always smaller than N_e. Therefore, the simulations were performed under the assumption of a constant population size and N_e was set to 10,000, which is large enough to fulfill the requirements that $n^2 << N_e$. The mutation rate was set to 5×10^{-6} mutations per site per generation. This results in an overall value of Θ of 0.1 (for a haploid population), comparable to published values for HIV evolution (BROWN 1997; RODRIGO *et al.* 1999). The model of evolution used in the simulations was a simple Jukes-Cantor substitution model (JUKES and CANTOR 1969). The simulated genealogies were drawn from populations with no selection, recombination or subdivision.

The serial sample coalescent algorithm was implemented in a small Java program for the purpose of generating coalescent trees under a variety of different sampling strategies (Table 3.1). This allowed an appraisal of the effect of different sampling strategies on the accuracy of tree-building algorithms. For each sampling strategy tested, a range of inter-sample divergences was tested from 0.5% to 10% divergence, with an increment of 0.5%. For each sampling strategy and each divergence, 1000 simulated genealogies were constructed. All simulations resulted in time-ordered DNA sequences 1000 nucleotides in length. This length is comparable with lengths of many gene loci available for phylogenetic study and is not so long that assuming no recombination is untenable. For each simulation a pairwise Jukes-Cantor distance matrix was constructed. The ability of sUPGMA and UPGMA to correctly reconstruct the simulated genealogies using the pairwise distances was evaluated. The reconstructed trees of each method were compared to the real tree using the symmetric difference index (SDI) tree comparison metric (ROBINSON and FOULDS 1981). This metric counts the number of clades in each tree that are not present in the other tree as a proportion of the total number of clades in both trees.

Table 3.1 Sampling stategies under which phylogenetic reconstruction was tested.

Total sequences	Sampling strategies[a]						
20	2 x 10	4 x 5	5 x 4	10 x 2			
40	2 x 20	4 x 10	5 x 8	8 x 5	10 x 4	20 x 2	
80	2 x 40	4 x 20	5 x 16	8 x 10	10 x 8	16 x 5	20 x 4

[a] Sampling strategies are represented by the number of time points multiplied by the number of sequences per time point.

Figure 3.2 shows the performance of sUPGMA and UPGMA on serially sampled datasets with four serial samples. Essentially the same pattern was seen for all sampling strategies. The performance of sUPGMA generally increases with divergence while the performance of UPGMA generally decreases. Figure 3.2 indicates that once some low threshold of inter-sample divergence is exceeded, sUPGMA reconstructs the genealogy more accurately than UPGMA.

Figure 3.2 Phylogenetic reconstruction performances of sUPGMA and UPGMA.

All simulations were performed on four samples, with 5, 10 or 20 sequences in each sample.

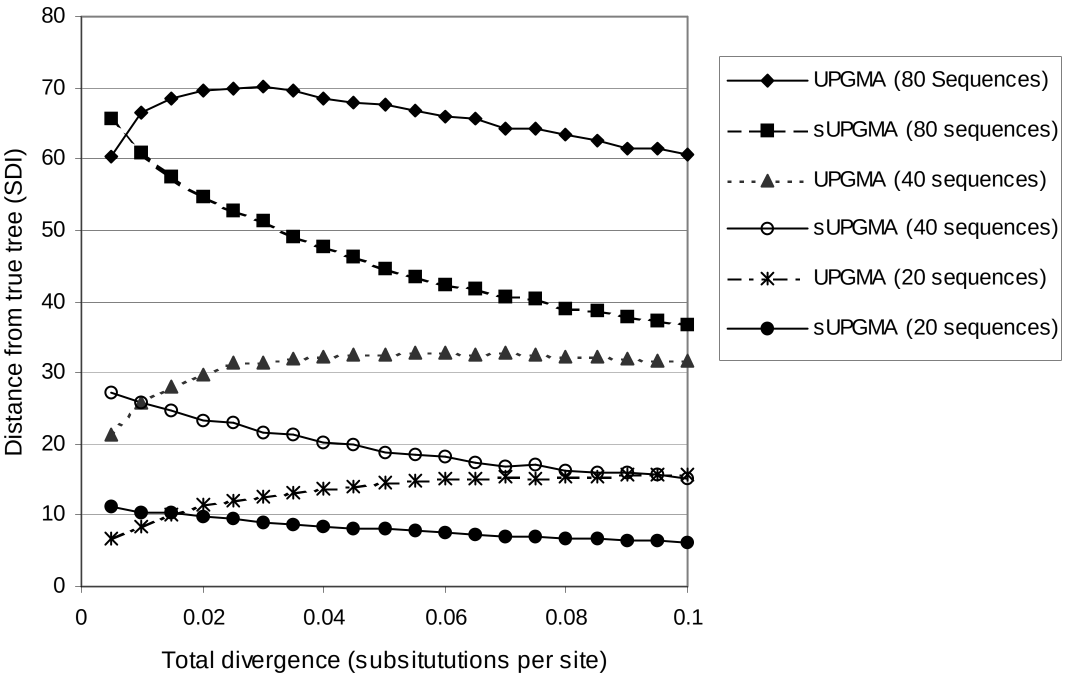


Table 3.2 shows the approximate threshold values for a variety of sampling strategies. Each threshold value was found by picking the lowest divergence for which sUPGMA performed better on average than UPGMA. In general, our simulations indicate that the divergence threshold decreases with an increase in the size of each sample. Therefore, collecting more sequences within each time point improves the ability of the least squares procedure to detect small divergences.

Table 3.2 Threshold values of total divergence over which sUPGMA outperforms UPGMA.

	Number of sampling occasions[a]						
Total sequences	**2**	**4**	**5**	**8**	**10**	**16**	**20**
20	0.01	0.02	0.02	-	0.035	-	-
40	0.005	0.01	0.015	0.02	0.025	-	0.05
80	0.005	0.01	0.01	0.015	0.02	0.035	0.04

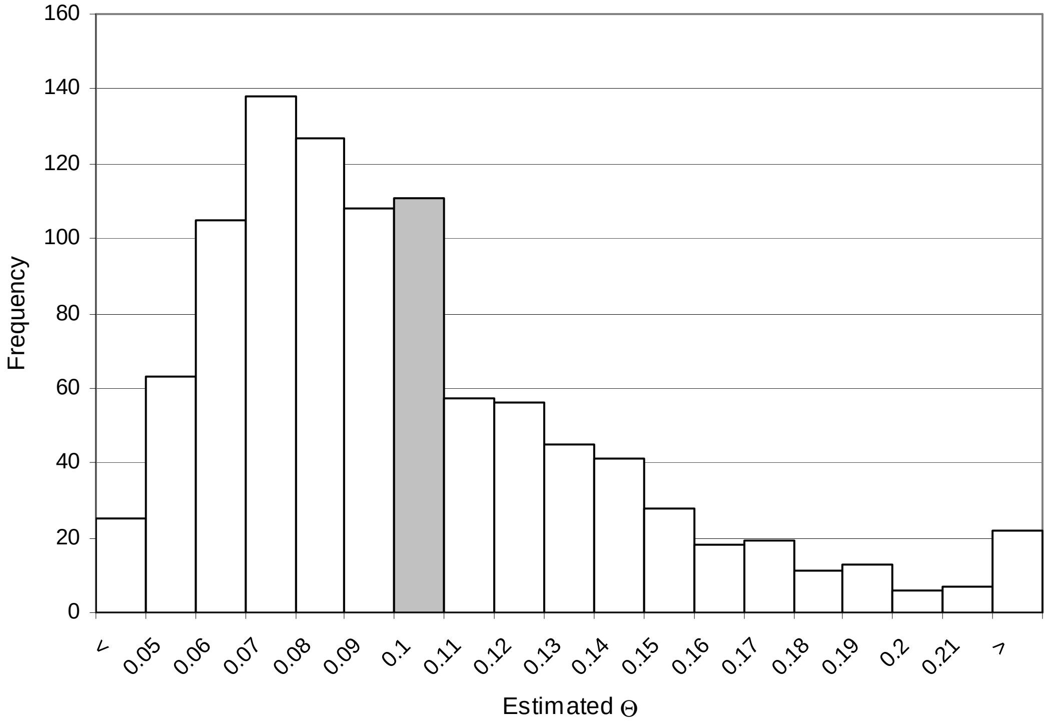
[a] The number of sequences in each time point is equal, and can be obtained from this table by dividing the number of timepoints by the total number of sequences.

3.6 Efficiency of Parameter Estimation

The efficiency of parameter estimation of Θ, ω and δs was measured by simulating two sets of 1000 serially sampled genealogies, one parameterised in accordance with Equation 3.3a, the other in accordance with Equation 5. One thousand genealogies of four samples, each with five sequences, were simulated under the Jukes-Cantor model of substitution, resulting in time-ordered sequences of 1000 nucleotides. Figure 3.3 shows the distribution of estimates of Θ (true value = 0.1) for the 1000 simulations with a total divergence over the four samples of 6%.

Figure 3.3 Θ estimates for four samples of five sequences for 1000 simulated trees.

The real value of Θ = 0.1, and the total divergence = 0.06 expected substitutions.



The mean estimate of Θ was 0.0986 with a skewness statistic of 1.753, showing that the least squares procedure produces estimates of Θ that are unbiased but which have a positively skewed distribution (Table 3.3 & Table 3.4). The least squares estimators of δ_1, δ_2, δ_3 and ω are also unbiased, although once again the distributions of the estimates are skewed. Figure 3.4 shows the empirical densities for estimates of δ_1, δ_2, δ_3 and Figure 3.5 shows the empirical densities of ω estimates.

Table 3.3 Parameter estimates under the *δ-parameterisation* for 1000 simulated datasets of four samples of five sequences.

	$\hat{\Theta}$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$
True value	0.1	0.02	0.02	0.02
Mean	0.0986	0.0203	0.0189	0.0205
standard deviation	0.0432	0.0229	0.0269	0.0477
Skewness	1.753	1.807	2.356	2.452
97.5th percentile	0.207189	0.079969	0.094122	0.141634
2.5th percentile	0.045227	-0.01973	-0.01408	-0.03477

Table 3.4 Parameter estimates under the *ω-parameterisation* for 1000 simulated datasets of four samples of five sequences.

	$\hat{\Theta}$	$\hat{\omega}$
true value	0.1	5×10^{-6}
Mean	0.09996	4.95×10^{-6}
standard deviation	0.0454	3.88×10^{-6}
skewness	1.797884	2.186
97.5th percentile	0.2224	1.56×10^{-5}
2.5th percentile	0.0440	2.71×10^{-7}

Figure 3.4 δ_1, δ_2 and δ_3 estimates for four samples of five sequences.

The results were obtained for 1000 simulated trees and the real values were: $\delta_1 = \delta_2 = \delta_3 = 0.02$.

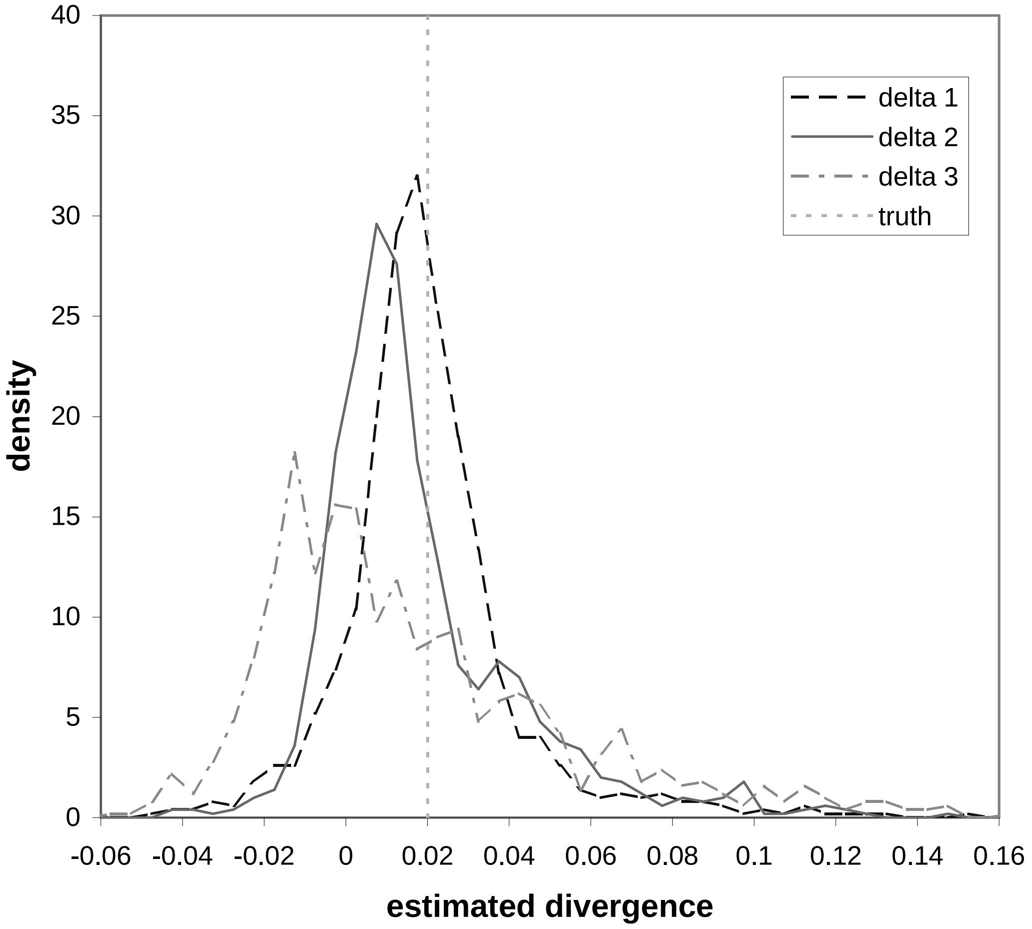
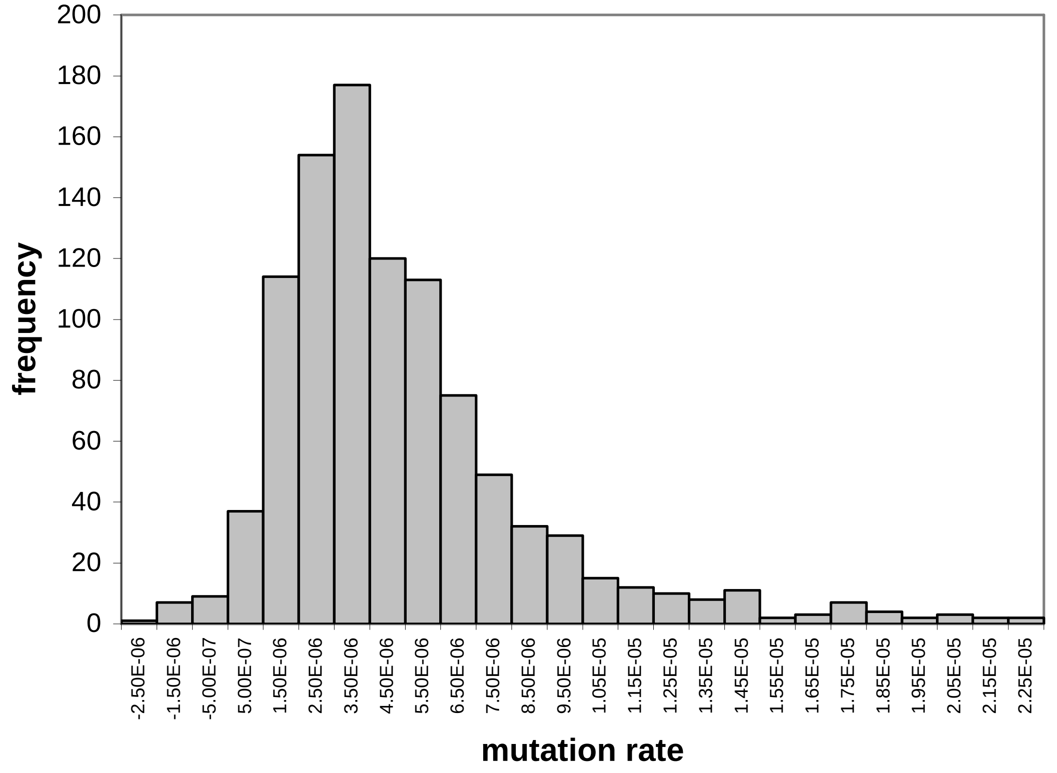


Figure 3.5 Estimated mutation rate, ω, from 1000 simulations of four samples of five sequences. The real value of mutation rate was 5×10^{-6} substitutions per site per generation.



3.7 An Example Dataset

In this section, we illustrate the use of sUPGMA with a dataset of serially sampled partial envelope (*env*) gene sequences of cell-associated HIV-1 DNA, obtained from a long-term asymptomatic individual over five sampling occasions. These samples and the patient history have been described previously (RODRIGO *et al.* 1999). In total, there are 60 sequences in this dataset. Pairwise distances were constructed using a general time-reversible model allowing for unequal nucleotide frequencies and relative rates of substitutions. Substitution and frequency parameters of the substitution model were estimated with PAUP* 4.0b4 (SWOFFORD 1999). sUPGMA was applied to the pairwise distance matrix to reconstruct the serial genealogy of the sequences, allowing different values of Θ and δs. We also reconstructed the genealogy by assuming a constant Θ and mutation rate, ω, and used parametric bootstrapping of 1000 simulated trees to obtain

95% confidence intervals for these parameters. The reconstructed trees are shown in Figure 3.6, and the associated parameter estimates are given in Table 3.5.

Figure 3.6 Two sUPGMA trees constructed from an example dataset.

Tree A was constructed under the assumption of a constant population size and a constant mutation rate. Tree B was constructed allowing a different population size at each sampling point and allowing the varying rate model, in which each time interval has a different mutation rate.

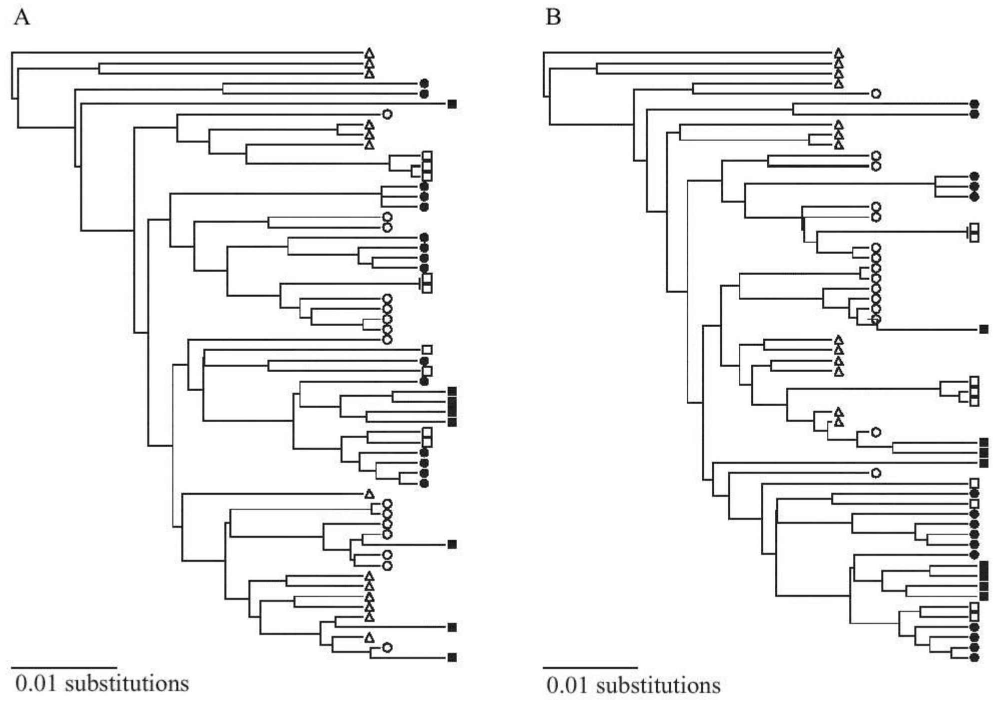


Table 3.5 Estimated parameters for example dataset.

Sample	Days from earliest (5) sample	No. of sequences	Θ estimates	δ estimates[a]
1	0	13	0.0410	0.00386 (1.80 x 10^{-5})[b]
2	214	15	0.0388	0.01054 (2.31 x 10^{-5})
3	671	15	0.0519	0.0 (0.0)
4	699	9	0.0452	9.54 x 10^{-4} (3.12 x 10^{-6})
5	1005	8	0.0410	N/a

[a] Measured in expected substitutions per site between the given sample and the sample immediately following it.
[b] Corresponding mutation rates are shown in brackets in mutations per site per day.

It is instructive to consider some of the main points of these results. When Θ and δ are allowed to vary, sUPGMA is unable to distinguish between Samples 3 and 4, i.e., for this interval $\delta = 0$. In fact, these two samples were obtained only 1 month apart, so this result is reasonable. When Θ is held constant, and ω is estimated, the values obtained are $\Theta = 0.0446$ (95% confidence interval: [0.0184, 0.1016]) and $\omega = 7.8 \times 10^{-6}$ substitutions per site per day (95% confidence interval: [-3.47×10^{-6}, 3.87×10^{-5}]). This estimate of ω translates into an annual substitution rate of 0.3%. This is certainly lower than other estimates of HIV-1 *env* gene substitution rate that have been obtained previously, which are on the order of 1% per year (SHANKARAPPA *et al.* 1999). It is not clear why our estimate of substitution rate is three times lower than other estimates. However, it is pertinent to note that with this patient, antiretroviral therapy was initiated at an early stage of the study, and this in turn may have lengthened the average generation time of infected cells (see below) and consequently lowered the substitution rate. When a varying rate of substitution is allowed, the average rate obtained over the entire 1005 days of the study is 1.53×10^{-5} substitutions per site per day (0.6% per year), which is closer to previously obtained results. However, this mean rate is still deflated by the very slow substitution rate observed in the last 306 days of the study (see Table 3.5).

Interestingly, the 95% confidence interval of our estimate of mutation rate encloses zero. While this can mean that there is no evidence that there has been a detectable accumulation of substitution over time, it can also mean that there are some sequences obtained at a later time point which appear more closely related to those from an earlier time point. In fact, in the original tree published by Rodrigo *et al* (1999), this appears to be the case.

3.8 Discussion

sUPGMA is a variant of UPGMA which constructs genealogies of samples of sequences obtained at different times, under the assumption of a molecular clock. sUPGMA is a two-step procedure. The first step involves estimating the expected sequence divergence between samples obtained at different times. The second step requires the construction of a corrected distance matrix adjusted to take account of these expected divergences, and subsequent clustering using UPGMA. Given a more accurate estimation procedure for the divergences, the accuracy of sUPGMA tree reconstruction can be improved. For example, given a perfect estimate of divergences, the sUPGMA procedure will perform better than UPGMA under all sampling strategies and divergences (simulations not shown). Therefore the threshold divergences required for sUPGMA to outperform UPGMA will be reduced by the use of better estimators of δs and/or ω.

When a molecular clock does not apply, UPGMA is known to perform poorly as a tree reconstruction method. However, in the case of clock-like data that have experienced large amounts of evolution, the accuracy of UPGMA in reconstructing clock-like genealogies has been favourably compared to methods such as maximum-likelihood phylogenetic reconstruction (PYBUS *et al.* 2000). Our results demonstrate that the accuracy of UPGMA for phylogenetic reconstruction can be improved, by modifying the distances between longitudinally sampled sequences to correct for the extra divergence expected between earlier time points and the most recent time point. The rationale behind using sUPGMA as a basis for a tree reconstruction procedure for serial samples is to provide an accurate and rapid estimation of a serially sampled genealogy. Both the criteria for large divergences and clock-like evolution are fulfilled in at least some virus populations (GOJOBORI *et al.* 1990; LEITNER and ALBERT 1999; SHANKARAPPA *et al.* 1999). In addition, and perhaps most importantly, the speed of sUPGMA allows very large datasets (with hundreds or thousands of sequences) to be analyzed with relative ease. This is an important feature when taking into account the size of genealogies already under consideration (SHANKARAPPA *et al.* 1999). The distance-corrected matrix that is constructed as part of sUPGMA can also be used with other members of the family of hierarchical algorithmic clustering methods such as WPGMA, Complete Linkage and Single Linkage clustering.

As part of our parameter estimation procedures, we also introduce two parameterisations of expected inter-sample sequence divergence. In one case – ω-parameterisation – divergence is expressed as a product of the sampling interval and mutation rate (the latter scaled to the same units of time as the sampling interval). A second parameterisation that we use, δ-parameterisation, is less constrained. With δ-parameterisation, the i^{th} interval between two sampling occasions is effectively allowed to have its own mutation rate, ω_i, so that $\delta_i = \omega_i t_i$, where t_i is the length of the interval. In a sense, δ-parameterisation provides a new intermediate model of evolution between the two extremes of a strict molecular clock and the absence of a molecular clock. We call this intermediate model the varying clock model. With HIV, for instance, the application of antiretroviral therapy leads to changes in the relative frequencies of different infected cell types (PERELSON *et al.* 1996). Since each cell type has a different mean generation time, a change in population structure will lead to a change in mean generation time, and consequently, a change in the average mutation rate. This has already been alluded to above when we analysed our example dataset. Under such conditions, a varying clock model may be appropriate. [Note that the varying clock model we propose is different from lineage-specific models of variable mutation rates. In the latter, mutation rate is assumed to change independently along different branches of the tree (HUELSENBECK *et al.* 2000; THORNE *et al.* 1998).]

Although we have focused on rapidly evolving viral populations here, it should be obvious that sUPGMA and its associated procedures of parameter estimation apply equally well to eukaryotic populations from which ancient and/or archival DNA is available. We anticipate that the search for better methods to analyse such populations will only become more important with the increasing frequency of longitudinal sampling strategies and the acquisition of DNA samples from ancient or archival material.

A computer program called PEBBLE that implements sUPGMA and other related methods, written in the Java programming language can be obtained from http://www.cebl.auckland.ac.nz/. This software will run on all computer platforms that support the Java Virtual Machine version 1.1 (JVM 1.1). This includes Microsoft Windows, Linux and MacOS.

3.9 Acknowledgements

We thank Matthew Goode and Greg Ewing for assistance in developing the PEBBLE software package and the distributed programming techniques used for the simulations. For helpful comments and discussion we thank Roald Forsberg and David Nickle. This work was supported by NIH Grant GM59174.

4 Inference of step-wise changes in HIV-1 *env* mutation rate using maximum likelihood

This chapter is based on a leading-author paper published in *Molecular Biology and Evolution* entitled “The inference of stepwise changes in substitution rates using serial sequence samples” by A.J. Drummond, R. Forsberg & A.G. Rodrigo (2001).

4.1 Overview

The molecular clock hypothesis, although a useful null hypothesis, is often rejected by statistical tests on real sequence data. Molecular sequences do not always evolve in a strictly clock-like manner. Substitution rate may vary for a number of reasons, including changes in selection pressure and effective population size, as well as changes in mean generation time. Here we present two new methods for estimating stepwise changes in substitution rates when serially sampled molecular sequences are available. These methods are based on "multiple rates with dated tips" (MRDT) models, and allow different rates to be estimated for different intervals of time. These intervals may correspond to the sampling intervals, or to intervals defined *a priori* not coincident with the times the serial samples are obtained. Two methods for obtaining estimates of multiple rates are described. The first is an extension of the phylogeny-based maximum likelihood estimation procedure introduced by Rambaut (RAMBAUT 2000), and the second is a new parameterisation of the pair-wise distance least-squares procedure used by Drummond & Rodrigo (2000). The utility of these methods is demonstrated on a genealogy of HIV-1 sequences obtained at five different sampling times from a single patient over a period of 34 months.

4.2 Introduction

Although molecular sequences accumulate substitutions over time, the rate at which this occurs may not be constant. The rate of substitution is dependent on biological processes including the intensity of selection, changes in effective population size (when selection is present) and changes in the dynamics of the population, say, a shift in mean generation time. These effects can change substitution rate (i) over time, (ii) in different lineages and (iii) at different positions along the sequence. We present methods that model the substitution rate of molecular sequences obtained serially from individuals within a population or between species (and higher taxa) by allowing the rate to change over time in a stepwise fashion.

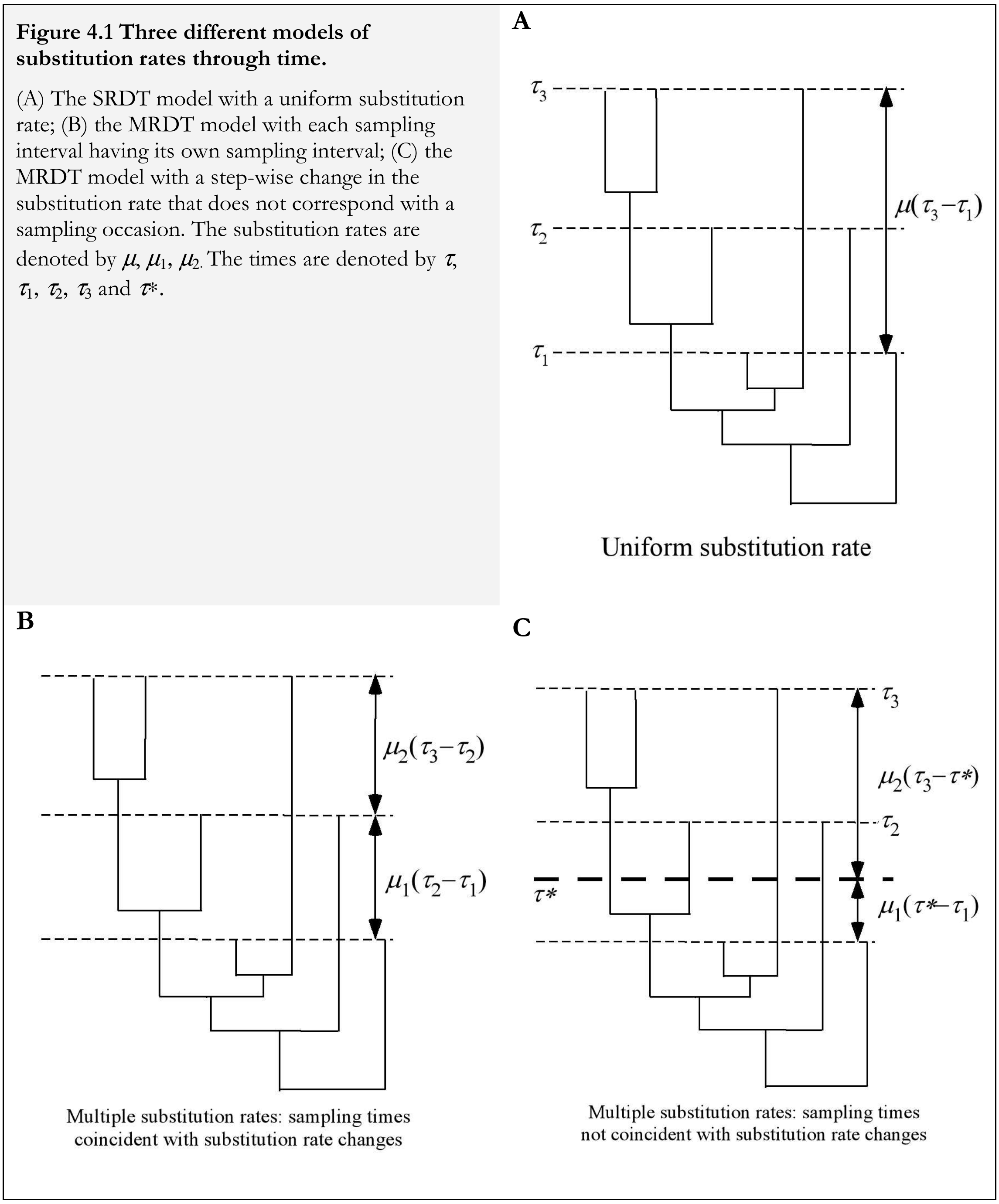
As mentioned in Chapter 1, population genetic studies that utilize molecular sequences typically rely on samples of sequences that have been obtained contemporaneously (FELSENSTEIN 1992b; FU 1994; NEE *et al.* 1995; PYBUS *et al.* 2000). However, recently there has been increased interest in the analysis of samples that are gathered serially, each at a different time. For example, in Chapter 2 the use of samples from rapidly evolving viral populations such as HIV-1 (LEITNER and ALBERT 1999; RODRIGO *et al.* 1999; SHANKARAPPA *et al.* 1999) and samples of ancient DNA from fossilized remains (LAMBERT *et al.* 2002) was discussed. It is our aim to derive estimates of substitutional parameters from this type of data, using biologically relevant models.

Recently, two papers independently described methods to estimate substitution rate, μ, from serial samples, under the assumption of a molecular clock. Rambaut (RAMBAUT 2000) shows how a phylogeny-based maximum-likelihood estimate (MLE) of the constant substitution rate, μ, expressing the divergence between dated sequences as a product of μ and the time interval, can be obtained (Figure 4.1A). Drummond & Rodrigo (2000), using a distance-matrix least-squares (LS) approach, parameterise inter-sample divergence in two ways. First, analogous to Rambaut's "single rate with dated tips" (SRDT) model, μ-parameterisation estimates only a single substitution rate, μ using μt_i as the inter-sample divergence for the i^{th} interval with elapsed time, t_i, (μ is the number of substitutions per unit time). Second, with δ-parameterization, each inter-sample interval is allowed to have its own substitution rate, μ_i, i.e. for the i^{th} interval with elapsed time, t_i, $\mu_i t_i = \delta_i$ (Figure 4.1B). In keeping with Rambaut's terminology, we will refer to this as the "multiple rates with dated tips" (MRDT) model. Drummond & Rodrigo (2000) go on to use these estimates of substitution rate in a phylogenetic reconstruction procedure called serial sample UPGMA (sUPGMA) which recovers a tree with lineages that terminate in the order of sampling (see Chapter 3).

Here, we extend Rambaut's tree-based SRDT likelihood estimation procedure to include the MRDT model. In addition, we show that there are two forms of the MRDT model, one where the rates are estimated differently for each sampling interval (corresponding to Drummond & Rodrigo's δ-parameterization above), and another where the rates are different for different *a priori*-defined intervals that do not necessarily coincide with sampling intervals (Figure 4.1C). ML and LS estimators can be constructed for both forms of the MRDT model. Finally, we illustrate the use of these methods on an example dataset of HIV-1 partial envelope (*env*) sequences obtained serially from an individual who was treated with Zidovudine midway through the sampling program.

Figure 4.1 Three different models of substitution rates through time.

(A) The SRDT model with a uniform substitution rate; (B) the MRDT model with each sampling interval having its own sampling interval; (C) the MRDT model with a step-wise change in the substitution rate that does not correspond with a sampling occasion. The substitution rates are denoted by μ, μ_1, μ_2. The times are denoted by τ, τ_1, τ_2, τ_3 and τ*.


A
τ_3
τ_2
τ_1
$\mu(\tau_3-\tau_1)$
Uniform substitution rate
B
$\mu_2(\tau_3-\tau_2)$
$\mu_1(\tau_2-\tau_1)$
Multiple substitution rates: sampling times coincident with substitution rate changes
C
τ_3
$\mu_2(\tau_3-\tau^*)$
τ_2
τ^*
$\mu_1(\tau^*-\tau_1)$
Multiple substitution rates: sampling times not coincident with substitution rate changes

4.3 Likelihood model

Let us consider the case of sequence data for which there is exact time information and a known phylogeny. Our generalisation allows for the rate of substitution to have step-wise changes over time, and gives rise to a multiple rate model. The MRDT model is constructed by dividing the one substitution rate of the SRDT model (RAMBAUT 2000) into a vector $\mathbf{\Omega}=\{\mu_1,\mu_2,\ldots,\mu_k\}$, where μ_i is the i^{th} substitution rate in the model (Figure 4.1). Hence this μ-parameterisation allows substitution rate to have a number of step-wise changes between the most recent and most ancient sampling times. As in the SRDT model, branch lengths from the root of the tree are no longer required to be equal. Instead, branch lengths must sum to values determined by the temporal spacing of the tip in question and the different substitution rates of the time periods that the tip traverses. Since the information about substitution rates comes from the relative positioning of tips in the tree, it is evident that rate parameters can only be estimated for time intervals where there exists at least one sequence sample. Hence the maximum number of μ parameters is given by the number of sampling points minus one, as one time point is needed as reference. However, this maximum number of rate parameters cannot be estimated for every tree topology. For example, take the simplest case of two sequences sampled at different times. In this situation, the uncertainty of the root confounds rate and time parameters, and the sequence data only holds information about the upper limit of the rate (set by the branch length between the two sequences).

The parameters of the tree are thus the substitution rates $\mathbf{\Omega}$ and the vector of times corresponding to the dated tips and the (n-1 for a bifurcating tree) internal node heights (h) measured in units of substitutions (following RAMBAUT 2000). Note that the tip times may be measured either in generations or some calendar unit, and a simple rescaling allows one to move between the two. Our framework estimates a series of substitution rates only within the interval bounded by the first and last samples. Specifically, no assumptions are made with regard to the rate between the earliest sampling time and the root of the tree. Over this interval, there are no chronological calibration points, and the branch lengths are free to be optimised in the standard manner as composite parameters of time and substitution rate. This rate may of course be of interest, for example in dating the most recent common ancestor (MRCA). In this case additional assumptions must be made: a natural assumption in the case of step-wise changes is that the earliest estimated rate remains constant when extrapolated back to the time of the root.

For a given tree, T, the likelihood of $\mathbf{\Omega}$ is the conditional probability of obtaining the sequence data, S, given $\mathbf{\Omega}$, T and τ, the vector of times, as well as the instantaneous substitution rate matrix, Q (also assumed to be known):

$$L(\mathbf{\Omega}) = Prob(S \mid \mathbf{\Omega}, T, \tau, Q) \qquad (4.1)$$

This likelihood is calculated in the standard manner (FELSENSTEIN 1981; GOLDMAN 1990; RODRIGUEZ *et al.* 1990) for phylogenetic trees; the addition of $\mathbf{\Omega}$ and τ enters the calculations as constraints on the branch tip positions (Figure 4.1B & C). The MLEs of the rates, $\hat{\mu}_i$, are jointly chosen such that $L(\hat{\mathbf{\Omega}})$ is maximised. The only remaining constraint in place is that each estimated substitution rate must not be less than zero.

When considering multiple substitution rates, confidence interval estimation is less straightforward than for a single rate. There are at least two ways of computing confidence intervals for multiple rates. First, multivariate upper and lower $(1-\alpha)$ confidence limits may be obtained by locating rates that correspond to log-likelihood values differing from the maximum-log-likelihood value by $\chi^2_{k,\alpha}/2$, where k is the number of rates estimated. If unbiased, these confidence intervals have a $(1-\alpha)$ probability of enclosing the true $\mathbf{\Omega}$. Second, a profile confidence likelihood interval may be obtained for each μ as follows. Over a range of μ_i, locate the upper and lower values of μ_i such that

$$-2\,|\ln L(\mu_1^*,\mu_2^*,...,\mu_i^*,\mu_{i+1}^*,...,\mu_k^*)-\ln L(\hat{\mu}_1,\hat{\mu}_2,...,\hat{\mu}_i,\hat{\mu}_{i+1},...,\hat{\mu}_k)\,|=\chi^2_{1,\alpha} \qquad (4.2)$$

where $\hat{\mu}_j$ is the MLE of the j^{th} rate, and μ_j^* is the maximum-likelihood estimate of the j^{th} rate when μ_i is fixed at a given value.

In the case where all elements of $\mathbf{\Omega}$ are equal, the MRDT model collapses to the SRDT model of a uniform molecular clock. If all μ parameters are set to zero, the MRDT model reduces to the standard contemporaneous tips clock model (GOLDMAN 1993; RAMBAUT 2000). In fact, under the likelihood framework, one is able to test whether the MRDT model is a significantly better model for the data than the SRDT model. Since the SRDT model is simply a constrained MRDT model, the standard asymptotic likelihood ratio test can be applied. In this case, the test statistic,

$$\Delta = 2(\ln L(\mathbf{\Omega},\text{ not all } \mu \in \mathbf{\Omega} \text{ equal}) - \ln L(\mathbf{\Omega},\text{ all } \mu \in \mathbf{\Omega} \text{ equal})) \qquad (4.3)$$

is asymptotically distributed χ^2 with k-1 degrees of freedom under the null hypothesis that the two models are not significantly different, where k is the number of μ parameters.

When testing the SRDT model against the SR model, the null and alternative hypotheses are of the form:

H_0: $\mu = 0$

H_1: $\mu > 0$

The test is a one-tailed test. Let α is chosen as the level of significance, then the null hypothesis should be rejected if

$$\Delta = 2(\ln L(\mu > 0) - \ln L(\mu = 0)) > \chi^2_{1,2\alpha} \qquad (4.4)$$

Incidentally, this result can also be derived by treating the constraint that μ has to be greater than or equal to zero as a boundary-value problem (OTA *et al.* 2000).

4.4 Least-squares model

With the distance matrix LS estimate of $\mathbf{\Omega}$ described by Drummond & Rodrigo (2000), the expected evolutionary distance, $d(m_i, n_j)$, between a pair of sequences m_i (of the i^{th} sample; assume this is the earlier timepoint) and n_j, is equal to the expected pairwise distance, Θ_i, for sequences from sample i plus the added substitutions accruing between sequences from sample i and sample j in the interval τ_j-τ_i. If there exist times, $\tau_{i+1}, \tau_{i+2}, ..., \tau_{j-1}$, in this interval that correspond to changes in substitution rate, then

$$\begin{aligned} d(m_i, n_j) = \Theta_i + \\ \mu_{i \to i+1}(\tau_{i+1} - \tau_i) + \mu_{i+1 \to i+2}(\tau_{i+2} - \tau_{i+1}) + ... + \mu_{j-1 \to j}(\tau_j - \tau_{j-1}) + \varepsilon_{m_i, n_j} \end{aligned} \qquad (4.5)$$

The parameter estimates $\hat{\mathbf{P}} = \{\hat{\mathbf{\Theta}}, \hat{\mathbf{\Omega}}\}$ are obtained by the standard LS solution:

$$\hat{\mathbf{P}} = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{d} \tag{4.6}$$

where **d** is the vector of the pair-wise distances, and **t** is the matrix of time intervals and [0,1]-values signifying the absence or presence, respectively, of the Θ's associated with each of the samples. Unlike the MLE, LS rate estimates obtained using Equation 4.6 are not constrained to be non-negative. Such a constraint can be added with appropriate linear programming strategies.

The standard error of the LS estimates of μ cannot be calculated easily because of the non-independence of the pairwise distances. Drummond & Rodrigo (2000) advocate the use of the parametric bootstrap (EFRON and TIBSHIRANI 1993; GOLDMAN 1993) to generate confidence intervals of the estimates. Parametric bootstrapping requires specification of a model and subsequent simulation of pseudoreplicate datasets with the same number of sequences and sites as the original data, assuming that the estimates recovered using the observed data are the "true" values of the parameters. With the SRDT model, and an assumption of a constant Θ over time, it is easy to generate pseudoreplicate datasets under a coalescent model in which population size is held constant (Drummond & Rodrigo, 2000). However, under the MRDT model, parametric bootstrapping is not simple, since any resampling procedure must accommodate changing substitution rates and multiple Θ. This is one drawback of the distance-based LS method – procedures for variance estimation are often elusive.

4.5 Example

In this section, we illustrate the use of the MRDT model on an HIV-1 dataset previously published (RODRIGO *et al.* 1999), where the onset of drug therapy is shown to coincide with a significant reduction in substitution rate. This dataset is the same analysed in Chapter 3.

Before the advent of potent combination therapy against HIV, drugs were less effective in lowering viral load and hindering progression towards AIDS. To investigate the affect of a one-drug therapy regime on the evolutionary progression of HIV, we analysed previously published data consisting of serially sampled partial HIV-1 envelope (*env*) sequences from an infected individual who began Zidovudine treatment partway through the sampling period (RODRIGO *et al.* 1999). Complete details of the dataset are given elsewhere (RODRIGO *et al.* 1999); briefly, the dataset contains an initial sample followed by additional samples at day 214, day 671, day 699 and day 1005. Monotherapy with Zidovudine was initiated after day 409 (DRUMMOND *et al.* 2001) and continued during the remaining time of study. Therefore the data set contains two samples before and three samples after treatment began.

It has been suggested that highly active combination antiretroviral therapy leads to a cessation of viral replication (FINZI *et al.* 1997; WONG *et al.* 1997). A natural question is whether monotherapy with Zidovudine had the effect of slowing or halting viral replication in this particular individual from whom samples were available. If viral replication does in fact cease (or slow down), this will be reflected in the rate at which substitutions accumulate, since it is during the process of viral replication that this occurs. This corresponds to testing whether a MRDT model that allows for a change in substitution rate after the onset of therapy provides a better fit to the data than a SRDT model, and if so, whether the estimated substitution rate after drug therapy is significantly different from zero.

The dataset consists of 60 sequences from five time points and the length of the alignment is 660 nucleotides. Gapped columns were included in the analysis. To begin with, the dataset was first split into two subsets, one containing all sequences before therapy (28 sequences; henceforth called pre-treatment) and the other containing all sequences after therapy commenced (32 sequences; henceforth called post-treatment). For each of these datasets, a neighbour-joining tree was built and a maximum likelihood general-time reversible (GTR) model was estimated using PAUP* 4.0b4 (SWOFFORD 1999).

Each tree was used to estimate a uniform substitution rate using the SRDT likelihood model as implemented in the computer program TIPDATE (RAMBAUT 2000). TIPDATE was also used to find the maximum-likelihood roots for the two trees. This is achieved by rooting the tree at every branch on the unrooted topology. For each root the branch-lengths are optimised while constraining tips in accordance with dates. The rooted topology with the maximum-likelihood (greatest *support*) is used to estimate the substitution rate. All estimated rates are reported in Table 4.1. A rate (μ_{before}) of 5.034×10^{-5} substitutions per site per day (1.84% per year, 95%-confidence limits = [1.02%, 2.73%]) was obtained for the pre-treatment sequences and a rate (μ_{after}) of 5.8×10^{-7} substitutions per site per day (0.021% per year, [0.0%, 0.77%]) for the post-treatment sequences. As μ_{after} has a confidence interval that encloses zero, we cannot show that significant substitutions have occurred since therapy commenced.

Table 4.1 ML and LS estimates of substitution rates under the SR, SRDT and MRDT models.

Dataset	Model	-lnL	MLE[a] substitutions/site/day	Hypothesis tests	Δ	*p*-value	LS estimates[a] substitutions/site/day
Complete	SR	4082.50	-				
	SRDT	4080.83	1.36×10^{-5} $[7.14\times10^{-6}, 2.09\times10^{-5}]$	SRDT vs SR	3.34	0.034	7.8×10^{-6} $[-3.47\times10^{-6}, 3.87\times10^{-5}]$[c]
	MRDT	4075.41	$\mu_{before} = 4.15\times10^{-5}$ $[2.6\times10^{-5}, 5.8\times10^{-5}]$[b] $\mu_{after} = 0.0$ $[0.0, 0.8\times10^{-5}]$[b]	MRDT vs SRDT	10.84	0.001	$\mu_{before} = 3.87\times10^{-5}$ $\mu_{after} = -3.35\times10^{-6}$
Before therapy	SR	2441.16	-				
	SRDT	2430.90	5.03×10^{-5} $[2.81\times10^{-5}, 7.49\times10^{-5}]$	SRDT vs SR	20.52	3×10^{-6}	2.69×10^{-5} $[-8.28\times10^{-6}, 1.22\times10^{-4}]$[c]
After therapy	SR	2542.34	.				
	SRDT	2542.10	5.8×10^{-7} $[0.0, 2.12\times10^{-5}]$	SRDT vs SR	0.48	0.24	4.51×10^{-6} $[-2.51\times10^{-5}, 6.59\times10^{-5}]$[c]

a Confidence intervals are presented in square brackets

b Profile likelihood confidence intervals.

c The 95% confidence intervals were obtained by parametric bootstrap using 1000 replicates. Parameter Θ was kept constant.

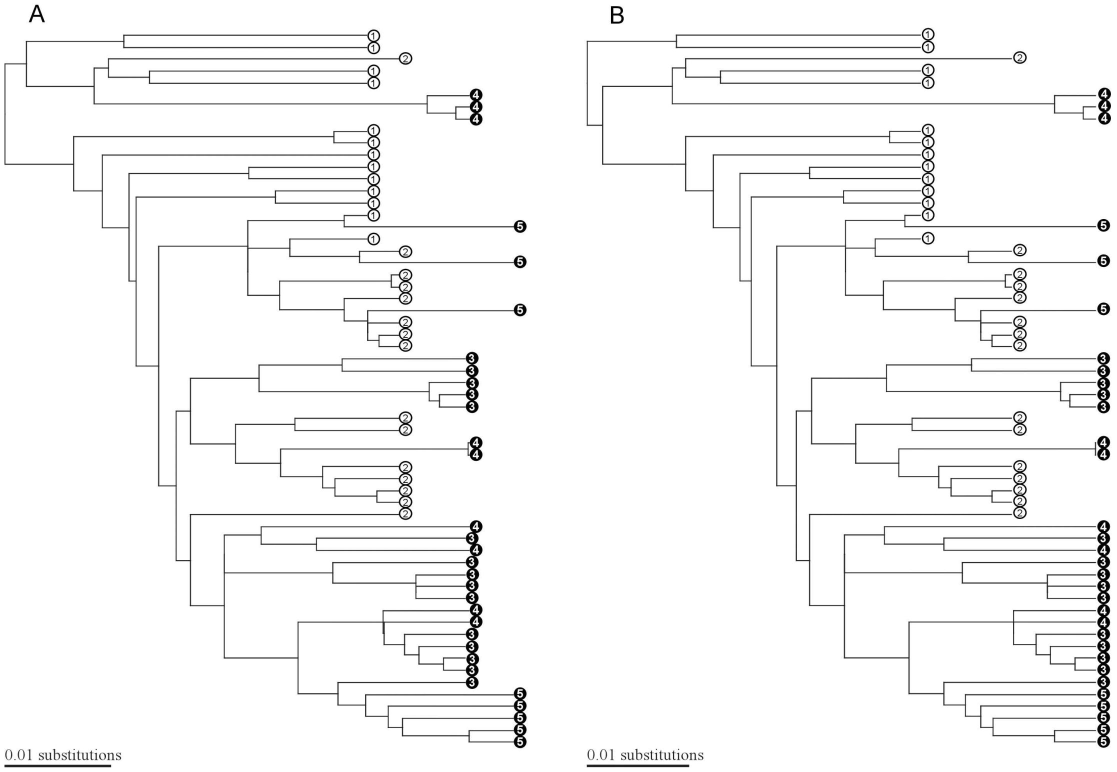
The complete dataset consisting of sequences obtained pre- and post-treatment was then used to obtain an unconstrained and unrooted neighbour-joining tree, once again using the GTR substitution model. Once again, an SRDT model was fitted to the tree (after the ML root was found) and a uniform substitution rate of 1.346×10^{-5} substitutions per site per day (0.49% per year [0.26%, 0.76%]) was estimated. An MRDT model was then fitted to the full dataset, allowing two substitution rates, the first up to the time of therapy (i.e., 409 days from the first sample), and the second after this time. Rates of 4.145×10^{-5} substitutions per site per day (1.51%) and 0.0 substitutions per site per day were estimated simultaneously for μ_{before} and μ_{after}, respectively. The maximum likelihood trees for the SRDT model and the MRDT model on the full dataset are shown in Figure 4.2.

To obtain the 95%-confidence intervals for both substitution rates, a grid search of the two parameters was undertaken. The rate μ_{before} was allowed to vary from 0 to 10^{-4} substitutions per site per day, while μ_{after} was allowed to vary from 0 to 5×10^{-5}

substitutions per site per day, both in steps of 10^{-6} substitutions per site per day. The likelihood surface resulting from this search is shown in Figure 4.3 as a contour plot.

Figure 4.2 Maximum-likelihood solutions for the full example dataset.

The (A) SRDT model and (B) MRDT model are shown. Open and filled circles represent pre- and post-treatment sequences, respectively. Sample numbers are given within circles.

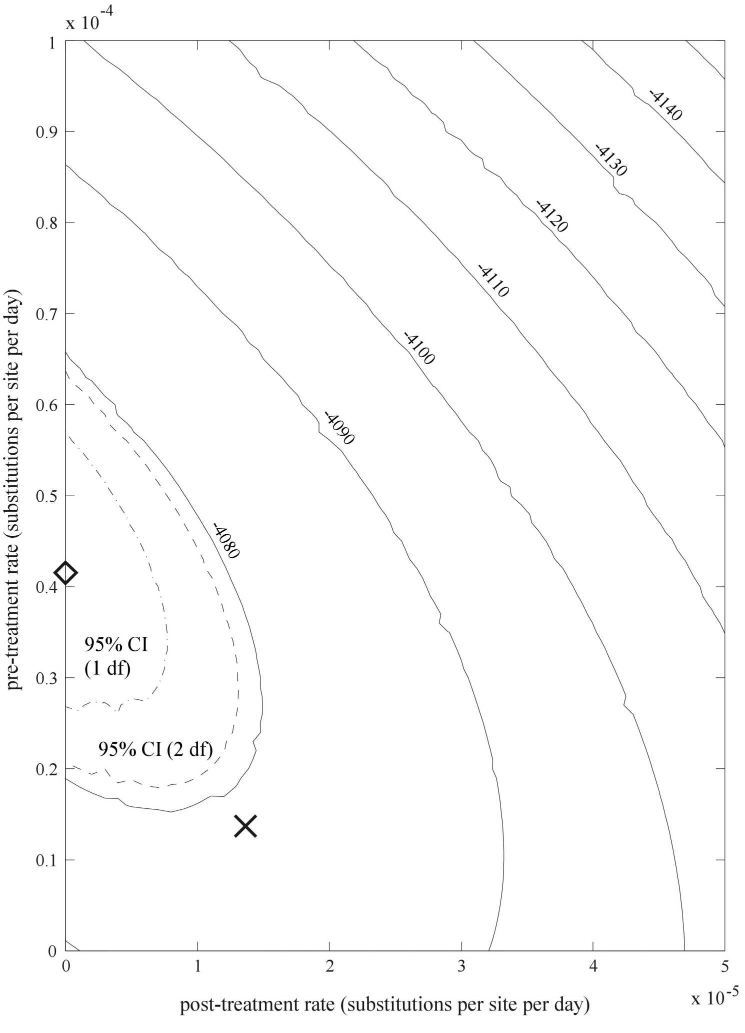


The resulting 95% profile confidence intervals were obtained by taking the maximum and minimum values of μ_{before} and μ_{after} on the contour demarcating $\chi^2_{1,0.05}/2$ (= 1.92) log-likelihood units from the maximum log-likelihood. For μ_{before} the profile-likelihood confidence interval is (2.6×10^{-5}, 5.8×10^{-5}), whereas for μ_{after} it is (0, 0.8×10^{-5}). The bivariate confidence interval for $\hat{\mathbf{\Omega}} = \{\hat{\mu}_{before}, \hat{\mu}_{after}\}$ is also outlined on the likelihood surface contour plot by the contour demarcating $\chi^2_{2,0.05}/2$ (=2.99) log-likelihood units from the maximum log-likelihood. The upper and lower values of μ_{before} and μ_{after} on this bivariate confidence interval contour are (2.1×10^{-5}, 6.1×10^{-5}) for μ_{before} and (0, 1.1×10^{-5}) for μ_{after}. Of course, these intervals are larger than the profile likelihood confidence intervals, but only marginally so.

Table 4.1 gives the log-likelihood scores obtained using the different models described above. For the complete dataset with samples pre- and post-treatment included, the most general clock-like model is the MRDT model. As explained above, the SRDT model is constrained so that all μ's are equal. The SR model with contemporaneous tips is a further constraint on the SRDT with all μ's equal and set to zero. In Table 4.1, likelihood ratio test statistics have been computed for MRDT vs. SRDT, and SRDT vs. SR models. The SRDT model is significantly better than the SR model ($p < 0.05$), and the MRDT model is significantly better than the SRDT model ($p < 0.01$).

Figure 4.3 The likelihood surface of μ parameters.

Both the 95% profile confidence region and the 95% bivariate confidence region are shown. A cross (×) marks the maximum likelihood point for equal rates, located outside of both confidence regions. A diamond (◊) marks the peak of the surface.



Similar analyses were performed for pre- and post-treatment samples, except that in these instances, the only comparison made was between the SRDT and SR models. For the pre-treatment samples, the SRDT model has a statistically better fit to the data than the SR model ($p < 0.01$). However, for the post-treatment sequence subset, the SRDT model cannot be distinguished statistically from the SR model. Taken on its own, this suggests that there is little or no accumulation of substitutions over this period. (Note that caution must be taken with this interpretation: as we discuss in the next section, the MRDT model is significantly worse than a model that assumes no consistent clocklike pattern of evolution amongst the sequences).

Equivalent estimates were also derived with the LS method. Table 4.1 summarises the results. Both the ML and LS procedures consistently estimate a higher pre-treatment substitution rate; approximately an order of magnitude greater than the estimated post-treatment rate.

4.6 Discussion

The framework presented allows for the modelling of complex evolutionary scenarios, such as the evolution of HIV-1 sequences undergoing drug therapy. Application of the MRDT model to samples obtained from an individual treated with Zidovudine appears to indicate a reduction in substitution rate after the commencement of therapy. Our results are consistent with those obtained elsewhere (CHUN *et al.* 1997; WONG *et al.* 1997). Independent estimates of rate from samples pre- and post-treatment have non-overlapping 95% confidence intervals and are therefore significantly different at α=0.05. This poses a problem for any SRDT estimation procedure. TIPDATE, for instance, returns a rate of 0.5% per year for the entire genealogy. This rate is lower than previously published rates of HIV-1 evolution (SHANKARAPPA *et al.* 1999), however it is similar to other published estimates for this dataset that assume a single rate (DRUMMOND and RODRIGO 2000). Here, we outlined a likelihood framework that addresses this discrepancy as well as providing a pair-wise distance least-squares estimation approach. There are, nonetheless, several features of these analyses that bear mention, and indicate that more work in this area is required.

For any analysis that involves the inference of some kind of clock-like behaviour, whether it be a constant clock or a changing clock, a first step should be a test of whether such a model is significantly worse than an unconstrained non-clock model (also called a "different rate" or DR model). The DR model is the standard used in phylogenetic tree reconstruction, and effectively allows every branch to have its own substitution rate. By doing this, the length of the i^{th} branch is an estimate of the composite parameter $\mu_i t_i$. If an SRDT model or MRDT model is significantly worse than the DR model, it means that at least some lineages are not evolving in a clock-like manner. In fact, where appropriate, we recommend a hierarchy of nested likelihood ratio tests: DR vs. MRDT, MRDT vs. SRDT, and SRDT vs. SR. For our example dataset, the DR model was always significantly better than the SRDT and MRDT models (data not shown). Our primary intention with the use of the dataset was simply to illustrate the methods described, rather than to make substantive statements about the effects of monotherapy on substitution rates. However, it is important to note this here for completeness.

The ML estimation procedures presented here (RAMBAUT 2000) assumes that the evolutionary history of the sequences i.e. the topology of the genealogy, is known or can be reconstructed exactly. The bias introduced into parameter estimation and hypothesis-testing procedures by using incorrect genealogies is largely unknown. On the other hand, the LS estimation procedure is not based on a reconstructed topology and therefore may not suffer from this possible source of bias. For example, for a single rate model the LS estimator has been shown to be an unbiased estimator (DRUMMOND and RODRIGO 2000). However, distance-based LS methods do not take into account the correlations induced by shared history, thus making variance estimation difficult.

Ultimately, the best approach would be to incorporate the uncertainty of the genealogy explicitly into a probabilistic framework. One way of taking the uncertainty of the topology into consideration in the likelihood model is to integrate over a number of topologies. A natural way to do this is to use a Markov chain Monte Carlo (MCMC) sampling procedure to sample tree space in proportion to the likelihood of the data (KUHNER *et al.* 1995). This has been used for example to incorporate the uncertainty in the tree topology into estimates of population size and growth rates (KUHNER *et al.* 1995; KUHNER *et al.* 1998), as discussed briefly in Chapters 1 and 2. This method has a natural extension to the estimation of substitution rates, and can also be used to find confidence intervals in topology space under the SRDT or MRDT models of evolution. Chapter 5 describes the development of an MCMC approach that enables estimation of single mutation rates while taking the uncertainty of genealogy into account. The extension to multiple rates is simple, but computationally expensive.

One of the interesting observations of this study is that different models (SR, SRDT, MRDT) can have different maximum likelihood tree topologies. This may turn out to be a common occurrence. For example, for a 45 sequence subset of the data, 729 strictly bifurcating maximum likelihood tree topologies were found. Although these trees had identical likelihood scores under an unconstrained (non-clock) model, they have a range of likelihood scores under the SR, SRDT, and MRDT models. Furthermore, no single strictly bifurcating topology represented the maximum likelihood topology under all three models. If one chooses to use different topologies for each model, then the asymptotic approximation to the likelihood ratio test cannot be used. Instead, some alternative procedure (GOLDMAN *et al.* 2000) should be used. A sampling method such as MCMC would also be useful in this case, as the sampling procedure integrates over tree space in proportion to the likelihood of the data. Thus, for two competing models, a null and alternative distribution can be compared.

In the previous section, we also alluded to the fact that different rootings of an unrooted tree can have different likelihood scores under a given model of substitution. By extension, this also means that different models may require the tree to be rooted differently. This does not change the mechanics of any likelihood ratio test, since no new free parameters are added to the model. However, if the root of the tree is not known, an extra step needs to be added to any analysis to find the appropriate root.

Serial molecular samples add a new dimension to population genetic studies. Since it is possible to estimate substitution (or mutation) rate independently of other parameters, it is also possible to decouple composite population parameters like $\Theta = 2N_e\mu$ (where N_e is the effective population size) into their component parts. The models we introduce here go one step further, and allow these parameters to be expressed as functions of time. Although we have only described stepwise changes in substitution rates, these models can be generalized to allow substitution rate to vary as any parametric function of time. With viral populations such as HIV-1, this becomes especially interesting since it allows us to study changes in average generation time and substitution rate during disease progression, or under different therapeutic regimes. In conjunction with the estimation of demographic functions of time (PYBUS *et al.* 2000) it also means that we can decompose $\Theta(t) = 2N_e(t)\mu(t)$ into the component functions of $N_e(t)$ and $\mu(t)$, where $\mu(t)$ is a stepwise function of time.

We have assumed that the times corresponding to changes in substitution rates are fixed either to the sampling times, or some time point known *a priori*. Similarly, we have also assumed that the phylogeny is known. However, since these times and the phylogeny are parameters embedded in the model, they can also be jointly estimated within the likelihood framework.

The models we have described apply to any set of molecular sequences of sufficient length, or obtained sufficiently far apart in time, that an appreciable number of substitutions have accumulated. These include ancient DNA sequences as well as rapidly evolving viral sequences. In conjunction with efforts to model lineage-specific rates (HUELSENBECK *et al.* 2000; THORNE *et al.* 1998), and other time- or lineage- dependent processes, the models presented here go some way towards a more realistic description of the evolution of molecular sequences.

MLE and LS estimates under the SR, SRDT and MRDT models can be obtained using the computer program PEBBLE, available from the website http://www.cebl.auckland.ac.nz/, or from the authors.

4.7 Acknowledgements

We thank Matthew Goode and Greg Ewing for assistance in developing the PEBBLE software package and the distributed programming techniques used for likelihood surface calculations. For helpful comments and discussion we thank David Nickle, Andrew Rambaut and Jeffrey Thorne.

5 Bayesian evolutionary inference of measurably evolving populations

This chapter is based on a leading-author paper published in *Genetics* entitled "Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data" by A.J. Drummond, G.K. Nicholls, A.G. Rodrigo & W. Solomon (2002).

5.1 Overview

Molecular sequences obtained at different sampling times from populations of rapidly evolving pathogens and from ancient sub-fossil and fossil sources are increasingly available with modern sequencing technology. Presented here is a Bayesian statistical inference approach to the joint estimation of mutation rate and population size that incorporates the uncertainty in the genealogy of such temporally spaced sequences by using Markov chain Monte Carlo (MCMC) integration. The Kingman coalescent model is used to describe the time structure of the ancestral tree. Information is recovered about the unknown true ancestral coalescent tree, population size and the overall mutation rate from temporally spaced data, that is, from nucleotide sequences gathered at different times, from different individuals, in an evolving haploid population. The methodological implications and what can be inferred, in various practically relevant states of prior knowledge, are discussed. Extensions for exponentially growing population size and joint estimation of substitution model parameters are also developed. The important features are illustrated on a genealogy of HIV-1 envelope (*env*) partial sequences and 400 synthetic data sets.

5.2 Introduction

As discussed in Chapter 1, there have been a significant number of developments in both population genetics inference and phylogenetic inference as a result of increase in computational power. This has given birth to a new field of computational evolutionary inference.

Here, we contribute to this growing field. We estimate population and mutation parameters, dates of divergence and tree topology from temporally spaced sequence data, using sample-based Bayesian inference. The important novelties in the inference are the data type (i.e. temporally sampled sequences), the relatively large number of unknown model parameters, and the MCMC sampling procedures, not the Bayesian framework itself. The coalescent gives the expected frequency with which any particular genealogy arises under the Fisher-Wright population model. The coalescent may then be treated, either as part of the observation process defining the likelihood of population parameters, or as the prior distribution for the unknown true genealogy. In either case we must integrate the likelihood over the state space of the coalescent. Both Bayesian and purely likelihood-based population genetic inference use the same reasoning, and software, up to the point where prior distributions are given for the parameters of the coalescent and mutation processes.

Are there then any important difficulties or advantages in a Bayesian approach over a purely likelihood-based approach? The principle advantage is the possibility of quantifying the impact of prior information on parameter estimates and their uncertainties. The new difficulty is to represent different states of prior knowledge of the parameters of the coalescent and mutation processes as probability densities. However, such prior elicitation is often instructive. In the absence of prior information, researchers frequently choose to use non-informative or improper priors for the parameters of interest. Such an approach may be problematic and can result in improper posterior distributions. There exist a number of important cases in the literature in which knowledgeable authors inadvertently analyse a meaningless, improper posterior distribution. Why then do we choose to treat improper priors? We are developing and testing inferential and sampling methods. These methods become more difficult as the amount of information in the prior is reduced. The sampling problem becomes significantly more difficult. We therefore treat the "worst case" prior that might naturally arise. Since this prior is improper, we are obliged to check that the posterior is proper. However, when confronted with a specific analysis, detailed biological knowledge should be encoded in the prior distributions wherever possible.

This work builds on previous contributions that developed Bayesian phylogenetic inference (HUELSENBECK *et al.* 2000; MAU *et al.* 1999; THORNE *et al.* 1998; YANG and RANNALA 1997) and Bayesian population genetic inference (WILSON and BALDING 1998). We begin with a description of the models we use, and then give the overall structure of the inferential framework followed by an overview of how MCMC is carried out. We mention extensions of the basic inference that allow for (i) deterministically varying populations and (ii) estimation of substitution parameters. Finally, we illustrate our methods with a group of studies of a sample of HIV-1 envelope (*env*) sequences, and a second group of studies of synthetic sequence data.

5.2.1 Kingman coalescent with temporally offset leaves

In this section we recapitulate and expand on the description of the coalescent density for the constant-sized Fisher-Wright population model described in Chapter 2. Later we will give the corresponding density for the case of a population with deterministic exponential growth. It is assumed genealogies are realised by the Kingman coalescent process. Our time units are 'calendar units before the present' (for example, days before present, or days BP), where the present is the time of the most recent leaf and set to zero. Let ρ denote the number of calendar units per generation and $\theta = N_e\rho$. The scale factor θ converts "coalescent time" to calendar time, and is one of two key objects of our inference. Notice that we will not estimate ρ and N_e separately, but only their product.

Consider a rooted binary tree g with n leaf nodes and n - 1 ancestral nodes. For node i, let t_i denote the age of that node in calendar units. Node labels are numerically increasing with age so $i > j$ implies $t_i \geq t_j$. Let I denote the set of leaf node labels and let Y denote the set of ancestral node labels. There is one leaf node $i \in I$ associated with each individual in the data. These individuals are selected, possibly at different times, from a large background population. An edge $\langle i, j \rangle$, $i > j$ of g represents an ancestral lineage. Going back in time, an ancestral node $i \in Y$ corresponds to a *coalescence* of two ancestral lineages. The root node, with label $i = 2n\text{-}1$, represents the most recent common ancestor (MRCA) of all leaves. Let t_I be the times of the leaves and t_Y be the divergence times of the ancestral nodes, and let E_g denote the edge set of g, so that $g = (E_g, t_Y)$ specifies a realisation of the coalescent process. For given n and t_I, let Γ denote the class of all coalescent trees (E_g, t_Y) with n leaf nodes having fixed ages t_I. The ages t_Y are subject to the obvious parent-child age order constraint. The element of measure in Γ is $dg = dt_{n+1}...dt_{2n-1}$ with counting measure over distinct topologies associated with the distinguishable leaves.

The probability density for a tree, $f_G(g \mid \theta)$, $g \in \Gamma$ is computed as follows. Let k_i denote the number of lineages present in the interval of time between the node i-1 and the node i. The coalescent process generates $g = (E_g, t_Y)$ with probability density

$$f_G(g \mid \theta) = \frac{1}{\theta^{n-1}} \cdot \prod_{i=2}^{2n-1} \exp\left(\frac{-k_i(k_i - 1)}{2\theta}(t_i - t_{i-1}) \right) \tag{5.1}$$

The interpretation is as follows. Fix a time t and suppose k lineages are present at that time. A coalescence event between any of the $k(k\text{-}1)/2$ pairs of distinguished lineages occurs at instantaneous rate $1/\theta$. Given that two lineages coalesce at time t, the probability it was some particular pair is 2/k(k-1). It follows that, in the time interval of length $t_i - t_{i-1}$ preceding the time of a leaf node $i \in I$, 'nothing' happens with probability $\exp(-k_i(k_i-1)(t_i-t_{i-1})/2\theta)$, and that the length of time, $t_i - t_{i-1}$, preceding coalescent node $i \in Y$ is a random variable with density $\frac{-k_i(k_i-1)}{2\theta}\exp(-k_i(k_i-1)(t_i-t_{i-1})/2\theta)$. Taking the product of these factors over all intervals $[t_{i-1}, t_i], i = 2, 3, ..., 2n-1$, we obtain Equation 5.1 (RODRIGO and FELSENSTEIN 1999).

5.2.2 DNA Substitution Model

We use the standard finite-sites selection-neutral likelihood framework (FELSENSTEIN 1981) with a general time-reversible (GTR) substitution model (RODRIGUEZ *et al.* 1990). However, as we are considering genealogies in calendar units (or generations) as opposed to mutations we take some space to develop notation.

Associated with each leaf node $i \in I$ there is a nucleotide sequence $D_i = (D_{i,1}, D_{i,2}, ..., D_{i,s}, ..., D_{i,L})$ of some fixed length L. Nucleotide base characters $D_{i,s}, i \in I, s = 1, 2, ..., L$ take values in the set $C = \{A, C, G, T\}$. An additional gap character, ϕ, indicates missing data, however this is treated as missing data rather than a fifth state. Let $D = (D_1, D_2, ..., D_n)^T$ denote the $n \times L$ matrix of sequences associated with the tree leaves, and let D_A denote the $(n-1) \times L$ matrix of unknown sequences associated with the ancestral nodes. The data is D together with t_I, that is, the n sequences observed in the leaf-individuals and the n ages at which those individual sequences were taken. Let $\boldsymbol{D} = C^{(n-1)L}$ denote the set of all possible ancestral sequences. Consider a site $s=1,2,...,L$ in the nucleotide sequence of an individual. The character at site s mutates in forward time according to a Poisson jump process with 4×4 rate matrix Q. Here, $Q_{i,j}$ is the instantaneous rate for the transition from character i to character j, and $A \leftarrow 1, C \leftarrow 2, G \leftarrow 3, T \leftarrow 4$. We assume mutations are independent between sites. Let $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ be a 1×4 vector of base frequencies, corresponding to the stationary distribution of the mutation process, $\pi Q = (0,0,0,0)$.

The matrix Q is parameterised in terms of a symmetric 'relative rate' matrix R,

$$R = \begin{bmatrix} & R_{A\leftrightarrow C} & R_{A\leftrightarrow G} & R_{A\leftrightarrow T} \\ R_{A\leftrightarrow C} & & R_{C\leftrightarrow G} & R_{C\leftrightarrow T} \\ R_{A\leftrightarrow G} & R_{C\leftrightarrow G} & & 1 \\ R_{A\leftrightarrow T} & R_{C\leftrightarrow T} & 1 & \end{bmatrix} \tag{5.2}$$

as

$$\begin{aligned} Q_{i,j} &= \frac{\pi_i R_{i,j}}{\sum_k \pi_k \sum_{l\neq k} \pi_l R_{k,l}},\ i \neq j \\ Q_{i,i} &= -\sum_{j\neq i} Q_{i,j} \end{aligned} \tag{5.3}$$

The time units of the rate $Q_{i,j}$ have been chosen so that the mean number of mutations per unit time occurring at a site is equal to one. Let μ give the mean number of mutations per calendar unit (for example, mutations per year) at a site.

The conversion factor μ is the second of the two principal objects of our inference. In addition to μ, the relative rates R may be estimated. We have found that wherever it is feasible to estimate the scale parameters μ and θ, our data is informative about the elements of R. We return to inference of relative rates in section 5.4.

We now write down the likelihood for μ. Consider an edge $\langle i, j\rangle \in E_g$ of tree g. The individual associated with node j is a direct descendant of the individual associated with node i. However, the sequences D_i and D_j may differ if mutations have occurred in the interval. Let e^Q denote the 4×4 matrix exponential of Q. In the standard finite-sites selection-neutral likelihood framework $\Pr\{D_{j,s} = c' \mid D_{i,s} = c\} = \left[e^{-Q\mu(t_i-t_j)}\right]_{c,c'}$ for $c \in C$. The probability for any particular set of sequences D, D_A to be realised at the nodes of a given tree is

$$\Pr\{D, D_A \mid g, \mu\} = \prod_{\langle i,j\rangle\in E_g} \prod_{\substack{s=1 \\ D_{j,s}\neq\phi}}^{L} \left[e^{Q\mu(t_i-t_j)}\right]_{D_{i,s},D_{j,s}} \tag{5.4}$$

(in the above formula, compact notation is obtained by including in the product over edges an edge terminating at the root from an ancestor of infinite age). We may eliminate the unknown ancestral sequences D_A from the above expression, by simply summing all $D_A \in \boldsymbol{D}$,

$$\Pr\{D \mid g, \mu\} = \sum_{D_A \in \boldsymbol{D}} \Pr\{D, D_A \mid g, \mu\} \tag{5.5}$$

It is feasible to evaluate this sum using a pruning algorithm (FELSENSTEIN 1981).

5.2.3 Bayesian Inference for scale parameters

We now consider Bayesian inference for scale parameters μ and θ. Each of these quantities takes a real positive value. The joint posterior density, $h_{\mathrm{M}\Theta G}(\mu,\theta,g \mid D)$, for the scale parameters and genealogy, is given in terms of the likelihood and coalescent densities above and two additional densities, $f_{\mathrm{M}}(\mu)$ and $f_{\Theta}(\theta)$. These functions quantify prior information about the scale parameters. Let Z be an unknown normalising constant. The posterior is then

$$h_{\mathrm{M}\Theta G}(\mu,\theta,g \mid D) = \frac{1}{Z}\Pr\{D \mid \mu, g\}\, f_G(g \mid \theta)\, f_{\mathrm{M}}(\mu)\, f_{\Theta}(\theta) \tag{5.6}$$

We are interested in the marginal density, $h_{\mathrm{M}\Theta}(\mu,\theta \mid D)$. We summarise this density using samples $(\mu,\theta,g) \sim h_{\mathrm{M}\Theta G}$. The sampled genealogies can be thought of as uninteresting “missing data”.

Consider now the densities $f_{\mathrm{M}}(\mu)$ and $f_{\Theta}(\theta)$. In any particular application these functions will be chosen to summarise available prior knowledge of scale parameters. It is common practice to avoid the problem of prior elicitation, and attempt to construct a ‘non-informative’ prior. This notion is poorly defined, since a prior may be non-informative with respect to some hypotheses but informative with respect to others. Nevertheless, we will illustrate sample-based Bayesian inference under a prior that contains little information. We do this for two reasons. First, we wish to give our sampling instruments a thorough workout. From this point of view an improper prior is the best choice. Second, when carrying out Bayesian inference, it is necessary to test the sensitivity of conclusions to changes in the state of prior knowledge. What conclusions would a person in a state close to ignorance reach from this data? The improper prior we consider represents ignorance of a rather natural kind. People using our methods will very likely want to consider this particular state of knowledge, along with others more representative of their own.

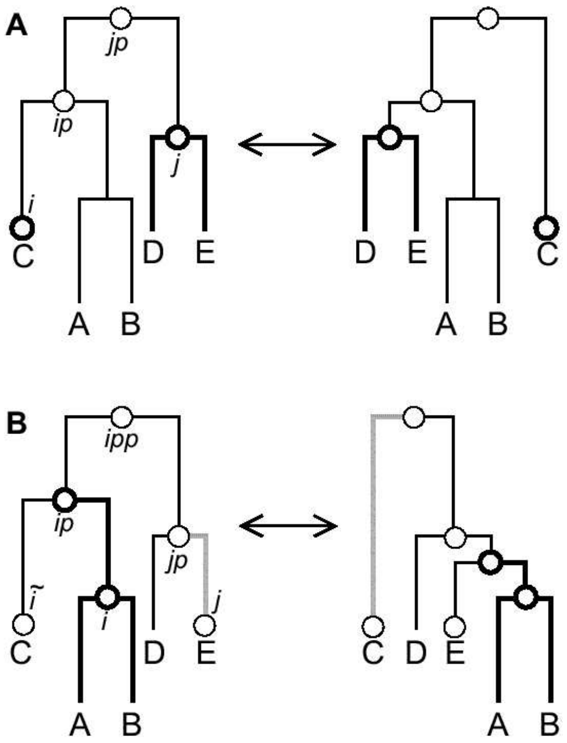
In our case μ and θ are both scale parameters (for time). Jeffreys' prior, $f(z) \propto 1/z,\ z > 0$, invariant under scale transformations $z \to az$, and the uniform prior on $z > 0$, are candidates for $f_{\mathrm{M}}(\mu)$ and $f_{\Theta}(\theta)$. If $f_{\mathrm{M}} \propto 1/\mu, f_{\Theta} \propto 1/\theta$ and $f_G(g\,|\,\theta)$ and $\Pr\{D\,|\,g,\mu\}$ are as given in Equation 5.1 and Equation 5.5, then it may be shown that the posterior density in Equation 5.6 is not finitely normalisable. We may nevertheless consider ratios of posterior densities, but that means the only feasible Bayesian inference, at least under the uniform, improper prior, is exactly frequentist inference. We cannot treat the parameters of interest as random variables. Suppose fixed upper limits of $\mu \le \mu^*$ and $t_{root} \le t^*_{root}$ may be set, along with a lower limit $\theta \ge \theta^*$. For the problems we use to illustrate our methods in section 5.5, conservative limits of this kind determine a state of knowledge that arises quite naturally. Moreover, it may be shown that the posterior density is finitely normalisable under uniform priors on the restricted state space, even though the prior on θ remains improper.

5.3 Markov chain Monte Carlo for evolutionary parameters

The posterior density $h_{\mathrm{M}\Theta G}$ is a complicated function defined on a space of high dimension (between 30 and 40 in the examples which follow). We summarise the information it contains by computing the expectations, over $h_{\mathrm{M}\Theta G}$, of various statistics of interest. These expectations are estimated using samples distributed according to $h_{\mathrm{M}\Theta G}$, and we use MCMC to gather the samples we need. MCMC and importance sampling are part of a family of Monte Carlo methods that may be used either individually or in concert to solve the difficult integration problems that arise in population genetic inference, and earlier work on this subject is cited in the introduction. Figure 5.1 shows a diagram of two proposal mechanisms used.

Figure 5.1 Diagrams of two proposal mechanisms used to modify tree topology during an MCMC analysis.

(A) This move called "narrow exchange", is similar to a nearest-neighbour interchange (NNI). It picks two subtrees at random under the constraint that they have an aunt-niece relationship, i.e. the parent of one is the grandparent of the other, but neither is the parent of the other. Once picked, these two subtrees are swapped so that doing so does not require any changes to the node heights to maintain parent-child order constraints. (B) The second move is similar to one proposed by Wilson and Balding (WILSON and BALDING 1998)8). It involves removing a subtree and reattaching on a new parent branch.



As always in MCMC, it is not feasible to test for convergence to equilibrium. MCMC users are obliged to test for stationarity as a proxy. We make three basic tests. First, we check that results are independent of the starting state, using ten independent runs with very widely dispersed initialisations. Secondly, we visually inspect output traces. These should contain no obvious trend. Thirdly, we check that the MCMC output contains a large number of segments that are effectively independent of one another, at least in the distribution determined empirically by the MCMC output. Let $\rho_f(k)$ give the autocorrelation at lag k for some function f of the MCMC output. Let γ_f denote the asymptotic standard deviation of some estimate of $\rho_f(k)$, formed from the MCMC output. Large lag autocorrelations should fall off to zero, and remain within $O(\gamma_f)$ of zero, as discussed by Geyer (1992). Note that in the examples that follow in section 5.5, these standards are not uniformly applied. The examples in sections 5.5.1.1 and 5.5.1.2 pass all three checks, but the examples in Section 5.5.1.3 only pass the first test. Here we are displaying the limitations of our MCMC algorithm, however we believe the convergence is adequate for the points we make.

The MCMC algorithm was implemented twice; by myself in JAVA, and by Dr Geoff Nicholls in MatLab. This allowed us to compare results and proved very useful in debugging some of the more complex proposal mechanism combinations. To minimise programming burden, one of our implementations (Dr Geoff Nicholls' in MatLab) was partial, allowing only fixed population size and fixed R to be compared.

We will now describe a Markov chain Monte Carlo algorithm for temporally spaced sequence data including proposal mechanisms used. Denote by $\Omega_{\mathrm{M\Theta}G}$ the space $[0,\infty)\times[0,\infty)\times\Gamma$ of all possible (μ, θ, g) values. Let

$$\Omega^*_{\mathrm{M\Theta}G} = \{(\mu,\theta,(E_g,t_Y)) \in \Omega_{\mathrm{M\Theta}G} : \mu \le \mu^*, \theta \le \theta^*, t_{root} \le t^*_{root}\}.$$

We now describe a Monte Carlo algorithm realising a Markov chain X_n, n=0,1,2,... with states $x = (\mu, \theta, g)$, $x \in \Omega^*_{\mathrm{M\Theta}G}$, and equilibrium $h_X = h_{\mathrm{M\Theta}G}$.

Suppose $X_n = x$. A value for X_{n+1} is computed using a Metropolis-Hastings algorithm. Define a set of random operations on the state. A given move may alter one or more of μ, θ and g. Label the different move types m=1,2,...,M. The random operation with label m, acting on state x, generates state x', with probability density $q_m(x' \mid x)$ say. Let $(a \wedge b)$ equal a if $a<b$ and b otherwise, let $(a \vee b)$ equal a if $a>b$ and b otherwise, let

$$P(x,x') = h_X(x'|D) / h_X(x|D)$$

stand for the ratio of posterior densities, and let

$$Q_m(x,x') = q_m(x|x') / q_m(x'|x)$$

give the ratio of the densities for proposals $x' \rightarrow x$, and $x \rightarrow x'$. The algorithm determining X_{n+1} given X_n can be described as follows. First, a label m is chosen according to some arbitrary fixed probability distribution on the M move types. A value for the candidate state x' is drawn according to the density $q_m(x' \mid x)$. Secondly, we accept the candidate, and set $X_{n+1} = x'$ with probability

$$\alpha_m(x, x') = 1 \wedge \left(P(x, x')\, Q_m(x, x') \right) . \tag{5.7}$$

Otherwise, with probability $1-\alpha_m(x, x')$, the candidate is rejected and we set $X_{n+1} = x$.

5.3.1 Proposal mechanisms

In this section we describe the proposal mechanisms (moves) and their acceptance probabilities. In each move $X_n = x$, with $x = (\mu, \theta, (E_g, t_Y))$. For each node i let $parent(i) \in Y$ denote the label of the node ancestral to i, and connected to i by an edge. We get a compact notation if we treat Y, and g, as if Y contained a notional $parent(root)$ node with $t_{parent(root)} = \infty$, as we did in Equation 5.4. Also, we now drop the convention that node labels increase with age.

Let $dx = d\mu\, d\theta\, dg$ in $\Omega^*_{\mathrm{M}\Theta G}$ and

$$H_X(dx \mid D) = h_X(x \mid D)dx .$$

The moves listed below determine an H_X-irreducible aperiodic Metropolis-Hastings kernel.

5.3.1.1 Scaling move

Label this move m=1. Let a real constant β>1 be given. For $\beta^{-1} \le \delta \le \beta$, let $x \rightarrow \delta x$ denote the transformation

$$(\mu, \theta, (E_g, t_Y)) \rightarrow (\mu/\delta, \delta\theta, (E_g, \delta t_Y)) .$$

If $x' = \delta x$ then $x = \delta' x'$ with $\delta' = 1/\delta$. The change of variables in the product measure is

$$H_X(dx' \mid D)d\delta' = \delta^{n-3}H_X(dx \mid D)d\delta .$$

Notice that this transformation is not simply a change of units. The times t_i associated with ancestral nodes $i \in Y$ are scaled, while leaf node times t_i, $i \in I$ (which are part of the data) are left unchanged.

The move is as follows. Choose a $\delta \sim \text{Unif}(\beta^{-1}, \beta)$ and set $x' = \delta x$. If $x \notin \Omega^{*}_{\mathrm{M\Theta}G}$ (for example if $\mu/\delta > \mu^{*}$, or the parent child age order constraint is violated at the unscaled leaves in the scaled tree) then the move fails and we set $X_{n+1} = x$. In a slight abuse of notation we set $Q_1(x, x') = 1/\delta^{n-3}$ in the formula for $\alpha_1(x, x')$ in Equation 5.7 (Green (1995) explains how this scale factor arises in Metropolis-Hastings MCMC). The choice β=1.2 gave reasonable acceptance rates in our simulations.

5.3.1.2 Wilson-Balding move

Label this move m=2. A random sub-tree is moved to a new branch. This move is based on the branch-swapping move of Wilson and Balding (1998). The SPR move in PAUP* (SWOFFORD 1999) is similar. However the move below acts on a rooted-tree and maintains all node ages except one.

Two nodes, $i, j \in I \cup Y$ are chosen uniformly at random without replacement. Let $jp = parent(j)$ and $ip = parent(i)$. If $t_{jp} \le t_i$, if $ip = j$ or $ip = jp$, then the move fails and we set $X_{n+1} = x$. Given i and j, the candidate state $x' = (\mu, \theta, g')$ is generated in the following way. Let $\tilde{i}$ denote the child of ip that is not i, and let $ipp = parent(ip)$, the grandparent of i. Reconnect node ip so that it is a child of jp and a parent of j, that is, set

$$E'_g = \{ \langle jp, ip \rangle, \langle ip, j \rangle, \langle ipp, \tilde{i} \rangle \cup E_g \setminus \langle jp, j \rangle, \langle ip, \tilde{i} \rangle, \langle ipp, ip \rangle \}$$

If node j is not the root, assign to node ip a new time t'_{ip} chosen uniformly at random in the interval $[(t_i \vee t_j), t_{jp}]$. If node j is the root, choose $\delta \sim \text{Exp}(\theta)$ and set $t'_{ip} = t_j + \delta$. Let

t'_Y denote the set of node times with t_{ip} replaced by t'_{ip}. Let $x' = (\mu, \theta, (E'_g , t'_Y))$. If node j and node ip are not root, the ratio $Q_2(x, x')$ in Equation 5.7 is

$$Q_2(x, x') = (t_{jp} - (t_i \vee t_j)) / (t_{ipp} - (t_i \vee t_{\tilde{i}})) .$$

If node j is the root,

$$Q_2(x, x') = \theta / (\exp(-\delta / \theta) (t_{ipp} - (t_i \vee t_{\tilde{i}})) ,$$

and if ip is the root,

$$Q_2(x, x') = (t_{jp} - (t_i \vee t_j)) \exp(- (t_{ip} - t_{\tilde{i}}) / \theta) / \theta .$$

5.3.1.3 Sub-tree exchange

Label this move m=3. Choose a node $i \in I \cup Y$. Let $ip = parent(i)$, $jp = parent(ip)$, and let j denote the child of jp that is not ip. If node i is either the root or a direct child of the root, or $t_{ip} < t_j$ then the move fails and we set $X_{n+1} = x$. Given i and j, the candidate state $x' = (\mu, \theta, g')$ is generated in the following way. Swap nodes i and j, setting

$$E'_g = \{ \langle ip, j \rangle, \langle jp, i \rangle \} \cup E_g \setminus \{ \langle jp, j \rangle, \langle ip, i \rangle \}$$

Let $x' = (\mu, \theta, (E'_g , t_Y))$. The ratio $Q_3(x, x') = 1$ in Equation 5.7.

The sub-tree exchange above is a local operation. In a second version of this move, we chose node j uniformly at random over the whole tree.

5.3.1.4 Node age move

Label this move $m=4$. Choose an internal node, $i \in Y$, uniformly at random. Let ip = *parent*(i) and let j and k be the two children of i (so i=*parent*(j) and i=*parent*(k), $j \neq k$). If i is not the root, choose a new time t_i' uniformly at random in $[(t_j \vee t_k), t_{ip}]$, otherwise, if i is the root, choose $\delta \sim \mathrm{Unif}(\beta^{-1}, \beta)$ (see move $m=1$) and set $t_i' = (t_j \vee t_k) + \delta(t_i - (t_j \vee t_k))$. Let t_Y' denote the set of ancestral node times, t_Y, with t_i replaced by t_i'. Let $x' = (\mu, \theta, (E_g, t_Y'))$. If i is not the root, then $Q_4(x, x') = 1$ in Equation 5.7. If i is the root then $Q_4(x, x') = 1/\delta$.

5.3.1.5 Random walk moves for θ and μ

Label this move $m=5$. The random-walk update to θ is as follows. Let a real constant $w_\theta > 0$ be given. Choose $\delta \sim \mathrm{Unif}(-w_\theta, w_\theta)$ and set $x' = (\mu, \theta+\delta, g)$. If $x \notin \Omega^*_{\mathrm{M\Theta}G}$, then the move fails and we set $X_{n+1} = x$. Since the candidate generation process is symmetric, $Q_5(x, x') = 1$, in the formula for $\alpha_5(x, x')$ in Equation 5.7. The random walk move for μ, with random-walk window parameter w_μ say, is similar to the move just described for θ. The window sizes w_θ and w_μ must be adjusted in order to get reasonable sampling efficiency.

5.3.2 Implementation, convergence checking and debugging

5.3.2.1 Convergence and standard errors

The efficiency of our Markov sampler, as a tool for estimating the mean of a given function f, is measured by calculating from the output $\tau_f = 1 + 2\sum \rho_f(k)$, the integrated autocorrelation time (IACT) of f. Dividing the run length by τ_f, we get the number of "effective independent" samples in the run (the number of independent samples required to get the same precision for estimation of the mean of f). We will call this the effective sample size (ESS). Better MCMC algorithms have smaller IACTs and thus larger ESSs, though it may be necessary to measure τ_f in units of CPU time in order to make a really useful comparison. One will typically want to run the Markov chain at least a few hundred times the IACT, in order to test convergence and get reasonably stable marginal histograms. Notice first, that we do not know the IACT when we set the MCMC running. Exploratory runs are needed. Secondly, a statement like "We ran the MCMC for 10^6 updates discarding the first 10^4" is worthless without some accompanying measurement of an IACT or equivalent. This point is made by Sokal (Sokal 1989). The summation cutoff in the estimate for the IACT, τ_f, is determined using a monotone sequence estimator (Geyer 1992). The IACTs we get for our MCMC algorithms suggest that analysis of large datasets (50-100 sequences and 500-1000 nucleotides) is feasible with current desktop computers. Examples may be found in Section 5.5 (Table 5.2) and in the Appendix.

The inverse of the IACT of a given statistic is the "mixing rate". Statistics with small mixing rates are called the "slow modes" of a MCMC algorithm. The mutation rate μ was the slowest mode among those we checked, and we therefore present IACTs for that statistic in Section 5.5.

5.3.2.2 Implementation issues

In this section we discuss debugging and MCMC efficiency of our two implementations. We compare expectations computed in the coalescent with estimates obtained from MCMC output. Standard errors are obtained from estimates of the corresponding IACT. Consider a tree with four leaves, two at time zero and two offset τ time units to greater age, and consider simulation in the coalescent, with no data. The expectation of t_{root} is

$$E_G\{\ t_{root}\ \} = (\ \tau + 4\theta/3\ )\ (\ 1 - e^{-\tau/\theta}) + (\ \tau + 3\theta/2\ )\ e^{-\tau/\theta}$$

A number of other expectations may also be computed.

For problems involving data, expectations are not available. However, an MCMC algorithm with several different move types may be tested for consistency. The equilibrium is the posterior distribution of μ, θ and g, and should not alter as we vary the proportions in which move types are used to generate candidate states. For example, move 2 (Wilson-Balding) is irreducible on its own, whilst moves 3 and 4 (Sub-tree exchange and Node-age move) form another irreducible group. We fix a small synthetic data set and compare the output of two MCMC runs: one generated using move 2 alone, and the other using only moves 3 and 4.

We now turn to questions of MCMC efficiency. Each update has a number of parameters, and these are adjusted by trial and error for each analysis, so that the MCMC is reasonably efficient. An *ad hoc* adaptive scheme, based on monitoring acceptance rates and akin to that described in Larget and Simon (1999), was used. The samples used in output analysis are taken from the final portion of the run, in which these parameters are fixed. The scaling and Wilson-Balding updates are particularly effective.

We have experimented with a range of other moves. However, whilst it is easy to think up computationally demanding updates with good mixing rates per MCMC update, we have focused on developing a set of primitive moves with good mixing rate per CPU second. In our experience simple moves may have low acceptance rates, but they are easy to implement accurately, and are rapidly evaluated. They may give good mixing rates when we measure in CPU-seconds. Larget and Simon (1999) have given an effective MCMC scheme for a similar problem. We did not use their scheme, as its natural data structure did not fit well with our other operators. A second update, which may be useful to us in future, would use the importance sampling process of Stephens and Donnelly (2000) to determine an independence sampling update.

Because of the explicit nature of MCMC inference, the details of a particular analysis, including the proposal mechanisms, the chain length, the evolutionary model and the prior distributions can be quite difficult to keep track of. An XML data format was developed to describe phylogenetic/population genetic analyses, which enables the user to write down the details of an analysis in a human-readable format that can also be used as the input for the computer program. The XML data format is described in Chapter 10. For the more visually inclined a graphical user interface (GUI) was developed that can generate the XML input files, given a NEXUS or PHYLIP alignment. This software is called MEPI and is also described in Chapter 10.

5.4 Exponential growth and relative rates of substitution

Extending the framework of Sections 5.2 and 5.3 to include deterministically varying models of population history and estimation of relative rate parameters is straightforward. Let $\Phi = (0,\infty)^5$ be the state space for the relative rates of R above the diagonal and excluding $R_{G\leftrightarrow T}$. Let $s = (\mu, \theta, g, r, R)$, and let $h_S(s|D)$ denote the posterior density for $S \in \Omega^*_S$ where $\Omega^*_S = \Omega^*_{M\Theta G} \times \Re \times \Phi$. The posterior probability density has the form

$$h_S(s \mid D) = \frac{1}{Z} \Pr\{D \mid \mu, g, R\}\, f_G(g \mid \theta, r)\, f_M(\mu)\, f_\Theta(\theta)\, f_r(r)\, f_R(R) \qquad (5.8)$$

Let T denote the age of the most recent leaf, i.e. $T = \min_{i \in I} t_i$. Here, $T = 0$. Let $t \geq T$ be a generic age. In this model $N_e = N_e(t)$. Recall that ρ, the number of calendar units per generation, is an unknown constant. Define a constant $\theta = N_e(T)\rho$ and a growth rate parameter r. The density $f_G(g \mid \theta, r)$ is the density determined by the coalescent process with a population growing as $N_e(t) = \frac{\theta}{\rho} e^{-r(t-T)}$ (SLATKIN and HUDSON 1991). In terms of the notation defined in Section 5.2 in connection with Equation 5.1, for genealogies with temporally spaced tips the density is

$$f_G(g \mid \theta, r) = \frac{1}{\theta^{n-1}} \cdot \prod_{i=2}^{2n-1} e^{rt_i} \exp\left(\frac{-k_i(k_i - 1)(e^{rt_i} - e^{rt_{i-1}})}{2\theta r} \right) \tag{5.9}$$

If all of the relative rates in R except $R_{G \leftrightarrow T}$ are estimated, we are fitting a general time-reversible model of substitution. However, it is sometimes useful to consider simpler nested models. One such model is the HKY model (HASEGAWA *et al.* 1985). In the HKY model transitions occur at rate κ relative to transversions. Thus $R_{A \leftrightarrow G} = R_{C \leftrightarrow T} = \kappa$ and $R_{A \leftrightarrow C} = R_{A \leftrightarrow T} = R_{C \leftrightarrow G} = R_{G \leftrightarrow T} = 1$. Either a Jeffreys' prior or a uniform prior can be used for the relative rates. However, as a result of our parameterisation, the Jeffreys' prior provides more accurate estimates. In the examples that follow, a uniform prior is used for R and κ as this represents the most ignorant state of knowledge and is more than adequate for the purpose of illustrating the methodology. In the same spirit $f_r(r)$ is set uniform on r, and this also proves acceptable.

5.5 Examples

In this section, we illustrate our methods on two HIV-1 *env* data sets and a series of synthetic data sets of comparable size.

5.5.1 HIV-1 *env* data

The method was first tested on HIV-1 partial envelope sequences obtained from a single patient over five sampling occasions spanning approximately 3 years: an initial sample (day 0) followed by additional samples after 214 days, 671 days, 699 days and 1005 days.

This data is the same dataset analysed in Chapters 3 and 4. An important feature of this data is that monotherapy with Zidovudine was initiated on day 409 (DRUMMOND *et al.* 2001) and continued during the remainder of the study. The total dataset consists of 60 sequences from these five time points and the length of the alignment is 660 nucleotides. Gapped columns were included in the analysis. The evidence for recombination seems to be negligible in this dataset (RODRIGO *et al.* 1999) and recombination is ignored for the purposes of illustrating our method. Rough estimates of N_e may be obtained by assuming a generation length of $\rho = 1$ day per generation (RODRIGO *et al.* 1999). However, we emphasize that we estimate $N_e\rho$ only in this work. The dataset was split into two subsets for separate analysis. One contained all pre-treatment sequences (28 sequences), and the other contained all sequences after treatment commenced (32 sequences; henceforth called post-treatment). The rationale behind this split is that a replication inhibitor such as Ziduvodine may affect both population size and mutation rate per unit time. In all of the analyses, base frequencies were fixed to empirically determined values, however, inference of these would have been trivial. Two analyses were undertaken on each dataset. The pre-treatment data is strongly informative for all parameters estimated. The results are robust to the choice of priors and MCMC convergence is quick. In contrast, the post-treatment data is only weakly informative for μ, θ and t_{root} parameters, the results are sensitive to the choice of prior and MCMC convergence is very slow.

5.5.1.1 Pre-treatment data, constant population size, HKY substitution

In this first analysis of the pre-treatment dataset, we fit the HKY substitution model and assume a constant population size. We are estimating μ, θ, g, and κ. The methods are illustrated using uniform prior distributions on μ and θ, an upper limit on mutation rate of $\mu^*=1$, a lower limit on $N_e\rho$ of $\theta^*=1$ and a very conservative upper limit on t_{root} of $t^*=10^7$ days. Ten MCMC runs were made, with starting values for mutation rate distributed on a log scale from 5×10^{-3} down to 10^{-7} mutations per site per day. This range greatly exceeds the range of values supported by the posterior. In order to test MCMC convergence on tree topologies, each of the ten MCMC runs was started on a random tree drawn from a coalescent distribution with population size equal to one thousand (in exploratory work we initialize on a sUPGMA or neighbour-joining topology). The 10 Markov chain simulations were run for 2,000,000 steps and the first 100,000 steps were discarded as burn-in. Each run took about four hours on a machine with a 700MHz Pentium III processor. The mean integrated autocorrelation time (IACT) of the mutation rate parameter was 4190, giving an effective sample size (ESS) of approximately 450 per simulation. Table 5.1 presents parameter estimates for all ten runs, illustrating close concordance between runs. Note also that the variability of estimated means between runs is in line with standard errors estimated within runs. This is a consistency check on our estimation of the IACT. Figure 5.2 shows the marginal posterior density of μ and θ for each of the ten runs. In all ten runs the consensus tree computed from the MCMC output was the same, despite the fact that the starting trees were drawn randomly (data not shown). Combining the output of all ten runs, the 95% HPD (highest posterior density) intervals for the mutation rate and t_{root} are respectively (4.20, 8.28)×10^{-5} mutations per site per day, and (580, 1040) days.

Figure 5.2 Marginal posterior densities for the pre-treatment dataset.

Ten independent runs are shown. Each run was started on a random topology and the initial mutation rates ranged from 5e-3 to 1e-7. (A) The mutation rate densities and (B) the densities for the parameter θ are both shown.

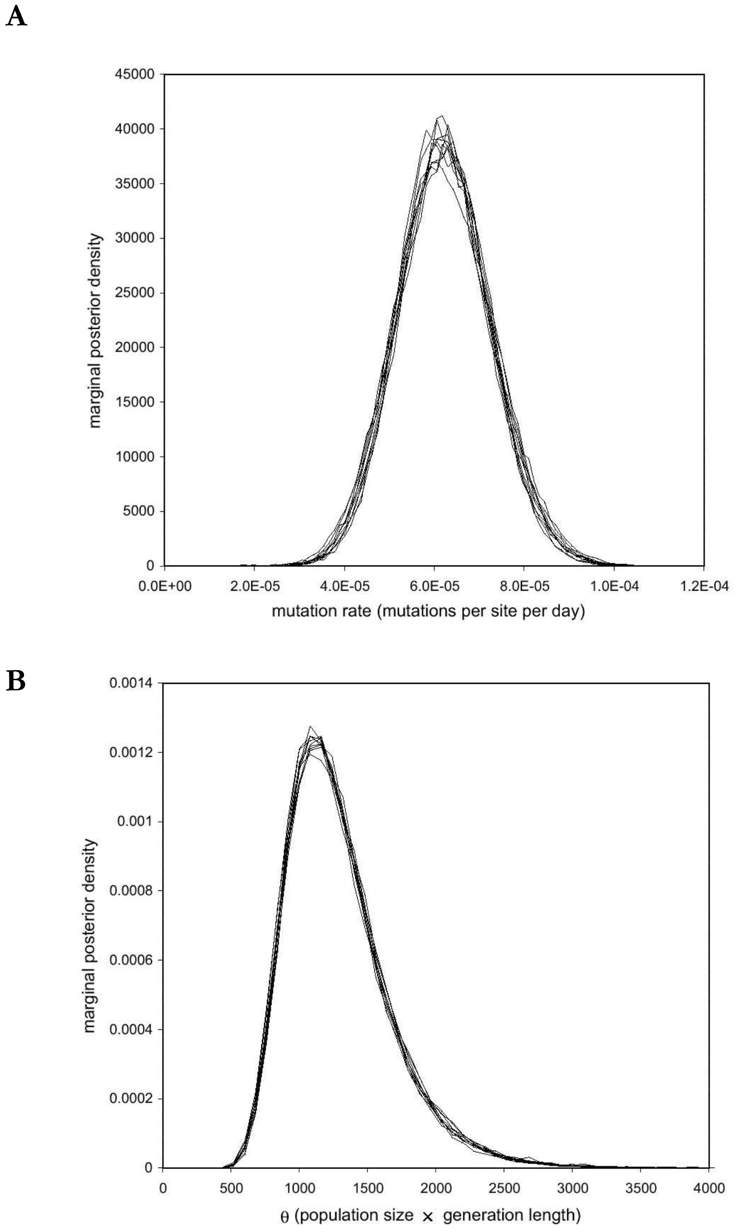


Table 5.1 Parameter estimates for ten independent analyses of the pre-treatment dataset with a simple model.

A constant population size and HKY model of mutation was assumed.

Run	Mutation rate (mutations generation^{-1} site^{-1} $\times 10^5$)	Population size $\times$ generation length (θ)	Age of root (days)	Transition/ transversion bias parameter (κ)
1	6.238 (0.0517)[a]	1284 (13.0)	796 (6.03)	4.132 (0.00634)
2	6.173 (0.0498)	1304 (12.7)	799 (5.99)	4.141 (0.00599)
3	6.218 (0.0466)	1291 (12.7)	794 (5.45)	4.124 (0.00631)
4	6.168 (0.0434)	1303 (14.0)	797 (5.65)	4.138 (0.00629)
5	6.297 (0.0474)	1269 (12.8)	784 (5.45)	4.134 (0.00640)
6	6.159 (0.0458)	1309 (12.4)	802 (6.21)	4.135 (0.00630)
7	6.308 (0.0539)	1270 (13.9)	784 (5.90)	4.130 (0.00678)
8	6.256 (0.0463)	1279 (11.5)	790 (5.63)	4.133 (0.00674)
9	6.247 (0.0474)	1283 (13.1)	791 (5.75)	4.122 (0.00661)
10	6.201 (0.0578)	1291 (15.4)	801 (7.54)	4.123 (0.00736)
Overall	6.227	1288	794	4.131
95% HPD interval	(4.20, 8.28)	(660, 2050)	(580, 1040)	(3.07, 5.31)

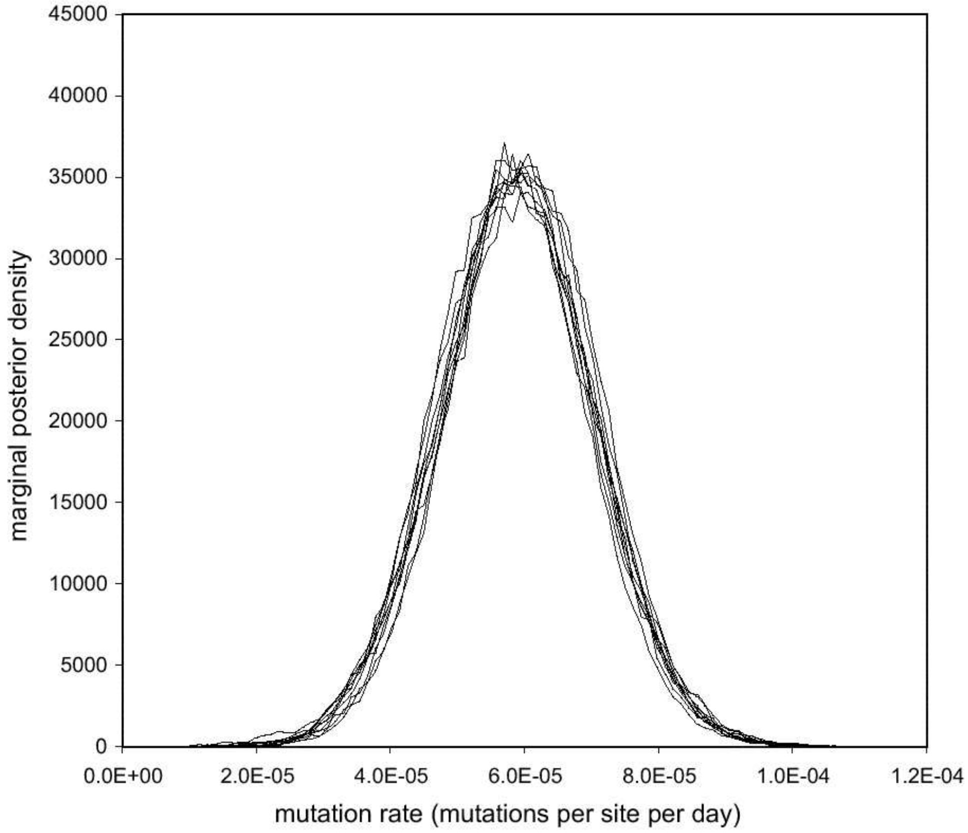
[a]Numbers in brackets are the standard errors of the means calculated using IACT statistic.

5.5.1.2 Pre-treatment data, exponential growth, general substitution model

In this second analysis of the pre-treatment dataset, we fit the general-time reversible substitution model, with exponential growth of population size. We are estimating μ, θ, g, r, $R_{A\leftrightarrow C}$, $R_{A\leftrightarrow G}$, $R_{A\leftrightarrow T}$, $R_{C\leftrightarrow G}$ and $R_{C\leftrightarrow T}$. This is the most parameter-rich model we fit. To assess the convergence characteristics of this analysis we ran ten independent runs of 3,000,000 cycles, each starting with an independent random tree topology (the mean IACT for μ was 7955 giving an ESS of 358 per run). Figure 5.3 shows the ten estimates of the marginal posterior density of mutation rate. Table 5.2 shows parameter estimates for each of the ten runs. Convergence is still achieved with the extra parameters.

Figure 5.3 Marginal posterior density for the pre-treatment dataset assuming exponential growth rate and a GTR model of substitution.

Ten independent runs are shown. Each run was started on a random topology and the initial mutation rates ranged from 5e-3 to 1e-7.



Compare the distribution of summary statistics under the two models - described here and in section 5.5.1.1. Given the nature of infection of HIV-1, it seems likely that an exponential growth rate assumption is more accurate. Estimated 95% HPD intervals for the growth rate $r = (1.09\times10^{-3}, 6.65\times10^{-3})$ exclude small growth rates, corroborating this view. The 95% HPD intervals for the mutation rate and t_{root} are respectively (3.61, 8.11) $\times 10^{-5}$ mutations per site per day, and (570, 1090) days. Compare these with the model in section 5.5.1.1. The change in model has minimal effect (< 10%) on the posterior mean mutation rate.

Table 5.2 Parameter estimates for ten independent analyses of the pre-treatment dataset with a complex model.

An exponential growth model and GTR model of substitution are assumed.

Run	Mutation rate (mutations generation^{-1} site^{-1} ×10^{5})	Population size × generation length (θ)	Age of root (days)	Growth rate (r×10^{3})
1	5.910 (0.0623)[a]	5404 (127)	800 (7.43)	3.815 (0.0407)
2	5.761 (0.0526)	5321 (125)	821 (7.05)	3.719 (0.0436)
3	6.045 (0.0550)	5089 (123)	786 (6.85)	3.832 (0.0418)
4	5.891 (0.0708)	5443 (172)	806 (8.56)	3.839 (0.0377)
5	5.849 (0.0609)	5338 (113)	812 (8.05)	3.815 (0.0423)
6	5.930 (0.0615)	5242 (170)	804 (8.66)	3.748 (0.0409)
7	5.857 (0.0589)	5318 (148)	806 (7.33)	3.780 (0.0388)
8	5.809 (0.0605)	5236 (123)	817 (7.51)	3.696 (0.0382)
9	5.982 (0.0542)	5064 (127)	795 (5.63)	3.786 (0.0382)
10	5.859 (0.0692)	5306 (188)	813 (10.2)	3.708 (0.0400)
Overall	5.889	5276	806	3.774
95% HPD interval	(3.61, 8.11)	(920, 12450)	(570, 1090)	(1.09, 6.65)

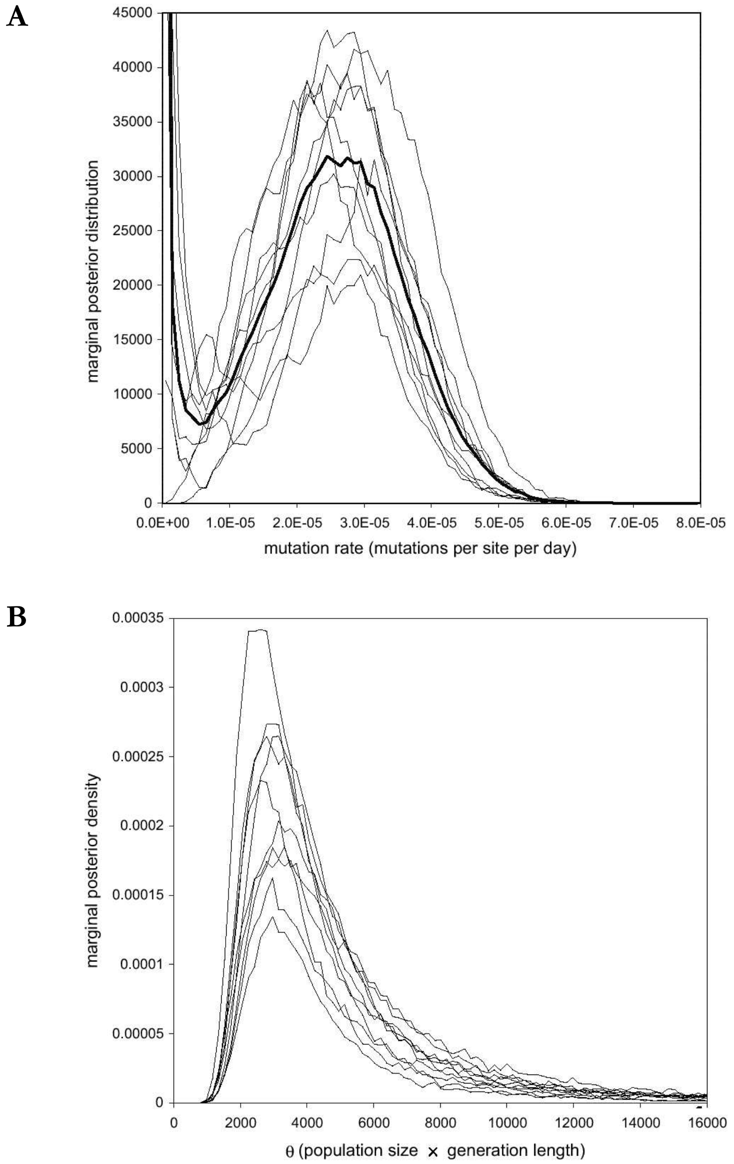
[a] Numbers in brackets are the standard errors of the means calculated using IACT statistic.

5.5.1.3 Post-treatment

The post treatment data is analysed twice under the HKY substitution model with constant population size. The first analysis uses the same priors as the first pre-treatment analysis. In contrast to the pre-treatment dataset, the mutation rate of the post-treatment dataset is difficult to estimate. This is illustrated in Figure 5.4, in which the marginal posterior densities of μ and θ estimated from ten independent MCMC runs, each 5,000,000 cycles long, are shown. We were unable to compute an IACT for each run, so we are unable to compare within and between run variability. However, the between run concordance visible in Figure 5.4 justifies the following statement. The post-treatment mutation rate shows one mode at about 2.8×10^{-5} mutations per site per day, with a second mode on the lower boundary. The data determines a diffuse, and bimodal, marginal posterior on μ. One of the modes is associated with states (μ, θ, g) with physically unrealistic root times (greater than the age of the patient). These are allowed, if we are not prepared to assert some restriction on t_{root}. This behaviour also occurs when we use a Jeffreys' prior on the mutation rate (data not shown). It reflects a real property of the data, namely that states of low μ and large t_{root} are not well distinguished from otherwise identical states of larger μ and smaller t_{root}.

Figure 5.4 Marginal posterior densities for the post-treatment dataset.

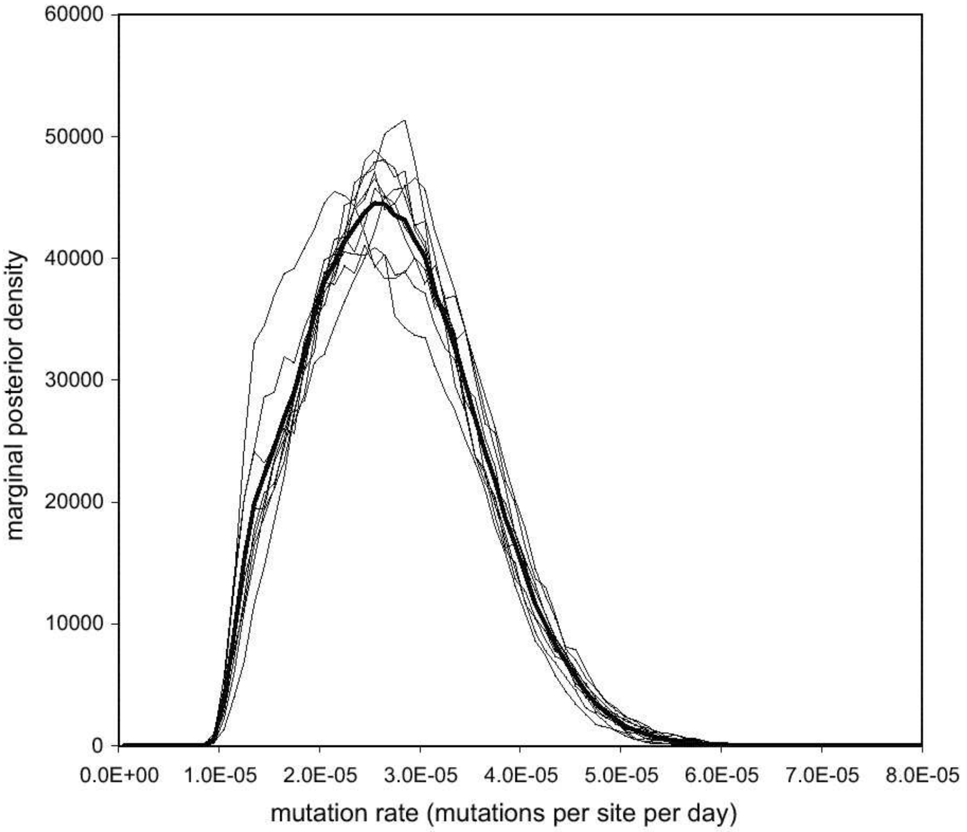
Ten independent runs are shown. Each run was started on a random topology and the initial mutation rates ranged from 5e-3 to 1e-7. (A) The mutation rate densities (dark line is mean) and (B) the densities for the parameter θ are both shown.



In the second post-treatment analysis, we revise the upper limit on t_{root} downwards from 10^7 to t^*=3650, a value more representative of actual prior knowledge for this data set. The new limit, set 3 years before seroconversion occurred in the infected patient, is still conservative. Here we explored the prior belief that HIV infection most often originates from a small, homogenous population and then subsequently accumulates variation. This prior effectively assumes that all viruses in an infected individual share a common ancestor at most as old as the time of infection of the host. Estimated 95% HPD interval for the mutation rate was (1.16, 4.27) × 10^{-5} mutations per site per day, markedly down on the pre-treatment mutation rate. Figure 5.5 depicts the resulting uni-modal marginal posterior density for mutation rate, showing that the spurious mode has been eliminated. Again, no IACT was computed. However, between run variability was much improved over Figure 5.4. Information about t_{root} has been converted into information about mutation rates and population size.

Figure 5.5 Marginal posterior density for the post-treatment mutation rate assuming a upper limit on t_{root}.

Ten independent runs are shown. Each run was started on a random topology and the initial mutation rates ranged from 5e-3 to 1e-7. The dark line shows the mean density.



5.5.2 Simulated sequence data

To test the ability of our inference procedure to recover accurate estimates of parameters from the above HIV-1 dataset, we undertook four simulation studies. In each experiment we generated 100 synthetic datasets. For experiment 1, the posterior estimates of θ, μ and κ obtained from the pre-treatment dataset in section 5.5.1.1 were used to generate 100 coalescent trees and then simulate sequences on each of the resulting trees. The synthetic data was generated under a constant-size population model with HKY mutation model, but analysed under an exponentially growing population model and a GTR mutation model.

In the second experiment, 100 synthetic datasets were generated using the pre-treatment parameter estimates in section 5.5.1.2 as the true values. In this case, the models for simulation and inference are matched. Synthetic data was generated under an exponentially growing population model and a GTR mutation model. In both experiments 1 and 2, uniform bounded priors were used for all parameters. Experiments 3 and 4 differed from experiments 1 and 2 only in that we used Jeffreys' prior for scale parameters (mutation rate, population size and relative rates).

All datasets had the same number of sequences (28), the same sampling times (0 and 214 days) and the same sequence length (660) as the pre-treatment dataset. Table 5.3 shows that the true values are successfully recovered (i.e. fall within the 95% HPD interval) ≥ 90% of the time in all cases except for the relative rate parameters in experiment 1. In the most complex model we fit, we recover true parameter values. The over-parameterisation present in experiments 1 and 3 does not seem problematic for estimating mutation rate, θ or growth rate. These results suggest that inference of biologically realistic growth rates is quite feasible. The relative rates performed the most poorly of the parameters of interest. This is caused predominantly because the uniform prior on relative rates introduces metric factors that inflate the densities. In experiment 1, when the true value of a relative rate parameter was not within the 95% HPD interval (which occurred 75 times out of 500), it was almost always over estimated (74 out of 75 times).

However, experiments 3 and 4 demonstrate that the use of a Jeffreys' prior for these and other scale parameters results in > 90% recovery in all parameters. We are not aiming to prescribe any particular non-informative prior. Our choice of uniform prior in earlier experiments is deliberately crude, but it allows us to describe the methodology with as little emphasis as possible on prior elicitation. The reader should undertake this process for their specific problem.

Table 5.3 Simulation studies to assess the performance of Bayesian inference.

This table shows the percentage of times that the true parameter was found in the 95% HPD region of the marginal posterior density.

Parameter	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Mutation rate	92	96	96	97
θ	98	99	96	97
Growth rate	91	92	94	92
$R_{A\to C}$	87*	93	96	92
$R_{A\to G}$	79*	90	96	94
$R_{A\to T}$	83*	90	94	96
$R_{C\to G}$	88*	96	98	91
$R_{C\to T}$	88*	92	98	94

* indicates success rate significantly lower than 95%.

5.6 Discussion

We have described Bayesian coalescent-based methods to estimate and assess the uncertainty in mutation parameters, population parameters, tree topology and dates of divergence from aligned temporally spaced sequence data. The sample-based Bayesian framework allows us to bring together information of different kinds, in order to reduce uncertainty in the objects of the inference. Much of the hard work is in designing, implementing and testing a suitable Monte Carlo algorithm. We found a suite of MCMC updates that perform satisfactorily.

We have analysed two contrasting HIV-1 datasets and 400 synthetic datasets to illustrate the main features of our methods. The results of section 5.5.2 show that a robust summary of parameter-rich models, including the joint estimation of mutation rate and population size, is possible for some moderate-sized datasets. The pre-treatment data restricts the set of plausible parameter values to a comparatively small range. For this dataset, useful results can be obtained from a state of ignorance about physically plausible outcomes. This situation is in contrast to the situation illustrated in Section 5.5.1.3 by the post-treatment data. For this data set, prior ignorance implies posterior ambiguity, in the form of a bimodal posterior distribution for the mutation rate. One of these modes is supported by genealogies conflicting with very basic current ideas about HIV population dynamics. We modify the coalescent prior on genealogies to account for this prior knowledge, restricting the most recent common ancestor to physically realistic values, and thus the ambiguity in mutation rate is removed. Similar results could be obtained in a likelihood-based analysis of the post-treatment data, since the prior information amounts to an additional hard constraint on the root time of the coalescent genealogy.

There is some redundancy in the set of MCMC updates we used, in the sense that the limiting distribution of the MCMC is unaltered if we remove the scaling update (move m=1) or the Wilson-Balding update (move m=2). However, these two updates types are needed in practice. There are two time scales in MCMC, time to equilibrium, and mixing time in equilibrium. The scaling move sharply reduces mixing time in equilibrium. The Wilson-Balding update is needed to bring the equilibrium time to acceptable values. We have seen MCMC simulations, minus the Wilson-Balding move, in which an apparently stationary Monte Carlo process undergoes a sudden and unheralded mean shift at around two million updates. This problem was picked up at the debugging stage, in comparisons between our two MCMC implementations. Subsequent simulation has shown that the genealogies explored in the first two million updates of that simulation were just one of the tree-clusters supported by the target distribution.

The methods presented here reduce to those of Felsenstein and co-workers (KUHNER *et al.* 1995) in the case of a uniform prior on $\Theta = 2N_e\mu$, a fixed R, a fixed μ and contemporaneous data, if instead of summarizing results using 95% HPD interval estimates, we use the mode and curvature of the posterior density for Θ to recover the MLE estimate and its associated confidence interval.

A distinction can be made between a dataset like the pre-treatment dataset, for which there is strong statistical information about mutation rates (we refer to populations from which such datasets may be obtained as "measurably evolving" in reference to considerations in Chapter 2) and a dataset like the post-treatment data, in which the statistical signal is weak. In both of these datasets the familiar parameter $\Theta = 2N_e\mu$ is in fact well determined by the data (not shown above), so that MCMC convergence in Θ is quick. However, it is only in the pre-treatment data that this parameter can easily be separated into its two factors. This is related to the well-known problem of identifiability for population size and mutation rate. We can see that temporally spaced data may or may not contain information that allows us to separate these two factors.

In this particular example, lineages of the post-treatment viruses branch from those of the pre-treatment viral population. Consequently a more appropriate analysis for this dataset would allow for a change of mutation rate and/or population size over the genealogy of the entire set of sequences. In the case of mutation rate, this has already been demonstrated within a likelihood framework (DRUMMOND *et al.* 2001). In a Bayesian analysis, coalescence of post-treatment lineages with pre-treatment lineages will tend to limit the age of the most recent common ancestor of the post-treatment data, so that the pre-treatment lineages will play the role of the reduced upper bound t_{root}^{*} in section 5.5.1.3.

A description of the software package called MEPI (Molecular Evolutionary Population Inference), developed using the Phylogenetic Analysis Library (PAL, DRUMMOND and STRIMMER 2001) is given in Chapter 10. MEPI implements the MCMC sampler described in this chapter (including extensions such as codon position rate heterogeneity and more complex population demographic models described in the next chapter).

6 Extending Bayesian Evolutionary Inference By Example

6.1 Overview

This chapter contains three examples/case studies, each of which extends, in some small way, the Bayesian inference framework developed in the previous chapter. The extensions considered are (i) gamma-distributed rate heterogeneity among sites, (ii) inference of the age of fossil remains by *molecular dating*, (iii) estimation of piecewise logistic growth and (iv) codon-position rate heterogeneity among sites. All of these extensions are implemented in the MEPI software described in Chapter 10.

6.2 Ancient DNA of Beringian Brown Bears: A case study

The Bayesian inference framework described in Chapter 5 has been previously used to uncover an apparently elevated mtDNA mutation rate in Adelie penguins (LAMBERT *et al.* 2002). Here, a second dataset was investigated to test if this pattern is a general characteristic of vertebrate mitochondrial evolution over short time frames (hundreds of thousands of years). A set of 30 ancient sequences collected from fossil brown bear bones in eastern Beringia (BARNES *et al.* 2002) was analysed, demonstrating a similarly elevated rate to that found in Adelie penguins. The rates presented here for Beringian bears are about 2-8 times faster than the earlier estimates of Waits *et al* (1998) of 11-14% per million years. A second experiment, which included an additional 17 modern sequences, yielded similar results, indicating a 2-5 fold larger rate then conventional wisdom would suggest. These findings have a potentially large impact on dating methods that use mitochondrial DNA and bring into question basic assumptions about the time-scale invariance of evolutionary processes. The question posed is, "Is the rate of evolution the same over different time scales?"

6.2.1 Data

Two sets of data were analysed to ascertain the rate of mtDNA evolution in brown bears. Experiment 1 comprised of two sections of the mitochondrial (mt) control region, 135 and 60 bases pairs (bp), respectively, from 30 radiocarbon-dated bones ranging in age from 9995 to >59000 years old. The dataset is fully described by Barnes *et al* (2002). The six sequences reported by Barnes *et al* (2002) that were recovered from undated bones were not used in this analysis.

The second dataset (Experiment 2) was analysed to assess the consistency of ancient and modern sequence material. Experiment 2 was constructed by adding an additional 17 modern sequences to Experiment 1. These additional sequences included modern brown bears (*Ursus arctos*), polar bears (*Ursus maritimus*) and a single black bear outgroup (*Ursus americanus*). All of these sequences have been previously used in Barnes *et al* (2002).

In both datasets the raw radiocarbon dates were used as surrogates for the real age (in years before present; years BP) of the bones. Although calibration curves exist for converting from radiocarbon dates to years BP, these calibration curves are currently only accurate as far back as about 20,000 years (STUIVER *et al.* 1998). As eleven of the sequences analysed had radiocarbon ages between 35,000 and 60,000, all ages were left uncalibrated. The extent to which this affects the analysis is partially investigated in section 6.3.

On the larger dataset, a further experiment (Experiment 3) was undertaken to estimate the amount of rate heterogeneity among sites and to assess the influence of rate heterogeneity on overall rate estimates and estimates of divergence times.

6.2.2 Results

Markov chain Monte Carlo integration (as described in Chapter 5 and implemented in the software package MEPI) was used to jointly estimate the divergence times, tree topology, mutation rate and transition/transversion ratio of two sets of mtDNA partial control region sequences. The posterior probability density under consideration in Experiments 1 and 2 is:

$$h(\mu,\theta,r,\kappa,g \mid D) = \frac{1}{Z}\Pr\{D \mid \mu,\kappa,g\}\, f_G(g \mid \theta,r)\, f_{\mathrm{M}}(\mu)\, f_{\Theta r}(\theta,r)\, f_\kappa(\kappa) \qquad (6.1)$$

Where:

μ is the mutation rate in mutations per year.

θ is the product of effective population size and generation length in years

r is the exponential growth rate of the population

κ is the transition/transversion bias

g is the genealogy g = (E_g, t_Y), the branching topology E_g and ancestral ages.

D is the sequence alignment data

$\Pr\{D|\mu, \kappa, g\}$ is the likelihood

$f_G(g\,|\,\theta, r)$ is the coalescent probability density

$f_M(\mu)$ is a uniform prior density on mutation rate.

$f_{\Theta r}(\theta, r)$ is a uniform prior density on demographic parameters.

$f_\kappa(\kappa)$ is a uniform prior density on transition/transversion bias.

In addition a novel extension to the MCMC method was developed to allow for the estimation of rate heterogeneity among sites in Experiment 3.

6.2.2.1 Experiment 1: ancient sequences only

The results of four independent MCMC runs, each starting from a random tree topology, are shown in Table 6.1. The first two runs assume an exponentially growing (or declining) population, while the second two runs assume a constant population size through time. Assuming an exponential growth rate, the estimated mutation rate was about 6.4×10^{-7}, with 95% highest posterior density (95% HPD) upper and lower limits of 2.5×10^{-7} and 10.3×10^{-7} respectively. The assumption of a constant population size gave very similar results. This estimate is 2-8 times larger than recent estimates of HVR1 substitution rate in brown bears (WAITS *et al.* 1998) using standard fossil-calibration techniques.

Table 6.1 Experiment 1: Analysis of 30 ancient sequences.

Results from four independent MCMC runs. The first two assume an exponentially growing population and the second two assume a constant population. The numbers in parentheses are the 95% highest posterior densities (95% HPD).

Parameter	**Run 1**	**Run 2**	**Run 3**	**Run 4**
Mutation rate (s/s/10^7 years)	6.34 (2.65-10.4)	6.37 (2.44 – 10.2)	6.14 (2.62-10.4)	6.11 (2.13-10.2)
Kappa	25 (8-49)	25 (8-49)	25 (9-50)	25 (8-49)
t_{MRCA} (1000's of years BP)	152 (91-233)	153 (90-233)	160 (93-249)	161 (90-260)
$N_e\rho$ (1000's)	67 (6-180)	66 (6-175)	65 (17-131)	66 (18-144)
Growth rate (x 10^{-6})	-1.9 (-22-17)	-2.1 (-22-17)	0.0*	0.0*
Effective sample size	516	599	589	574

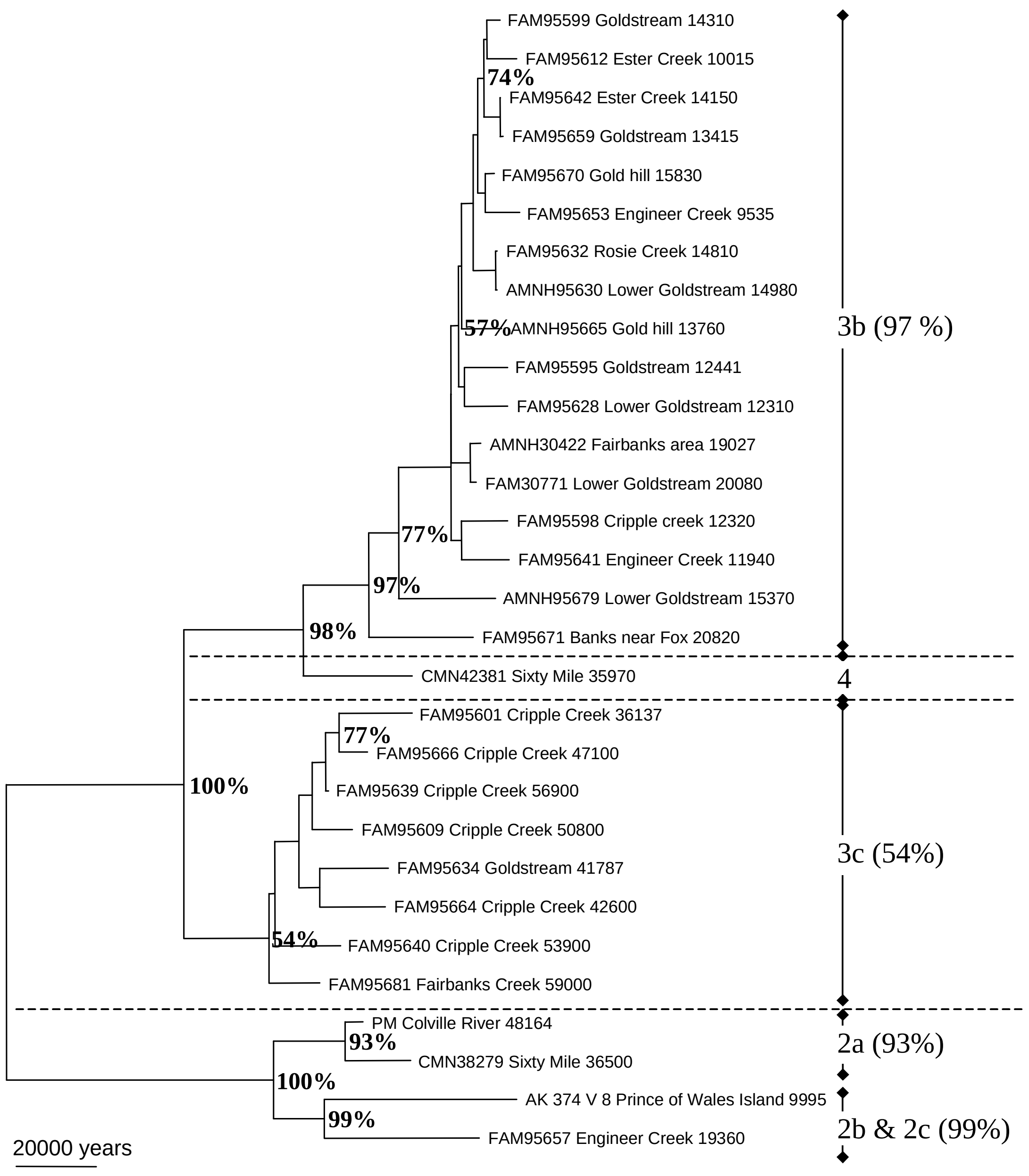
*The growth rate was constrained to zero in runs 3 and 4.

Unsurprisingly for mtDNA, the transition/transversion ratio is heavily biased towards transitions, estimated at $\kappa \approx 25$ (95% HPD: 8-49). The age of the most recent common ancestor (MRCA) of this group of ancient sequences was estimated at 150,000 years BP (90,000 – 230,000), much more recent than would be suggested by conventional estimates of mtDNA rates in this region (WAITS *et al.* 1998).

A sample tree from the posterior distribution of Run 2 is displayed in Figure 6.1. The posterior probabilities of individual clades are presented, showing that the basic backbone of the tree is well determined, but there is significant uncertainty in the order of branching events within each major clade. This is unsurprising as only 21 of the 30 sequences are unique. In fact, identical sequences don't always form monophyletic groups! This is due to back mutations (reversions). Even though a particular reversion has a low probability of occurring, there are a lot more possible sequences with one or more reversions than with none, so the observation of at least some reversions in a set of sequences becomes more likely with more sequences. In general, the structure of the tree, despite being constrained to a molecular clock, is concordant with the neighbour-joining (NJ) tree presented by Barnes *et al* (2002). Both exponential growth and constant population size models gave similar results for all estimated parameters.

Figure 6.1 Sample tree from Experiment 1.

Clades that have a posterior probability of greater than 50% are labelled. The clade designation used by Barnes *et al* (2002) is also shown.


FAM95599 Goldstream 14310
FAM95612 Ester Creek 10015
74%
FAM95642 Ester Creek 14150
FAM95659 Goldstream 13415
FAM95670 Gold hill 15830
FAM95653 Engineer Creek 9535
FAM95632 Rosie Creek 14810
AMNH95630 Lower Goldstream 14980
57%
AMNH95665 Gold hill 13760
FAM95595 Goldstream 12441
FAM95628 Lower Goldstream 12310
AMNH30422 Fairbanks area 19027
FAM30771 Lower Goldstream 20080
FAM95598 Cripple creek 12320
77%
FAM95641 Engineer Creek 11940
97%
AMNH95679 Lower Goldstream 15370
98%
FAM95671 Banks near Fox 20820
CMN42381 Sixty Mile 35970
FAM95601 Cripple Creek 36137
77%
FAM95666 Cripple Creek 47100
100%
FAM95639 Cripple Creek 56900
FAM95609 Cripple Creek 50800
FAM95634 Goldstream 41787
FAM95664 Cripple Creek 42600
54%
FAM95640 Cripple Creek 53900
FAM95681 Fairbanks Creek 59000
PM Colville River 48164
93%
CMN38279 Sixty Mile 36500
100%
AK 374 V 8 Prince of Wales Island 9995
99%
FAM95657 Engineer Creek 19360
20000 years
3b (97 %)
4
3c (54%)
2a (93%)
2b & 2c (99%)

6.2.2.2 Experiment 2: ancient and modern bear sequences

The addition of 17 modern sequences has a significant impact on the estimated mutation rate. The estimate is revised downward to around 4.4×10^{-7} (95% HPD: 2.5×10^{-7}-6.4×10^{-7}). The results of four independent MCMC runs, each starting from a random tree topology, are shown in Table 6.2. As with Experiment 1, the first two runs assume an exponential growth (or decline) in the Beringian bear population and the second two runs assume a constant population size through time. The assumption of a constant population size results in a slightly lower estimated mutation rate, of 4.24×10^{-7}, averaged over two runs.

When compared with Experiment 1, the lower limit is about the same, but the upper limit is significantly reduced. The estimated rate in Experiment 2 is still 2-5 times faster then previous estimates of 1.1×10^{-7}–1.4×10^{-7} (WAITS *et al.* 1998). The transition/transversion ratio is still very uncertain but concordant with the estimate established in Experiment 1. The 95% HPD interval of the transition/transversion bias is about 15-75.

The estimated age of the most recent common ancestor of black, brown and polar bears is 120,000 – 320,000 years ago, under the exponential model. In both the first two runs, the estimated growth rate spans zero, suggesting little support for concerted growth or decline of bear populations over the last hundred thousand years.

Table 6.2 Experiment 2: Analysis of 30 ancient sequences and 17 modern sequences.

Results from four independent MCMC runs. The first two assume an exponentially growing population and the second two assume a constant population. The numbers in parentheses are the 95% highest posterior densities (95% HPD).

Parameter	Run 1	Run 2	Run 3	Run 4
Mutation rate (s/s/10^7 years)	4.44 (2.56-6.33)	4.44 (2.52 – 6.41)	4.21 (2.38-6.18)	4.27 (2.51-5.98)
Kappa	40 (14-76)	40 (15-75)	41 (15-76)	41 (14-75)
t_{MRCA} (1000s of years BP)	210 (120-320)	210 (120-320)	235 (130-365)	230 (130-350)
$N_e\rho$ (1000s)	290 (115-525)	285 (100-520)	200 (100-335)	190 (95-310)
Growth rate (×10^{-6})	8.4 (-1.6-19)	8.3 (-1.3-19)	0.0*	0.0*
Effective sample size	611	624	618	741

*The growth rate was constrained to zero in runs 3 and 4.

The posterior clade probabilities support many details of the tree presented in Barnes *et al* (2002), as shown in Figure 6.2. One important detail is different; in this tree, clade 4 falls within clade 3.

6.2.2.3 Experiment 3: ancient and modern bear sequences with rate heterogeneity

Recently it has been suggested that the elevated mutation rates observed in mtDNA over short time frames is due to a large amount of rate heterogeneity among sites in the control region of mtDNA (HEYER *et al.* 2001). To test this hypothesis, a third experiment was undertaken, in which rate heterogeneity among sites was modelled using a discrete approximation of the gamma distribution (YANG 1994) with four rate categories. The shape parameter of the gamma distribution was estimated using MCMC.

Let α be the shape parameter of the gamma distribution. Following equation 5.6 in Chapter 5, the posterior probability density under examination here is:

$$
\begin{aligned}
&h_{\Gamma}(\mu,\theta,r,\kappa,\alpha,g \mid D) = \\
&\qquad \frac{1}{Z}\Pr\{D \mid \mu,\kappa,g,\alpha\}\, f_G(g \mid \theta,r)\, f_{\mathrm{M}}(\mu)\, f_{\Theta r}(\theta,r)\, f_{\alpha}(\alpha)\, f_{\kappa}(\kappa)
\end{aligned}
\tag{6.2}
$$

$f_{\alpha}(\alpha)$ is a uniform prior density on the shape parameter. The probability $\Pr\{D \mid \mu,\kappa,g,\alpha\}$ was calculated by the method of Yang (YANG 1994), using a discretization of the gamma distribution and integrating over all rate categories at all sites, as implemented in the open-source programming library PAL (DRUMMOND and STRIMMER 2001). It should be noted that as the shape parameter of the gamma distribution approaches infinity, the gamma distribution approaches a singleton distribution at which all sites evolve at the same rate. Conversely, as the shape parameter approaches zero, most sites evolve extremely slowly while a few sites have widely distributed rates. Thus a uniform prior density on the shape parameter was chosen to allow for arbitrarily large values, corresponding to exactly equal rates across sites.

Table 6.3 shows the results of two independent MCMC runs, estimating the gamma shape parameter from 47 Beringian brown bear sequences. It is evident that there is significant rate heterogeneity in the Beringian Brown bear sequences. The slowest of the four (equally represented) rate categories is estimated to evolve ~2000 times slower than the fastest. Furthermore, the introduction of a gamma distribution of rates increases our estimate of the mean mutation rate, to about 5.6×10^{-7} (3.1-8.4).

Table 6.3 Experiment 3: Analysis of 30 ancient sequences and 17 modern sequences with gamma-distributed rate heterogeneity among sites.

Results from two independent MCMC runs. Both analyses assume an exponentially growing population. The numbers in parentheses are the 95% highest posterior densities (95% HPD).

Parameter	Run 1	Run 2
Mutation rate (s/s/10^7 years)	5.53 (3.0-8.35)	5.61 (3.14-8.53)
Gamma shape parameter	0.27 (0.16-0.41)	0.27 (0.16-0.41)
Kappa	48 (16-89)	47 (17-88)
$N_e\rho$ (1000s)	235 (95-410)	230 (85-400)
Growth rate (×10^{-6})	6.3 (-3.0-17)	6.3 (-3.4-17)
Effective sample size	686	669

It should be apparent that the contribution of the fastest sites (with rates of approximately 18.2 x 10^{-7} [10.1-27.3]) would be quickly masked by saturation of the sites involved as the timeframe increases. The time frame over which these sites are informative can be estimated as 0.5 / 18.2×10^{-7} ≈ 275,000 years. The value of 0.5 is the diversity of saturated sequences of binary characters. This is appropriate for mtDNA in which transitions vastly outnumber transversions. Even the more generous estimate of 0.75 / 10.1×10^{-7} (assuming *no* transition/transversion bias *and* taking the lower limit of the rate estimate) will still lead to a maximum timeframe of 750,000 years over which an accurate estimate of mean mutation rate can be obtained. Table 6.4 shows the maximum time frame over which different rate categories can provide phylogenetic and molecular rate information, based on a discrete approximation of the gamma distribution with a shape parameter of 0.27 and four rate categories. The calculated saturation times for the slowest rate category are probably unreliable because over long enough time frames the rate category of a site is likely to change due to shifting structural constraints and molecular sequence context. Thus most sites will not remain in the slowest rate category for tens of millions of years. Evidence to support this conclusion exists in the inability to align control region sequences from distantly related vertebrates.

Table 6.4 The maximum time frame over which different rate categories can provide phylogenetic and molecular rate information.

The estimates are based on a discrete approximation of the gamma distribution with a shape parameter of 0.27 and four rate categories.

Category (25% each)	Mean rate (×10^{-7})	Saturation time (binary characters)	Saturation time (4 characters)
1	0.0089 (0.0049-0.013)	560,000,000	843,000,000
2	0.51 (0.28-0.77)	9,800,000	14,700,000
3	3.7 (2.0-5.5)	1,350,000	2,030,000
4	18.2 (10.1-27.3)	275,000	412,000

6.2.3 Discussion

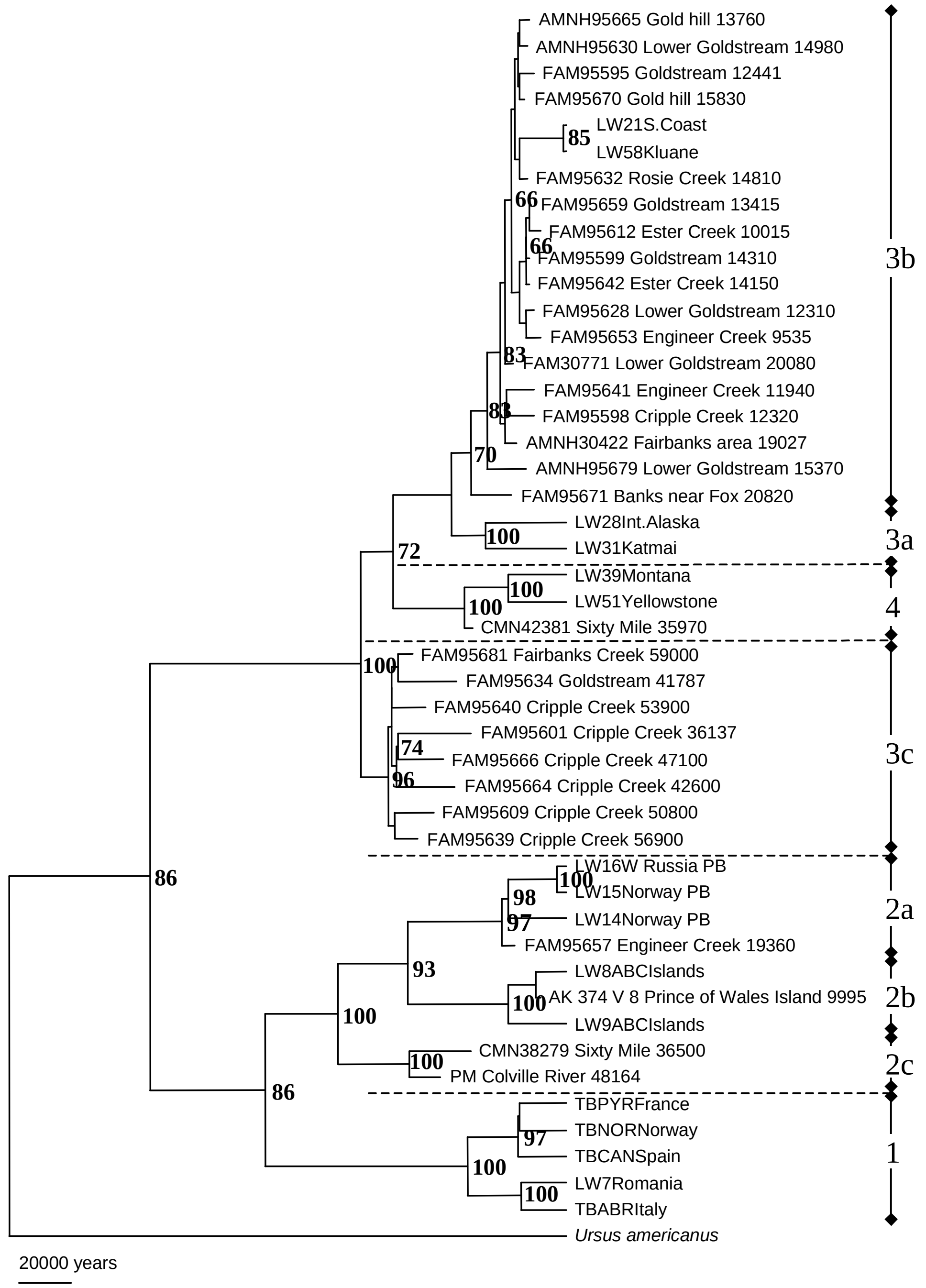
The recent use of both pedigree material (HEYER *et al.* 2001) and ancient DNA from sub-fossil bones (LAMBERT *et al.* 2002) to study the rate of mitochondrial evolution over relatively short periods (<1,000,000 years) has suggested that the rate of evolution in the control region in vertebrates may be much faster than previously thought. Here, an analysis of Beringian brown bear sequences ranging up to at least as old as 59,000 years supports these recent studies and suggests the mitochondrial rates are at least 2 times faster than previously thought when measured over short (in geological terms) time frames.

Radiocarbon dating starts to become inaccurate for dates greater than 20,000 years and fails completely to discriminate between different ages beyond an upper limit of about 50,000-60,000 years. Three of the bones used in this analysis are dated to this upper range of 50,000-60,000 years, and thus these ages should be regarded as *minimum* estimates of the actual ages. What happens if we take into account the uncertainty in their ages? One possibility is doubling the ages of the three oldest bones and repeating experiments 1, 2 and 3. A better approach might be to let the ages of these bones become parameters of the model, with a lower limit of the radiocarbon age, and then integrate over all possible ages to estimate the mutation rate. In the next section on *molecular dating*, this solution is attempted, without any large change in the estimate of mutation rate.

Why are the mutation rates estimated here faster than those estimated by Waits *et al*? A recent study of human mtDNA (HEYER *et al.* 2001) has suggested that this observation is due to extreme rate heterogeneity in the mitochondrial control region. Thus, short timeframes permit measurements of the rate of mutation at highly mutable sites, while measurements of rates over long time frames are dominated by slowly evolving sites. This effect is further exacerbated by the high transition/transversion bias in mtDNA, which causes most sites to be effectively binary. Thus the fastest sites are saturated very quickly and provide phylogenetic information only over very short time frames. The above analysis of rate heterogeneity in Beringian brown bear sequences supports this hypothesis, by uncovering extremely rapid evolution in some sites of the molecule that will be completely saturated over long time periods, resulting in an apparent reduction in evolutionary rate. If this hypothesis is correct then it has important consequences for phylogenetic dating of divergences. Rates of evolution calibrated from the fossil record over long periods (such as tens of millions of years) will not accurately reflect the rate of evolution observed over tens, or hundreds of thousands of years. Thus conservation genetic and population genetic studies that use mtDNA to date intra-specific divergences or divergences between closely related species, should revise the mutation rates used for calibration upwards. In terms of the current literature this would bring many hundreds of estimates of the age of common ancestors in mammal and bird populations nearer to the present.

Figure 6.2 Sample tree from Experiment 2.

All sub-clades in Barnes *et al* (2002) (1, 2a, 2b, 2c, 4, 3a, 3b, 3c) are monophyletic in this representative tree. In addition all have >70% posterior probabilities. Unlike Barnes *et al*, clade 3 and clade 4 are not reciprocally monophyletic. Clade 4 falls inside clade 3 and the posterior density most strongly supports its placement between 3a and 3c, however alternative positions within clade 3 have appreciable probability.


AMNH95665 Gold hill 13760
AMNH95630 Lower Goldstream 14980
FAM95595 Goldstream 12441
FAM95670 Gold hill 15830
85
LW21S.Coast
LW58Kluane
FAM95632 Rosie Creek 14810
66
FAM95659 Goldstream 13415
FAM95612 Ester Creek 10015
66
FAM95599 Goldstream 14310
FAM95642 Ester Creek 14150
FAM95628 Lower Goldstream 12310
FAM95653 Engineer Creek 9535
83
FAM30771 Lower Goldstream 20080
FAM95641 Engineer Creek 11940
83
FAM95598 Cripple Creek 12320
AMNH30422 Fairbanks area 19027
70
AMNH95679 Lower Goldstream 15370
FAM95671 Banks near Fox 20820
3b
100
LW28Int.Alaska
LW31Katmai
3a
72
100
LW39Montana
100
LW51Yellowstone
CMN42381 Sixty Mile 35970
4
100
FAM95681 Fairbanks Creek 59000
FAM95634 Goldstream 41787
FAM95640 Cripple Creek 53900
FAM95601 Cripple Creek 36137
74
FAM95666 Cripple Creek 47100
96
FAM95664 Cripple Creek 42600
FAM95609 Cripple Creek 50800
FAM95639 Cripple Creek 56900
3c
86
100
LW16W Russia PB
98
LW15Norway PB
97
LW14Norway PB
FAM95657 Engineer Creek 19360
2a
93
LW8ABCIslands
100
AK 374 V 8 Prince of Wales Island 9995
LW9ABCIslands
2b
100
100
CMN38279 Sixty Mile 36500
PM Colville River 48164
2c
86
TBPYRFrance
97
TBNORNorway
TBCANSpain
100
LW7Romania
100
TBABRItaly
1
Ursus americanus
20000 years

6.3 Molecular dating of undated and old sub-fossil bones

In section 6.2 the problem of estimating mutation rate was considered for the situation of temporally spaced molecular sequences of known ages. In viral populations, where the evolutionary rate is very fast, this temporal spacing can be achieved by simply sampling the population of interest over a relatively short time period (for example, a couple of years). In this case the ages of the sequences are known exactly. However in the case of vertebrate populations, the mutation rate being much lower, it is necessary to rely on sub-fossil remains to obtain a sufficient temporal depth to estimate evolutionary rates. In this case the ages of the sequences are *not* known exactly. For bones aged less than 10,000 years this may not be an issue as the calibration curve between radiocarbon dates and calendar dates is quite well determined (albeit non-linear). However for older bones, the radiocarbon dates are a very rough estimator, and for ages above 60,000 years radiocarbon dating is not possible.

In the situation of radiocarbon-dated material, the problem can be turned on its head somewhat. Given a number of accurately dated molecular sequences and a small number of inaccurately (or un-) dated sequences, it may be possible to gain information about the ages of the undated material from their sequence relationships to the dated material. This is a kind of *molecular dating*. In loose terms, the most likely placement and branch length of an undated sequence, in a rooted and dated tree, will provide information about how likely various ages of the said sequence are.

A set of Beringian brown bear sequences is analysed. Of the 53 sequences analysed 44 sequences are treated as having known ages, and nine sequences are treated as unknown, with or without a lower bound.

Let t_u be the set of unknown (or partially known) ages and a strict subset of the set of ages of the leaf nodes (t_I). Following equation 5.6 in Chapter 5, the posterior probability density under examination is:

$$\begin{aligned} h_{md}(\mu,\theta,r,\kappa,g,t_u \mid D) = \\ \frac{1}{Z}\Pr\{D \mid \mu,\kappa,g,t_u\}\, f_G(g,t_u \mid \theta,r)\, f_{\mathrm{M}}(\mu)\, f_{\Theta r}(\theta,r)\, f_{\kappa}(\kappa)\, f_T(t_u) \end{aligned} \tag{6.3}$$

The prior $f_T(t_u)$ can accommodate various upper and lower limits on the ages of undated (or partially dated) sequence. This prior information could be based on radiocarbon dating, stratigraphic information and other auxiliary sources of information. In the context of this analysis, it consists of an upper limit on all nine bones of uncertain age, and a non-trivial lower limit on three of them. Both uniform and Jeffrey's prior densities were investigated, within the upper and lower boundaries.

6.3.1 Methods and materials

The dataset was comprised of the 47 sequences analysed in Experiment 2 above, along with an additional six undated sequences from Barnes *et al* (2002). This represents all of the ancient material published in Barnes *et al* (2002) and most of the same modern material. The three oldest sequences in Experiment 1 and 2 (FAM95640, FAM95639 and FAM 95681) were treated as undated with lower limits of 53900, 56900 and 59000 respectively.

Simple random walk and scaling proposal mechanisms were tested, and both performed adequately for the purposes, although the scaling proposal mechanisms required less fine-tuning of parameters, in all cases performing adequately with a simple 0.5-2.0 scaling range. Each proposal of a new age for an undated sequence was made independently in the MCMC chain. This represents the simplest possible proposal scheme, and while proving adequate for the problem at hand, could undoubtedly be improved on.

6.3.2 Priors

A lower limit of 1 year was assumed for the six undated sequences. The three old sequences were given lower limits of 53900, 56900 and 59000 as mentioned earlier. An upper limit on age was also introduced as a prior for all undated material. The upper age was set at 150,000 years. The rationale for this selection is based on the fact that sequence data was recovered from the bones in question. With current techniques, the upper age limit at which recovery of a 135 bp fragment is still feasible is probably about 80,000-100,000 years (Alan Cooper, *personal communication*). This serves as a natural *a priori* belief, based on expertise. In order to be conservative the upper limit was extended out to 150,000 years BP. Within the upper and lower boundaries, both Jeffreys' and uniform distributions were investigated as priors.

6.3.3 Results

Firstly, a naïve analysis of the dataset was undertaken, in which all nine sequences were given a uniform prior on age (with appropriate lower limits in the case of the three old sequences), with an upper limit of 150,000 years BP. Only two of the sequences had posterior densities with considerably lower upper limits than the prior (KU23034, FAM95597). This suggests that there was not enough sequence data to estimate precise posterior densities for the other seven sequences. Table 6.5 shows the estimated ages of the nine sequences with unknown dates.

A second analysis was also undertaken, in which a Jeffreys' prior on sequence age was used. This implies that an age ten times younger is ten times more likely and thus strongly favours younger ages. One of the six undated bones (IB_Duvanny_Yar) and the three old bones (FAM95640, FAM95639 and FAM 95681) maintained a strong signal for antiquity under this prior. Low estimates of age were established for the remaining five bones under this prior. The estimates of ages were therefore sensitive to selection of a prior as can be seen in Table 6.5.

Table 6.5 Estimates of bone ages based on sequence comparison to known ages using MCMC.

Estimates with uniform prior and Jeffreys' priorgive are shown. In all cases the upper limit on age was assumed to by 150,000 years BP. IB_Duvanny_Yar was the only undated bone that had strong sequence signal suggesting that it was an old bone. KU23034 and FAM95597 both had molecular signals indicating young bones. The three old radiocarbon dated bones (FAM95640, FAM95639 and FAM 95681) all have molecular signals that indicate they could easily be ~20,000 years older than the minimum estimates established by radiocarbon dating.

Sequence Id	Location	Radiocarbon date	Estimated age (uniform prior) rounded to the nearest 1000 years	Estimated age (Jeffreys' prior) rounded to the nearest 1000 years
RSM1962/63	Scotland	-	48000 (0-130000)	3000 (0-20000)
IB	Duvanny Yar	-	70000 (12000-136000)	37000 (0-87000)
KU23034	Ice Cave	-	16000 (0-45000)	2000 (0-10000)
FAM95596	Goldstream	-	67000 (0-136000)	7000 (0-37000)
FAM95597	Gold hill	-	25000 (0-63000)	6000 (0-22000)
AMNH30421	Fairbanks	-	45000 (0-114000)	2000 (0-11000)
FAM95640	Cripple Creek	>53900	84000 (53900-130000)	69000 (53900-94000)
FAM95639	Cripple Creek	>56900	97000 (56900-140000)	79000 (56900-108000)
FAM95681	Fairbanks Creek	>59000	92000 (59000-138000)	77000 (59000-108000)

6.3.4 Discussion

The use of uniform and Jeffreys' priors gave very different results for the estimated absolute ages of undated bones, suggesting that there is not enough sequence information to determine precise ages of the undated bones by the *molecular dating* technique described. Despite this, molecular dating still provided useful information. IB_Duvanny_Yar was the only undated bone that had strong sequence signal suggesting that it was an old bone, even under the Jeffreys' prior. KU23034 and FAM95597 both had strong molecular signals for being young bones. The three old radiocarbon dated bones (FAM95640, FAM95639 and FAM 95681) all had molecular signals that indicate they could easily be ~20,000 years older than the minimum estimates established by radiocarbon dating. This kind of qualitative information could help to make decisions as to which undated, but sequenced, bones should be radiocarbon dated. However this analysis also suggests that precise molecular dating is not possible without obtaining much longer sequences (>>200 bp) from undated bones.

6.4 Hepatitis C in Egypt: A non-MEP analysis

The number of people infected by Hepatitis C virus (HCV) worldwide has been estimated at about 170 million (World Health Organisation; 1997). HCV is a positive, single-stranded RNA virus and is a member of the *Flaviviridae* family of viruses. The HCV genome is ~10,000 nucleotides long, encoding a single polyprotein of about 3,000 amino acids. Egypt has a high prevalence of HCV, estimated at about 10-20% of the total population. It has been suggested that the high incidence of HCV in Egypt has resulted from the use of unsterile injection equipment during widespread treatment of the population with parenteral antischistosomal therapy (PAT) from the 1920s to the 1980s (FRANK *et al.* 2000; HABIB *et al.* 2001; NAFEH *et al.* 2000). In this section a computationally intensive sample-based Bayesian inference approach is used to investigate the demographic signal contained in modern Hepatitis C viral sequences obtained from infected members of the Egyptian population. This dataset provides two interesting methodological challenges: (i) piecewise logistic demographic model and (ii) codon-position specific rate heterogeneity.

6.4.1 The piecewise logistic demographic model

Coalescent theory can be used to calculate the expected distribution of times of coalescent events given a parametric demographic model. In chapter 5 two simple parametric models of demography were investigated: the constant population size model, $N(t) = N_C$, and the exponential growth model, $N(t) = N_C e^{-rt}$. Neither of these simple models of population size dynamics is an adequate description of the PAT-mediated population expansion in the Egyptian HCV epidemic over the last century. Therefore, to facilitate investigation of the HCV epidemic a new model called the *piecewise logistic* model of demography is introduced:

$$N(t) = \begin{cases} N_C & \text{if } t \le x \\ N_C e^{-r(t-x)} & \text{if } x < t < y \\ N_A & \text{if } t \ge y \end{cases} \tag{6.4}$$

This model allows for a small ancestral population to undergo exponential expansion over a finite period of time, resulting in a large modern population. N_C is the current population size, assumed to have been constant back in time until time x, and N_A is the ancestral population size, assumed constant at times early than time y. For the interval between x and y, exponential growth (with growth rate r) is assumed. For the current purposes the units of population size can be assumed to be *effective number of infections*. It should be noted that this model is completely specified with four parameters, although five are given. Any one of the five parameters given is fully determined by the values of the other four. For example, given values for the parameters N_C, x, y and r, the ancestral population size, N_A, can be calculated by, $N_A = N_C e^{-r(y-x)}$.

Given a known genealogy (including branching order and node heights), maximum-likelihood parameter estimates can be readily obtained for a parametric demographic model such as the one described above, by applying coalescent theory (HUDSON 1990; KINGMAN 1982a). However in most cases, exact knowledge of the genealogical relationships of sampled sequences (for example, the ages of ancestral nodes) is not available. One solution is to integrate over all possible genealogies that are supported by the data, and find the average estimates of the demographic parameters. One way to do this is using sampled-based Bayesian inference. Chapter 5 described an MCMC method for the joint estimation of genealogy, demography (population history) and mutation rate. This inferential framework can be readily extended to the problem at hand. MCMC can be used to weight the contribution of an individual genealogy to parameter estimates in proportion to its posterior probability. For a given mutation rate, μ, this is achieved by initially sampling the entire parameter space:

$$h_{\mathrm{plog}}(N_C, x, y, r, g \mid D) \propto \Pr\{D \mid \mu, g\}\, f_G(g \mid N_C, x, y, r)\, f_{Nxyr}(N_C, x, y, r) \qquad (6.5)$$

where $\Pr\{D \mid \mu, g\}$ is the likelihood, $f_G(g \mid N_C, x, y, r)$ is the coalescent density of the *piecewise logistic* model, and $f_{Nxyr}(N_C, x, y, r)$ is the prior density of the parameters. However, we are only interested in the marginal density, $h_{Nxyr}(N_C, x, y, r \mid D)$. We summarise this density by using samples $(N_C, x, y, r) \sim h_{Nxyrg}$ sampled from the full density. The genealogical information is discarded and can be thought of as uninteresting "missing data".

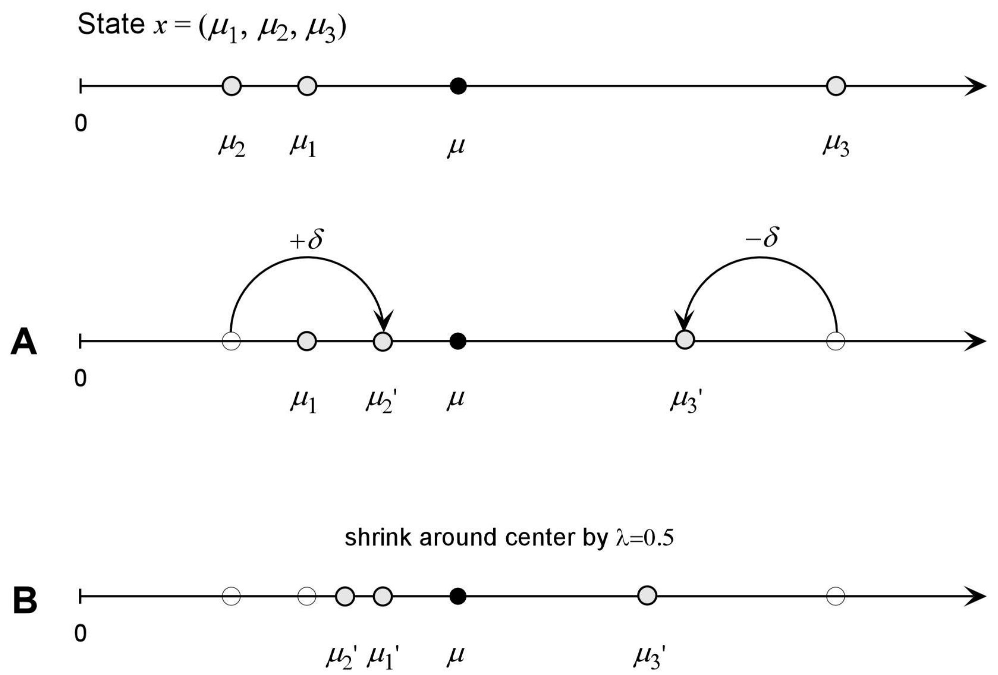
As discussed in early chapters, it is not possible to separate the contributions of substitution rate and population size to sequence diversity with contemporaneous sequences. Therefore, to calibrate the parameters of the demographic model in terms of calendar units, it was necessary to fix the overall mutation rate (μ) to the 'known' value of 5×10^{-3} substitutions per site per year.

6.4.2 Codon position rate heterogeneity

Although the mean substitution rate was fixed, the relative contributions of codon positions 1, 2 and 3 to the overall rate can be estimated using MCMC integration. To achieve this, two proposal mechanisms were employed (see Figure 6.3).

Figure 6.3 Two proposal mechanisms for estimating the relative contributions of codon positions 1, 2, and 3 to the overall rate of substitution.

(A) A random walk that preserves the mean rate of a pair of positions. This proposal mechanism can change the order of the rates as depicted in the figure. (B) A centred scaling proposal mechanism, that shrinks or expands the rates around the mean by a scale factor λ. This proposal mechanism preserves the order of the rates.



6.4.2.1 A random walk that preserves the mean rate of a pair of positions

First consider a set of codon position rates $x = (\mu_1, \mu_2, \mu_3)$ with a mean rate of $\mu = (\mu_1 + \mu_2 + \mu_3)/3$, which must be conserved. Consider a proposal mechanism in which a random pair of rates, say, μ_1 and μ_3, is selected and a small random number $\delta \sim Unif(-\omega, \omega)$ is added to one rate and subtracted from the other, so that the new state is $x' = (\mu_1 + \delta, \mu_2, \mu_3 - \delta)$. This proposal mechanism is symmetric, so that the Hastings ratio is 1.0, and maintains the mean rate of the selected pair of rates and thus the triple. It should be noted that the proposed state x' is immediately rejected if any of the mutation rates becomes negative. This proposal mechanism by itself is sufficient to sample the space of codon position contributions to overall substitution rate, however a second proposal mechanism was used to improve convergence rates in the MCMC chain.

6.4.2.2 A centred scaling proposal mechanism

Again, consider a set of codon position rates $x = (\mu_1, \mu_2, \mu_3)$ with a mean rate of μ. Chose a random scale factor $\lambda \sim \text{Unif}(0.5,2)$ and propose a candidate set of codon position rates, $x' = (\mu_1 + \lambda(\mu_1 - \mu), \mu_2 + \lambda(\mu_2 - \mu), \mu_3 + \lambda(\mu_3 - \mu))$. This proposal mechanism has the effect of expanding or contracting the three rates around their mean. By itself, this proposal mechanism cannot change the order of the rates and thus is not sufficient to sample the full parameter space of interest, however it is useful for sampling the variance of the rates.

6.4.2.3 The full posterior probability density

The full posterior probability density under consideration is:

$$h_{HCV}(N_C, x, y, r, g, \mu_1, \mu_2, \mu_3, \kappa \mid D, \mu) =$$

$$\frac{1}{Z}\Pr\{D \mid \mu_1, \mu_2, \mu_3, \kappa, g\} \times \qquad (6.6)$$

$$f_G(g \mid N_C, x, y, r) f_{Nxyr}(N_C, x, y, r) f_\mu(\mu_1, \mu_2, \mu_3 \mid \mu)$$

Where $f_\mu(\mu_1, \mu_2, \mu_3 \mid \mu) = \begin{cases} 1 & \text{iff } (\mu_1 + \mu_2 + \mu_3)/3 = \mu \\ 0 & \text{otherwise} \end{cases}$

The MCMC algorithm described in Chapter 5 was extended to allow sampling of the posterior probability density described in equation 6.3 above. The input data was a sample of modern contemporaneous Hepatitis C nucleotide sequences.

6.4.3 Data

Two datasets of partial E1 gene sequences, obtained from a comprehensive recent study (RAY *et al.* 2000), were analysed to estimate the demographic history of HCV in Egypt over the last century. Theses sequences were all sampled from the same year (1993), and they show no obvious correlation between geographic distance and genetic distance. In addition the sample is geographically diverse, with ample phylogenetic information. An independent estimate of nucleotide substitution rate for the partial E1 gene region is already available, of, 5 x 10^{-3} substitutions/site/year (PYBUS *et al.* 2001).

Dataset A consists of 68 E1 partial sequences of length 411 bp (63 type 4 sequences and 5 subtype 1g sequences). This dataset contains all E1 sequences published in the recent survey by Ray *et al* (2000) except for three isolates belonging to types 1a and 1b that were excluded because they are probably not representative of the endemic population of HCV. Dataset B contains only the 63 type 4 sequences of Dataset A.

6.4.3.1 Priors

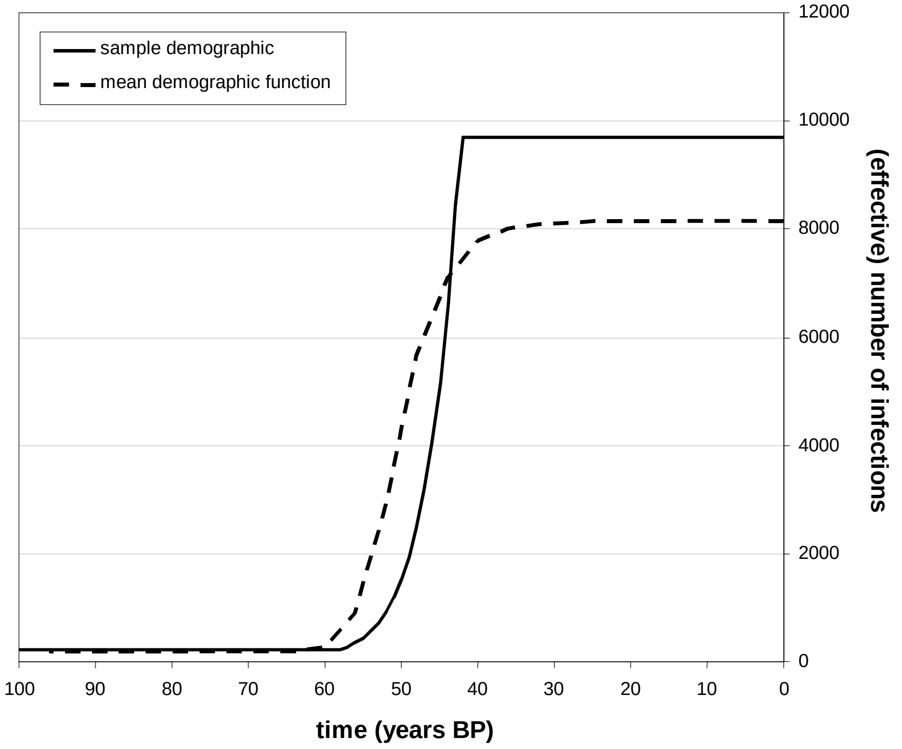
An upper limit on exponential growth rate of 0.75 was chosen. This corresponds to a doubling of the HCV population size every year and is far higher than previous estimates of HCV growth rates (PYBUS *et al.* 2001). This limit was impinged on, suggesting that extremely high growth rates are compatible with the data analysed. However the HPD was focused around substantially lower growth rates of about 0.1-0.2. An upper limit on x and y of three hundred years BP was chosen. These limits were never reached in sampling the posterior.

6.4.4 Results

Table 6.6 shows the estimated marginal posterior densities of some parameters of interest. These estimates suggest a pronounced growth in effective infection size of the Egyptian HCV epidemic of about two orders of magnitude over a period of about 20 years, from mid 1930s to the mid 1950s. Figure 6.4 depicts a sample *piecewise logistic* demographic function from the MCMC analysis of Dataset A, along with a mean demographic function obtained by averaging the demographic height at each time across all of the demographic parameter values sampled by MCMC.

Figure 6.4 Demography of the Egypt HCV epidemic.

Both a sample demographic function from the MCMC chain of an analysis of Egypt HCV sequences and the mean demographic function resulting from averaging all demographic functions in the MCMC chain for Dataset B. The sequence data analysed was thus comprised of E1 gene sequences from HCV type 4.



Finally, we found that allowing for codon position rate heterogeneity had a significant effect on the estimated time of the most recent common ancestor (t_{MRCA}), estimating an older age of the root than a model assuming uniform rates across sites (data not shown).

Table 6.6 Parameter estimates for two datasets of partial E1 genes.

These estimates suggest a pronounced growth in effective infection size of about two orders of magnitude over a period of about 20 years from the mid 1930s to the mid 1950s.

Parameter	Dataset A	Dataset B
r (exponential growth rate)	0.264 (0.075, 0.620)	0.237 (0.072, 0.564)
y (exponential growth start date)	1934 (1924, 1943)	1932 (1922, 1940)
x (exponential growth end date)	1953 (1941, 1966)	1953 (1941, 1966)
N_C (current number of infections)	10310 (4095, 18960)	8779 (3323, 15780)
N_A (ancestral number of infections)	245 (153, 345)	170 (99.6, 251)
codon position 1	0.45e-3 (0.40e-3, 0.49e-3)	0.45e-3 (0.39e-3, 0.49e-3)
codon position 2	0.23e-3 (0.1.98e-3, 0.27e-3)	0.25e-3 (0.21e-3, 0.28e-3)
codon position 3	1.69e-3 (1.63e-3, 1.75e-3)	1.68e-3 (1.62e-3, 1.74e-3)
transition/transversion rate ratio	7.71 (6.75, 8.69)	8.26 (7.13, 9.38)
date of most recent common ancestor	1374 (1258, 1481)	1710 (1673, 1747)

6.4.5 Conclusion

A *piecewise logistic model* of epidemiology was investigated as a tool for analysis of 68 partial E1 gene sequences from HCV strains isolated in a comprehensive survey of Hepatitis C in Egypt. Parameter estimates of the model were obtained by MCMC integration to take into account uncertainty in the genealogical relationships of the sampled sequenced. A strong signal for a rapid growth of HCV in Egypt in the middle of the last century was detected. The data could not eliminate the possibility that HCV prevalence experienced an extremely rapid increase, but could reject the hypothesis that there has been no change in HCV prevalence over the last century. Finally a strong signal for rate heterogeneity among sites by codon position category was observed in the data, and the incorporation of this knowledge into the inference has a marked effect on estimates of the age of the most recent common ancestor of HCV in Egypt.

6.5 Discussion

Section 6.2 provided further support for the recent observation of higher evolutionary rates in mtDNA control region sequences than originally thought (LAMBERT *et al.* 2002). This analysis also showed that a relatively small number of short sequences (~200 bp) could still give rise to accurate estimates of substitution rates in mtDNA if ancient material is distributed over a large temporal scale (in this case, at least 60,000 years). This should be compared to other recent studies in which a large number of longer sequences (~350 bp) over a short temporal scale (LAMBERT *et al.* 2002) was used to yield similar estimates of the substitution rate of the mitochondrial control region. The question was asked, "Is the rate of evolution the same over different time scales?" It seems apparent that extreme rate heterogeneity (the fastest 25% of the sites appeared to be accumulating substitutions 2000 times the slowest 25% of the sites) could provide an explanation for an *apparent* discrepancy of overall substitution rate over different time scales. This provides an obvious *prediction*: data from longer timescales (for example across all *Carnivora*) should have both *apparently* reduced overall substitutions rates and *apparently* reduced rate heterogeneity because of complete saturation of the fastest sites.

Another possible explanation for the variation of rates over different time scales is that the effects of purifying selection will be more apparent at larger time scales. This would seem more likely to occur in the case of protein-coding sequences, in which a small amount of genetic drift over short time frames may be tolerated, however over long time frames, proportionally larger shifts in sequence may be more strongly resisted by purifying selection. This argument is compelling, however little is known of the particular functional and structural constraints on the DNA sequence in the mitochondrial control loop.

Section 6.3 investigated the possibility of dating sequences of unknown ages by comparing them to sequences with known ages in an MCMC framework. This analysis resulted in marginal posterior distributions for the ages of nine undated Beringian bear bones. The marginal densities were diffuse and highly sensitive to choices of prior distributions, suggesting that only a very weak signal existed for molecular dating with the sequence lengths available. Despite this, qualitative statements about the relative ages of bones were fairly robust to changes in prior information.

The two case studies described in this chapter demonstrated a number of simple yet powerful extensions that can be made to the basic framework described in Chapter 5.

The Bayesian inference of gamma-distributed rate heterogeneity among sites was demonstrated using the Beringian bear mtDNA dataset. Although the estimated rate heterogeneity was very high, it had only a modest effect on the estimated substitution rate (~25% increase) and divergence times over the time scale considered. An alternative approach to rate heterogeneity among sites was investigated in the analysis of a set of HCV partial E1 gene sequences, where a separate substitution rate was estimated for each codon position. Again, substantial rate heterogeneity was observed and the wobble position was estimated to be evolving 7 times faster than the 2nd codon position. However, unlike the Beringian bear data, the assumption of rate heterogeneity in the HCV dataset was associated with a larger estimate (~2 times larger) of the age of the root than similar analyses that assumed a single rate across sites (data not shown). It is also worth noting that these two models of rate heterogeneity could easily be used in concert, so that a gamma shape parameter is estimated for each codon position.

The evolutionary models investigated in this chapter are just a few of the modifications that can be made to the Bayesian framework described in Chapter 5. Programmatically, they were made easier by the existence of open-source programming libraries such as PAL (DRUMMOND and STRIMMER 2001). Nonetheless, the ease with which they were implemented was largely due to the nature of the MCMC algorithm. If a model can be simulated then it can be implemented in MCMC.

In some instances (for example, the *piecewise logistic growth* model of demography) it remains to be seen under what prior conditions the models investigated here represent meaningful posterior distributions. Despite this, the results of this chapter again demonstrate that, as an exploratory tool with statistical rigor, MCMC is a powerful alternative to more classical approaches such as maximum likelihood and least-squares regression.

7 Evolution of RNA secondary structure

7.1 Introduction

Evolutionary inference is often conducted on gene sequences that code for functional molecules such as proteins or functional RNA molecules. The structure of these molecules is often important to their function. In chapter 8 I investigate the ability of molecular structure information to improve on two problems in phylogenetics: multiple sequence alignment and phylogenetic reconstruction.

Understanding evolutionary processes requires an understanding of the ways in which mutation affects phenotypic fitness. In a molecular context this question can be rephrased: How do mutations in a gene affect the fitness of the gene's product? To answer this question, knowledge of the relation between the primary sequence of the gene (genotype) and the active conformation of the product (phenotype) is required. This relation is called the sequence-structure (or genotype-phenotype) mapping. A sequence-structure mapping is a function that maps a (DNA) sequence into a molecular structure. Obtaining a sequence-structure mapping requires a solution to the structure prediction problem. Molecular evolution results from the combined action of mutational processes acting on the *genotype* and 'selective' processes (purifying and adaptive) acting on the *phenotype*, and thus its full understanding requires knowledge of the sequence-structure mapping.

This chapter focuses on RNA secondary structure of the small-subunit ribosomal RNA molecule as a phenotype. This molecule is found in all living organisms and plays a central role in biology, including viral replication, as part of the translation machinery of the ribosome. As a result its structure is highly conserved across all three domains of life. Furthermore, RNA secondary structure is a computationally tractable intermediate phenotype that can be determined directly from its genotype. RNA structure prediction from sequence data is possible by thermodynamic, kinetic and phylogenetic techniques.

The aim of this research is to compare the predicted structures of evolutionarily related molecules, and in doing so ascertain if structural information can improve our understanding of molecular evolution. The goal is to use structural information to accurately align sequences and infer phylogenetic relationships.

7.2 Secondary structure prediction

As mentioned above, obtaining a sequence-structure mapping requires a solution to the structure prediction problem. Structure prediction is based on the hypothesis that the 3D biologically active structure of a gene's product (RNA or protein) can be predicted from the corresponding nucleotide or amino acid sequence. Approximate structure prediction techniques have been developed for both protein and RNA molecules.

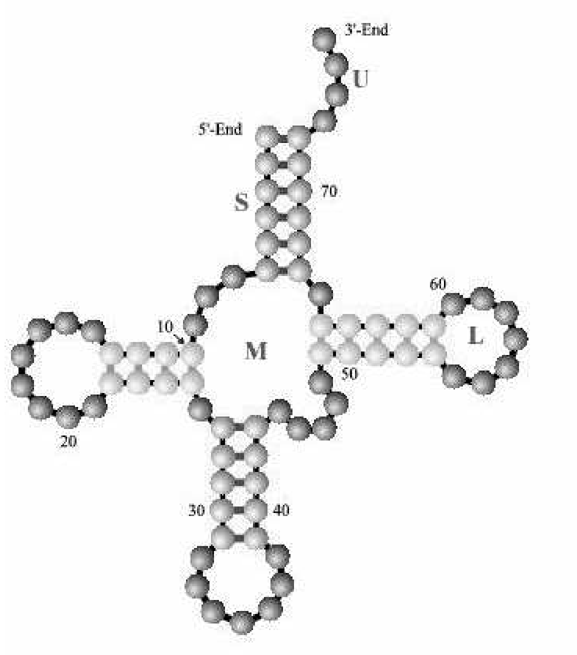
Obtaining any mapping from genotype to phenotype requires environmental factors to be fixed. At the level of the organism this constraint can be very difficult to justify. Phenotypes are the product of the interaction between the developing organism and its environment, both of which vary over evolutionary time frames. Likewise, in a molecular context, the phenotype is the product of the genotype and its kinetic development in the cellular environment. However, structure prediction techniques must fix the cellular environment as a constant. The implications of this assumption will not be considered here, except to say that phylogenies containing organisms living in very different environments should be treated with care.

The biologically active conformation of an RNA or protein molecule can be described by its three dimensional structure (i.e. the average spatial positioning of its constituent atoms). However, accurately deriving this information from a primary sequence is currently not feasible. For RNA molecules a convenient intermediate description of molecular structure exists, called the secondary structure. The secondary structure can be described by a list of Watson-Crick (**AU** and **GC**) and **GU** paired nucleotide positions (base-pairs) in the RNA sequence (REIDYS *et al.* 1997). These base pairs bring complementary sub-sections of the molecule together to form helices (or stems or stacks). A number of coarse structural elements can be defined for secondary structures; i.e. stack, loop, bulge, unpaired regions and multi-stem loop (see Figure 7.1).

In a mathematical sense, secondary structures are contact graphs with an associated adjacency matrix (REIDYS *et al.* 1997). This simple definition of secondary structures allows statistical analysis by conventional combinatorics in a way difficult or impossible to achieve for the full three dimensional structure of RNA. For instance, straightforward estimates of the number of possible secondary structures of a given chain length are readily obtainable. Three approaches to RNA secondary structure prediction - thermodynamic, kinetic and phylogenetic techniques – are discussed below.

Figure 7.1 A transfer RNA (tRNA) secondary structure.

This diagram of a transfer RNA (tRNA) illustrates most of the major features of RNA secondary structure: S = stem/helix, M = multi-stem loop, L = loop, U = unpaired region. (Reproduced and modified from Fontana, 1998).



7.2.1 Thermodynamic structure prediction

Thermodynamic prediction techniques are based on the assumption that an RNA molecule is in its thermodynamic ground state in the cell. This ground state can be calculated by minimising the free energy of the molecule. Efficient algorithms to calculate the minimum free energy secondary structure have been developed (WUCHTY *et al.* 1999; ZUKER and STIEGLER 1981). In these algorithms the free energy of a secondary structure is approximated by the sum of the free energies of simple sub-structures such as hairpin loops, bulges, multi-stem loops and unpaired regions. The free energies of these sub-structures are experimentally determined.

The thermodynamic stability (and ground state) of an RNA molecule is affected by variables such as the temperature, salt content and pH of the environment (SERRA *et al.* 1997). These variables are fixed during experiments that measure the stability of small RNA fragments (SERRA *et al.* 1997; SINGH and KOLLMAN 1996; XIA *et al.* 1997). From these analyses the thermodynamic stability at different pH, temperature and salinity can be extrapolated. Thus thermodynamic RNA secondary structure prediction algorithms require not only the base sequence, but also an approximation of the cellular environment of the molecule being studied.

Thermodynamic secondary structure prediction is valid because the base pairing and base pair stacking interactions in an RNA molecule utilise the majority of the free energy of the molecule, and the individual contributions are reasonably independent (WALTER *et al.* 1994). Because of this, secondary structure prediction is also a valid intermediate step for determining the tertiary interactions and complete 3D structure of the molecule (REIDYS *et al.* 1997).

It can be argued that for large molecules the minimum free energy conformation is not as important as some sub-optimal conformations (WUCHTY *et al.* 1999; ZUKER 1989; ZUKER *et al.* 1991). Many close-to-optimal conformations may have significant biological functions. Furthermore, as we will see later, the minimum free energy structure might not be attained *in vivo*, for reasons to do with kinetics. Two improvements on the original algorithm have been described and implemented to address this inadequacy.

McCaskill developed a description of RNA secondary structure based on statistical mechanical theory (MCCASKILL 1990). By considering the complete ensemble of all permissible secondary structure conformations for a given sequence and their associated free energies, McCaskill was able to describe an RNA molecule by its base-pairing probability matrix (BPPM). Each entry (p_{ij}) in the matrix is the probability of the base-pair (i, j) occurring in the molecule at equilibrium. The value of p_{ij} is determined by the proportion of permissible conformations that contain base-pair (i, j) weighted by the thermodynamic stability (Boltzmann factor) of those conformations.

A slightly different technique has been employed to find a subset of the conformations represented by the statistical mechanical ensemble (WUCHTY *et al.* 1999; ZUKER 1989). It predicts all structures with free energies within a threshold of the minimum free energy (δ). Thus taking the minimum free energy of an RNA molecule to be ε, these algorithms determine all conformations with free energies within the range $\varepsilon \leq x \leq \varepsilon + \delta$. This technique is particularly useful in two important ways. It only generates the relatively stable conformations likely to be biologically important and it is easier to interpret the molecule as a group of stable conformations, than as a matrix of contingent base-pairing probabilities.

However all of these thermodynamic methods suffer from some shared inadequacies. They are unable to successfully predict the secondary structure of molecules that are modified by interaction with other molecules, have a significant proportion of non-canonical base pairs, or include pseudoknots or more complex structures (FIELDS and GUTELL 1996; HUYNEN *et al.* 1997; KONINGS and GUTELL 1995). Despite these weaknesses, the general agreement of thermodynamic prediction with other techniques suggests that thermodynamic assumptions are reasonably robust (FIELDS and GUTELL 1996; HUYNEN *et al.* 1997; KONINGS and GUTELL 1995; ZUKER and JACOBSON 1995; ZUKER *et al.* 1991).

7.2.2 Kinetic structure prediction

High-resolution kinetic experiments of RNA molecules have shown that RNA folding is an ordered process (BATEY and DOUDNA 1998). Kinetic secondary structure prediction is based on the fact that secondary structure is the outcome of a controlled process of folding. The problem of folding can be visualised as the relaxation of the molecule on a free energy landscape (GULTYAEV *et al.* 1995; VAN BATENBURG *et al.* 1995). The landscape is one in which the height of the terrain represents the free energy of the conformation at that point, and neighbouring points in the terrain are similar structural conformations. The process of folding a complete molecule can then be understood as a downhill path from the initial conformation to a nearby basin, where the basin represents a stable conformation with low free energy. The deepest basin in the landscape thus corresponds to the minimum free energy structure determined by a pure thermodynamic algorithm.

Kinetic prediction methods (for both proteins and RNA) attempt to simulate a walk in a free energy (or folding) landscape (FLAMM *et al.* 2000; GULTYAEV *et al.* 1995; VAN BATENBURG *et al.* 1995). The folding (energy relaxation) process will not always lead to the minimum free energy structure (FLAMM *et al.* 2000; GULTYAEV *et al.* 1995; VAN BATENBURG *et al.* 1995). However, it will lead to a meta-stable structure. RNA folding simulations generally incorporate two different kinetic properties of RNA maturation. The first is the folding that occurs during transcription of the RNA from the DNA template. The second is the creation and destruction of helices (stacks) in an orderly way during maturation, leading to more and more stable conformations. Algorithms that employ both of these techniques have been developed (FLAMM *et al.* 2000; GULTYAEV *et al.* 1995; VAN BATENBURG *et al.* 1995).

Kinetic methods could conceivably be used to select the most likely candidates for biologically important structures from the large class of relatively stable sub-optimal structures that exists for most sequences. They also have the advantage that they can predict important tertiary interactions such as those involved in pseudo-knots[7] (GULTYAEV *et al.* 1995; VAN BATENBURG *et al.* 1995). Unfortunately, kinetic prediction methods are very computationally expensive and do not yet appear to give significantly better predictions than their more mature purely thermodynamic counterparts.

7.2.3 Phylogenetic (or comparative) structure prediction

Both of the previous classes of secondary structure prediction methods have as a central principle the concept of thermodynamic molecular equilibrium. Phylogenetic prediction methods are unique in being completely independent of thermodynamic considerations.

Whereas a thermodynamic approach predicts an entire structure (or ensemble of probable conformations) from a single sequence, phylogenetic methods use many related sequences to arrive at a secondary structure. The basic procedure involves the detection of compensating substitutions in sequence alignments of closely related organisms. This technique has been used extensively in the development of secondary structure of rRNA (SCHNARE *et al.* 1996). Phylogenetic secondary structure prediction relies on two main assumptions. Firstly, if two points in a sequence co-vary in a sequence alignment they are probably closely coupled in the folded molecule. Secondly, the secondary structure of the molecule is conserved over evolutionary time-scales (SCHNARE *et al.* 1996).

Notwithstanding these assumptions, direct comparison of thermodynamic and phylogenetic prediction methods can be misleading for a number of other reasons. For example, phylogenetic methods necessarily predict only those base-pairs that are evolutionarily conserved. Base-pair modifications that do not affect the fitness of the molecule will not necessarily be detected by compiling secondary structure databases. Furthermore, phylogenetically derived secondary structures may represent a mosaic of different conformations that are all evolutionarily important. This is of interest because more than one of the sub-optimal structures of the molecule may be biologically significant (ROSENBAUM *et al.* 1993).

[7] Pseudoknots are a condition of RNA base pairing in which base-pairs are interleaved (Stadler & Haslinger, 1997). For example base pairs (i, j) and (k, l) form a pseudoknot if $i<k<j<l$ or $k<i<l<j$. These structures are not permitted in most thermodynamic structure prediction algorithms because they require much more complex algorithms. Non-nested pseudoknots require an extension of the mathematical formalism of secondary structures to a class of entities known as planar graphs (Stadler & Haslinger, 1997).

Having said this, phylogenetic techniques also have a number of distinct advantages over thermodynamic techniques. They allow the detection of non-canonical base-pairs. In some organisms the proportion of non-canonical base pairs is quite significant and correlates negatively with accuracy of thermodynamic-based prediction algorithms (FIELDS and GUTELL 1996). Comparative sequence analysis can also be used to identify tertiary interactions such as base-triples (GAUTHERET *et al.* 1995) and pseudoknots (SCHNARE *et al.* 1996; WILLS 1992).

The existence of independent methods for predicting secondary structure from sequence (thermodynamic, kinetic and phylogenetic) is very useful in objectively examining each method's strengths and weaknesses (FIELDS and GUTELL 1996; HUYNEN *et al.* 1997; KONINGS and GUTELL 1995; ZUKER and JACOBSON 1995; ZUKER *et al.* 1991).

7.2.4 Hybrid techniques

Recently the possibility of estimating the stabilities of small RNA sub-structures from populations of phylogenetically predicted secondary structures has been investigated (MATHEWS *et al.* 1999). For example the pseudo-energy[8] of a particular hairpin structure can be calculated by its prevalence in a large pool of secondary structures. If it occurs often in predicted structures then it is probably a stable sub-structure. This technique might provide a means to rapidly improve the thermodynamic parameters that are difficult to obtain empirically. However, it also creates the danger of reducing the independence of the two techniques.

Another technique of interest uses interactive constraint satisfaction to manipulate an energy minimisation procedure (GASPIN and WESTHOF 1995). This technique is unique in that it allows the user to interactively define constraints on the RNA folding both before and during the folding process, thus allowing the investigation of different assumptions about constraints. Using this technique, tertiary interactions within the molecule and interactions with other molecules can be investigated by imposing constraints that simulate these interactions.

7.3 Properties of RNA sequence-structure maps

A number of researchers have investigated the general properties of RNA sequence-structure maps. In particular the properties of the mapping defined by thermodynamic prediction have been rigorously investigated (BASKARAN *et al.* 1996; FONTANA *et al.* 1993;

[8] These statistically derived 'energies' are called pseudo-energies to differentiate them from experimentally derived energies.

SCHUSTER *et al.* 1994). A number of these properties have important implications for the evolutionary dynamics of RNA molecules (FONTANA and SCHUSTER 1998; HUYNEN 1996; HUYNEN and HOGEWEG 1994; HUYNEN *et al.* 1996). More general studies on the properties of redundant sequence-structure mapping functions have also been undertaken (REIDYS *et al.* 1997).

RNA folding is a many-to-one mapping. For a given length molecule, there are many more possible sequences than secondary structures. Thus many sequences fold into the same structure. Furthermore, structures can be partitioned into common and rare categories based on the number of sequences that fold into them. The vast majority of sequences fold into the small minority of structures that are common (refer to Figure 7.2) These properties of the RNA sequence-structure mapping have been investigated by exhaustive search techniques and inverse folding[9] (SCHUSTER *et al.* 1997).

Through a mathematical model (REIDYS *et al.* 1997) and neutral path simulations it has been demonstrated that neutral networks exist for common structures in both RNA (SCHUSTER and STADLER 1998) and proteins (BABAJIDE *et al.* 1997).

A neutral network is a network of neighbouring sequences in sequence space[10] that all fold into the same secondary structure. The neutral network of a common structure extends across large distances in sequence space. For example, it has been predicted that by a series of neutral mutations a tRNA molecule can be changed at every nucleotide position but at all times maintain its secondary structure (HUYNEN *et al.* 1996). Figure 7.2 shows a simplified sequence space and neutral network in which each sequence has only 6 neighbours.

The existence of large neutral networks in sequence space has led to an exciting new view of the evolutionary dynamics of RNA molecules. Simulations show that these neutral networks are exploited by evolving populations of molecules to 'search' large areas of sequence space (HUYNEN 1996; HUYNEN *et al.* 1996). In these simulations, RNA evolution is manifested as a series of extended periods of neutral drift punctuated by

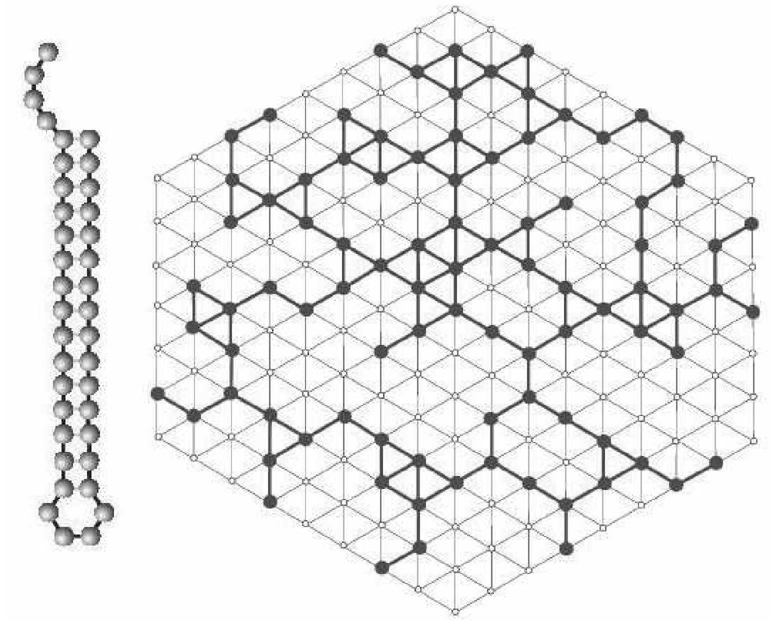
[9] Inverse folding is the reverse of structure prediction. From a structure, the possible sequences that produced the sequence are predicted.

[10] A sequence space is an abstract space that represents all sequences of a certain length (N). Thus an N=100 RNA sequence space contains all possible 100 nucleotide RNA sequences. Each point in a sequence space represents a unique sequence, and is neighboured by all of its one base-pair mutants. Thus for an N=100 sequence space, each sequence has 300 one-mutant neighbours. A move in sequence space from one point to a neighbouring point thus represents a base-pair mutation in a sequence. For long sequences there are obviously many more directions to move in sequence space than can be easily represented in two dimensions.

rapid adaptive improvements. The periods of neutral drift are characterised by diffusion of the population across the dominant neutral network[11]. In each generation the sequence population produces new mutants, and these mutants are either members of the dominant neutral network or represent new shapes. Mutants that represent new shapes constitute a 'search' for fitter candidates in shape space[12] to replace the dominant structure. The adaptive improvements observed in simulations thus represent a Darwinian relocation of the sequence population onto a neutral network of greater fitness (HUYNEN *et al.* 1996). These simulations show that neutral evolution potentially plays an important role in adaptation of RNA molecules (FONTANA and SCHUSTER 1998; HUYNEN 1996; HUYNEN *et al.* 1996).

Figure 7.2 The concept of common secondary structures.

This is indicated by the (simplified) sequence space and neutral network alongside the structure. Common secondary structures are characterised by extensive connected networks in sequence space. All of the nodes connected by heavy lines represent sequences that fold into the structure illustrated. (Reproduced and modified from Fontana, 1998).



7.3.1 Structural Discontinuity and Punctuated Equilibrium

Two characteristics of RNA evolution simulations are of great interest in terms of evolutionary theory. The first is the important role that neutral drift appears to have. This observation appears to be more compatible with the neutral theory of evolution

[11] The dominant neutral network represents the structure that is currently the fittest in the population.

[12] Shape is used synonymously with secondary structure. Shape space is used here to mean an intuitive structure-based parallel to sequence space. A more formal definition has been developed (Fontana & Schuster; 1998), and will be discussed in later sections.

(KIMURA 1968; KIMURA 1983) than with Darwinian orthodoxy. The second characteristic is the abruptness of the (infrequent) adaptive improvements. At the organismic level the debate over punctuated equilibrium has continued since first hypothesised by Eldredge & Gould (1972). The debate has often focused on the question of what constitutes a punctuated change (ELDREDGE and GOULD 1997; GOULD and ELDREDGE 1993). A recent addition to this debate was developed from observations of evolutionary punctuations in RNA evolution simulations (FONTANA and SCHUSTER 1998).

Fontana and Schuster focussed on providing a formal definition of secondary structure neighbourhood (or nearness). With a neighbourhood function for RNA secondary structures, a precise definition of shape space (a structural analogue to sequence space) was described (FONTANA and SCHUSTER 1998). This definition of shape space allowed the classification of continuous and discontinuous[13] shape transitions. Continuous transitions occur between *neighbours* in shape space whereas discontinuous transitions occur between *non-neighbours*.

Fontana and Schuster defined the nearness of two structures in terms of the extent to which their corresponding neutral networks come into close proximity in sequence space[14] (FONTANA and SCHUSTER 1998). For a given structure (A), the closest neighbouring structure is that structure that most often results from non-neutral mutants of sequences folding into A.

Fontana and Schuster observed that although continuous transitions are much more common (by definition), it was the discontinuous transitions that most often coincided with adaptive improvements in their simulations. Therefore they demonstrated that their classification of structural transitions coincides with their observations of sudden fitness improvements. When a fitness improvement occurs, it most often results from a rare discontinuous change.

7.4 Prokaryotic systematics

Tens of thousands of ribosomal RNA small sub-unit (rRNA SSU) gene loci have been sequenced from microbial organisms. A large proportion of these sequences exist

[13] In this proposal structural *continuity* and dis*continuity* are meant in the sense defined by Fontana and Schuster (1998).

[14] The structural nearness relationship can be asymmetric because neutral networks vary greatly in size. So the accessibility of the transition A → B may not equal the accessibility of the transition B → A (A and B are neutral networks).

without a corresponding identification of the organism from which they came. The advent of PCR techniques has allowed microbiologists to obtain large numbers of DNA fragments from the environment without isolating the source organisms. This has enabled the classification of species and groups of bacteria that have not been individually isolated or cultured.

With this growing class of microorganisms that cannot be easily classified by standard numerical taxonomy, the importance of molecular phylogenetics has increased. The most important molecule used to infer microbial phylogenies is rRNA SSU. However, a number of properties of RNA molecules make the analysis of these genes problematic given current molecular systematic techniques. For example, many alignments of RNA are edited manually to correct for improper placement of insertions/deletions (indels) by standard alignment algorithms. When an indel occurs in a region of the molecule that is paired in the secondary structure, a compensating indel also generally occurs on the paired portion. However, gap creation penalties in standard alignment algorithms do not take this into account (O'BRIEN *et al.* 1998).

Sequence alignments of rRNA genes can be partitioned into regions of high conservation and regions of high variability (O'BRIEN *et al.* 1998). The conserved regions have very few informative sites (polymorphisms). In contrast, the variable regions are often highly polymorphic and very difficult to align, due to ambiguous positioning of indels. This makes the determination of homologous nucleotide positions difficult. A comparison of secondary structure similarity instead of sequence similarity may help determine true homology more accurately. In Chapter 8, the use of secondary structure information to improve both sequence alignment procedures and phylogenetic inference is investigated.

7.5 Discussion

This chapter has described the developments in RNA secondary structure prediction and the current theoretical understanding of RNA secondary structure evolution. How does this information impact the fields of molecular evolution and phylogenetic inference? A number of specific questions are apparent:

1. How does conservation of RNA secondary structure impact the patterns of substitution of the RNA sequence?
2. How can RNA secondary structure be used to assist in sequence alignment?
3. How can RNA secondary structure be used to assist in phylogenetic inference?

All of these questions are addressed in the following chapter.

8 RNA-based evolutionary inference

8.1 Introduction

This chapter is presented in three parts. In section 8.2, RNA secondary structure is assumed to be known (or predicted) and conserved across the taxa of interest. Under these assumptions an empirical substitution model for RNA is derived from a large alignment of sequences, and simulation studies are undertaken to investigate the properties of the model. In sections 8.3 and 8.4, RNA secondary structure is assumed known (or predicted), and *mostly* conserved. Under these assumptions, both multiple sequence alignment and Bayesian phylogenetic inference are investigated.

8.2 Structurally-specific RNA substitution models

The question posed in this section is: 'Does the secondary structure of a nucleotide site in a functional RNA molecule impose structure-specific substitution patterns on that site?' This question arises from the notion that if selective processes affect substitution patterns, then they will do so through the function and therefore structure of the RNA molecule. In this simple model, the phenotype is the structure of the RNA molecule and the genotype is the sequence of the gene encoding the molecule. This model is appropriate for structural RNA molecules that do not also encode for a protein product. RNA secondary structure describes the base-pairing of the molecule, and is a topological rather than 3D description of the molecule's shape. One of the useful properties of RNA secondary structure is that the hydrogen bonds involved in base pairing account for the majority of the free energy of the molecule.

The approach we have taken to answer the question posed is to partition an RNA molecule into different structural categories and determine if different categories have different substitution patterns. A number of papers discussing the influence of protein secondary structure on the replacement patterns in protein evolution have been previously published (GOLDMAN *et al.* 1998; JONES *et al.* 1994; THORNE *et al.* 1996). Differential patterns of nucleotide substitution can be interpreted similarly as the result of selective constraints imposed by different secondary structures. Thorne, Goldman and Jones (TGJ) used an empirical approach to estimate the model of evolution (amino acid replacement) in different 'structural environments'. This previous work on amino acid replacement models provides a basis for the techniques used here. We extend this kind of analysis by attempting to understand what selective processes impose these patterns on the gene sequences. This is achieved through computer simulations of thermodynamic and mutation stability of random sequences mutated under different substitution models.

8.2.1 Methods

Following Thorne, Goldman and Jones (1996) an empirical method for determining the substitution process in different rRNA structure categories was used.

For this purpose an alignment of all eubacterial 16S rRNA sequences was extracted from the RDP database (MAIDAK *et al.* 2001). This alignment was then trimmed to contain only Domain 2 of the molecule, and all records in the database that did not contain the complete sequence over this region or that contained ambiguous nucleotides were removed. The remaining sequences were then passed through a final filtering stage, where the thermodynamically favourable structures of each sequence were predicted using MFOLD 3.0 (copyright 1996, Dr M Zuker). If, for a given sequence, no structures were found that matched (in terms of the branching structure of helices in the published *Escherichia coli* secondary structure of domain 2), then that sequence was also removed.

It has been shown that the overall structure of the 16S molecule is highly conserved over the entire eubacterial tree (GUTELL 1994), especially in terms of the presence or absence of helices and hence helix branching patterns. However, as more models have been produced for specific species, it has been found that within the constraints of the overall structure there is flexibility, especially in the length of the helices and placement of small bulges and internal loops. The use of MFOLD 3.0 and the filtering step of the above protocol were designed to account for this. The methodology outlined here departs from TGJ in this respect.

TGJ used 'known' structures of individual proteins within each protein family and imposed that structure on all sequences within the family. They considered all pairs of sequences that shared at least 85% similarity and contained at least one sequence that was the closest sequence in the database to the other. Here the same general criteria for comparisons are used, but with a more stringent 95% similarity threshold. This stringency was made possible by the availability of a large number of sequences and a comparatively small number of parameters requiring empirical estimation.

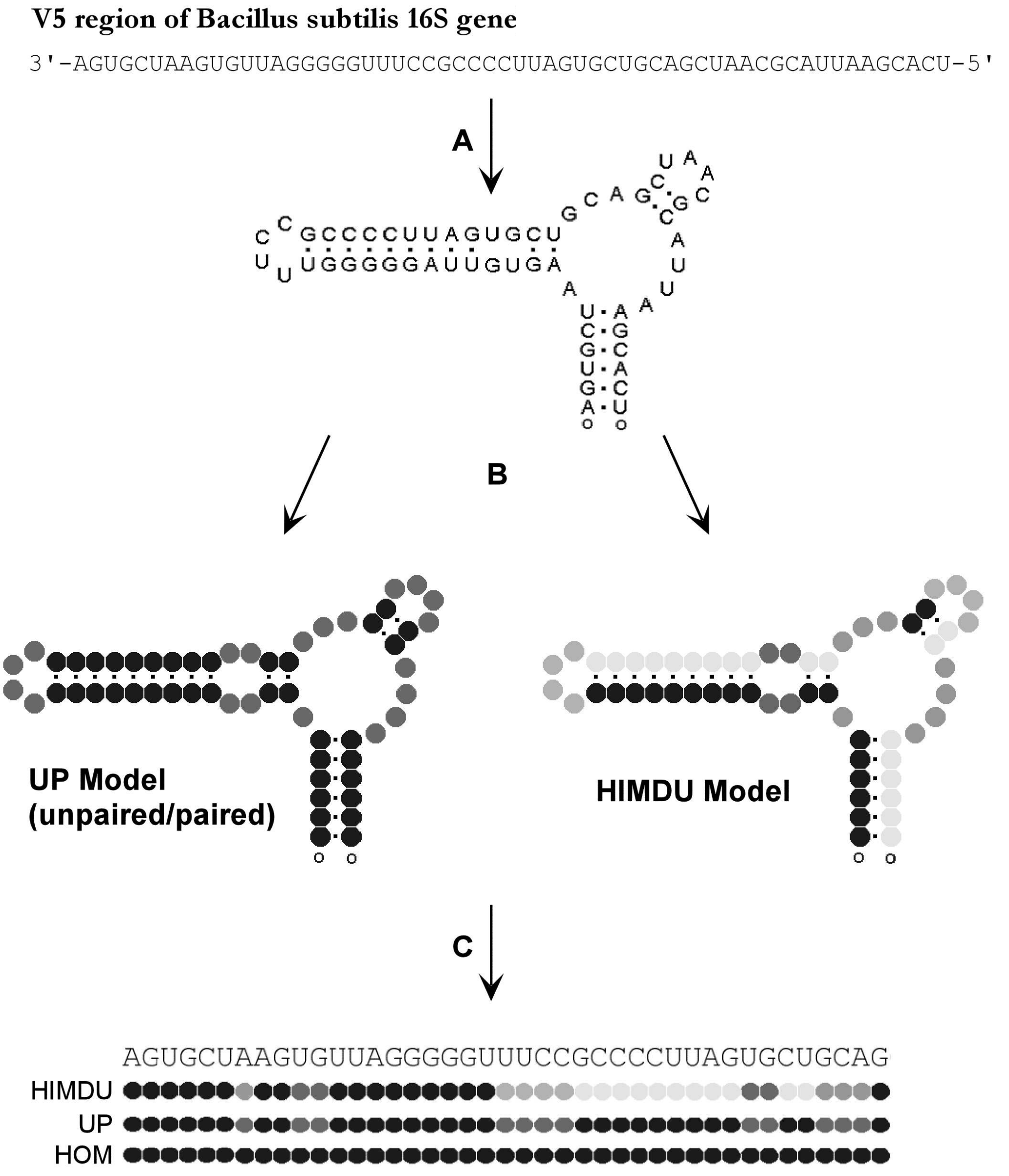
Finally, in each pair of sequences compared, only nucleotide positions that had the same predicted structural category were used (see Figure 8.2).

8.2.2 Models for structural environment assignment

Using the predicted secondary structures produced by MFOLD 3.0, three strategies were employed to assign structural environments to nucleotide positions. The first strategy distinguished two structure categories – (U)npaired and (P)aired (henceforth referred to as the UP model), while the second distinguished a total of 5 structure categories – (H)airpin loop, (I)nternal loop or bulge, (M)ulti-stem loop, (D)ownstream paired and (U)pstream paired (the HIMDU model). A null hypothesis assignment strategy in which all nucleotide positions were assigned the same category was also used. This represents a homogeneous (HOM) model in which secondary structure is ignored. The HOM model is the one usually employed in maximum likelihood phylogenetic reconstruction using this gene. Examples of the two non-trivial assignment strategies are depicted in Figure 8.1.

Figure 8.1 Annotation of rRNA-encoding sequences with structure category information.

(A) Predict secondary structure using thermodynamic criteria (minimum free energy). (B) Determine the structural environments under different models. Two models are shown here: the UP model (unpaired/paired) and the HIMDU model (Hairpin, Internal bulge, Multi-stem loop, Downstream-paired and Upstream-paired) (C) Annotate the sequence with structure categories**.** The third model is a homogeneous model (HOM) in which structural information is ignored.

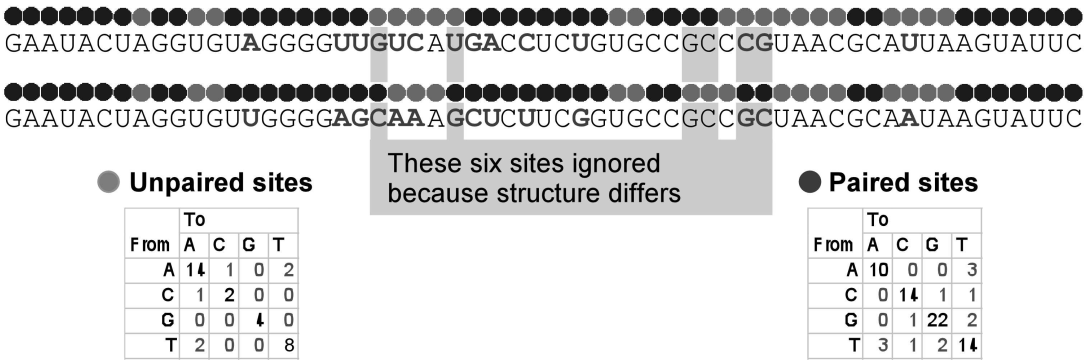

V5 region of Bacillus subtilis 16S gene
3'-AGUGCUAAGUGUUAGGGGGUUUCCGCCCCUUAGUGCUGCAGCUAACGCAUUAAGCACU-5'
A
B
UP Model (unpaired/paired)
HIMDU Model
C
AGUGCUAAGUGUUAGGGGGUUUCCGCCCCUUAGUGCUGCAG
HIMDU
UP
HOM

8.2.3 Empirically-derived substitution models constructed from a large sequence alignment

An alignment of 2487 SSU domain 2 sequences extracted from the RDP database was used to derive substitution parameters under the UP and HIMDU structure category models. An example of substitution statistics collected from a pair of annotated sequence fragments is illustrated in Figure 8.2. Each pair of annotated sequences in which (i) one sequence is the closest match to the other, and (ii) the overall difference is within some threshold, is used to collect substitution statistics. Nucleotide differences are categorized by type (A↔C, A↔G, A↔T, C↔G, C↔T, G↔T) and by structure category (for example, unpaired/paired in the UP model). If the structure categories at a site don't match in the two sequences, then the site is ignored, as these sites do not meet the assumption of structure conservation.

Figure 8.2 The collection of substitution statistics from a pair of sequences.

Each pair of annotated sequences in which (i) one sequence is the closest match to the other, and (ii) the overall difference is within some threshold, is used to collect substitution statistics. Nucleotide differences (in red) are categorized by type (A↔C, A↔G, A↔T, C↔G, C↔T, G↔T) and by structure category (unpaired or paired, in this example). If the structure categories at a site don't match in the two sequences, then the site is ignored, as these sites do not meet the assumption of structure conservation.


GAAUACUAGGUGUAGGGGUUGUCAUGACCUCUGUGCCGCCCGUAACGCAUUAAGUAUUC
GAAUACUAGGUGUUGGGGAGCAAAGCUCUUCGGUGCCGCCGCUAACGCAAUAAGUAUUC
These six sites ignored because structure differs
● Unpaired sites
● Paired sites

	To			
From	A	C	G	T
A	14	1	0	2
C	1	2	0	0
G	0	0	4	0
T	2	0	0	8

	To			
From	A	C	G	T
A	10	0	0	3
C	0	14	1	1
G	0	1	22	2
T	3	1	2	14

8.2.3.1 Base frequency heterogeneity

Both the UP and HIMDU models exhibit distinct partitioning of base composition into different structural categories.

One of the more striking patterns immediately observable in this dataset is the difference in base frequency patterns in different structure categories. Adenosine makes up over 40% of unpaired regions and less than 15% of paired regions. Even within paired regions, the proportions of Guanine and Cytosine are reversed, and as one would expect, this results in a complementarity of base frequencies in paired regions. The question this begs is, why are the upstream-paired regions so much more G-rich (>40%) than the downstream-paired (<30%)? Figure 8.3A and Figure 8.3B show the empirical base frequencies for the UP and HIMDU models respectively.

8.2.3.2 Instantaneous rate matrices

Figure 8.3C and Figure 8.3D shows the empirically derived instantaneous rate matrices for the HOM model and the UP model respectively, calculated from the 2487 sequences analysed. The UP model reveals a strong transition/transversion bias in the paired regions that is not as apparent in the unpaired parts of the molecules. However the unpaired regions seem to have a bias towards C↔T.

Although the rate matrices for the HIMDU model were calculated, there were not enough sites to get accurate estimates because some structure categories had only a few representative sites in the region looked at. For this reason, only the UP model was pursued in the next section on simulations. However, in both models, a bias in the transition/transversion ratio was most clearly seen in the paired regions (data not shown for HIMDU model).

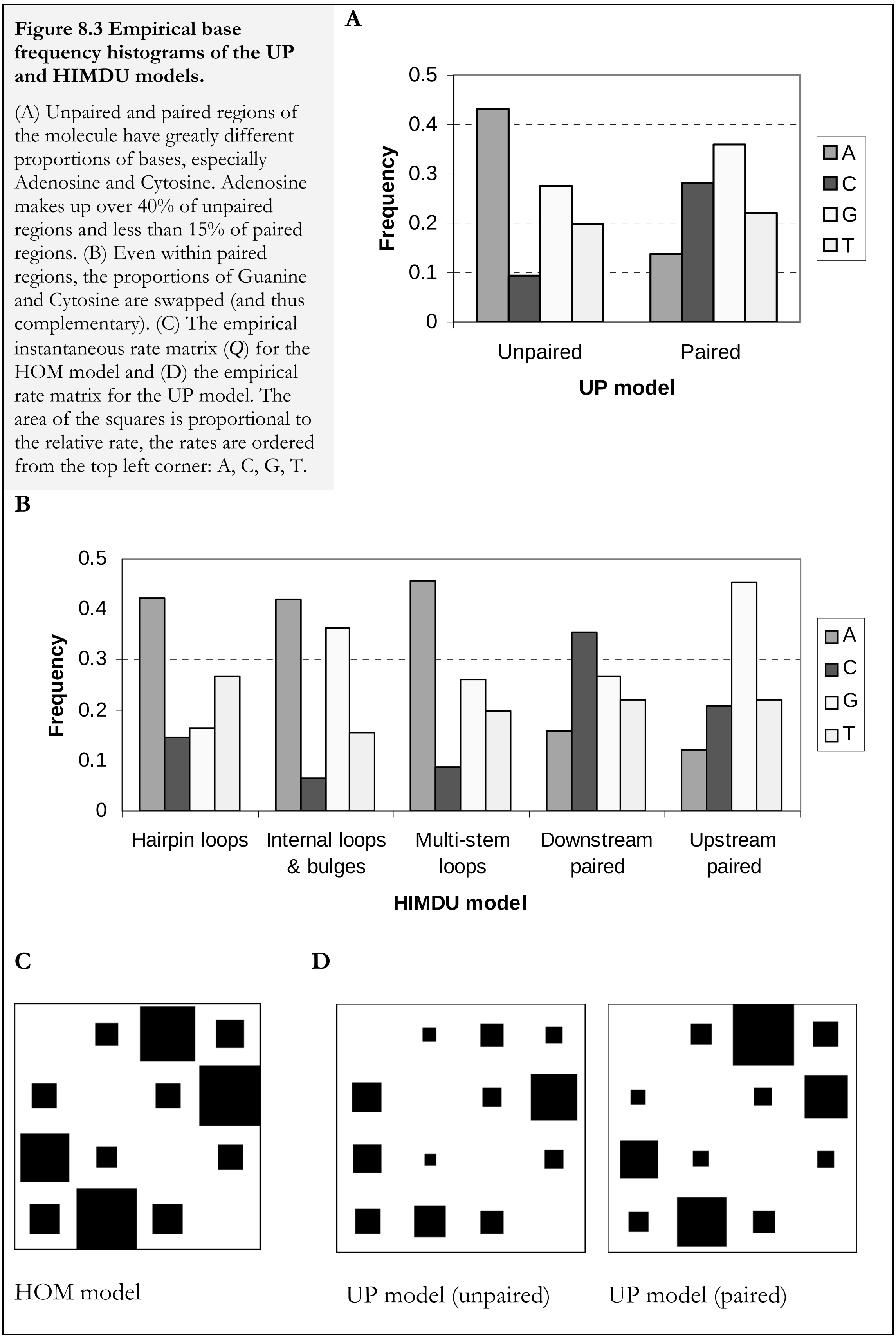


Figure 8.3 Empirical base frequency histograms of the UP and HIMDU models.

(A) Unpaired and paired regions of the molecule have greatly different proportions of bases, especially Adenosine and Cytosine. Adenosine makes up over 40% of unpaired regions and less than 15% of paired regions. (B) Even within paired regions, the proportions of Guanine and Cytosine are swapped (and thus complementary). (C) The empirical instantaneous rate matrix (*Q*) for the HOM model and (D) the empirical rate matrix for the UP model. The area of the squares is proportional to the relative rate, the rates are ordered from the top left corner: A, C, G, T.

8.2.4 Simulations of RNA substitution models

The UP and HOM substitution processes were compared by simulation. Both of the models were used to simulate the evolution of 2000 random 100-nucleotide sequences through 250 structure-consistent point mutations per sequence. An initial structure for each sequence was determined by free energy minimization techniques (ZUKER 1989). An *a priori* assumption of the simulations was that the initial structure of each sequence had to be conserved. This assumption is based on the biological evidence of the extremely high conservation of 16S secondary structure over the entire eubacterial tree. Thus for each sequence only mutations that conserved its structure were 'accepted'.

After each of 250 accepted point mutations, the mean mutational stability and mean thermodynamic stability of the populations were measured. In the first experiment the HOM model of substitution was used as the source of mutation, whereas in the second the UP model was used. In this way, the extent to which each of these models encodes mutational and/or thermodynamic stability was assessed.

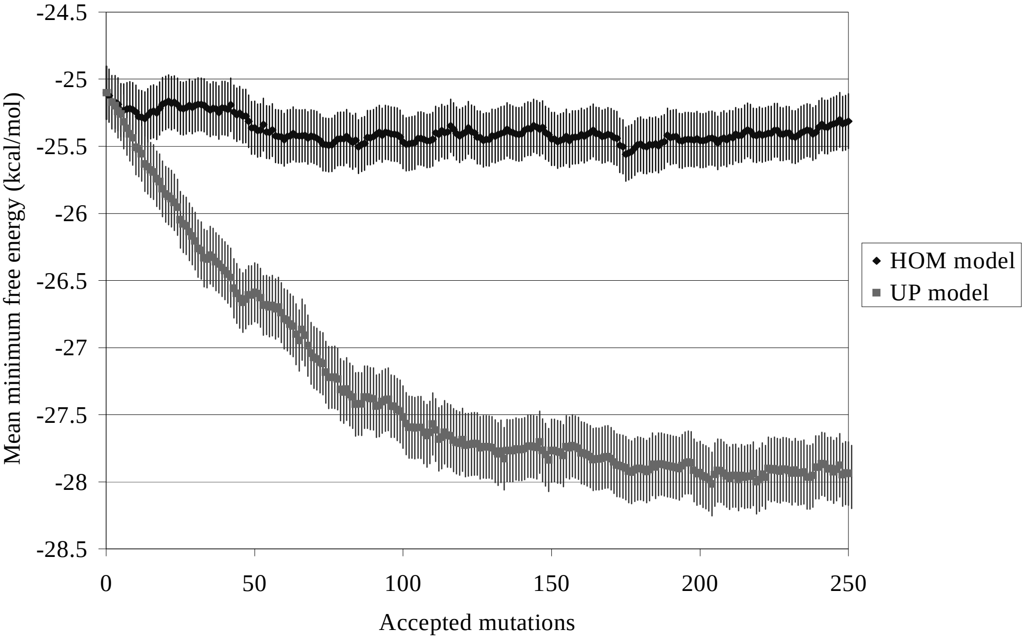
Figure 8.4 demonstrates that under the UP model both the mutational and thermodynamic stability of the simulated sequences increase over time.

Figure 8.4 A comparison of the structural stability of the HOM and UP models.

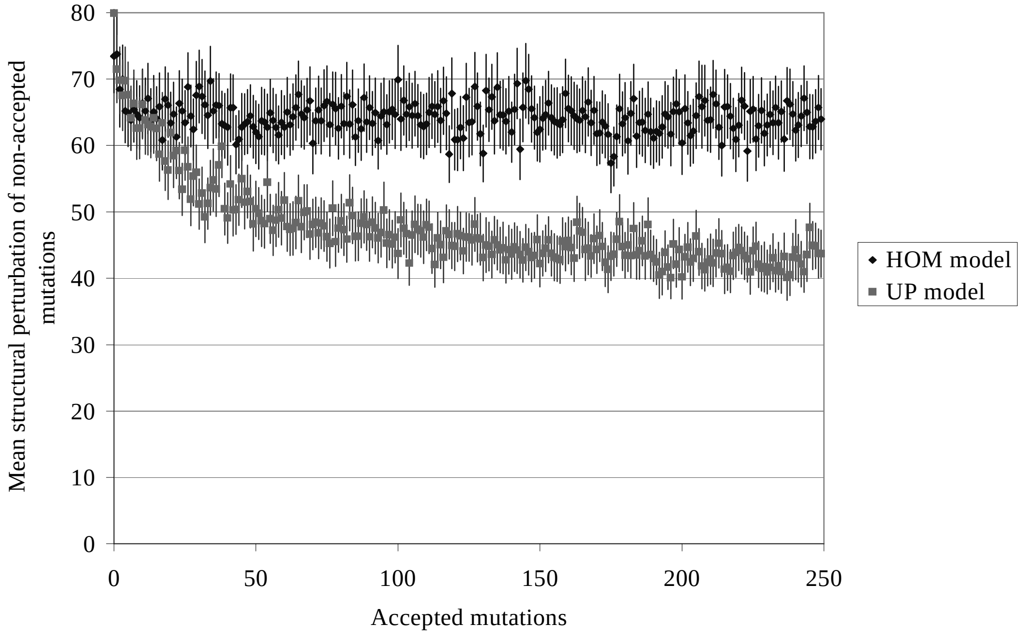
The UP model encodes structural stability, both in terms of (A) thermodynamic stability and (B) stability in the face of mutational pressure. These figures show the result of simulations of sequence under the UP model, starting from a non-partitioned initial sequence.

.

A. thermodynamic structural stability



B. structural stability to mutation



8.2.4.1 Results

Direct examination of the resulting sequences reveals that this increase in stability coincides with the sequences becoming partitioned into different base compositions in different structural categories. Thus, The UP model encodes structural stability, both in terms of (A) thermodynamic stability and (B) stability in the face of mutational pressure. If one remembers that the substitutions *observed* arise from the combination of mutation and s*election* than the results depicted in Figure 8.4 provide strong evidence that the acceptance, by selection, of mutations was determined predominantly by the criterion of thermodynamic and/or mutational stability.

8.2.5 A maximum-likelihood comparison on a small dataset

A likelihood ratio test was employed to compare the fit of the HOM model and the UP model for a set of *Pseudomonas* 16S rRNA sequences on a given tree (AISLABIE *et al.* 2000). The published tree topology was used. To assign the UP categories to the nucleotides of this data set, an *Escherichia coli* 16S rRNA sequence was aligned to the *Pseudomonas* sequences using PILEUP in the GCG package. This alignment was then used to overlay the *Escherichia coli* secondary structure (GUTELL 1994) onto the *Pseudomonas* sequences. The sequence data was then split into two sets and the likelihood of each was calculated based on the empirically derived substitution models for the respective structure categories, described in section 8.2.2. The two log-likelihood values were then combined and compared with the log-likelihood of the combined dataset estimated under a single homogeneous substitution model.

Using PAUP* (SWOFFORD 1999) this likelihood ratio test was also undertaken with substitution models optimised for the data sets under analysis.

8.2.5.1 Results

Table 8.1 shows that the empirically-derived structural model UP outperformed the HOM model, even when the homogeneous model was optimised (HOM*) and the UP model was not (UP).

Table 8.1 Log-likehood values of HOM and UP models of RNA evolution.

Both optimised and un-optimised the UP model proved significantly better than the HOM model as a description of the Pseudomonas dataset analysed.

Model	**Log Likelihood**	**Likelihood ratio tests**		
		Df	**HOM**	**HOM***
HOM	-3082.14	-	-	-
HOM*	-3024.58	-	-	-
UP	-3001.69	10	P<0.00001	P<0.00001
UP*	-2947.28	10	P<0.00001	P<0.00001

8.2.6 Conclusions

Maximum likelihood methods of phylogenetic reconstruction have traditionally assumed that all sites along a gene evolve under the same model of substitution and independently of other sites. While rate heterogeneity has been addressed using Hidden Markov Models (Felsenstein and Churchill 1996; Yang 1995), heterogeneity of substitution model based on RNA secondary structure has not. The assumption of pattern homogeneity has been one of convenience, as the interactions between different sites were poorly understood. However as more and more data about the secondary structure of structural RNA molecules have become available, it has become possible to incorporate this information into phylogenetic reconstruction methods (Muse 1995; Rzhetsky 1995; Tillier and Collins 1995; 1998).

Assigning a 'structural environment' to each nucleotide takes into account different selective pressures but does not take into account the contingency of changes at different sites. As a result the interaction of sites with each other (perhaps due to contact in three-dimensional structure) is not modelled in this analysis. Sites are treated independently and are linked only in that some sites share the same structural environment. Nevertheless, the lack of independence between sites becomes readily apparent when considering the secondary structure of a structural RNA molecule. Pairing between distant sites is the rule rather than the exception and the maintenance of these pairs is often crucial to the function of the molecule. Therefore a substitution in one half of a pair will almost invariably lead to a concomitant substitution in the other. This correlation between distant sites cannot be captured within the standard phylogenetic model of maximum likelihood estimation, without a reassessment of the definition of a site. For example instead of treating every nucleotide as a separate evolutionary unit, each *pair* of nucleotides in the secondary structure helix could be treated as an evolutionary unit. Instead of 4 states (A, C, G, T), these new units would have 6 states (A-T, T-A, C-G, G-C, T-G, G-T). Tillier and Collins (1995; 1998) have implemented a version of this method. Despite not incorporating information about the non-independence of sites, the method described here is a significant improvement on conventional models of substitution and it sheds light on how base composition is partitioned in different secondary structure categories.

The eventual goal is the understanding of how the complete 3D structure and function of a gene product affects the evolutionary patterns of the nucleotides in the gene's coding region. This goal, while still distant, has been assisted in recent times by increased knowledge of the structures of molecules of interest as well as an improved understanding of the general properties of the genotype to phenotype mapping. A successful incorporation of structural information into a model of molecular evolution will require a precise understanding of the relationship between structural homology and sequence homology. The research presented herein is a small step towards that goal.

8.3 Multiple sequence alignment of RNA

Secondary structure is more highly conserved than primary sequence in molecules such as the ribosomal RNA small- and large- subunits (SSU and LSU). Therefore secondary structure prediction methods could potentially be used to improve RNA sequence alignment techniques. This could be particularly important, for example, in bacterial systematics where SSU (also known as 16S-like) rRNA is the main molecule for phylogenetic classification.

Phylogenetic inference techniques rely on the correct identification of homologous characters. In terms of molecular data such as rRNA sequences, the predominant molecular character is the nucleotide position. The act of assigning nucleotides sites to classes of homologues is called *sequence alignment*. Homologous (comparable) nucleotide positions are determined by attempting to optimise the overall *sequence similarity* over two or more related sequences, through the insertion of gaps into the sequences. However structure is more highly conserved than sequence in both RNAs (LÜCK *et al.* 1996) and proteins (LEVITT and GERSTEIN 1998). Therefore it can be argued that *structural similarity* may be a more powerful technique for ascertaining homology. The use of secondary structure prediction might therefore improve the accuracy of sequence alignment.

One of the least rigorously treated problems in molecular evolution is the problem of multiple sequence alignment. This seems to stem from a general inability to come up with concrete statistical models of nucleotide indels that is mathematically tractable. The focus in this section is on an *ad hoc* method by which RNA structural information can be used to enhance the performance of traditional "model-free" methods of sequence alignment.

8.3.1 Pair-wise alignments

In the absence of an evolutionary model, it would seem that a good alignment of two sequences is one that maximises the apparent similarity of the two sequences. Needleman and Wunsch (1970) introduced a dynamic programming approach to solve the problem of pair-wise alignment. This dynamic programming method is guaranteed to produce an optimal alignment of two given sequences for a number of simple alignment-scoring schemes. A popular scheme for scoring alignments is the *affine gap costs* scheme, in which a gap of length k is penalized $g + e(k\text{-}1)$, where g is a fixed "gap-opening penalty" and e is a "gap-extension penalty". The dynamic programming solution to the *affine gap costs* scheme of pair-wise alignment was first described by Gotoh (1982). The algorithm runs in quadratic time (i.e. O(L_1L_2), where L_1 and L_2 are the lengths of the sequences). The *affine gap costs* scoring scheme will be the focus of this section.

8.3.2 Multiple sequence alignments

Given that an exact O(L_1L_2) dynamic programming solution exists for the alignment of two sequences, it seems intuitive that an exact O($L_1L_2 \ldots L_n$) dynamic programming solution should exist for n sequences. This intuition is correct and a program that performs exact multiple sequence alignment is available (LIPMAN *et al.* 1989)[15]. Unfortunately, this method is extremely slow and is currently only feasible for very small alignments (<6 sequences).

By far the most widely used method of multiple sequence alignment is progressive pair-wise alignment (FENG and DOOLITTLE 1987). This is the method implemented in ClustalW and ClustalX (THOMPSON *et al.* 1997; THOMPSON *et al.* 1994), both of which are highly popular automated alignment tools. Progressive pair-wise alignment is a heuristic method. The first step involves calculation of an alignment between all pairs of sequences. This pair-wise alignment uses standard dynamic programming methods to find an exact optimal alignment (given an *affine gap costs* scoring scheme) between two sequences. The resulting pair alignments are used to calculate evolutionary distances between the sequences. ClustalX uses the Hamming distance (*H*) or a simple correction thereof. These pair-wise distances are then used to construct a *guide tree* between the sequences to be aligned. The tree is used to guide a hierarchical algorithm of successive pair-wise alignments between clusters. The last two clusters are aligned to form the final alignment at the root of the guide tree.

8.3.3 Statistical alignment

The alignment methods outlined in section 8.2.1 and 8.2.2 are based on *affine gap cost* scoring schemes. These models are in some sense "model-free" as they are not based on any explicit statistical model of the evolutionary process of indels.

Thorne, Kishino & Felsenstein (1991) introduced a method to calculate the maximum likelihood alignment between a pair of sequences under the assumption that insertions and deletions occur one nucleotide at a time, governed by a birth-death process. Their work has since been extended to the problem of multiple sequence alignments (HEIN 2001; STEEL and HEIN 2000), however it is fair to say that the field of statistical

[15] However, it should be noted that since this method does not use a phylogeny the penalties is uses are not 'correct' in the sense that they don't approximate an evolutionary process.

alignment is still in its infancy. Statistical multiple sequence alignment is still not feasible for most practical situations, especially when the precise phylogenetic relationships between the sequences are not known. Current methods assume a fixed tree topology and overly simple models of insertion and deletion. For these reasons, less sophisticated "model-free" methods will be investigated in this section.

8.3.4 Sequence alignment of RNA sequences

An alignment based on secondary structure alone can be produced by representing each nucleotide in the molecule as one of three characters: unpaired, paired upstream or paired downstream (see Figure 8.2). These new character sequences can then be subjected to sequence alignment (and tree-building) in the same manner as nucleotide sequences. It should be noted that information is lost in this coding of the molecule. This is a manifestation of the neutrality (or redundancy) in the mapping from base sequence to secondary structure. This is because the application of this conversion discards neutral substitutions that do not change the structure of the molecule. A more powerful method might be to combine both sequence and structure information into a single "combined" character string. Because structure is more highly conserved than sequence we expect to see greater structural homology than sequence homology in the variable regions of the alignment.

8.3.4.1 Combined sequence-structure characters

For each site in a RNA-encoding sequence, both the sequence information (A, C, G, T) and the structure information (unpaired, paired-upstream, paired-downstream) are combined into a single state, so that each site can be one of 12 states {**A·**, **C·**, **G·**, **T·**, **A(**, **C(**, **G(**, **T(**, **A)**, **C)**, **G)**, **T)**}, where structure categories {·, **(**, **)**} represent unpaired, paired-upstream and paired-downstream respectively. This secondary structure model is intermediate between the UP and HIMDU models discussed in section 8.2, and captures the two main empirical features of secondary structure discovered in that section: (i) unpaired-paired differences and (ii) upstream-downstream complementarity.

8.3.4.2 Non-independence of secondary structure paired regions

Unfortunately both of the above methods of structure alignment still treat each nucleotide independently. While this is an obvious shortfall of these methods, it doesn't preclude their usefulness. The "combined" character method suggested here has the advantage of ease of use with (i) available software and (ii) existing statistical techniques for alignment and phylogenetic inference. In fact is has been demonstrated that the assumption of independence of sites made by maximum-likelihood methods does not adversely bias phylogenetic reconstruction from RNA encoding gene sequences (TILLIER and COLLINS 1995).

Here, alignments based on nucleotide sequences are compared with alignments based on "combined" character sequences for a group of 28 bacterial nucleotide sequences for which the 16S rRNA secondary structure is known. The results suggest that without strong statistical information on nucleotide indels, RNA structure information can improve the performance of standard alignment algorithms.

8.3.5 Combining RNA secondary structure and sequence information

In this section an *affine gap costs* alignment scoring system that incorporates both sequence and secondary structure information is investigated. It is shown to be more robust to uncertainty in insertion and deletion rates than a scoring scheme that ignores structure information. To demonstrate the scoring scheme, twenty-eight 16S sequences representative of bacterial diversity were aligned under a variety of gap costs.

8.3.5.1 Sequence info only

A set of sequences representative of the major lineages of eubacteria was compiled from the ribosomal database of Gutell *et al* (CANNONE *et al.* 2002). Because it is not obvious what appropriate values for gap costs g and e are, a set of 1600 candidate alignments were generated. The ClustalX software was used to generate an alignment for all combinations of g_G and e_G between 0.05 and 2.0, in step sizes of 0.05[16]. Each of the 1600 multiple sequence alignments generated was evaluated by considering the sum of the scores of all pair-wise alignments implied by it. Different *evaluation schemes* (i.e. different values of gap costs g_E and e_E) favoured different candidate alignments (see Figure 8.5). This is not surprising, as each alignment was generated using a different *generation scheme* (i.e. different values of gap costs g_G and e_G). In fact, had it been feasible to generate the alignments using *exact* multiple sequence alignment (e.g. with MSA), each *evaluation scheme* would have chosen the alignment from the corresponding *generation scheme* as optimal; meaning that all 1600 generated alignments would have been optimal under at least one *evaluation scheme*. However, this is not the case with progressive pair-wise alignment. In fact, a very small number of the candidate alignments were optimal over a large range of *evaluation schemes*. This is of practical interest, and suggests that progressive pair-wise alignment is very

[16] To differentiate between gap costs used to *generate* alignments (*generation schemes*) and gap costs used to *evaluate* alignments (*evaluation schemes*) subscripts G=generate and E=evaluate are used.

sensitive to the *generation scheme* used. For example the alignment generated using g_G=0.9 and e_G=0.25 was the best alignment (of the 1600 candidate alignments) for 36.5% of the *evaluation schemes* considered (see Figure 8.5). Perhaps more surprisingly, this alignment was *not* the optimal alignment for its corresponding *evaluation scheme*. Instead the alignment generated using g_G=0.65 and e_G=0.15 was the best alignment under the g_E=0.9 and e_E=0.25 *evaluation scheme*.

How do we pick which of the generated alignments is the best? It depends on the *evaluation scheme* we consider the best. Which *evaluation scheme* do we consider the best? The approach taken here was to choose a set of plausible *evaluation schemes* (the white triangle in Figure 8.5) and pick the alignment that is the optimal for the largest proportion of them. The *evaluation schemes* considered plausible are those that penalize longer indels proportionally equal or less than an equivalent number of independent single-nucleotide indels ($e_E \leq g_E$). In addition only *evaluation schemes* that penalize an insertion no more than twice a substitution were considered. By this method we chose the alignment generated with parameter values of g_G=0.9 and e_G=0.25 as the best alignment of the 28 bacterial sequences.

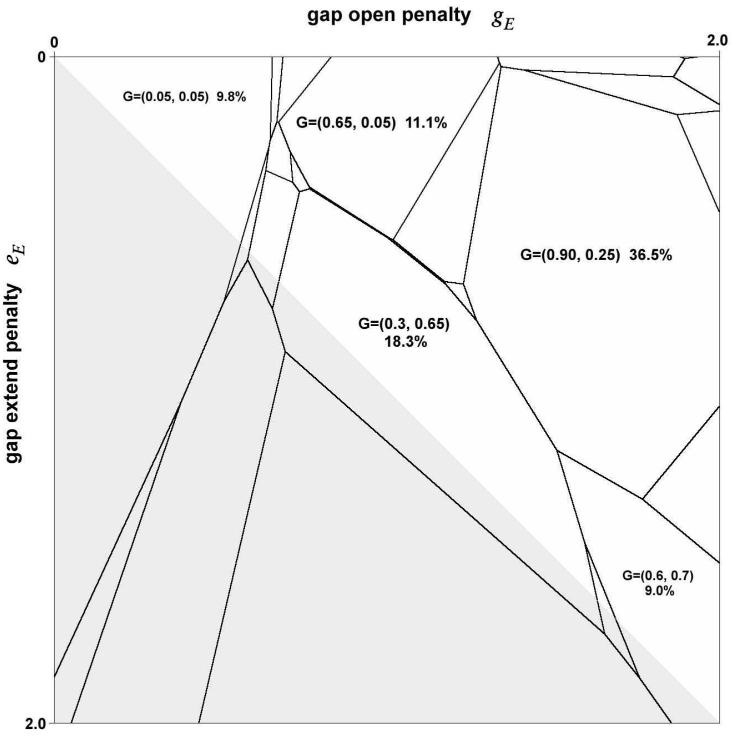


Figure 8.5 Comparison of different *evaluation schemes* where only sequence information was considered.

Each polygon represents a set of *evaluation schemes* that unanimously pick a single candidate alignment as the best. For the largest regions, the gap-opening penalty and gap-extension penalty of the *generation scheme* are shown. *Evaluation schemes* in grey are non-biological and ignored for this analysis. The percentages represent the proportion of the white triangle occupied by the five largest regions. Diagrams of this type, involving just sequence information have been previously investigated by Fitch and Smith (FITCH and SMITH 1983)

8.3.5.2 Adding structural information

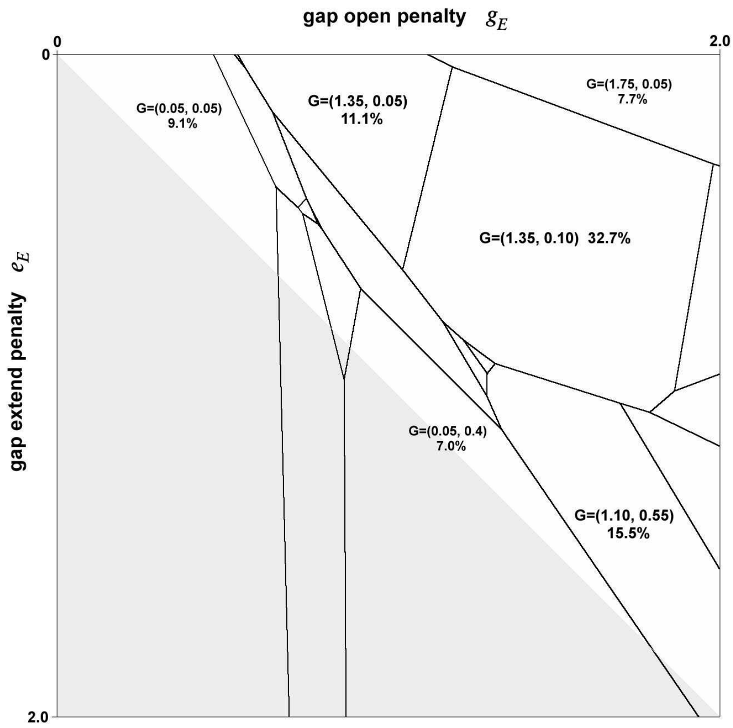
By including structural information in the *generation scheme* a second set of 1600 candidate alignments was generated for the same 28 eubacterial sequences. The structural information was added by penalizing aligned positions that had a mismatch in structure category. The mismatch penalties used were:

	Unpaired	Paired upstream	Paired downstream
Unpaired	-	0.5	0.5
Paired upstream		-	1.0
Paired downstream			-

The above penalties were chosen so that a change from paired to unpaired was penalized less than the more radical change that would result in a paired-upstream changing to a paired-downstream or *vice versa*. The absolute weights were chosen to give approximately similar weights to sequence and structure information. The generation scheme was the same as in section 8.3.5.1 apart from these added penalties. The nucleotide evaluation scheme was retained, so that structural similarity was not part of the *evaluation* criteria. Again there were a small number of candidate alignments that were optimal for a wide range of evaluation schemes (Figure 8.6).

Figure 8.6 Comparison of different *evaluation schemes* where both sequence and structure information was considered.

Each polygon represents a set of *evaluation schemes* that unanimously pick a single candidate alignment as the best. For the largest regions, the gap-opening penalty and gap-extension penalty of the *generation scheme* are shown. *Evaluation schemes* in grey are non-biological and ignored for this analysis. The percentages represent the proportion of the white triangle occupied by the six largest regions.

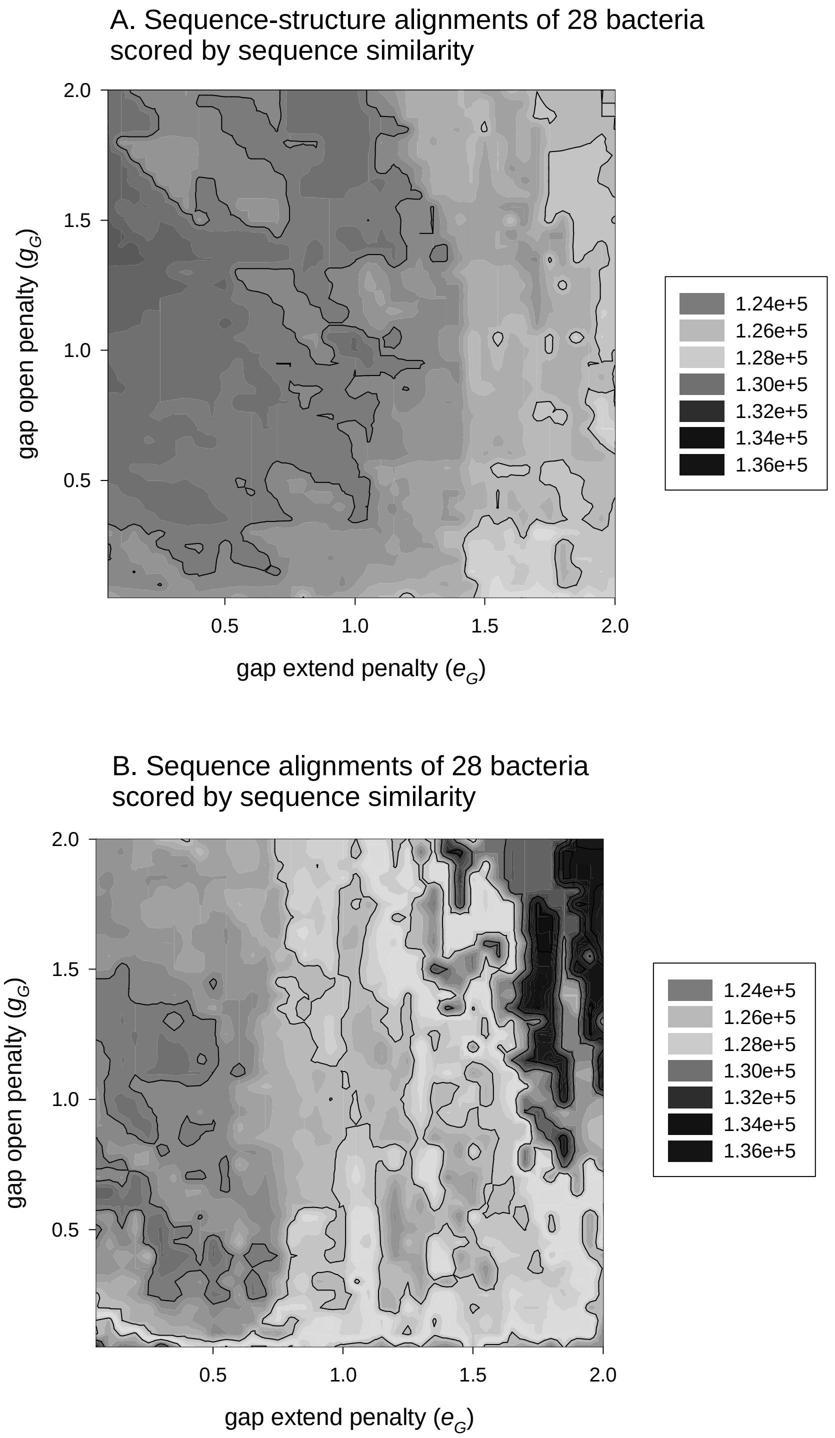


Another way to compare alignments constructed with and without RNA structure information is to consider the distribution of alignment scores under a single *evaluation scheme*. This comparison provides information about how stable alignment scores are, when different gap *generation schemes* are used, with and without structural information. In Figure 8.7 a graphical representation of 1600 alignments generated with only sequence information and 1600 alignments generated with both sequence and structure information is presented. This figure shows that the use of structural information provides alignments that are robust to misspecification of gap penalties in the *generation schemes*. The alignment score surface of alignments generated with structural information is flatter and has a higher average than the alignment surface generated from sequence data alone.

Figure 8.7 Comparison of different alignment *generation schemes*.

(A) alignments produced using both sequence and structure penalties and (B) alignments produced using only sequence penalties.

Red areas of the plots are good alignments, yellow are average and green and blue are poor. The alignments that use structure information are much more robust to different gap penalties. In determining the scores, penalties were: transversions 1.0, transitions 0.5, upstream-downstream 1.0, unpaired-paired 0.5, gap open 1.0, gap extend 0.25.



A. Sequence-structure alignments of 28 bacteria scored by sequence similarity

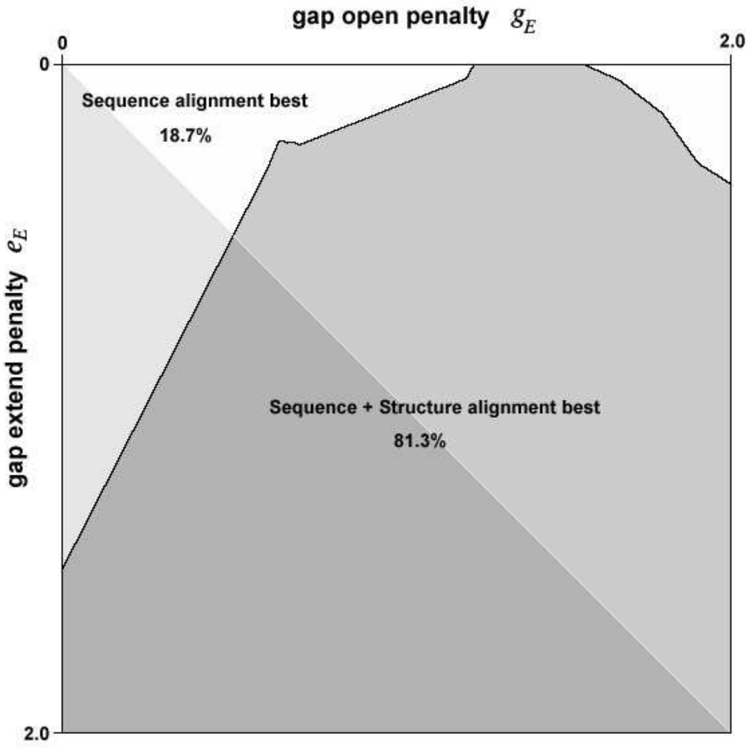
B. Sequence alignments of 28 bacteria scored by sequence similarity

8.3.5.3 Is structure information always better?

Will the addition of structure information always improve progressive pair-wise alignment? To investigate this question, evaluation schemes were categorized by whether the best alignment of the 3200 candidates did or did not use structure information. Figure 8.8 shows that most (~80%) but not all *evaluation schemes* picked an alignment generated with the aid of RNA secondary structure information. Furthermore the *evaluation schemes* that preferred sequence-only alignments were generally those schemes with lower gap-opening penalties. These *evaluation schemes* lead to 'gappy' subjectively poor optimal alignments that do not conform to expectations consistent with biological principles.

Figure 8.8 Comparison of sequence alignments versus sequence-structure alignments.

Most (~80%) but not all *evaluation schemes* picked an alignment generated with the aid of RNA secondary structure information.



8.3.6 Conclusions

The use of information about RNA secondary structure improves the robustness of progressive pair-wise alignment to uncertainty in the choice of *affine gap costs*. For many particular *evaluation schemes* use of structural information enables the generation of higher scoring alignments than alignments generated by using sequence information alone. Ultimately, a model of secondary structure evolution, based on the work of Fontana and Schuster should provide the means for statistical alignment of RNA-encoding genes. However until then, it seems that the use of RNA secondary structure in standard alignment packages such as ClustalX, may be a useful alternative to naïve sequence-only alignment.

8.4 Bayesian inference of substitution parameters and phylogeny from sequence-structure data

In section 8.2, rRNA evolution was considered under the assumption that the secondary structure of the molecule was completely conserved across eubacteria. However, in reality, small local rearrangements of RNA secondary structure do accumulate over long periods of time. In this section a model of evolution that permits changes in both structure and sequence is investigated. The combined sequence-structure character states described above in section 8.3.4.1 where used within a sample-based Bayesian inference framework.

8.4.1 Bayesian inference of combined sequence-structure characters

The posterior probability density under consideration is:

$$h_{comb}(T, R_{comb} \mid D) = \frac{1}{Z} \Pr\{D \mid T, R_{comb}\}\, f_T(T)\, f_{Rcomb}(R_{comb}) \qquad (8.1)$$

Where:

T is an unrooted tree topology with branch lengths in mutations

R_{comb} is a relative rate matrix of combined sequence-structure characters

$f_T(T)$ is a uniform prior density on trees

$f_{Rcomb}(R_{comb})$ is a Jeffreys' prior on relative rates

$\Pr\{D| T, R_{comb}\}$ is the likelihood

The equilibrium frequencies of combined character states in the data were fixed to values derived empirically from the input data and were not part of the inference. Two proposal mechanisms where used to sample the posterior density described in 7.1. The first is the LOCAL move described by Wilson and Balding (1998) for unrooted trees (also used in MrBayes). In addition a simple scaling move on each relative rate was employed as described for the case of nucleotides relative rates in Chapter 5.

8.4.2 Data

An alignment of 62 16S-like ribosomal RNA sequences was analysed. The *generation scheme* that produced the 'best' alignment of eubacteria in the previous section was used to align a larger set of 16S sequences, that included 12 eukaryotic sequences, 22 archaeal sequences in addition to the 28 bacterial sequences analysed in section 8.3. The resulting alignment was then assumed 'known' and used as input to an MCMC analysis to jointly estimate the relative rates and phylogeny of the sequence-structure alignment. The secondary structure used for these sequences was as published in the Gutell online database (CANNONE *et al.* 2002).

8.4.3 Results

Figure 8.9 and Figure 8.10 show the estimated relative rate matrix and a sample genealogy respectively. The relative rate matrix bears striking resemblance, along the diagonal, to the models developed in section 8.2. In addition, the off-diagonal rates, in which structures change, are much lower, empirically confirming statements of structural conservation across 16S-like sequences.

Figure 8.9 Estimated relative rate matrix (*R_{comb}*) for combined characters.

Empirical frequencies are factored out, so the *Q* matrix can be obtained by scaling each column by the frequency of the corresponding state. The rates were estimated relative to A·↔ C· (the gray circle in top left). The inner circle corresponds to the lower bound on the estimate of the relative rate, and the area enclosed by the outside of the circle corresponds the upper bound on the estimate. The thickness of the circle thus represents the uncertainty of the estimated rate.

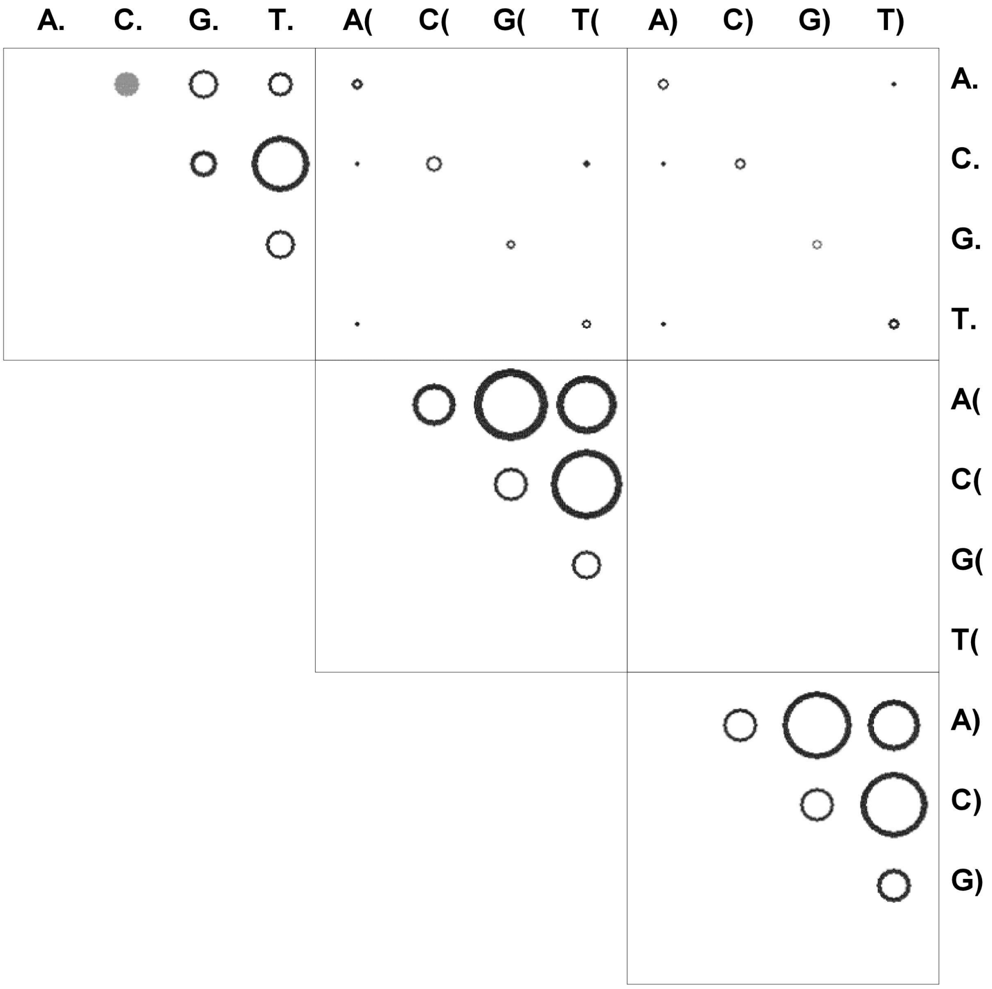
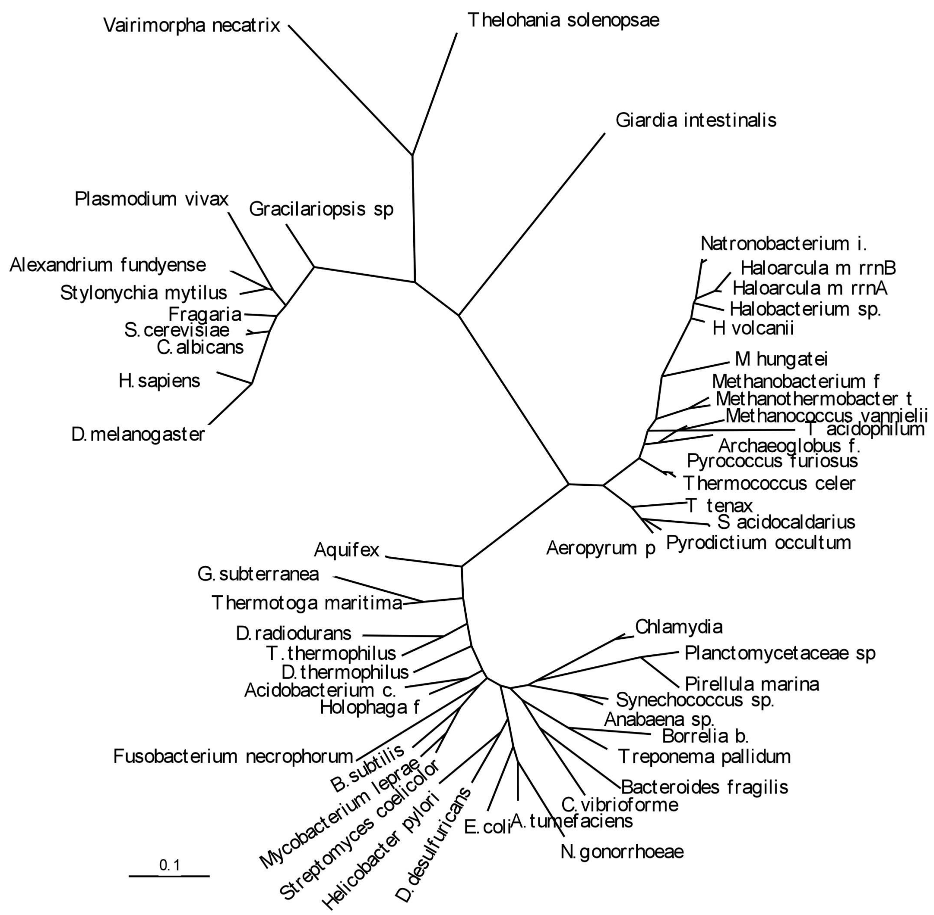


Figure 8.10 Sample tree of life established using sequence and structure information.



8.5 Discussion

The research presented in this chapter provides examples of the use of large data sets and computationally intensive methods in the investigation of molecular evolution and molecular systematics. With the growing databases of DNA and protein sequences, there is an increasing demand for detailed examination and analysis of this information. This type of research is essential to the continued progress of all areas of molecular biology. Molecular biologists are faced with an ever-increasing need to use computational techniques in their research. A simple example of this is the fundamental importance of comparison of molecules for similarities both in sequence and structure. By improving these computational techniques the work done by molecular biologists is also improved.

It is for this same reason that theoretical biology is important. The testing and adoption of theoretical evolutionary hypotheses has direct implications on many important computational tools. For example, almost all character-based phylogenetic tree-building algorithms have a model of molecular evolution as a basis. By testing a new theoretical hypothesis of molecular evolution this research tests the validity of the models of evolution employed in current tree-building algorithms for RNA-based phylogenies.

9 A Tangent: Spatial Population Genetics

9.1 Overview

Most classical population genetics models treat populations as either panmictic (perfect mixing), and thus exhibiting no spatial structure, or highly structured into (perfectly mixing) subpopulations with a migration matrix defining flux between them. Here, an alternative model, based on spatial diffusion in n-dimensions will be introduced and a Bayesian inference framework will be described. The inference method is related to (i) comparative methods for Brownian characters and (ii) maximum likelihood methods on continuous characters. Spherical geometry is considered for the practical purposes of using latitude and longitude information in real populations. Finally, a simple class of lattice models are used to carry out simulations that demonstrate some of the characteristics of spatially distributed populations. The aim of this chapter is to provoke thought and demonstrate future directions that might be fruitful in Bayesian population genetics, and no attempt is made to validate the methods described.

9.2 A simple model of movement for inference

The *panmixia* assumption of the Wright-Fisher population model amounts to a statement that the progeny of one individual can potentially displace the progeny of *any* other individual in the population, no matter how far away. This is obviously not true if the area covered by the population is large in comparison to the average area accessible to an organism in its lifetime. The diffusion process is a simple and general characterization of random movement that can account for the localization of individuals in a population. Almost any jump process (such as a random walk in one dimension) can be represented as a diffusion process in some kind of limit. Gaussian diffusion in n-dimensions is the continuous limit of a random (Brownian) walk on an n-dimensional lattice. In this section, a simple model of movement based on diffusion in n-dimensions (MALECOT 1948) will be discussed in the context of inference from a small sample of sequence data collected at known times and places.

9.2.1 Bayesian estimation of diffusion in n-dimensions on a tree

The rate of a diffusion process is governed by the diffusion coefficient, D, which corresponds to the rate of increase in area accessible from some starting point. Here, geographical diffusion of organisms and their haplotypes is considered. If time is measured in years and area is measured in square kilometres, then $D = 10$, say, implies that on average the area accessible to a haplotype/organism, starting from a given point at time *t*=0, increases at a rate of 10 km^2/year. In this section we develop a formulation of the posterior density of the rate of diffusion of a group of organisms given a known genealogical topology and known ages and positions of a sample of sequences. First we will consider the simple problem of diffusion in a Euclidean space of *n* dimensions:

$$\frac{\partial p}{\partial t} = D\,\frac{\partial^2 p}{\partial \mathbf{x}^2} \tag{9.1}$$

where position **x** has *n* dimensions. The probability density function for a haplotype/organism travelling Euclidean distance *d*, in any direction from a defined starting point, in a given time *t* is:

$$p(d,t) = \left(\frac{1}{\sqrt{\pi D t}}\right)^n \exp\left(\frac{-d^2}{Dt}\right) \tag{9.2}$$

An edge (or alternatively 'branch'), joining two nodes on an evolutionary tree represents a lineage of organisms through time (going back in time: child, parent, grandparent, great-grandparent, et cetera). The genetic information of these organisms can be thought of as the baton carried by runners in a relay race that is passed from one runner to the next during the race. For a given lineage, the baton is passed on each time a new organism is born. Here we are considering a 'race' in which the runners are running about in a Brownian fashion. The average diffusive properties of this birth/diffusion/birth process along an edge of an evolutionary tree can be analysed by regarding the lineage of organisms as a single, long-lived organism, moving in a Brownian fashion, with an average diffusion rate, *D*. It then follows from equation 9.2 that the log-likelihood (dropping constants) of two nodes in a tree, joined by an edge of length *t*, being separated in space by Euclidean distance *d*, is:

$$\ln(p(d,t)) \propto \frac{-d^2}{Dt} + \frac{n}{2}\ln(Dt) \tag{9.3}$$

One of the first formulations of diffusion on a tree was by Cavalli-Sforza & Edwards (1967). Cavalli-Sforza & Edwards used this formulation for modelling genetic drift of gene frequencies. Here, diffusion will be used to model movement in *physical* space rather than genetic space. First, recall from chapter 5 that E_g denotes the edge set of g, so that $g = (E_g, t_Y)$ specifies a genealogy (the branching topology E_g, and the node times t_Y). For a given genealogy, g, a given diffusion rate D, and modern positions in space $\mathbf{x}_I$, the probability density function for ancestral nodes having positions $\mathbf{x}_Y$ is:

$$P(\mathbf{x}_Y \mid g, D, \mathbf{x}_I) = \prod_{\langle i,j \rangle \in E_g} \left(\frac{1}{\sqrt{\pi D(t_i - t_j)}} \right)^n \exp\left(\frac{-\left|\mathbf{x}_i - \mathbf{x}_j\right|^2}{D(t_i - t_j)} \right) \tag{9.4}$$

where $|\mathbf{x}_i - \mathbf{x}_j|$ is defined as the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Equation 9.4 gives us a way to obtain the most probable set of physical positions of ancestral nodes in a tree with fixed topology, fixed node ages and fixed diffusion rate. Unfortunately it is usually the case that the rate of diffusion is actually a parameter of interest. When the rate of diffusion is not known we can use MCMC integration. Assuming a fixed genealogy g, the posterior probability density of interest is:

$$P(\mathbf{x}_Y, D \mid g, \mathbf{x}_I) = \frac{1}{Z} P(\mathbf{x}_Y \mid g, D, \mathbf{x}_I) f_D(D) \tag{9.5}$$

Where Z is the unknown normalizing constant and $f_D(D)$ is the prior distribution of D. Furthermore it is often the case that the ages of the ancestral nodes are not known either. For a fixed topology and unknown times we can adapt and extend the posterior density given by Equation 5.6 in Chapter 5. Using S for sequence data we are interested in the posterior probability density:

$$P(\mathbf{x}_Y, D, \mu, \theta, t_Y \mid S, E_g) = \frac{1}{Z} \Pr\{S \mid \mu, g\} P(\mathbf{x}_Y \mid g, D, \mathbf{x}_I) f_G(g \mid \theta) f_{\mathrm{M}}(\mu) f_{\Theta}(\theta) f_D(D)$$

(9.6)

Where:

$\mathbf{x}_Y$	the physical positions in Euclidean space of ancestral nodes Y.
D	the diffusion coefficient (in units area / time)
μ	the mutation rate (in units of time)
θ	the product of effective population size and generation length
t_Y	the times (ages) of ancestral nodes.
$f_G(g \mid \theta)$	the coalescent density of genealogy, g, given population size θ.

The posterior probability density described in equation 9.6 can be sampled using MCMC by adding two simple symmetric proposal mechanisms to the MCMC kernel described in Chapter 5: A random walk on the diffusion coefficient and a random walk in Euclidean space of each of the ancestral node positions.

9.2.2 Ecological data and spherical geometry

Often, geographical information (in the form of latitude and longitude coordinates) is available in conjunction with sequence data. For widely distributed populations and species, geographic distances and genetic distances are frequently correlated (BARNES *et al.* 2002), suggesting that isolation by distance is a common mechanism in natural populations. If the area considered is large, or far away from the equator, longitude and latitude data cannot be treated as Euclidean coordinates. However, it should be noticed that, in equations 9.1 and 9.2 the geographical information appears only as a distance in n-dimensional space. The surface of the earth is, though not Euclidean, still approximately locally flat. Additionally, a number of methods exist for estimating the distance between two points on the earth. The simplest, called the Great Circle distance, is a basic result of spherical trigonometry. It works by simply approximating the shape of the earth with a sphere. A reasonably accurate approximation of diffusion on the surface of a sphere can be modelled, by simply replacing the Euclidean distances in equations 9.1 and 9.2 with Great Circle distances. This provides the potential for inference of latitude and longitude coordinates of ancestral nodes in a tree, when coordinates of the sequences at the leaves are known. Let r_E be the mean radius of the earth in kilometres (according to NASA the volumetric mean radius is 6371.0 kilometres; http://nssdc.gsfc.nasa.gov/planetary/factsheet/earthfact.html).

The Great Circle distance (in kilometres) between two (latitude, longitude) points $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x}_2 = (x_2, y_2)$ on the Earth, can be calculated by:

$$d_{GC}(\mathbf{x}_1, \mathbf{x}_2) = r_E \arccos(\sin(x_1)\sin(x_2) + \cos(x_1)\cos(x_2)\cos(y_2 - y_1)) \tag{9.7}$$

Equation 9.4 can be rewritten as:

$$P(\mathbf{x}_Y \mid g, D, \mathbf{x}_I) = \prod_{\langle i,j \rangle \in E_g} \left(\frac{1}{\sqrt{\pi D(t_i - t_j)}} \right)^2 \exp\left(\frac{-d_{GC}(\mathbf{x}_i, \mathbf{x}_j)^2}{D(t_i - t_j)} \right) \tag{9.8}$$

where $d_{GC}(\mathbf{x}_1, \mathbf{x}_2)$ is given by equation 9.7 and $\mathbf{x}_Y$ and $\mathbf{x}_I$ are now a set of unknown and known (latitude, longitude) pairs respectively. It should be noted, that this is only an approximate method, and its discussion is confined to geographical scales much smaller than the radius of the earth ($d_{GC} << r_E$), where the approximation is fairly accurate. From here we can develop an analogue of equation 9.5 that will allow us to simultaneously infer divergence times, rate of diffusion and the location on earth of the ancestral nodes in a tree for which sequence data and geographical position information is available.

9.3 Simple models of movement for simulation

For the purposes of simulation, a class of simple spatial models of evolving populations was investigated. The standard Wright-Fisher population model can be extended to include the concept of spatially localized interactions by placing the population on a two-dimensional lattice and requiring that each member of the population interact only with a local population of neighbouring individuals. This model, referred to as the *isolation by distance* model, has been investigated by simulation at least as far back as 1971 (ROHLF and SCHNELL 1971). This method of relaxing the panmictic assumption of the Wright-Fisher model was used to test of the robustness of the inference strategy described in section 9.2.

9.3.1 Lattice models

Many models of spatial dynamics use the simplifying concept of a lattice to discretize space. All the models considered in this section are described in terms of a finite sized two-dimensional lattice. We will focus on "constant-organisation" models in which the population size is exactly fixed, so that there are no population size fluctuations from generation to generation. In the simplest case each point on the lattice contains exactly one individual (for example, a 20 x 20 lattice has 400 individuals). Generations are discrete and synchronized, so that in each new generation, all individuals are replaced. In each generation, each point on the lattice is filled with the progeny of a parent from the previous generation. The parent is chosen from a neighbourhood surrounding the focal point. Two different neighbourhoods will be considered below. A number of boundary conditions of the lattice at the edges are conceivable including periodic, absorbing or reflective. In this study, periodic boundary conditions were used, which have the simplifying property that all cells are dynamically equivalent.

9.3.2 The box neighbourhood

The first neighbourhood investigated is a simple neighbourhood with a positive integer parameter K, such that the point (i, j) is in the neighbourhood of (x, y) iff $\max(|x-i|,|y-j|) \le K$. This simply defines a square centred on the point (x, y) enclosing $(2K+1)^2$ points. As K increases, the size of the neighbourhood increases. Each generation, each point on the lattice is filled with the progeny of a randomly selected parent from its neighbourhood. In this way, each member of the population in generation G will be represented by, at most, $(2K+1)^2$ progeny in generation $G+1$.

Related progeny will be clustered around the original parent position Figure 9.1 illustrates a population with some highlighted neighbourhoods of size *K*=2.

Figure 9.1 The neighbourhood model.

N = 400 (20x20) and *K* = 2. Four individuals and their corresponding neighbourhoods are highlighted. The progeny of only two of the highlighted individuals are able to compete in the next generation.

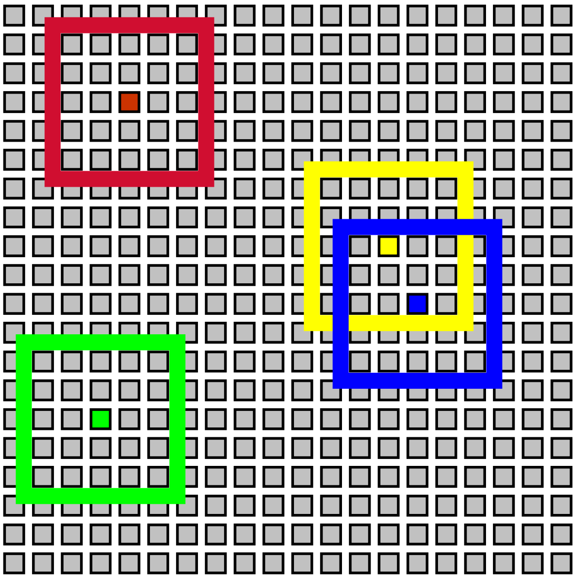


Figure 9.2 shows three different populations of 40,000 individuals that each started from a uniform field of genomes before evolving for 2000 generations. Mutation rates of 1×10^{-5}, 5×10^{-5}, 1×10^{-4} mutations per site per generation were used. In addition, Figure 9.3 shows the spatial evolution of a single nucleotide position over four orders of magnitude of mutation rate. All of these examples are for a neighbourhood size of *K*=1.

Figure 9.2 Three different populations evolving spatially under the neighbourhood model.

Each population has 40,000 individuals on a 200 x 200 lattice. Each population started from a uniform field of genomes and was evolved for 2000 generations. Mutation rates of (A) $1\text{x}10^{-5}$, (B) $5\text{x}10^{-5}$, (C) $1\text{x}10^{-4}$ mutations per site per generation are shown. In all three cases K = 1. For the purpose of representation, each genome was a binary string of length 24 and was interpreted as a 3-byte RGB encoded colour for visual representation. This means that some mutations have a much larger impact on the colour than others, which, while nicely capturing the qualitative nature of real biology is not exactly a representation of neutral evolution!

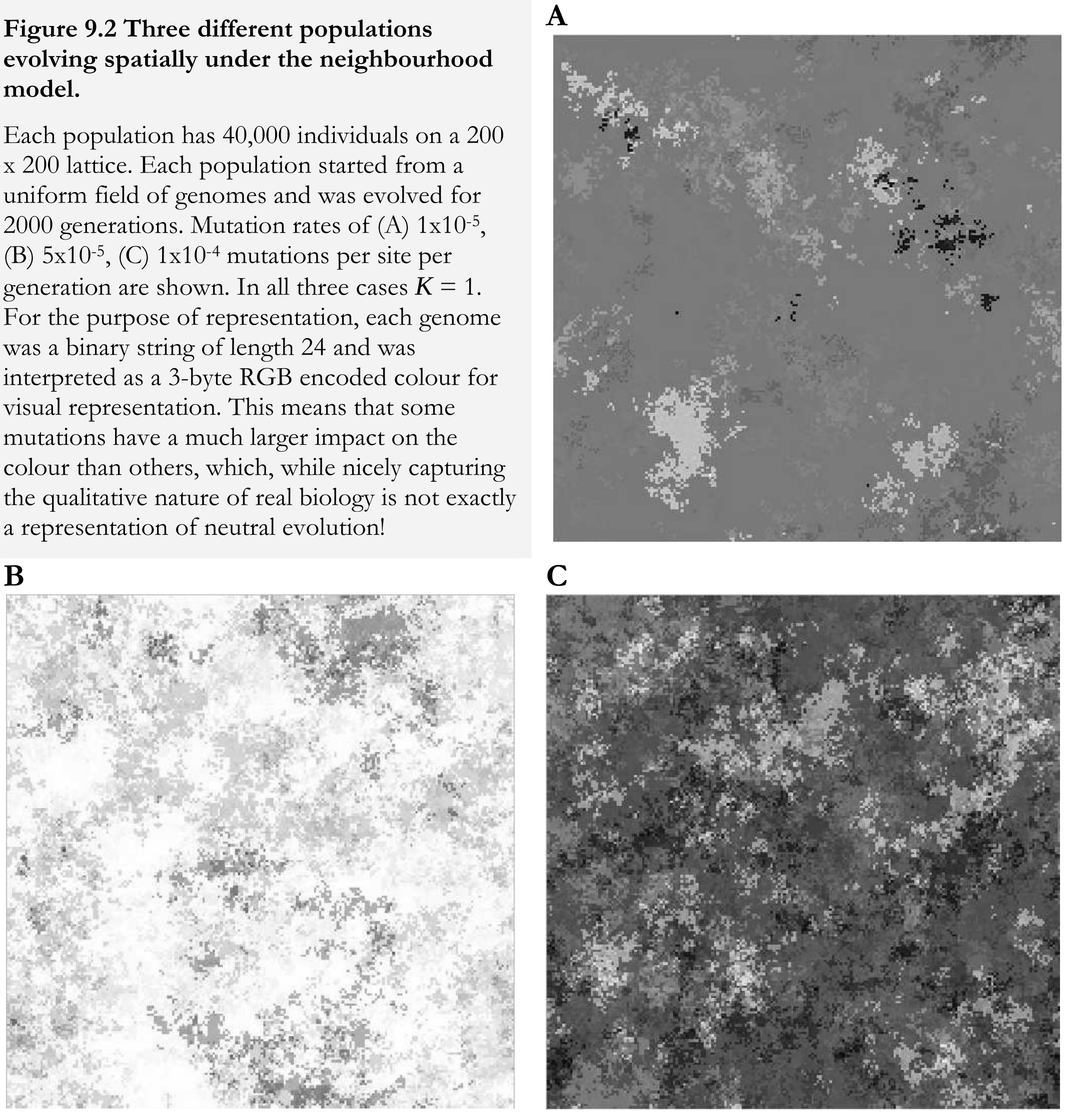
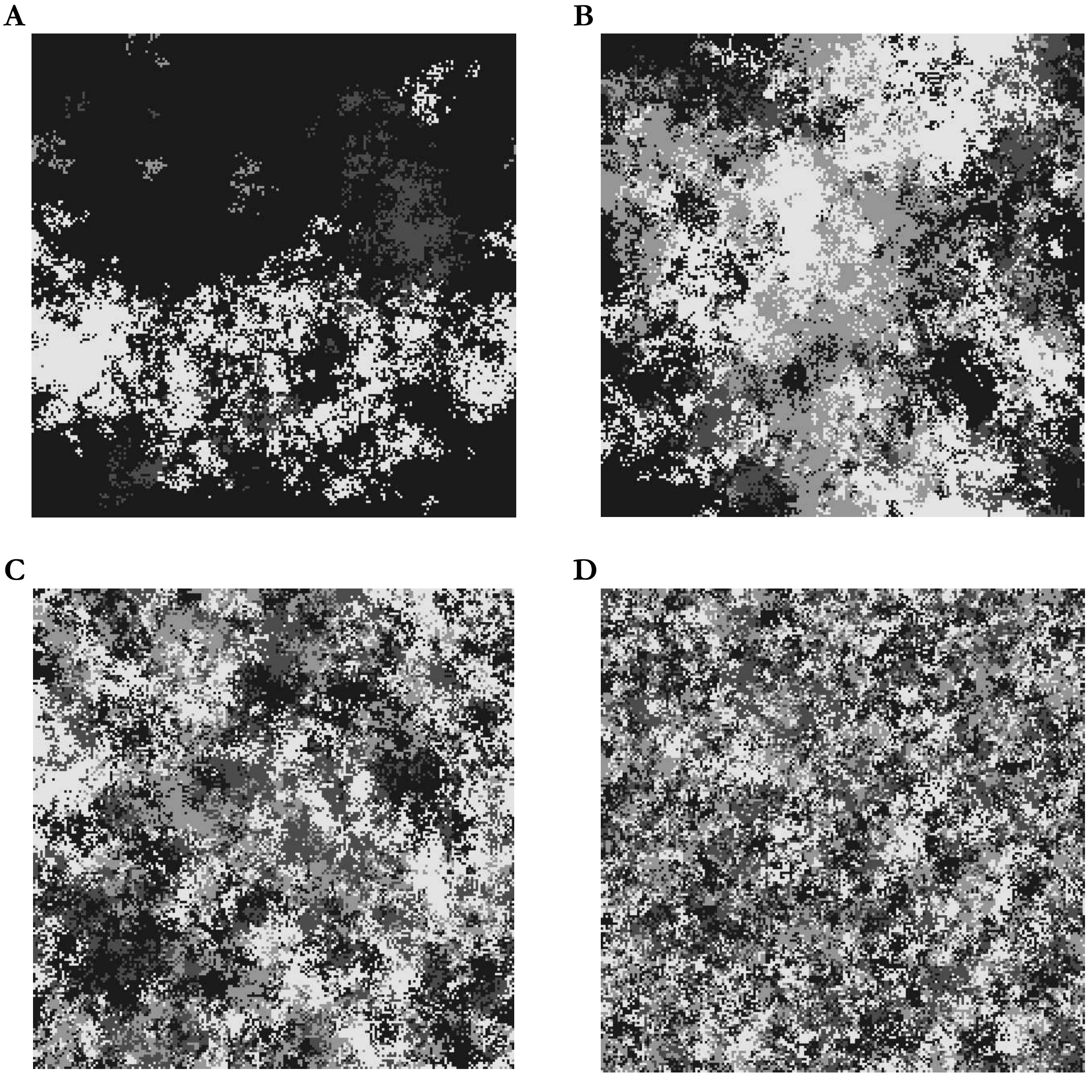


Figure 9.3 The effect of mutation rate on spatial patterns of a single nucleotide site under the neighbourhood model.

Each population has 40,000 individuals on a 200 x 200 lattice. Mutation rates of (A) 1x10^{-5}, (B) 1x10^{-4}, (C) 1x10^{-3} and (D) 1x10^{-2} mutations per site per generation are shown. In all four cases K = 1. Each colour represents one of the four nucleotides A, C, G and T.



9.3.3 The Gaussian neighbourhood

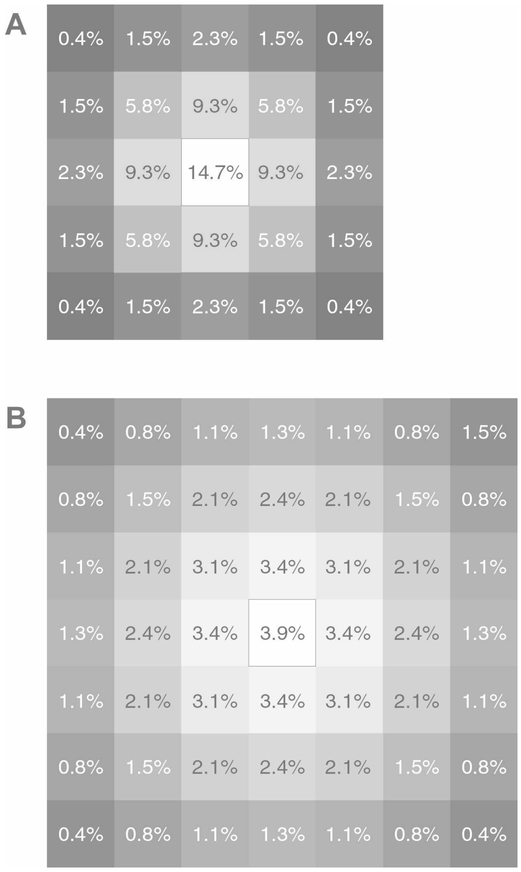
The Gaussian neighbour uses a discretized bivariate normal distribution centred around the focal point on the lattice to choose a parent. As described in equations 9.1 and 9.2 the diffusion coefficient, D, determines the rate of diffusion. To simulate diffusion of an individual on a lattice, a random Gaussian number with a standard deviation of $\sqrt{D}$ is picked for each direction x and y each generation. Figure 9.4 shows the percentage probabilities of picking each neighbouring cell as the parent of the central cell in the next generation for two different values of the diffusion coefficient, D.

9.3.4 Restrictions of "constant-organisation" models

The models described above assume a fixed density of one individual per unit area. An interesting alternative is to allow each cell on the lattice to have a carrying capacity, *C*, so that at most, *C* individuals can occupy a cell at any one time. This model suggests a new method of simulation. Let us now consider time to be measured in discrete calendar units (for example, years) rather than generations. Each year, an organism picks a new position in its neighbourhood to *move* to. Each organism then has a small probability of dying. Each organism then has a (usually slightly larger) probability of giving birth to a single offspring in the same spot. Finally, for every cell that has more individuals in it than the carrying capacity, individuals are randomly removed until there are *C* individuals remaining. These models will not be investigated further here, apart from saying that; if possible, an inference method should be robust to changes in spatial models of this kind.

Figure 9.4 The Gaussian neighbour model on a lattice.

This figure shows the percentage chance that the parent of the central position comes from each neighbouring position, for (A) *D* = 1 and (B) *D* = 4.



A

0.4%	1.5%	2.3%	1.5%	0.4%
1.5%	5.8%	9.3%	5.8%	1.5%
2.3%	9.3%	14.7%	9.3%	2.3%
1.5%	5.8%	9.3%	5.8%	1.5%
0.4%	1.5%	2.3%	1.5%	0.4%

B

0.4%	0.8%	1.1%	1.3%	1.1%	0.8%	1.5%
0.8%	1.5%	2.1%	2.4%	2.1%	1.5%	0.8%
1.1%	2.1%	3.1%	3.4%	3.1%	2.1%	1.1%
1.3%	2.4%	3.4%	3.9%	3.4%	2.4%	1.3%
1.1%	2.1%	3.1%	3.4%	3.1%	2.1%	1.1%
0.8%	1.5%	2.1%	2.4%	2.1%	1.5%	0.8%
0.4%	0.8%	1.1%	1.3%	1.1%	0.8%	0.4%

9.4 Discussion

This chapter describes an alternative statistical inference strategy for spatially resolved populations to models that assume discrete subpopulation structure. In addition, some simple simulations are undertaken to develop an intuition about the properties of such populations. Further simulation work is required to validate the inference method described. In addition the relationship between the models presented in this chapter and the geographically resolved models recently investigated by Epperson (EPPERSON 1999) need to be established.

10 MEPI

Molecular Evolution and Population Inference

10.1 Introduction

This chapter describes the MEPI (Molecular Evolution and Population Inference) software package for evolutionary inference and does not describe in detail the merits and pitfalls of either Bayesian inference or genealogy-based evolutionary inference. For a detailed discussion of technical aspects of this software see Chapter 5 or the paper it is based on (DRUMMOND *et al.* 2002).

MEPI is a software package developed for the Bayesian inference of molecular evolution and population genetics using molecular sequence data. The dynamics of a population over time and the action of mutation and selection at the molecular level leave their traces in the pattern of nucleotide sequences observed in a sample of individuals taken from the population. The combination of these forces leads to a particular shape of the (unknown) phylogenetic tree or genealogy of these samples as well as in the pattern of mutations/substitutions seen in an alignment of sequences. Given an explicit probabilistic model of molecular evolution and population dynamics and a set of (possibly woefully uninformative) prior beliefs about the parameters of interest, Bayesian inference can be used to jointly estimate the combination of model parameters that are most probable given the observed data. MEPI provides population inference based on the coalescent (DRUMMOND *et al.* 2002; HUDSON 1990; KINGMAN 1982a; RODRIGO and FELSENSTEIN 1999) and molecular evolutionary inference based on independent-sites neutrally evolving likelihood models (FELSENSTEIN 1981; HASEGAWA *et al.* 1985; RODRIGUEZ *et al.* 1990). The details of the inference engine are described in Drummond *et al* (2002).

With MEPI, a researcher can obtain estimates and joint probability densities for mutation parameters, population parameters, dates of divergences, and genealogies (or phylogenetic trees). Because of the rich variety of models that can be investigated using MEPI, researchers should focus attention on the parameters of most importance and vary other assumptions to test the sensitivity to prior and model selection.

10.2 Programs

The MEPI software package is made up of four main programs: `mepi`, `mepix`, `tracer` and `treesummary`. All of these programs are written in the Java programming language. To run them, Java Runtime Environment (JRE) version 1.4.0 or greater must be installed on your system. These programs are distributed using the Java

Network Launching Protocol (JNLP). This enables (among other things) automatic updating of MEPI over the web. Currently two programs support JNLP: Java Web Start and OpenJNLP. You must have one of these programs installed to use MEPI. Both are freely available for Microsoft Windows, Mac OS X, Linux and SunOS. Java Web Start comes as part of MacOS X version 10.1 and later versions. If you have Java Web Start installed the mepi programs can be run by simply clicking on a link at http://www.cebl.auckland.ac.nz/mepi/index.html. After the software has been downloaded once it can be used offline indefinitely. Furthermore newer versions of the software are automatically downloaded.

Program	Brief description
`mepi`	Performs an MCMC analysis.
`tracer`	Plots and summarizes the log file that is generated by `mepi`.
`mepix`	Easy-to-use GUI for creating mepix analysis files.
`treesummary`	Displays and summarized the trees file that is generated by `mepi`.

10.2.1 mepi

The program `mepi` takes as input an analysis file that describes an MCMC analysis, including all aspects of raw data, priors, models in use and technical details of MCMC operators. The analysis file is written in an XML language called `mepix`. This language is described in detail in section 0.

The output of the program `mepi` is a log file and a trees file. The log file holds a set of samples of the states that the MCMC chain has visited. The trees file likewise contains a sample of trees that the MCMC chain visited. The log file can be analysed by the program `tracer` and the trees file can be view by TreeView (PAGE 1996) or `treesummary`.

10.2.2 mepix

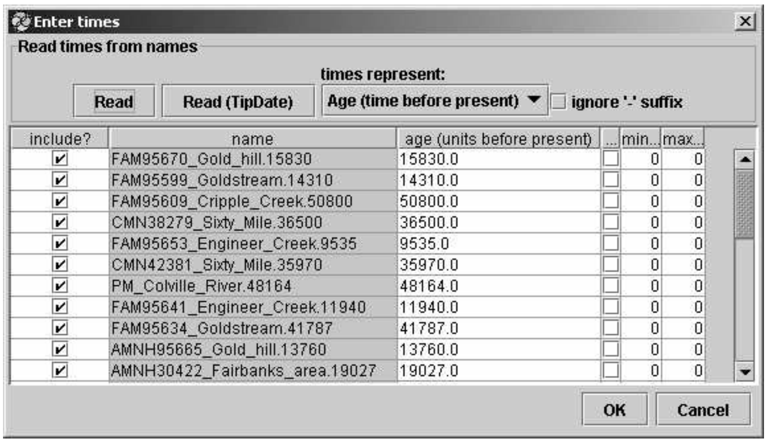
The program `mepix` provides a simple way for users to create a large number of different analysis files for input into `mepi`. It has a graphical user interface (GUI) and can read Phylip format alignments and interleaved Nexus format alignments. Some complex simulations will require modification of the output file by hand before the `mepix` file is ready for input into `mepi`.

10.2.2.1 "Enter times" dialog

This dialog lets you enter the ages of the sequences in units before present (for example, years before present). If the names of the sequences have the age (or time) of the sequence as a suffix then the times can be directly read by pressing one of the two Read buttons at the top of the dialog. The "Read" button assumes that the times are integer and separated from the rest of the name by a '.' as shown in Figure 10.1. The "Read (TipDate)" button assumes the age/time could be a decimal number and uses as much of the end of the name as can validly be interpreted as a number. The option box labelled "times represent:" can be used to specify whether the suffix represents an age (bigger is older) or a time (bigger is younger).

If the names do not have times as a suffix, then the times can be directly entered in to the age column of the table. The "include?" column can be used to exclude some taxa. If your data is contemporaneous you can just click on "OK".

Figure 10.1 "Enter times" dialog.



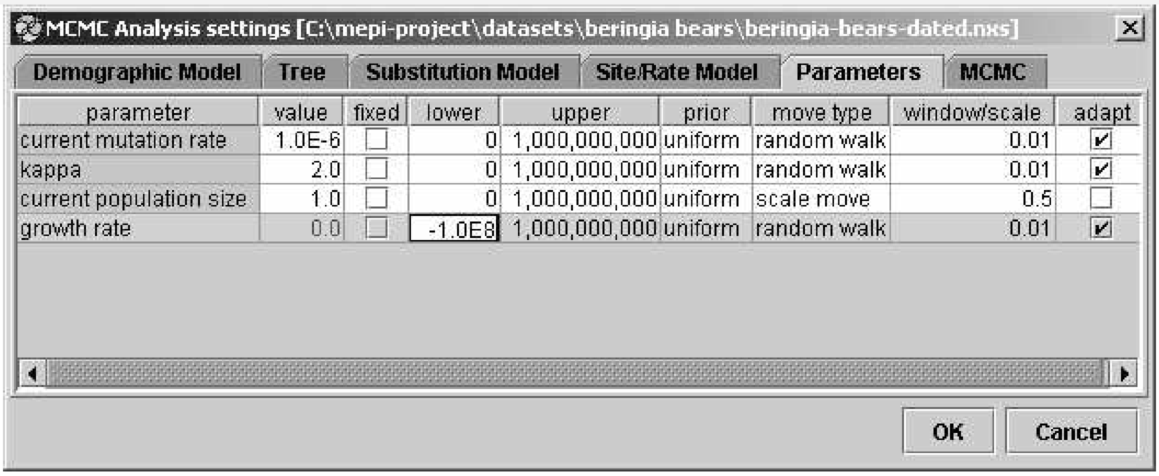
include?	name	age (units before present)	...	min...	max...
✓	FAM95670_Gold_hill.15830	15830.0	□	0	0
✓	FAM95599_Goldstream.14310	14310.0	□	0	0
✓	FAM95609_Cripple_Creek.50800	50800.0	□	0	0
✓	CMN38279_Sixty_Mile.36500	36500.0	□	0	0
✓	FAM95653_Engineer_Creek.9535	9535.0	□	0	0
✓	CMN42381_Sixty_Mile.35970	35970.0	□	0	0
✓	PM_Colville_River.48164	48164.0	□	0	0
✓	FAM95641_Engineer_Creek.11940	11940.0	□	0	0
✓	FAM95634_Goldstream.41787	41787.0	□	0	0
✓	AMNH95665_Gold_hill.13760	13760.0	□	0	0
✓	AMNH30422_Fairbanks_area.19027	19027.0	□	0	0

10.2.2.2 "MCMC Analysis settings" dialog

The settings dialog allows the user to select the evolutionary model to be analysed, the operators used in the MCMC analysis and the length of the MCMC run. It consists of a number of tabs for different aspects of the evolutionary model (for example, demographic model, tree, substitution model, site/rate model), a "Parameters" tab and an "MCMC" tab.

The "Parameters" tab contains a list of all the parameters that can be pontentially estimated (not including tree node heights and topologies) based on the selections in the other panels (see Figure 10.2). In this tab each parameter can be fixed at a user-specified value or an MCMC operator can be tailored for the sampling of it. The "lower" and "upper" columns determine the prior limits on each parameter. The "adapt" column allows the user to specify if the operators should be automatically fine-tuned (this should be used sparingly as it is not always guaranteed to work!). Operator tuning parameters (specified in the “window/scale” column) should be chosen so that the acceptance rate of the parameter is between 10-40%. For current population size a good choice is a “scale move” with a scale parameter of 0.5.

Figure 10.2 “Parameters” tab of the “MCMC Analysis settings” dialog.


MCMC Analysis settings [C:\mepi-project\datasets\beringia bears\beringia-bears-dated.nxs]
Demographic Model
Tree
Substitution Model
Site/Rate Model
Parameters
MCMC
OK
Cancel

parameter	value	fixed	lower	upper	prior	move type	window/scale	adapt
current mutation rate	1.0E-6	□	0	1,000,000,000	uniform	random walk	0.01	☑
kappa	2.0	□	0	1,000,000,000	uniform	random walk	0.01	☑
current population size	1.0	□	0	1,000,000,000	uniform	scale move	0.5	□
growth rate	0.0	□	-1.0E8	1,000,000,000	uniform	random walk	0.01	☑

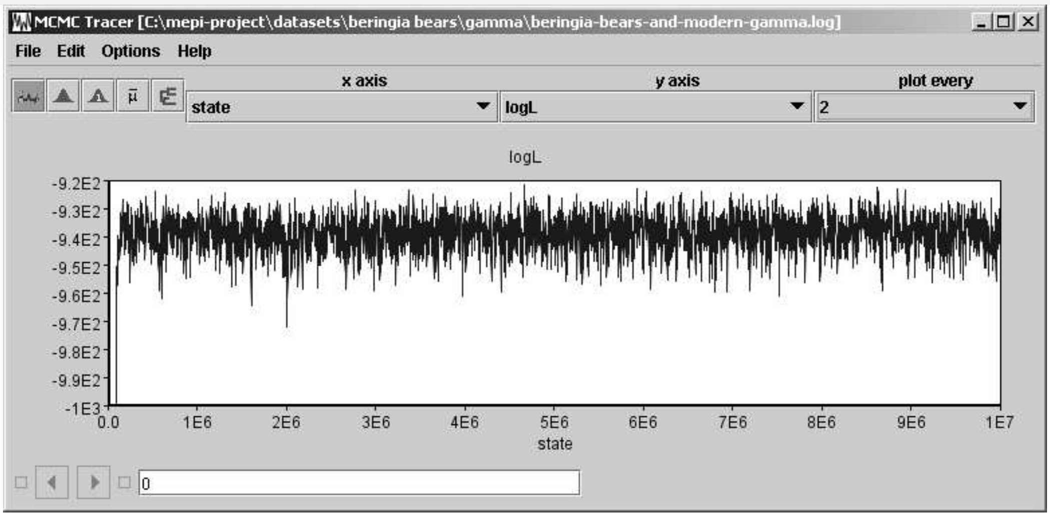
After selecting the appropriate analysis settings and clicking “OK” the user can save the generated mepix file using the “Save As…” menu item in the “File” menu.

10.2.3 tracer

The program `tracer` allows the user to plot the MCMC traces from the log file output of `mepi` and do simple analyses to calculate mean estimates, highest posterior density (HPD) intervals and autocorrelation times (ACT). A trace of the log-likelihood of an analysis of 47 Beringian bear sequences is shown in Figure 10.3.

Figure 10.3 A screenshot of tracer program.

A trace of log-likelihood in a 10,000,000 state Markov chain.



MCMC Tracer [C:\mepi-project\datasets\beringia bears\gamma\beringia-bears-and-modern-gamma.log]

x axis: state

y axis: logL

plot every: 2

10.2.4 treesummary

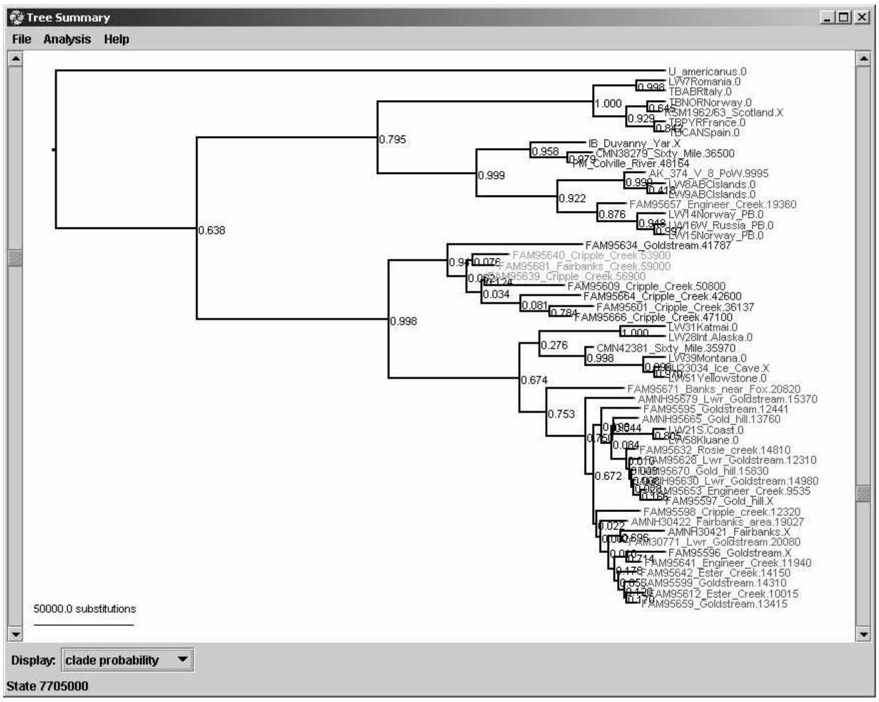
The program `treesummary` allows the user to view trees and do some simple analyses on the tree file output of `mepi`. The MCMC state of the tree is displayed in the status bar at the bottom of the window. The posterior probabilities of the clades present in the current tree can be calculated by selecting "clade probability" from the option box. The left scrollbar adjusts the scale of the trees and the right scrollbar can be used to select the tree to view. This program currently only reads tree files outputted by mepi. Figure 10.4 shows a screen shot of treesummary.

10.2.4.1 Printing trees

There is currently no method for printing the trees in treesummary. TreeView is able to print trees. However a quick way to get an image of your tree under Microsoft Windows is to capture the window to the clipboard by pressing Alt+PrintScreen and then pasting the captured image into a Word document or image processing software.

Figure 10.4 A screenshot of treesummary program.

A sample tree with posterior probabilities of the clades displayed.


Tree Summary
File
Analysis
Help
U_americanus.0
LW7Romania.0
0.998
TBABRItaly.0
1.000
TBNORNorway.0
0.645
KSM1962/63_Scotland.X
0.929
TBPYRFrance.0
0.847
TBCANSpain.0
0.795
IB_Duvanny_Yar.X
0.958
CMN38279_Sixty_Mile.36500
0.979
PM_Colville_River.48164
0.999
AK_374_V_8_PoW.9995
0.999
LW8ABCIslands.0
0.418
LW9ABCIslands.0
0.922
FAM95657_Engineer_Creek.19360
0.876
LW14Norway_PB.0
0.948
LW16W_Russia_PB.0
0.997
LW15Norway_PB.0
0.638
FAM95634_Goldstream.41787
FAM95640_Cripple_Creek.53900
0.076
FAM95681_Fairbanks_Creek.59000
FAM95639_Cripple_Creek.56900
0.124
FAM95609_Cripple_Creek.50800
0.034
FAM95664_Cripple_Creek.42600
0.081
FAM95601_Cripple_Creek.36137
0.784
FAM95666_Cripple_Creek.47100
0.998
LW31Katmai.0
1.000
LW28Int.Alaska.0
0.276
CMN42381_Sixty_Mile.35970
0.998
LW39Montana.0
U23034_Ice_Cave.X
0.970
LW51Yellowstone.0
0.674
FAM95671_Banks_near_Fox.20820
AMNH95679_Lwr_Goldstream.15370
0.753
FAM95595_Goldstream.12441
AMNH95665_Gold_hill.13760
0.044
LW21S.Coast.0
0.805
LW58Kluane.0
0.750
0.084
FAM95632_Rosie_creek.14810
AM95628_Lwr_Goldstream.12310
95670_Gold_hill.15830
0.672
H95630_Lwr_Goldstream.14980
M95653_Engineer_Creek.9535
FAM95597_Gold_hill.X
FAM95598_Cripple_creek.12320
0.022
AMNH30422_Fairbanks_area.19027
AMNH30421_Fairbanks.X
0.696
AM30771_Lwr_Goldstream.20080
FAM95596_Goldstream.X
0.714
FAM95641_Engineer_Creek.11940
0.178
FAM95642_Ester_Creek.14150
AM95599_Goldstream.14310
0.120
AM95612_Ester_Creek.10015
0.170
FAM95659_Goldstream.13415
50000.0 substitutions

Display: clade probability

State 7705000

10.2.4.2 The tree file format

The tree file format is a simple list of rooted newick format trees with branch lengths. Each tree is preceded on the line above it by a comment in square brackets containing the MCMC state of the tree. This format is directly readable by TreeView, although the state information is discarded (see Figure 10.5).

Figure 10.5 An example tree file.

This example file has 3 clock-constrained trees of four taxa A, B, C, and D, representing states 0, 100 and 200 of an MCMC analysis.

```
[0]
((A:0.5,B:0.5):0.75, (C:0.6, D:0.6):0.65);
[100]
((A:0.6,B:0.6):0.65, (C:0.65, D:0.65):0.6);
[200]
(((A:0.5,B:0.5):0.1, C:0.6):0.1, D:0.7);
```

10.3 The mepix file format

The input file for the mepi program is an XML document written in a language called mepix. In this section we will look at an example MCMC analysis written in mepix and look in detail at the various features of the mepix language.

The example file in section 10.3.1 describes an analysis that jointly estimates mutation rate, population size, transition/transversion ratio (kappa) and tree topologies and divergence times in 10 HIV-1 env sequence fragments. The analysis runs for 100,000 cycles and outputs the results to a file called hiv1.log. Sequences 06-10 were sampled 214 day before sequences 01-05. An HKY substitution model is assumed and the equilibrium frequencies are fixed to empirical values calculated from a larger dataset of the same region. By itself this file is quite readable and self-explanatory and a close examination of it is encouraged. However it fails to demonstrate the full range of variations available to users of mepi. To remedy this a more or less exhaustive list of valid elements and attributes that may appear within a mepix input file are given in the following sections.

10.3.1 Example file: hiv1.mepix

```
<?xml version="1.0" ?>
<!-- An example file containing 10 hypothetical HIV-1 env sequences -->
<!-- sequences 01-05 are 214 days older than sequences 06-10 -->
<mcmc chainlength="100000" keepevery="10" outfile="hiv1.log" >
   <data>
      <alignment datatype="nucleotide" datatypeid="0">
         <sequence name="01">AAAGAAGAGGTAGTAATTAGATCTGAAAAT</sequence>
         <sequence name="02">GAAGAAGAGGTAGTAATTAGATCTGAAAAT</sequence>
         <sequence name="03">AAAGAAGAGGTAGTGATTAGATCTGAAAAT</sequence>
         <sequence name="04">AAAGAAGAGGTAGTGATTAGATCTGAAAAT</sequence>
         <sequence name="05">AAAGAAGAGGTAGTAATTAGATCTGAAAAT</sequence>
         <sequence name="06">GAAGAAGAGGTAGTAATTAGATCTGAAGAT</sequence>
         <sequence name="07">AAAGAAGAGGTAGTGATTAGATCTGAAGAT</sequence>
         <sequence name="08">AAAGAAGAGGTAGTAATTAGATCTGAAGAT</sequence>
         <sequence name="09">AAAGAAGAGGTAGTAATTAGATCTGAAGAT</sequence>
         <sequence name="10">GAAGAAGAGGTAATAATTAGATCTGAAGAT</sequence>
      </alignment>
      <timedata units="days" origin="0" direction="backwards">
         <time value="214">01 02 03 04 05</time>
         <time value="0">06 07 08 09 10 </time>
      </timedata>
   </data>
   <evolutionmodel type="coalescent">
      <ratematrix model="HKY" datatype="nucleotide" datatypeid="0">
         <frequencies>0.406 0.152 0.212 0.230</frequencies>
         <parameter name="kappa" value="2.0" />
      </ratematrix>
      <sitemodel type="uniform">
         <mutationratemodel type="constant">
            <parameter name="current mutation rate" value="5.0E-3" />
         </mutationratemodel>
      </sitemodel>
      <demographicmodel type="constant" units="generations">
         <parameter name="current population size" value="1000.0" />
      </demographicmodel>
   </evolutionmodel>
   <operators>
      <operator paramname="current mutation rate" type="random walk" windowsize="1e-5"
            />
      <operator paramname="current population size" type="scale" scalefactor="0.5" />
      <operator paramname="kappa" type="scale" scalefactor="0.5" />
      <operator type="node height" topthreeonly="true" />
      <operator type="node height" />
      <operator type="narrow exchange" />
      <operator type="wide exchange" />
      <operator type="wilson-balding" />
      <operator type="scale tree" scalefactor="0.9" />
   </operators>
   <prior type="coalescent" />
   <prior paramname="current mutation rate" minimum="0.0" maximum="1.0" type="uniform"
         />
   <prior paramname="kappa" minimum="0.0" maximum="1e9" type="Jeffreys’" />
   <prior paramname="current population size" minimum="0" maximum="1e9" type="uniform"
         />
</mcmc>
```

10.3.2 mepix elements

In the following sections a brief description of the elements and attributes that make up a `mepix` file will be given. Each subsection will give details of a single element including valid attributes and subelements. The sections are ordered alphabetically. Section 10.3.8 describes the `mcmc` element, the parent of all other elements.

Attributes that are restricted to a finite set of literal strings have an attribute type of the form (`"literal1"`, `"literal2"`, …, `"literal3"`). The strings within quotes, including spaces, are the only valid attribute values for these *enumerated* attributes. Default values are given for attributes that may be omitted. If ‘`N/a`’ appears in the default value column of an attribute, then the attribute is required.

10.3.3 The alignment element

```
<alignment datatype="nucleotide" datatypeid="0">
   <sequence name="hiv-01">AAAGAAGAGGTAGTAATTAGATCTGAAAAT</sequence>
   <sequence name="hiv-02">GAAGAAGAGGTAGTAATTAGATCTGAAAAT</sequence>
   <sequence name="hiv-03">AAAGAAGAGGTAGTGATTAGATCTGAAAAT</sequence>
   <sequence name="hiv-04">AAAGAAGAGGTAGTGATTAGATCTGAAAAT</sequence>
   <sequence name="hiv-05">AAAGAAGAGGTAGTAATTAGATCTGAAAAT</sequence>
   <sequence name="hiv-06">GAAGAAGAGGTAGTAATTAGATCTGAAGAT</sequence>
   <sequence name="hiv-07">AAAGAAGAGGTAGTGATTAGATCTGAAGAT</sequence>
   <sequence name="hiv-08">AAAGAAGAGGTAGTAATTAGATCTGAAGAT</sequence>
   <sequence name="hiv-09">AAAGAAGAGGTAGTAATTAGATCTGAAGAT</sequence>
   <sequence name="hiv-10">GAAGAAGAGGTAATAATTAGATCTGAAGAT</sequence>
</alignment>
```

An example alignment of 10 nucleotide sequences.

The alignment element contains an alignment of sequences. The sequences can be of nucleotides, amino acids, codons or binary characters. It should be noted that strictly speaking the observation is the set of raw sequences (as the true alignment of the sequences is often not known), but mepi does not yet provide inference of alignments.

Attribute name	Attribute type	Default value	Description
datatype	("nucleotide", "amino acid", "codon", "binary")	"nucleotide"	The type of sequence data contained in this alignment.
datatypeid	Integer	0	An alternative method of specifying the datatype. 0 = nucleotide 1 = amino acids 2 = binary 4 = codons
missing	String	"-"	This string contains the characters that should be interpreted as missing data (i.e. gaps in the alignment and missing data at ends).

Element name	Required?	Multiple allowed?	Description
sequence	Yes	Yes	Contains a sequence string (including gaps) and name.

10.3.4 The data element

```
<data>
   <alignment datatype="nucleotide" datatypeid="0">
      <sequence name="01">AAAGAAGAGGTAGTAATTAGATCTGAAAAT</sequence>
      <sequence name="02">GAAGAAGAGGTAGTAATTAGATCTGAAAAT</sequence>
      <sequence name="03">AAAGAAGAGGTAGTGATTAGATCTGAAAAT</sequence>
      <sequence name="04">AAAGAAGAGGTAGTGATTAGATCTGAAAAT</sequence>
      <sequence name="05">AAAGAAGAGGTAGTAATTAGATCTGAAAAT</sequence>
      <sequence name="06">GAAGAAGAGGTAGTAATTAGATCTGAAGAT</sequence>
      <sequence name="07">AAAGAAGAGGTAGTGATTAGATCTGAAGAT</sequence>
      <sequence name="08">AAAGAAGAGGTAGTAATTAGATCTGAAGAT</sequence>
      <sequence name="09">AAAGAAGAGGTAGTAATTAGATCTGAAGAT</sequence>
      <sequence name="10">GAAGAAGAGGTAATAATTAGATCTGAAGAT</sequence>
   </alignment>
   <timedata units="days" origin="0" direction="backwards">
      <time value="214">01 02 03 04 05</time>
      <time value="0">06 07 08 09 10 </time>
   </timedata>
</data>
```

An example data element containing an alignment of 10 nucleotide sequences. The first 5 sequences are 214 days older than the second 5.

The data element contains all of the observation data on which an analysis is based. These observations currently include sequence alignments and sequence ages. It should be noted that strictly speaking the observation is the raw sequences (as the true alignment of the sequences is often not known), but mepi does not yet provide inference of alignments. Also in the case of ancient DNA the observation data is usually radiocarbon ages rather than calendar ages, but again, mepi does not currently allow inference of true ages from radiocarbon ages.

Element name	Required?	Multiple allowed?	Description
alignment	*	No	Contains information about sequence names and alignment.
timedata	*	No	Contains information about the ages of the sequences (i.e. the time structure). If no time data is given then all sequences are assumed to be contemporaneous (i.e. all times are set to zero).

*At least one of alignment and timedata is required.

10.3.5 The demographicmodel element

```
<demographicmodel type="exponential" units="generations">
   <parameter name="current population size" value="1000.0" />
   <parameter name="growth rate" value="0.1" />
</demographicmodel>
```

An example of an exponential demographic model.

The demographicmodel element describes a model of population (size) dynamics over time. The type and units attributes are both required.

Attribute name	Attribute type	Default value	Description
type	("constant", "exponential", "constexp", "constexpconst")	N/a	constant = constant population size exponential = exponentially growing population size constexp = constant ancestral population size followed by exponential growth. constexpconst = constant ancestral population followed by exponential growth phase, followed by a second constant phase at the current population size.
units	("days", "months", "years", "generations")	"generations"	This attribute specifies the units in which the parameters of this demographic model are specified. It is recommended that this is matched with the units of the timedata element.

Element name	Required?	Multiple allowed?	Description
parameter	No	Yes	Contains details of a parameter of the demographic model. See parameter element for details.

10.3.6 The evolutionmodel element

```
<evolutionmodel type="coalescent">
   <ratematrix model="HKY" datatype="nucleotide" datatypeid="0">
      <frequencies>0.406 0.152 0.212 0.230</frequencies>
      <parameter name="kappa" value="2.0" />
   </ratematrix>
   <sitemodel type="uniform">
      <mutationratemodel type="constant">
         <parameter name="current mutation rate" value="5.0E-3" />
      </mutationratemodel>
   </sitemodel>
   <demographicmodel type="constant" units="generations">
      <parameter name="current population size" value="1000.0" />
   </demographicmodel>
</evolutionmodel>
```

An example of an evolutionary model with a HKY model of substitution, uniform across sites, and a constant population size of 1000.

The evolutionmodel element provides a description of the evolutionary model used to analyze the given data. This element contains information about the mutation rates, the substitution process and the population models used in the analysis.

Attribute name	Attribute type	Default value	Description
initialtree	("supgma", "coalescent")	"coalescent"	The method used to generate the initial tree if one isn't given.

Element name	Required?	Multiple allowed?	Description
demographicmodel	No	No	Contains the demographic model used for coalescent prior. The mcmc element must have a coalescent prior element to make this meaningful!
ratematrix	*	No	Contains the rate matrix model type to be used/estimated.
sitemodel	*	No	Contains the site model used, parameters of this may be estimated if appropriate operators exist.
tree	No	No	Contains the starting tree. This tree may be modified during simulation if operators such as wilson-balding, narrow exchange and wide exchange are specified in operators element. If no tree is specified a tree is generated from the data.

* both are required if data is being analyzed. If the prior is being sampled a ratematrix and a sitemodel are not required.

10.3.7 The frequencies element

```
<frequencies>0.406 0.152 0.212 0.230</frequencies>
```

The frequencies element describes the equilibrium frequencies of a rate matrix. The frequencies are white-space-delimited in the body of the element. DNA frequencies are ordered A, C, G, T.

10.3.8 The mcmc element

```
<mcmc chainlength="100000" keepevery="10" outfile="hiv1.log" >
...
</mcmc>
```

The entire analysis must be embodied with a mcmc element, which is the first thing the mepi program looks for. There should be only one such element in the input file.

Attribute name	Attribute type	Default value	Description
chainlength	Integer	1000	The length of the chain to be run, excluding any adaptive optimization.
keepevery	Integer	1	This number determines how often a state is logged to the outfile. A value of 10 indicates that every tenth state is logged.
outfile	String	"mcmc.log"	The path of the output file.
verbose	Boolean	false	If true then tells all, else remains fairly silent.
repeats	Integer	1	the number of times to repeat the analysis.

Element name	Required?	Multiple allowed?	Description
data	*	No	contains information about sequence alignment and individual ages of sequences.
simulatedata	*	No	describes how to simulate data to be analyzed.
evolutionmodel	Yes	No	describes the mutation model, site model and population model used in the analysis
operators	Yes	No	contains all of the operators used by the mcmc algorithm.
prior	No	Yes	all the prior elements together describe the full prior. Each element typically describes the prior for a single parameter.

* Either data or simulatedata must be present, and only one is allowed.

10.3.9 The mutationratemodel element

```
<mutationratemodel type="constant">
   <parameter name="current mutation rate" value="2.5E-3" />
   <parameter name="ancestral mutation rate" value="5.0E-3" />
   <parameter name="step time" value="100" />
</mutationratemodel>
```

An example of a stepped mutation rate model that assumes an ancestral rate of 5.0e-3 that changed suddenly 100 time units ago to 2.5e-3.

The mutationratemodel element describes a mutation rate model over time.

Currently, there are two models, constant and stepped.

Attribute name	Attribute type	Default value	Description
type	("constant", "stepped")	N/a	constant = constant mutation rate. stepped = a mutation rate model with an instantaneous change (or step) in mutation rate at some time in the past.

Element name	Required?	Multiple allowed?	Description
parameter	No	Yes	Contains details of a parameter of the mutation rate model.

10.3.10 The node element

```
<node height="0.5" name="my archaea">
    <node height="0.4" name="">
        <node height="0.3" name="">
            <node height="0.0" name="Halobacterium"
                    />
            <node height="0.0" name="Haloferax" />
        </node>
        <node height="0.0" name="Thermoplasma" />
    </node>
    <node height="0.0" name="Methanobacterium" />
</node>
```

A fragment of a rooted tree of archaea exhibiting the nested structure of node elements used to define an evolutionary history.

The node element describes a node in a tree. The node can be a leaf (in which case it represents an actual sequence) or it can be an internal node (in which case it represents an ancestral divergence in the tree or genealogy). If it is an internal node, then it will have child nodes nested in it.

Attribute name	**Attribute type**	**Default value**	**Description**
name	String	""	This attribute specifies the name that is given to this node. The default name is the empty string.
height	Double	0.0	This attribute specifies the height of this node. The units of this height are determined by the units attribute in the parent tree element.

Element name	**Required?**	**Multiple allowed?**	**Description**
node	No	Yes	The children of this node, if any exist. In a strictly bifurcating tree, the number of these children nodes will be either 0 or 2.

10.3.11 The operator element

<operator paramname="**current population size**" type="**scale**" scalefactor="**0.5**" />

The operator element describes a single mcmc operator. Operators come in two flavours: simple and special. Simple operators act on a single parameter of interest by name and have two flavours: random walk and scale move. Special operators typically act on a number of parameters simultaneously and have usually been designed specifically to improve the performance of certain kinds of analyses.

Attribute name	Attribute type	Default value	Description
paramname	String	N/a	The name of the parameter to operate on. See standard parameters in section 10.3.13 for an explanation. This attribute is valid if the operator type is `random walk` or `scale`.
type	("random walk", "scale", "narrow exchange", "wide exchange", "wilson-balding", "node height", "scale tree", "stochastic SPR", "stochastic SPR", "local-unrooted", "centered rate scale", "relative site-rate")	N/a	The type of operator. If the type is `random walk` or `scale` then the operator is a simple operator, otherwise it is a special operator.
windowsize	Double	N/a	This attribute must appear if the operator type is `random walk`. It is ignored otherwise.
scalefactor	Double	N/a	This attribute must appear if the operator type is `scale`. It is ignored otherwise.
weight	Integer	1	This is the weighting that this operator gets when the next move is being picked in the mcmc chain. If the operator schedule is sequential, this is the number of times this move is called consecutively, otherwise this is the relative proportion of the total weight that this operator has.

10.3.12 The operators element

```
<operators sequential="false">
  <operator paramname="current mutation rate" type="random walk" windowsize="1e-5"
      />
  <operator paramname="current population size" type="scale" scalefactor="0.5" />
  <operator paramname="kappa" type="scale" scalefactor="0.5" />
  <operator type="node height" topthreeonly="true" />
  <operator type="node height" />
  <operator type="narrow exchange" />
  <operator type="wide exchange" />
  <operator type="wilson-balding" />
  <operator type="scale tree" scalefactor="0.9" />
  <operator type="centered rate scale" />
  <operator type="relative site-rate" windowsize="0.00025" adapt="false" />
</operators>
```

An example of an operators element used to describe the moves the MCMC sampler will use.

The operators element provides a description of all of the mcmc operators that are used to move around in the state space. The operators you specify will determine what parameters are integrated over and what parameters are fixed. For example if an operator acting on the variable kappa is included then kappa will be integrated over (included as part of the inference) rather than being conditioned on. Extreme care should be taken when selecting the operators to be used. Not all combinations of operators describe valid MCMC samplers. If you are unsure then don't guess! Results of an analysis may be meaningless for certain combinations of parameters.

Attribute name	Attribute type	Default value	Description
sequential	Boolean	false	If true then the operators are used sequentially, otherwise an operator is chosen randomly (perhaps with a weighting) in each cycle of the mcmc simulation.

Element name	Required?	Multiple allowed?	Description
operator	Yes	Yes	One of perhaps many operators used by the mcmc algorithm to integrate over the state space of interest.

10.3.13 The parameter element

```
<parameter name="kappa" value="2.0" />
```

The parameter element describes a single parameter of interest (be it an object of inference or conditioned on). The parameter is generally associated to an aspect of an evolutionary model by its nested position in the XML document. Parameters are referred to by name in operators and prior elements.

Attribute name	Attribute type	Default value	Description
name	String	N/a	The name of the parameter.
value	Double	N/a	The value of the parameter.

Parameter name	Model(s)	Natural Range	Description
current population size	demographicmodel	(0, ∞)	The population size at time 0.0. This parameter is required for all demographic models (constant, exponential, constexp and constexpconst).
growth rate	demographicmodel	(-∞, ∞)	The exponential growth rate. A positive value for this parameter means a increasing population size going *forward* in time (decreasing back in time). This parameter is required for the exponential, constexp and constexpconst models.
alpha	demographicmodel	[0, ∞)	The size of the ancestral population (before exponential growth phase) relative to the current population size. A value of 1.0 means the ancestral and current populations are the same size. This parameter is used for the constexp and constexpconst models. **Note**: If the growth rate is negative then alpha must be greater than 1.0, else it must be smaller than 1.0!
ancestral population size	demographicmodel	[0, ∞)	This parameter is the absolute population size of the ancestral population before the exponential growth phase and can be used with constexp and constexpconst models. This is an alternative parameterization to alpha allowing (for example) for one of the population sizes to be fixed and the other to vary. **Note**: If the growth rate is negative then the ancestral population size must be greater than the current population size, else it must be smaller!
tx	demographicmodel	[0, ∞)	This parameter is the duration of the current population size constant phase in constexpconst models.
tmrca	*Special*	(0, ∞)	This parameter can only be used to specify a prior. It is the height of the root of the tree. Setting a prior on this parameter restricts the

			values of the tree height.
`current mutation rate`	mutationratemodel	(0, ∞)	The mutation rate. This parameter is required for both constant and stepped mutation rate models. In a stepped model this is the most recent mutation rate.
`ancestral mutation rate`	mutationratemodel	(0, ∞)	The ancestral mutation rate in a stepped mutation rate model.
`step time`	mutationratemodel	[0, ∞)	The time at which the mutation rate changes from the `ancestral mutation rate` to the `current mutation rate` in a stepped mutation rate model.
`random branch length`	*special*	[0, ∞)	This parameter can be used only to create an operator. An operator that uses paramname="random branch length" will operate on a randomly selected branch length in the tree each time it is called.
`A-C`	ratematrix	[0, ∞)	The rate (ignoring equilibrium frequencies) of A↔C transversions relative to G↔T =1. Used in GTR rate matrix.
`A-G`	ratematrix	[0, ∞)	The rate (ignoring equilibrium frequencies) of A↔G transitions relative to G↔T = 1. Used in GTR rate matrix.
`A-T`	ratematrix	[0, ∞)	The rate (ignoring equilibrium frequencies) of A↔T transversions relative to G↔T = 1. Used in GTR rate matrix.
`C-G`	ratematrix	[0, ∞)	The rate (ignoring equilibrium frequencies) of C↔G transversions relative to G↔T = 1. Used in GTR rate matrix.
`C-T`	ratematrix	[0, ∞)	The rate (ignoring equilibrium frequencies) of C↔T transitions relative to G↔T = 1. Used in GTR rate matrix.
kappa	ratematrix	[0, ∞)	The rate (ignoring equilibrium frequencies) of transitions (A↔G, C↔T) relative to transversions (A↔C, A↔T, C↔G, G↔T). This parameter is used in the HKY rate matrix.

10.3.14 The prior element

```
<prior paramname="current mutation rate" minimum="0.0" maximum="1.0" type="uniform"/>
```

The prior element provides a description of a component of the prior, usually pertaining to a single parameter of interest. The full prior used in the Bayesian inference is the combination of all the prior elements.

Attribute name	**Attribute type**	**Default value**	**Description**
paramname	String	N/a	The name of the parameter to define a prior for. See standard parameters in section 10.3.13 for an explanation.
minimum	Double	0.0	The lowest value that the parameter is allowed to assume.
maximum	Double	∞	The highest value that the parameter is allowed to assume.
type*	("uniform", "Jeffreys'", "coalescent", "exponential tmrca prior")	N/a	Currently only two basic forms of prior distribution are provided: uniform and Jeffreys'. The uniform distribution weights every value (within the valid range) equally. The Jeffreys' disitrbution weights smaller values more highly (i.e. $f(x) \propto \frac{1}{x}$). The coalescent is a special prior that uses Kingman's coalescent to provide a prior distribution on genealogies based on the population model provided. The exponential tmrca prior is a special prior and should be accompanied with an attribute mean="x" where x is the mean of the exponential prior on tmrca.
mean	Double	N/a	The mean of the exponential tmrca prior. Ignored if type is not exponential tmrca prior.

*If coalescent or exponential tmrca prior is specified as the type then paramname, minimum and maximum are not valid and should not be present.

10.3.15 The ratematrix element

```
<ratematrix model="GTR" datatype="nucleotide" datatypeid="0">
   <frequencies>0.406 0.152 0.212 0.230</frequencies>
   <parameter name="A-C" value="1.0" />
   <parameter name="A-G" value="2.0" />
   <parameter name="A-T" value="1.0" />
   <parameter name="C-G" value="1.0" />
   <parameter name="C-T" value="2.0" />
</ratematrix>
```

An example GTR rate matrix that assumes fixed frequencies and starts with rate parameters corresponding with kappa=2.

The ratematrix element describes a rate matrix to be used during an mcmc simulation or alternatively to be used to generate sequence data.

Attribute name	Attribute type	Default value	Description
model	("JC", "F81", "F84", "HKY", "GTR")	N/a	This attribute specifies the rate matrix model being used.
datatype	("nucleotide", "amino acid", "codon", "binary")	"nucleotide"	The type of sequence data contained in this alignment.
datatypeid	Integer	0	An alternative method of specifying the datatype. 0 = nucleotide 1 = amino acids 2 = binary 4 = codons

Element name	Required?	Multiple allowed?	Description
parameter	No	Yes	Contains details of a parameter of the rate matrix model.
frequencies	No	No	Contains the equilibrium frequencies used in this rate matrix model. If this is not specified the frequencies are calculated empirically from the given alignment.

10.3.16 The sequence element

```
<sequence name="01">AAAGAAGAGGTAGTAATTAGATCTGAAAAT</sequence>
```

The sequence element contains a single (named and aligned) sequence fragment. The content of this element is the raw sequence.

Attribute name	Attribute type	Default value	Description
name	String	N/a	The name of sequence contained in this element.

10.3.17 The simulatedata element

```
<simulatedata length="660">
   <ratematrix model="GTR" datatype="nucleotide" datatypeid="0">
      <frequencies>0.406 0.152 0.212 0.230</frequencies>
      <parameter name="A-C" value="1.505" />
      <parameter name="A-G" value="3.418" />
      <parameter name="A-T" value="0.419" />
      <parameter name="C-G" value="0.477" />
      <parameter name="C-T" value="3.136" />
   </ratematrix>
   <mutationratemodel type="constant">
      <parameter name="current mutation rate" value="5.889E-5" />
   </mutationratemodel>
   <simulatetree>
      <demographicmodel type="constant" units="generations">
         <parameter name="current population size" value="5000" />
      </demographicmodel>
   </simulatetree>
   <timedata units="generations" origin="0" direction="backwards">
      <time value="214">01 02 03 04 05</time>
      <time value="0">06 07 08 09 10</time>
   </timedata>
</simulatedata>
```

An example simulatedata element that simulates sequences under a GTR model along a 10 taxa tree with 5 sequences 214 days old and 5 sequences 0 days old. The sequence length is 660 nucleotides. The tree is itself simulated using the coalescent with the assumption of an population size of 5000.

The simulatedata element provides a description of an evolutionary model and sampling strategy that can be used to generate a simulated dataset. This element can replace a data element in order to do parametric simulations. These observations include sequence alignments and sequence ages.

Attribute name	Attribute type	Default value	Description
length	Integer	500	The length of the sequences to be generated.

Element name	Required?	Multiple allowed?	Description
ratematrix	Yes	No	Contains the mutation rate matrix (*Q*) used to evolve sequences on a tree.
timedata	Yes	No	Describes the time structure, labels and number of sequences.
simulatetree	*	No	Describes the population model used to simulate a tree.
tree	*	No	Contains a tree used when simulating the sequence data.

* Either tree or simulatetree must be present, and only one is allowed.

10.3.18 The simulatetree element

```
<simulatetree>
   <demographicmodel type="exponential" units="generations">
      <parameter name="current population size" value="5000" />
      <parameter name="growth rate" value="0.1" />
   </demographicmodel>
</simulatetree>
```

A simulatetree element describing a demographic model that can be used to simulate a tree. The number of tips and their times are determined by the sibling timedata element.

The simulatetree element describes a population model used to simulate a coalescent tree. This element is currently only valid when placed within a simulatedata element. The timedata element within the same simulatedata element is used to decide the number and names of the taxa.

Element name	Required?	Multiple allowed?	Description
demographicmodel	Yes	No	The demographic model used to simulate a coalescent tree.

10.3.19 The sitemodel element

```
<sitemodel type="codon position">
   <!-- first codon position -->
   <mutationratemodel type="constant">
      <parameter name="current mutation rate" value="7.9E-4" />
   </mutationratemodel>
   <!-- second codon position -->
   <mutationratemodel type="constant">
      <parameter name="current mutation rate" value="7.9E-4" />
   </mutationratemodel>
   <!-- third codon position -->
   <mutationratemodel type="constant">
      <parameter name="current mutation rate" value="7.9E-4" />
   </mutationratemodel>
</sitemodel>
```

A codon position site model. All three codon positions start with the same rate in this example. However if the appropriate operators are used ("centered rate scale" and "relative site-rate") then these site specific rates will be estimated. Note that first mutationratemodel will be the first codon position and so on.

The sitemodel element describes a site model of mutation. Currently two models exist: `uniform` and `codon position`. For the codon position model, note that first mutationratemodel will be the first codon position and so on.

Attribute name	Attribute type	Default value	Description
type	("uniform", "codon position")	"codon position"	If uniform, then all sites in the alignment are treated identically. If codon position is selected then each of the three codon positions are given there own mutation rate model.

Element name	Required?	Multiple allowed?	Description
mutationratemodel	Yes	Yes	Contains details of a single mutation rate model. Only one of these elements is expected in a uniform model, and exactly three are expected in a codon position model.

10.3.20 The time element

```
<time value="214">01 02 03 04 05</time>
```

The time element contains a white-space-separated list of sequence names of the given time (age). In the case of ancient DNA the observation data is usually radiocarbon ages rather than true ages, but mepi does not currently allow inference of true ages from radiocarbon ages.

Attribute name	Attribute type	Default value	Description
value	Double	N/a	The time (age) of the sequence names enclosed by the time element.

10.3.21 The timedata element

```
<timedata units="days" origin="0" direction="backwards">
   <time value="214">01 02 03 04 05</time>
   <time value="0">06 07 08 09 10 </time>
</timedata>
```

An example timedata element with two samples of 5 sequences 214 days apart.

The timedata element contains information about a time line and the ages of a set of labels that relate to sequences in the alignment element. In the case of ancient DNA the observation data is usually radiocarbon ages rather than true ages, but mepi does not currently allow inference of true ages from radiocarbon ages.

Attribute name	**Attribute type**	**Default value**	**Description**
units	("days", "months", "years", "generations")	"generations"	This attribute specifies the units in which the time is specified.
origin	Double	0.0	An alternative method of specifying the datatype. 0 = nucleotide 1 = amino acids 2 = binary 4 = codons
direction	("backwards", "forwards")	"-"	This string contains the characters that should be interpreted as missing data (i.e. gaps in the alignment and missing data at ends).

Element name	**Required?**	**Multiple allowed?**	**Description**
time	Yes	Yes	Contains a list of sequence names that share this time/age.

10.3.22 The tree element

```
<tree units="mutations">
   <node height="0.267779" name="">
      <node height="0.182928" name="">
         <node height="0.138605">
            <node name="human"/>
            <node name="chimp"/>
         </node>
            <node name="gorilla"/>
      </node>
      <node name="orangutan"/>
   </node>
</tree>
```

An example tree element of 4 apes.

The tree element describes a tree to be used during an mcmc simulation or alternatively to be used to generate sequence data.

Attribute name	Attribute type	Default value	Description
units	("days", "months", "years", "generations")	"generations"	This attribute specifies the units in which the time is specified.

Element name	Required?	Multiple allowed?	Description
node	Yes	No	The root node of the tree. Only one node may be the direct child of the tree element.

10.4 Conclusion

The software package MEPI is a suite of programs for the Bayesian inference of molecular evolution and population genetics from molecular sequence data. It implements sampling of all of the posterior probability densities described in Chapters 5, 6 and 8. MEPI is developed in the programming language Java and benefits heavily from the open source project: Phylogenetic Analysis Library (PAL).

11 Personal Conclusion

Ironically, the two aspects of this thesis that accounted for the majority of time and effort are barely even spoken of: the programming and the talking. The implementation of the algorithms and the auxiliary software (such as graphical user interfaces and the XML language described in Chapter 10) used to conduct this research program was its most consuming aspect. A close second was plain old-fashioned scientific discourse. Loud discussions with my supervisors about questions like "What exactly is a species?" or "Is the concept even useful?" regularly echoed through the corridors, often to the amusement of nearby colleagues.

The field of biology as a whole, and evolutionary inference in particular, is now at a point of maturity where progress is often intimately linked with CPU cycles and computation. I must confess to wishing, at many stages, I could have conducted more of my research outdoors with pencil and paper. However I feel that evolutionary biology has long since passed the days of Fisher and Wright, when research could be conducted almost without data, and certainly without computers. By this, I do not mean that statistical and mathematical theories are neither interesting nor relevant, but that we must not ignore the masses of data in choosing between alternative hypotheses. We, as a scientific community, are drowning in a sea of data that demands to be listened to. For me this is an exhilarating prospect. The answers to many of the burning questions of evolutionary biology are out there just waiting to be found. A fellow theoretician once said to me that he was a filter feeder; at the bottom of the scientific food chain relying on others' data. I have a feeling that the tides are changing and a Kuhnian revolution is upon us.

11.1 Open problems

In many ways this thesis has served more as a starting point for me rather than any kind of an end. In developing the methods presented herein, some of the open problems facing evolutionary biology as a field have become more apparent. The explicit modelling of selection, both adaptive and purifying, at the molecular level is one such open problem. The challenge is to integrate the knowledge of molecular structure and function with models of mutation. Most current models of molecular evolution lump mutations and selection into a single bucket called "substitutions". I feel that both sides of the mutation-selection balance need more rigorous mechanistic and statistical treatment.

Models of mutation need to integrate information about the DNA polymerase apparatus and its biochemical properties with regard to slippage and (mis-) repair. More important, and more demanding is the development of models of selection that incorporate our knowledge of sequence-structure mapping. In the case of RNA secondary structure, the research presented in Chapter 8 falls well short of this second goal, at best merely highlighting the problems associated with current methods.

Another open problem is the joint estimation of phylogeny and alignment. Within the Bayesian framework described in Chapters 5, 6, 8 and 10, I would boldly say that this task is programmatically straightforward. The need is for more scientific discourse. The difficulty is in developing realistic models of the evolutionary process of insertions and deletions (indels). Only two years ago, upon mentioning at a conference my ambition to solve this problem, I received a chuckle, and was told it was unreachable because of computer limitations. However, a few months ago I again mentioned my aspirations at another meeting and heads nodded in almost bored agreement that it was the way to go. Such is the exciting pace of computational advance.

Rate heterogeneity both among sites and across lineages is currently treated as phenomenological. The statistical models used are justified by their empirical fit, rather than their mechanistic interpretation. Therefore, a third open problem is the integration of knowledge about (i) mutational hotspots and (ii) allometry and metabolic rate information in the analysis of rate heterogeneity, both among sites and across lineages.

11.1.1 The rate invariance problem

A fourth open problem is quite directly related to the research I conducted on *measurably evolving populations*. It relates to the internal and external conflict of molecular data with regard to time scales. The facts clash. One exemplar is "When did humans first get HIV-1?" Another is "When did the most recent common ancestor of modern humans exist?" Both of these questions require some knowledge of the rate of evolution, or else a calibration point. Specifically, recent examples seem to indicate that the rate of evolution is faster on shorter time scales.

11.1.1.1 Fast rates of mtDNA evolution in bears and penguins over short time frames

Data from both bears (see Chapter 6) and Adelie penguins (see Appendix) suggest a fast rate of evolution over short time frames (tens to hundreds of thousand years) in comparison with rates over longer time frames (tens to hundreds of million years). This discrepancy is a factor in the range of 2-7 times. In the case of bears, the comparison of estimated evolutionary rates over different time scales for the same genetic region and the same species still exhibits this discrepancy.

11.1.1.2 Fast rates of HIV-1 evolution within patients

Within an infected patient, HIV-1 regularly evolves at rates of about 1-2% sequence divergence per year (DRUMMOND *et al.* 2001; SHANKARAPPA *et al.* 1999). However a recent high-profile analysis of rates of evolution of HIV-1 over a much longer time frame (since its introduction into humans) estimated an overall rate of about 0.24% per year (KORBER *et al.* 2000). This discrepancy appears to have gone relatively unnoticed. However it represents a second exemplar of what I believe is an open problem in evolutionary biology.

11.1.1.3 The mouse and the elephant

Imagine if you will a mouse tied to an elephant by a fine stretchy lead and collar. The mouse is quite active, and from a position next to the elephant, it can run quickly in any direction for a short while, before the lead eventually drags him back to the elephant. The elephant on the other hand is much more sedate and contemplative, occasionally taking a step, this way or that, and thus moving the centre of the mouse's world. If one can only see mice then they appear to behave in a very odd manner. Over short periods, mice tend to move at a very rapid rate. However over long periods, the distance they travel does not correspond to the fast rate. The long timescale rate of mouse movement is much slower, because the slower movements of the unobserved elephant dominate it.

In this model, we might regard the mouse's movement as mutation and the tether as purifying selection. The elephant might then represent the archetypal sequence and her movement is the slowly changing background environment or changes of sequence context.

11.1.1.4 Weakly coupled models of molecular evolution

The rather fanciful imagery conjured up above leads us to a model of molecular evolution that may partially explain the observation that rates of substitution are faster over short times frames than long time frames. Hierarchical coupling of multiple rates that act on different time scales may be a general pattern generated by a number of different mechanisms, like (i) mutation-selection balance in shifting environments, (ii) context-dependent rate heterogeneity across sites, (iii) covarion models of molecular evolution and (iv) coupling of function of different gene products. It would seem to me, that all of these examples can be regarded as models of (weakly) coupled molecular evolution that have the capacity to present different evolutionary rates over different time scales. These concepts are by no means fully formed scientific theorems; rather they are 'merely' analogies/hypotheses for the beginnings of future directions.

11.2 Future directions

The use of Bayesian inference often leads to a great flexibility in model specification. This is because any model that can be simulated can be sampled using Metropolis-Hastings Markov chain Monte Carlo techniques. As a result the handful of models investigated in this thesis is just a beginning point. A theme in this research has been the incorporation of data other than molecular sequences, such as sampling times (Chapter 5), radiocarbon ages (Chapter 6), geographic position (Chapter 9) and RNA secondary structure (Chapter 8) into phylogenetic and population genetic inference. A partial list of possible directions to take from here would read something like:

1. Development of a likelihood function that specifically incorporates the sequence-structure (genotype-phenotype) mapping.
2. Development of statistical models of the insertion and deletion (indel) process.
3. Development of statistical models that take into account coupling of evolution of different gene fragments.
4. Extension of analysis of *measurably evolving populations* to include recombination.
5. Extension of analysis of *measurably evolving populations* to include standard subpopulation models of migration.
6. Extension of analysis of *measurably evolving populations* to include "relaxed molecular clock" models of evolution.
7. Validation of the diffusion model suggested in Chapter 9.

8. Incorporation of radiocarbon dating error into analyses of ancient DNA.

9. Incorporation of external information such as metabolic rate and allometric information into phylogenetic inference of evolutionary rate heterogeneity.

11.3 Conclusion

During the last three or four years of my academic life I have been thinking about how evolution works. In my mind a simulation of the entire evolutionary process is the ultimate goal. If that is so, then the goal is a long way off. But small steps can certainly be made swiftly. The historical crises of cladistics and phenetics seem remote and forgotten to me – abandoned arguments and no longer necessary. Here I have endeavoured to advocate a different approach, resoundingly statistically explicit. I have argued here for computational, data-rich methods that incorporate all sources of knowledge into a cohesive framework of inference. I am hopeful that in the near future, others and myself will tackle some of the open problems and travel in some of the directions I have outlined above.

12 References

ADCOCK, G. J., E. S. DENNIS, S. EASTEAL, G. A. HUTTLEY, L. S. JERMIIN *et al.*, 2001 Mitochondrial DNA sequences in ancient Australians: Implications for modern human origins. Proc Natl Acad Sci U S A **98:** 537-542.

AISLABIE, J., J. FOGHT and D. J. SAUL, 2000 Aromatic hydrocarbon-degrading bacteria from soil near Scott Base Antarctica. Polar Biology **23:** 183-188.

BABAJIDE, A., I. L. HOFACKER, M. J. SIPPL and P. F. STADLER, 1997 Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. Fold Des **2:** 261-269.

BAHLO, M., and R. C. GRIFFITHS, 2000 Inference from Gene Trees in a Subdivided Population. Theoretical Population Biology **57:** 79-95.

BARNES, I., P. MATHEUS, B. SHAPIRO, D. JENSEN and A. COOPER, 2002 Dynamics of Pleistocene population extinctions in Beringian brown bears. Science **295:** 2267-2270.

BASKARAN, S., P. F. STADLER and P. SCHUSTER, 1996 Approximate scaling properties of RNA free energy landscapes. J Theor Biol **181:** 299-310.

BATEY, R. T., and J. A. DOUDNA, 1998 The parallel universe of RNA folding. Nat Struct Biol **5:** 337-340.

BEERLI, P., and J. FELSENSTEIN, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152:** 763-773.

BEERLI, P., and J. FELSENSTEIN, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci USA **98:** 4563-4568.

BROWN, A. J., 1997 Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc Natl Acad Sci U S A **94:** 1862-1865.

BUONAGURIO, D. A., S. NAKADA, J. D. PARVIN, M. KRYSTAL, P. PALESE *et al.*, 1986 Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. Science **232:** 980-982.

CANNONE, J. J., S. SUBRAMANIAN, M. N. SCHNARE, J. R. COLLETT, L. M. D'SOUZA *et al.*, 2002 The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics **3:** 2.

CAVALLI-SFORZA, L. L., and A. W. F. EDWARDS, 1967 Phylogenetic analysis: models and estimation procedures. Evolution **32:** 550-570.

CHOR, B., M. D. HENDY, B. R. HOLLAND and D. PENNY, 2000 Multiple maxima of likelihood in phylogenetic trees: an analytic approach. Mol Biol Evol **17:** 1529-1541.

CHUN, T. W., L. STUYVER, S. B. MIZELL, L. A. EHLER, J. A. MICAN *et al.*, 1997 Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. Proc Natl Acad Sci U S A **94:** 13193-13197.

CONSTANTINO, R. F., R. A. DESHARNAIS, J. M. CUSHING and B. DENNIS, 1997 Chaotic dynamics in an insect population. Science **275:** 389-391.

DARWIN, C., 1859 *On the Origin of Species by means of natural selection*. John Murray, London.

DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN, 1977 Maximum likelihood from incomplete data via the EM algorithm. J Roy. Statist. Soc. B **39:** 1-38.

DRUMMOND, A., R. FORSBERG and A. G. RODRIGO, 2001 The inference of stepwise changes in substitution rates using serial sequence samples. Mol Biol Evol **18:** 1365-1371.

DRUMMOND, A., and A. G. RODRIGO, 2000 Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Mol Biol Evol **17:** 1807-1815.

DRUMMOND, A., and K. STRIMMER, 2001 PAL: an object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics **17:** 662-663.

DRUMMOND, A. J., G. K. NICHOLLS, A. G. RODRIGO and W. SOLOMON, 2002 Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **In press**.

EDWARDS, A. W. F., 1972 *Likelihood.* Cambridge University Press, Cambridge.

EDWARDS, A. W. F., and L. L. CAVALLI-SFORZA, 1964 Reconstruction of evolutionary trees in *Phenetic and phylogenetic classification*, edited by V. H. HEYWOOD and J. MCNEILL. Systematics Association Publication No. 6, London.

EFRON, B., and R. TIBSHIRANI, 1993 *An introduction to the bootstrap.* Chapman and Hall, London.

ELDREDGE, N., and S. J. GOULD, 1972 Punctuated equilibria: an alternative to phyletic gradualism in *Models in Paleobiology*, edited by T. M. J. SCHOPF. Freeman, San Francisco.

ELDREDGE, N., and S. J. GOULD, 1997 On punctuated equilibria. Science **276:** 338-341.

EPPERSON, B. K., 1999 Gene genealogies in geographically structured populations. Genetics **152:** 797-806.

FARRIS, J. S., 1973 A Probability Model for Inferring Evolutionary Trees. Systematic Zoology **22:** 250-256.

FEARNHEAD, P., and P. DONNELLY, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299-1318.

FELSENSTEIN, J., 1973a Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. Systematic Zoology **22:** 240-249.

FELSENSTEIN, J., 1973b Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet **25:** 471-492.

FELSENSTEIN, J., 1978 Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology **27:** 401-410.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368-376.

FELSENSTEIN, J., 1985 Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:** 783-791.

FELSENSTEIN, J., 1988 Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet **22:** 521-565.

FELSENSTEIN, J., 1992a Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. Genet Res **60:** 209-220.

FELSENSTEIN, J., 1992b Estimating effective population size from samples of sequences: Inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genetical Research Cambridge **59:** 139-147.

FELSENSTEIN, J., and G. A. CHURCHILL, 1996 A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol **13:** 93-104.

FELSENSTEIN, J., M. K. KUHNER, J. YAMATO and P. BEERLI, 1999 Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data., pp. 163-185 in *Statistics in Molecular Biology and Genetics*, edited by F. SEILLIER-MOISEIWITSCH. Institute of Mathematical Statistics and American Mathematical Society, Hayward, CA.

FENG, D. F., and R. F. DOOLITTLE, 1987 Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol **25:** 351-360.

FIELDS, D. S., and R. R. GUTELL, 1996 An analysis of large rRNA sequences folded by a thermodynamic method. Fold Des **1:** 419-430.

FINZI, D., M. HERMANKOVA, T. PIERSON, L. M. CARRUTH, C. BUCK *et al.*, 1997 Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. Science **278:** 1295-1300.

FISHER, R., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinburgh **52:** 399-433.

FISHER, R. A., 1921 On the 'Probable Error' of a coefficient of correlation deduced from a small sample. Metron **1:** 3-32.

FISHER, R. A., 1922a On the dominance ratio. Proc. Roy. Soc. Edinburgh **42:** 321-341.

FISHER, R. A., 1922b On the mathematical foundations of theoretical statistics. Philos Trans R Soc A **222:** 309-368.

FISHER, R. A., 1925 *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh.

FISHER, R. A., 1930 *The Genetical Theory of Natural Selection.* Oxford University Press, Oxford.

FITCH, W. M., R. M. BUSH, C. A. BENDER and N. J. COX, 1997 Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci U S A **94:** 7712-7718.

FITCH, W. M., J. M. LEITER, X. Q. LI and P. PALESE, 1991 Positive Darwinian evolution in human influenza A viruses. Proc Natl Acad Sci U S A **88:** 4270-4274.

FITCH, W. M., and E. MARGOLIASH, 1967 Construction of phylogenetic trees. Science **155:** 279-284.

FITCH, W. S., and T. F. SMITH, 1983 Optimal sequence alignments. Proc Natl Acad Sci USA **80:** 1382-1386.

FLAMM, C., W. FONTANA, I. L. HOFACKER and P. SCHUSTER, 2000 RNA folding at elementary step resolution. Rna **6:** 325-338.

FONTANA, W., D. A. KONINGS, P. F. STADLER and P. SCHUSTER, 1993 Statistics of RNA secondary structures. Biopolymers **33:** 1389-1404.

FONTANA, W., and P. SCHUSTER, 1998 Continuity in evolution: on the nature of transitions. Science **280:** 1451-1455.

FRANK, C., M. K. MOHAMED, G. T. STRICKLAND, D. LAVANCHY, R. R. ARTHUR *et al.*, 2000 The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. Lancet **355:** 887-891.

FU, Y. X., 1994 A phylogenetic estimator of effective population size or mutation rate. Genetics **136:** 685-692.

FU, Y. X., 2001 Estimating mutation rate and generation time from longitudinal samples of DNA sequences. Mol Biol Evol **18:** 620-626.

GASCUEL, O., 1997 BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol **14:** 685-695.

GASPIN, C., and E. WESTHOF, 1995 An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints. J Mol Biol **254:** 163-174.

GAUTHERET, D., S. H. DAMBERGER and R. R. GUTELL, 1995 Identification of base-triples in RNA using comparative sequence analysis. J Mol Biol **248:** 27-43.

GEYER, C. J., 1992 Practical Markov chain Monte Carlo. Statist. Sci. **7:** 473-511.

GILLESPIE, J. H., 1989 Could natural selection account for molecular evolution and polymorphism? Genome **31:** 311-315.

GILLESPIE, J. H., 1995 On Ohta's Hypothesis: Most Amino Acid Subsitutions Are Deleterious. J. Mol. Evol. **40:** 64-69.

GILLOOLY, J. F., J. H. BROWN, G. B. WEST, V. M. SAVAGE and E. L. CHARNOV, 2001 Effects of size and temperature on metabolic rate. Science **293:** 2248-2251.

GOJOBORI, T., E. N. MORIYAMA and M. KIMURA, 1990 Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad Sci U S A **87:** 10015-10018.

GOLDMAN, N., 1990 Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analysis. Syst. Zool. **39:** 345-361.

GOLDMAN, N., 1993 Statistical tests of models of DNA substitution. J Mol Evol **36:** 182-198.

GOLDMAN, N., J. P. ANDERSON and A. G. RODRIGO, 2000 Likelihood-based tests of topologies in phylogenetics. Syst. Biol. **49**.

GOLDMAN, N., J. L. THORNE and D. T. JONES, 1998 Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149:** 445-458.

GOLDMAN, N., and Z. YANG, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11:** 725-736.

GOULD, S. J., and N. ELDREDGE, 1993 Punctuated equilibrium comes of age. Nature **366:** 223-227.

GREEN, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711-732.

GRIFFITHS, R. C., 1989 Genealogical-tree probabilities in the infinitely-many-site model. J Math Biol **27:** 667-680.

GRIFFITHS, R. C., and P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. J Comput Biol **3:** 479-502.

GRIFFITHS, R. C., and S. TAVARE, 1994 Ancestral inference in population genetics. Statistical Science **9:** 307-319.

GULTYAEV, A. P., F. H. VAN BATENBURG and C. W. PLEIJ, 1995 The computer simulation of RNA folding pathways using a genetic algorithm. J Mol Biol **250:** 37-51.

GUTELL, R. R., 1994 Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994. Nucleic Acids Res **22:** 3502-3507.

HABIB, M., M. K. MOHAMED, F. ABDEL-AZIZ, L. S. MAGDER, M. ABDEL-HAMID *et al.*, 2001 Hepatitis C virus infection in a community in the Nile Delta: risk factors for seropositivity. Hepatology **33:** 248-253.

HANNI, C., V. LAUDET, D. STEHELIN and P. TABERLET, 1994 Tracking the origins of the cave bear (Ursus spelaeus) by mitochondrial DNA sequencing. Proc Natl Acad Sci U S A **91:** 12336-12340.

HASEGAWA, M., H. KISHINO and T. YANO, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol **22:** 160-174.

HASSELL, M. P., H. N. COMINS and R. M. MAY, 1994 Species coexistence and self-organizing spatial dynamics. Nature, London **370:** 290-292.

HASTINGS, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97-109.

HAYASHIDA, H., H. TOH, R. KIKUNO and T. MIYATA, 1985 Evolution of influenza virus genes. Mol Biol Evol **2:** 289-303.

HEIN, J., 2001 An algorithm for statistical alignment of sequences related by a binary tree. Pac Symp Biocomput**:** 179-190.

HENDY, M. D., and D. PENNY, 1982 Branch and bound algorithms to determine minimal evolutionary trees. Mathematical Biosciences **59:** 277-290.

HEYER, E., E. ZIETKIEWICZ, A. ROCHOWSKI, V. YOTOVA, J. PUYMIRAT *et al.*, 2001 Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. Am J Hum Genet **69:** 1113-1126.

HILLIS, D. M., J. P. HUELSENBECK and C. W. CUNNINGHAM, 1994 Application and accuracy of molecular phylogenies. Science **264:** 671-677.

HOLMES, E. C., L. Q. ZHANG, P. SIMMONDS, C. A. LUDLAM and A. J. LEIGH BROWN, 1992 Convergent and divergent sequence evolution in the surface envelope glycoprotein of HIV-1 within a single infected patient. Proc. Natl. Acad. Sci. USA **89:** 4835-4839.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology **7:** 1-14.

HUELSENBECK, J. P., B. LARGET and D. SWOFFORD, 2000 A compound poisson process for relaxing the molecular clock. Genetics **154:** 1879-1892.

HUELSENBECK, J. P., F. RONQUIST, R. NIELSEN and J. P. BOLLBACK, 2001 Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294:** 2310-2314.

HUYNEN, M., R. GUTELL and D. KONINGS, 1997 Assessing the reliability of RNA folding using statistical mechanics. J Mol Biol **267:** 1104-1112.

HUYNEN, M. A., 1996 Exploring phenotype space through neutral evolution. J Mol Evol **43:** 165-169.

HUYNEN, M. A., and P. HOGEWEG, 1994 Pattern generation in molecular evolution: exploitation of the variation in RNA landscapes. J Mol Evol **39:** 71-79.

HUYNEN, M. A., P. F. STADLER and W. FONTANA, 1996 Smoothness within ruggedness: the role of neutrality in adaptation. Proc Natl Acad Sci U S A **93:** 397-401.

JENKINS, G. M., A. RAMBAUT, O. G. PYBUS and E. C. HOLMES, 2002 Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol **54:** 156-165.

JONES, D. T., W. R. TAYLOR and J. M. THORNTON, 1994 A mutation data matrix for transmembrane proteins. FEBS Lett **339:** 269-275.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules., pp. 21-132 in *Mammalian Protein Metabolism*, edited by H. N. MUNRO. Academic Press.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624-626.

KIMURA, M., 1983 *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge.

KIMURA, M., and T. OHTA, 1971 Protein polymorphism as a phase of molecular evolution. Nature **229:** 467-469.

KIMURA, M., and T. OHTA, 1974 On some principles governing molecular evolution. Proc Natl Acad Sci U S A **71:** 2848-2852.

KINGMAN, J. F. C., 1982a The coalescent. Stochastic Processes and Their Applications **13:** 235-248.

KINGMAN, J. F. C., 1982b On the genealogy of large populations. J Appl. Probability **19A:** 27-43.

KISHINO, H., and M. HASEGAWA, 1989 Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol **29:** 170-179.

KONINGS, D. A., and R. R. GUTELL, 1995 A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. Rna **1:** 559-574.

KORBER, B., M. MULDOON, J. THEILER, F. GAO, R. GUPTA *et al.*, 2000 Timing the ancestor of the HIV-1 pandemic strains. Science **288:** 1789-1796.

KRYSTAL, M., D. BUONAGURIO, J. F. YOUNG and P. PALESE, 1983 Sequential mutations in the NS genes of influenza virus field strains. J Virol **45:** 547-554.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421-1430.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149:** 429-434.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393-1401.

KUMAR, S., and S. SUBRAMANIAN, 2002 Mutation rates in mammalian genomes. Proc Natl Acad Sci U S A **99:** 803-808.

LAMBERT, D. M., P. A. RITCHIE, C. D. MILLAR, B. HOLLAND, A. J. DRUMMOND *et al.*, 2002 Rates of evolution in ancient DNA from Adelie penguins. Science **295:** 2270-2273.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

LARGET, B., and D. SIMON, 1999 Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution **16:** 750-759.

LEITNER, T., and J. ALBERT, 1999 The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc Natl Acad Sci U S A **96:** 10752-10757.

LEONARD, J. A., R. K. WAYNE and A. COOPER, 2000 From the cover: population genetics of ice age brown bears. Proc Natl Acad Sci U S A **97:** 1651-1654.

LEVITT, M., and M. GERSTEIN, 1998 A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci U S A **95:** 5913-5920.

LEWIS, P. O., 2001 Phylogenetic systematics turns over a new leaf. Trends in Ecology and Evolution **16:** 30-37.

LEWONTIN, R. C., and J. L. HUBBY, 1966 A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics **54:** 595-609.

LI, W. H., M. TANIMURA and P. M. SHARP, 1988 Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol **5:** 313-330.

LIPMAN, D. J., S. F. ALTSCHUL and J. D. KECECIOGLU, 1989 A tool for multiple sequence alignment. Proc Natl Acad Sci U S A **86:** 4412-4415.

LOREILLE, O., L. ORLANDO, M. PATOU-MATHIS, M. PHILIPPE, P. TABERLET *et al.*, 2001 Ancient DNA analysis reveals divergence of the cave bear, Ursus spelaeus, and brown bear, Ursus arctos, lineages. Curr Biol **11:** 200-203.

LOTKA, A. J., 1925 *Elements of physical biology*. Williams & Wilkins Co., Baltimore.

LÜCK, R., G. STEGER and D. RIESNER, 1996 Thermodynamic prediction of conserved secondary structure: Application to RRE-element of HIV, tRNA-like element of CMV and mRNA of prion protein. J Mol Biol **258:** 813-826.

LUKASHOV, V. V., C. L. KUIKEN and J. GOUDSMIT, 1995 Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. J Virol **69:** 6911-6916.

MAIDAK, B. L., J. R. COLE, T. G. LILBURN, C. T. PARKER, JR., P. R. SAXMAN *et al.*, 2001 The RDP-II (Ribosomal Database Project). Nucleic Acids Res **29:** 173-174.

MALECOT, G., 1948 *Le Mathematique de l'heredite*. Masson & Cie., Paris.

MARTINEZ, C., L. DEL RIO, A. PORTELA, E. DOMINGO and J. ORTIN, 1983 Evolution of the influenza virus neuraminidase gene during drift of the N2 subtype. Virology **130:** 539-545.

MATHEWS, D. H., J. SABINA, M. ZUKER and D. H. TURNER, 1999 Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol **288:** 911-940.

MAU, B., M. A. NEWTON and B. LARGET, 1999 Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55:** 1-12.

MAY, R. M., 1976 Simple mathematical models with very complicated dynamics. Nature **261:** 459-467.

MCCASKILL, J. S., 1990 The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29:** 1105-1119.

METROPOLIS, N., A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER and E. TELLER, 1953 Equations of state calculations by fast computing machines. J Chem Phys **21:** 1087-1091.

MORAN, P. A. P., 1958 The effect of selection in a haploid genetic population. Proc. Camb. Phil. Soc. **54:** 463-467.

MUSE, S. V., 1995 Evolutionary analyses of DNA sequences subject to constraints of secondary structure. Genetics **139:** 1429-1439.

NAFEH, M. A., A. MEDHAT, M. SHEHATA, N. N. MIKHAIL, Y. SWIFEE *et al.*, 2000 Hepatitis C in a community in Upper Egypt: I. Cross-sectional survey. Am J Trop Med Hyg **63:** 236-241.

NEE, S., E. C. HOLMES, A. RAMBAUT and P. H. HARVEY, 1995 Inferring population history from molecular phylogenies. Philos Trans R Soc Lond B Biol Sci **349:** 25-31.

NIELSEN, R., and Z. YANG, 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:** 929-936.

O'BRIEN, E. A., C. NOTREDAME and D. G. HIGGINS, 1998 Optimization of ribosomal RNA profile alignments. Bioinformatics **14:** 332-341.

OHTA, T., and M. KIMURA, 1971 On the constancy of the evolutionary rate of cistrons. J Mol Evol **1:** 18-25.

OTA, R., P. J. WADDELL, M. HASEGAWA, H. SHIMODAIRA and H. KISHINO, 2000 Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol Biol Evol **17:** 798-803.

PAGE, R. D., 1996 TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci **12:** 357-358.

PAGEL, M., 1999 Inferring the historical patterns of biological evolution. Nature **401:** 877-884.

PEARSON, W. R., G. ROBINS and T. ZHANG, 1999 Generalized neighbor-joining: more reliable phylogenetic tree reconstruction. Mol Biol Evol **16:** 806-816.

PENNY, D., and M. D. HENDY, 1985 Testing methods of evolutionary tree construction. Cladistics **1:** 266-278.

PERELSON, A. S., A. U. NEUMANN, M. MARKOWITZ, J. M. LEONARD and D. D. HO, 1996 HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science **271:** 1582-1586.

PYBUS, O. G., M. A. CHARLESTON, S. GUPTA, A. RAMBAUT, E. C. HOLMES *et al.*, 2001 The epidemic behavior of the hepatitis C virus. Science **292:** 2323-2325.

PYBUS, O. G., A. RAMBAUT and P. H. HARVEY, 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155:** 1429-1437.

RAMBAUT, A., 2000 Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics **16:** 395-399.

RAY, S. C., R. R. ARTHUR, A. CARELLA, J. BUKH and D. L. THOMAS, 2000 Genetic epidemiology of hepatitis C virus throughout egypt. J Infect Dis **182:** 698-707.

REIDYS, C., P. F. STADLER and P. SCHUSTER, 1997 Generic properties of combinatory maps: neutral networks of RNA secondary structures. Bull Math Biol **59:** 339-397.

ROBINSON, D. F., and L. R. FOULDS, 1981 Comparison of phylogenetic trees. Mathematical Biosciences **53:** 131-148.

RODRIGO, A. G., and J. FELSENSTEIN, 1999 Coalescent approaches to HIV population genetics in *Molecular evolution of HIV*, edited by K. CRANDALL. Johns Hopkins University Press, Baltimore, MD.

RODRIGO, A. G., E. G. SHPAER, E. L. DELWART, A. K. IVERSEN, M. V. GALLO *et al.*, 1999 Coalescent estimates of HIV-1 generation time in vivo. Proceedings of the National Academy of Sciences of USA **96:** 2187-2191.

RODRIGUEZ, F., J. L. OLIVER, A. MARIN and J. R. MEDINA, 1990 The general stochastic model of nucleotide substitution. J Theor Biol **142:** 485-501.

ROGERS, J. S., and D. L. SWOFFORD, 1999 Multiple local maxima for likelihoods of phylogenetic trees: a simulation study. Mol Biol Evol **16:** 1079-1085.

ROHLF, F., and G. SCHNELL, 1971 An investigation of the isolation by distance model. American Naturalist **105:** 295-324.

ROOS, D. S., 2001 Computational biology. Bioinformatics--trying to swim in a sea of data. Science **291:** 1260-1261.

ROSENBAUM, V., T. KLAHN, U. LUNDBERG, E. HOLMGREN, A. VON GABAIN *et al.*, 1993 Co-existing structures of an mRNA stability determinant. The 5' region of the Escherichia coli and Serratia marcescens ompA mRNA. J Mol Biol **229:** 656-670.

RZHETSKY, A., 1995 Estimating substitution rates in ribosomal RNA genes. Genetics **141:** 771-783.

SAITOU, N., and M. NEI, 1986 Polymorphism and evolution of influenza A virus genes. Mol Biol Evol **3:** 57-74.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4:** 406-425.

SAVILL, N. J., D. C. HOYLE and P. G. HIGGS, 2001 RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics **157:** 399-411.

SCHNARE, M. N., S. H. DAMBERGER, M. W. GRAY and R. R. GUTELL, 1996 Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23 S-like) ribosomal RNA. J Mol Biol **256:** 701-719.

SCHUSTER, P., W. FONTANA, P. F. STADLER and I. L. HOFACKER, 1994 From sequences to shapes and back: a case study in RNA secondary structures. Proc R Soc Lond B Biol Sci **255:** 279-284.

SCHUSTER, P., and P. F. STADLER, 1998 Sequence Redundancy in Biopolymers: A Study on RNA and Protein Structures, pp. 163-186 in *Viral Regulatory Structures*, edited by G. MYERS. Addison-Wesley, Reading, MA.

SCHUSTER, P., P. F. STADLER and A. RENNER, 1997 RNA structures and folding: from conventional to new issues in structure predictions. Curr Opin Struct Biol **7:** 229-235.

SEO, T. K., J. L. THORNE, M. HASEGAWA and H. KISHINO, 2002a Estimation of Effective Population Size of HIV-1 Within a Host. A pseudomaximum-likelihood approach. Genetics **160:** 1283-1293.

SEO, T. K., J. L. THORNE, M. HASEGAWA and H. KISHINO, 2002b A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times. Bioinformatics **18:** 115-123.

SERRA, M. J., T. W. BARNES, K. BETSCHART, M. J. GUTIERREZ, K. J. SPROUSE *et al.*, 1997 Improved parameters for the prediction of RNA hairpin stability. Biochemistry **36:** 4844-4851.

SHANKARAPPA, R., J. B. MARGOLICK, S. J. GANGE, A. G. RODRIGO, D. UPCHURCH *et al.*, 1999 Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection. Journal of Virology **73:** 10489-10502.

SHIELDS, G. F., and A. C. WILSON, 1987 Calibration of mitochondrial DNA evolution in geese. J Mol Evol **24:** 212-217.

SHIMODAIRA, H., and M. HASEGAWA, 1999 Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol **16:** 1114-1116.

SHIMODAIRA, H., and M. HASEGAWA, 2001 CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics **17:** 1246-1247.

SINGH, S. B., and P. A. KOLLMAN, 1996 Understanding the thermodynamic stability of an RNA hairpin and its mutant. Biophys J **70:** 1940-1948.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555-562.

SNEATH, P. H. A., and R. R. SOKAL, 1973 *Numerical Taxonomy.* W.H. Freeman and Co., San Francisco, CA.

SOKAL, A., 1989 Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms. Cours de Troisieme Cycle de la Physique en Suisse Romande.

SOKAL, R. R., 1961 Distance as a Measure of Taxonomic Similarity. Systematic Zoology **10:** 70-79.

SOKAL, R. R., and C. D. MICHENER, 1958 A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. **38:** 1409-1438.

STEEL, M., 1994 The maximum likelihood point for a phylogenetic tree is not unique. Syst. Biol. **43:** 560-564.

STEEL, M., and J. HEIN, 2000 A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by a star tree. Letters in Applied Mathematics **in press**.

STEPHENS, M., and P. DONNELLY, 2000 Inference in Molecular Population Genetics. Journal of the Royal Statistical Society B **62:** 605-655.

STRIMMER, K., and A. RAMBAUT, 2002 Inferring confidence sets of possibly misspecified gene trees. Proc R Soc Lond B Biol Sci **269:** 137-142.

STUIVER, M., P. J. REIMER, E. BARD, J. W. BECK, G. S. BURR *et al.*, 1998 INTCAL 98 radiocarbon age calibration, 24,000-0 cal BP. Radiocarbon **40:** 1041-1083.

SWOFFORD, D. L., 1999 PAUP*. Phylogenetic analysis using parsimony (* and other methods). pp. Sinauer, Sunderland, Mass.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437-460.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN and D. G. HIGGINS, 1997 The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res **25:** 4876-4882.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673-4680.

THORNE, J. L., N. GOLDMAN and D. T. JONES, 1996 Combining protein evolution and secondary structure. Mol Biol Evol **13:** 666-673.

THORNE, J. L., H. KISHINO and J. FELSENSTEIN, 1991 An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol **33:** 114-124.

THORNE, J. L., H. KISHINO and I. S. PAINTER, 1998 Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. **15:** 1647-1657.

TILLIER, E. R., and R. A. COLLINS, 1995 Neighbour joining and maximum-likelihood with RNA sequences - addressing the interdependence of sites. Mol Biol Evol **12:** 1-15.

TILLIER, E. R., and R. A. COLLINS, 1998 High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics **148:** 1993-2002.

VAN BATENBURG, F. H., A. P. GULTYAEV and C. W. PLEIJ, 1995 An APL-programmed genetic algorithm for the prediction of RNA secondary structure. J Theor Biol **174:** 269-280.

VENTER, J. C., M. D. ADAMS, E. W. MYERS, P. W. LI, R. J. MURAL *et al.*, 2001 The sequence of the human genome. Science **291:** 1304-1351.

VOLTERRA, V., 1926 *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi.*

WAITS, L. P., S. L. TALBOT, R. H. WARD and G. F. SHIELDS, 1998 Phylogeography of the North American brown bear and implications for conservation. Conservation Biology **12:** 109-117.

WALTER, A. E., D. H. TURNER, J. KIM, M. H. LYTTLE, P. MULLER *et al.*, 1994 Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proc Natl Acad Sci U S A **91:** 9218-9222.

WATSON, J. D., and F. H. CRICK, 1953 A Structure for Deoxyribose Nucleic Acid. Nature **171:** 737-738.

WHELAN, S., and N. GOLDMAN, 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol **18:** 691-699.

WILLS, P. R., 1992 Potential pseudoknots in the PrP-encoding mRNA. J Theor Biol **159:** 523-527.

WILSON, I. J., and D. J. BALDING, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499-510.

WOLINSKY, S. M., B. T. M. KORBER, A. U. NEUMANN, M. DANIELS, K. J. KUNTSMAN *et al.*, 1996 Adaptive evolution of HIV-1 during the natural course of infection. Science **272:** 537-542.

WONG, J. K., M. HEZAREH, H. F. GUNTHARD, D. V. HAVLIR, C. C. IGNACIO *et al.*, 1997 Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. Science **278:** 1291-1295.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97-159.

WUCHTY, S., W. FONTANA, I. L. HOFACKER and P. SCHUSTER, 1999 Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers **49:** 145-165.

XIA, T., J. A. MCDOWELL and D. H. TURNER, 1997 Thermodynamics of nonsymmetric tandem mismatches adjacent to G.C base pairs in RNA. Biochemistry **36:** 12486-12497.

YANG, Z., 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol **10:** 1396-1401.

YANG, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39:** 306-314.

YANG, Z., 1995 A space-time process model for the evolution of DNA sequences. Genetics **139:** 993-1005.

YANG, Z., 1998 Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol **15:** 568-573.

YANG, Z., R. NIELSEN, N. GOLDMAN and A. M. PEDERSEN, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431-449.

YANG, Z., and B. RANNALA, 1997 Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol **14:** 717-724.

YULE, G. U., 1924 A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philos Trans R Soc Lond B **213:** 21-87.

ZUKER, M., 1989 On finding all suboptimal foldings of an RNA molecule. Science **244:** 48-52.

ZUKER, M., and A. B. JACOBSON, 1995 "Well-determined" regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA. Nucleic Acids Res **23:** 2791-2798.

ZUKER, M., J. A. JAEGER and D. H. TURNER, 1991 A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. Nucleic Acids Res **19:** 2707-2714.

ZUKER, M., and P. STIEGLER, 1981 Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res **9:** 133-148.

13 Appendix: Additional published papers