**Accounting for Dependent Evolution Among Sites: Phylogenetic and Population Genetic Approaches**

by

Chris Anthony Nasrallah

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Integrative Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John P. Huelsenbeck, Chair
Professor Rasmus Nielsen
Associate Professor Ian Holmes

Fall 2012

**Accounting for Dependent Evolution Among Sites: Phylogenetic and Population Genetic Approaches**



by

Chris Anthony Nasrallah

## Abstract

Accounting for Dependent Evolution Among Sites: Phylogenetic and Population Genetic Approaches

by

Chris Anthony Nasrallah

Doctor of Philosophy in Integrative Biology

and the Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor John P. Huelsenbeck, Chair

Models of the evolution of DNA sequences typically assume that each position of the sequence evolves independently of all others. This assumption is unrealistic in most cases and is made either for simplicity, computational tractability, or because the nature of the dependence may not be well understood. Proteins and RNAs present instances in which the three dimensional structure of the molecules are essential for function, and introduce dependence among sites in clearly defined ways. Here I explore models that can account for dependence among sites, use them to explore the evolution of DNA sequences containing dependence both within a population and between species, and develop a new substitution model that can be used to make inferences about the strength of natural selection acting on these sequences.

In the first chapter I demonstrate the importance of accounting for dependent evolution among sites for phylogenetic inference. Using a realistic model of the evolution of proteins and RNAs based on known structures, I simulate the evolution of DNA sequences in which the evolution at each site can depend on many other positions in the sequence. Using these simulated data I show that phylogenetic methods that assume sites evolve independently are impaired in their ability to infer the true topology relating the species, and I quantify the error in this estimation as a function of the strength of the dependence, the tree length, the topology, and the specific type of molecular structure. This underscores the importance of accounting for such dependent evolution among sites in studies of molecular evolution.

In the second chapter I explore the dynamics of the substitution process within a population rather than between species. One of the central questions when accounting for epistatic interactions among sites is how two changes, which when taken together are neutral, can spread in a population when a single change in isolation is deleterious. This process of compensatory evolution has been explored by population genetics theory in the case when natural selection acting against the intermediate state is very strong. Here I explore the case in which natural selection against the intermediate states is moderate to weak using forward

time population genetic simulations of the simplest possible case of two dependent sites. I show that when selection is weak the two substitutions can be made one at a time, that as selection increases the substitutions are made more frequently in tandem, and how these patterns are functions of population size, mutation rate, and recombination.

In the third chapter I utilize the insights about the dynamics of compensatory evolution within a population from the second chapter to reexamine the evolution of dependent sites between species. I develop a new substitution model for the analysis of RNA that accounts for the probability of the different pathways to compensatory substitution. This model is interpretive, in that parameters have direct meaning with respect to the strength of natural selection acting against deleterious intermediate states. I implement this model in a Bayesian framework for parameter estimation, and demonstrate its utility for making inferences about historical selective pressures on RNA sequences using a 5S ribosomal RNA dataset. This represents the first probabilistic evolutionary model that both accounts for dependent evolution among sites and connects population genetic dynamics with substitution patterns between species.

Taken together, these studies reveal a great deal about the nature of the evolutionary process when sites are not independent. They explore these processes both within a population and between species, and then use insights from one to better inform the other, attempting to connect these two historically separate approaches to the study of evolution. The advances here are not limited to RNA and proteins, but are generally applicable to any instance in which epistatic interactions can be found, from speciation genetics to the evolution of functional morphology.

In memory of Jido Elias

# Contents

## Acknowledgments

I would like to thank my committee members John Huelsenbeck, Rasmus Nielsen, and Ian Holmes for being consistently giving with their time and affording me their full attention when needed, and to Montgomery Slatkin who often seemed like an unofficial fourth committee member. A special thanks to John for always immediately putting down whatever he was doing anytime I came to him with a question.

My academic development owes every bit as much to other students and postdoctoral researchers as it does to professors. I would especially like to recognize Philip Johnson, Weiwei Zhai, Anna Sapfo-Malaspinas, Tracy Heath, Bastien Boussau, and Jeremy Brown for their advice, support, and for their friendship.

A very special thank you to Alan Turner, who first suggested that I might be interested in reading about phylogenetics. He was right.

I would also like to thank my tremendous community of family and friends, both academic and non-academic, for making my life in Berkeley vibrant and fulfilling in every way.

And perhaps most importantly, I would like to thank my partner Eva for being a constant reminder of what is important. For her boundless support and patience I can never thank her enough.

# Chapter 1

# Evaluating Phylogenetic Models via Phylogenetic Simulation of Dependent Evolution

## 1.1 Introduction

One of the fundamental assumptions made by most methods of phylogenetic inference is that characters evolve independently. This is of course not the case in reality, and there has in recent years been an effort to develop models that more accurately reflect the various types of dependence among sites that have been observed in a biological context.

One kind of dependence is the correlation of rates of substitution at adjacent sites. Yang (1995) and Felsenstein and Churchill (1996) developed methods that allowed the rate of substitution at a given site to depend on the rates of substitution at neighboring sites. But it is important to note that in these models it is only the overall *rate* of substitution that is correlated among sites; substitutions under the model remain independent at different sites. We will focus on methods in which both the rate and types of changes observed at one nucleotide position are dependent upon the nucleotide observed at another position in the sequence.

An example of this kind of dependence, in which adjacent sites can influence not only the rate but also the types of substitutions that occur, is found in the triplet codon structure in protein-coding DNA. Certain substitutions may be less frequent at one site because a change at that site would alter the amino acid encoded by the three sites taken together. Muse and Gaut (1994) and Goldman and Yang (1994) developed codon-based methods to address these concerns, and Nielsen and Yang (1998) expressed the codon model in the form most commonly used today.

Dependence can also arise due to the secondary structure of RNA molecules. Particular attention has been paid to developing methods that address the pairing of nucleotides in RNA stem formations (Schöniger and von Haeseler, 1994; Tillier, 1994; Tillier and Collins,

1995). Dependencies due to secondary structure are often more complicated than those at adjacent sites, as the dependent positions may be quite far from each other in terms of sequence position. It should be noted that these models, like the codon models, are one-substitution-at-a-time models.

Codon models for protein-coding DNA and doublet models for RNA share in common a general approach for accounting for dependence: they expand the basic evolutionary unit in the model from the nucleotide to the triplet or to the doublet, respectively. Robinson et al. (2003) took this approach to its logical endpoint, using the entire protein-coding DNA sequence as the unit of evolution. They considered dependencies resulting from amino acid interactions as well as those resulting from solvent accessibility, and in doing so they allowed the number of other sites on which a given site was dependent to vary across the sequence. Rodrigue et al. (2005) took a similar approach but using only the amino acid interactions, and Kleinman et al. (2006) showed that the model fit is much better when the solvent accessibility is included. See Anisimova and Kosiol (2009) for a recent review of several of these models of substitution.

Error in phylogenetic estimation due to dependent evolution has been detected in recent datasets as well. Castoe et al. (2009) identified thirteen mitochondrial protein-coding regions in squamates that they believe to be the result of strong non-neutral convergence. They argue that models of evolution that can account for convergence due to negative selection, such as those which consider the structure of a protein, might be useful for detecting similar cases that may otherwise strongly bias phylogenetic estimates.

Here we quantify how robust methods of phylogenetic inference are to violation of the assumption of independence. We use an evolutionary model similar to other sequence-based evolutionary models (Robinson et al., 2003; Rodrigue et al., 2005) to simulate sequence evolution under plausible dependent constraints based on RNA and protein structures, and we evaluate the performance of traditional phylogenetic methods on these simulated data sets. We find that even small amounts of dependence in the data can lead to significant error in estimation of the true topology, and that this is especially true for RNA.

## 1.2 Methods

### General Strategy

We are interested in testing whether or not methods of phylogenetic inference that assume independent evolution at each site are robust to violation of that assumption. We are specifically interested in the ability to recover the correct tree topology, rather than in accurately estimating branch lengths or other model parameters. The general strategy is as follows. 1) Simulate an alignment under a known tree topology and set of branch lengths, with a known model of dependence. 2) Estimate the tree from the simulated alignment using standard methods of phylogenetic inference, all of which assume independence of substitutions at different sites. 3) Assess the accuracy of the methods. The methods we will test are maximum

likelihood using the general time-reversible model of substitution with gamma-distributed rate variation (GTR+Γ; Tavaré, 1986; Yang, 1993, 1994), neighbor-joining (Saitou and Nei, 1987) using GTR+Γ distances, and parsimony as implemented in PAUP* 4.0b10x (Swofford, 1998). We will not be interested in comparing these methods to each other, rather in examining the effect of dependence in the data on all of these methods.

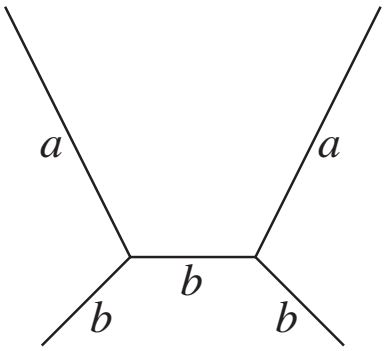



Figure 1.1: The four-taxon tree. The ratio of the branch lengths $a/b$ and the total tree length $V$ are the parameters of interest. As $a/b$ becomes large the inference problem becomes increasingly difficult.

The simplest case in which to study the effect of dependent evolution on phylogenetic inference is with the four-taxon tree, and has been well-studied previously (Felsenstein, 1978; Huelsenbeck and Hillis, 1993). We largely focus on the four-taxon case in order to obtain a thorough understanding for how the tree length, tree topology, and varying levels of dependence affect inference. The tree we consider is shown in Figure 1.1. Two opposing terminal branches share a common length $a$, while the other two terminal branches and the internal branch share a common length $b$. The proportion $a/b$ will be of great interest to us; when this quantity is larger the inference problem is increasingly difficult. We will also be interested in the total tree length ($V$), which allows the tree to be expanded or contracted while preserving the branch length proportions.

## Evolutionary Model

Calculation of the likelihood (or the parsimony score) of an alignment is greatly simplified by the independence assumption. If all sites are independent then the probability of an alignment is simply the product of the probability of each column in the alignment (or the sum of the parsimony scores). To calculate this probability the substitution process at a particular site is modeled as a continuous-time Markov chain. The process is governed by a rate matrix $\mathbf{Q} = \{q_{ij}\}$ where $q_{ij}$ is the rate of change from state $i$ to state $j$. This rate of change depends only on the current state $i$, and does not depend on what states may have been observed in the past (Markov property). Furthermore, the rate $q_{ij}$ is agnostic to what is

happening at every other site in the sequence. When dependence among sites is introduced it will not affect the Markov property, but the rate of change at a given site will depend on the state of the process at other sites.

The rate matrix $\mathbf{Q}$ can take many forms. For RNA-coding sequences the matrix $\mathbf{Q}$ might be described by anything from the Jukes-Cantor model (Jukes and Cantor, 1969) to the GTR model (Tavaré, 1986), and does not in principle need to be time-reversible. For protein-coding sequences $\mathbf{Q}$ could be described by various codon-based models (Muse and Gaut, 1994; Goldman and Yang, 1994) with different rates for synonomous/nonsynonomous sites as well as for transitions/transversions. These codon models typically restrict the possible changes from codon $i$ to only those codons $j$ that involve a single nucleotide substitution, and disallow stop codons.

Just as codon-based models expand the unit of evolution from the nucleotide to the codon, the model we consider further expands the unit of evolution from the nucleotide to the entire sequence. Consequently we will be interested in *sequence* transition probabilities. More formally, we will consider a continuous-time Markov chain in which the state space is the set of all possible sequences of length $N$ nucleotides. Let $x$ and $y$ be two such sequences. Then for all $x$ and $y$ the matrix of rates of change from $x$ to $y$ can be defined as

$$\mathbf{R} = \{r_{xy}\} = \begin{cases} 0 & \text{if } x \text{ and } y \text{ differ at 2 or more positions} \\ uq_{ij}E(x,y) & \text{if } x \text{ and } y \text{ differ at only 1 position} \\ -\sum_{x \neq y} r_{xy} & \text{if } x = y \end{cases}$$

where $q_{ij}$ is, as described above, the rate of substitution under an independent-sites model for the position in the sequence that is changing, $E(x, y)$ compares the relative structural fitness of sequences $x$ and $y$ and is described in detail below, and $u$ is a rate-scaling factor to ensure that branch lengths are interpretable in terms of average number of substitutions per site.

In the simulations of this paper we specify the underlying $\{q_{ij}\}$ as follows. For RNA, we assume the K80 model of substitution (Kimura, 1980) with the transition/transversion rate ratio $\kappa = 3$. For proteins, we assume a codon substitution model (Nielsen and Yang, 1998) with transition/transversion rate ratio $\kappa = 1$, nonsynonomous/synonomous substitution rate ratio $\omega = 1$ and equal codon frequencies. These relatively simple substitution models were chosen to better examine the effects introduced by the dependencies due to structure, described below.

## Energy of a Sequence

We will utilize a concept borrowed from structure prediction: a sequence folded into a particular structure will have a free energy associated with it. In structure prediction the sequence is fixed and the structure of lowest energy is sought; here we invert this problem by conditioning on the structure being fixed. We assume that all sequences share a common,

fixed structure that must be maintained to preserve functionality. It is this structure that determines the interactions among sites and their relative positions, and therefore determine their evolutionary interdependencies.

We will define the energy of a single sequence $x$ as $E(x)$, but we find it useful to conceive of $E(x)$ not as an actual energy, rather as a measure of how well sequence $x$ corresponds to the given structure, or as a kind of "structural fitness." If $x$ could reasonably fold into the given structure we expect $E(x)$ to be low, ideally negative. We can then calculate $E(y)$ for any sequence $y$ as well. $E(x, y)$ then takes on the meaning of a comparison of the relative structural fitness of the two sequences. The precise form of $E(x, y)$ can in principle vary, and we will define $E(x, y)$ differently for RNA and for proteins.

For RNA, what we call energies are folding free energy changes ($\Delta G$) predicted using the current nearest neighbor model of Turner and co-workers (Mathews et al., 2004). These free energy changes are predicted for a given base pairing structure using the efn2 model (Mathews et al., 1999). This approach utilizes information from both the base-pairing and the coaxial stacking of nucleotides, allowing the potential to incorporate more information than a simple doublet model that considers doublets to be independent of each other. For RNA, we then define

$$E(x, y) = e^{(E_z(x) - E_z(y))z}$$

where $z$ is a free parameter determining the degree to which the difference in structural fitness affects the rate of substitution. Note that when $z = 0$, $E(x, y) = 1$ for all $x$ and $y$, reducing the model to the independent-sites model specified by the single-site rate matrix $\mathbf{Q}$.

For proteins, we adopt the approach of Robinson et al. (2003) in simplifying the constraints governing the structure into two properties: energies due to pairwise interactions of amino acids and to solubility constraints (denoted $E_p(x)$ and $E_s(x)$ respectively). To do this we utilize statistical potentials, which are pseudo-energy values associated with plausibilities of some aspect of the structure estimated from protein sequences of known structure. For pairwise interactions of amino acids, we can from the protein structure determine the relative positions of all amino acids in three-dimensional space, and declare two amino acids to be "in contact" if any of their non-hydrogen atoms are less than 4.5 Å apart (Bastolla et al., 2001). Pairs of amino acids whose three dimensional proximity is due to sequential proximity (within three positions or less) are not considered to be in contact. Following Rodrigue et al. (2005), if our sequence is of length $N$ we can describe a contact map as an $N \times N$ matrix $\mathbf{C}$ where

$$\mathbf{C} = \{c_{lm}\} = \begin{cases} 1 & \text{if positions } l \text{ and } m \text{ are in contact} \\ 0 & \text{if positions } l \text{ and } m \text{ are not in contact} \end{cases}$$

where $l$ and $m$ are indices of sequence position (Rodrigue et al., 2005).

Two aspects of this formulation should be noted. First, unlike RNA where sites can potentially pair, here a single site can be considered in contact with multiple other sites. Second, these interactions are all weighted equally regardless of actual physical distance, as long as they are sufficiently close. It would be straightforward to alter the latter such that

the relative distance is preserved and certain interactions are more influential than others. As described by Rodrigue et al. (2005), we can now define the energy of the sequence $x$ with respect to pairwise potentials as the sum of the pair potentials for all pairs of amino acids in contact:

$$E_p(x) = \sum_{1 \le l \le m \le N} c_{lm} b_{x_l, x_m}$$

where $x_l$ and $x_m$ are the amino acids of sequence $x$ at positions $l$ and $m$ respectively, and $\mathbf{B} = \{b_{x_l,x_m}\}$ is the pair potential matrix of Bastolla et al. (2001).

To model solubility constraints on protein evolution we follow Robinson et al. (2003), who used an analysis of a large number of proteins to estimate how frequently a particular amino acid is observed at different degrees of solvent accessibility [see also Jones et al. (1992) and Jones (1999)]. From the protein structure we determine the solvent accessibility of a particular amino acid position. The energy with respect to solubility of sequence $x$, $E_s(x)$, is then the sum across all sites of the plausibility of seeing the observed amino acid at that accessibility level,

$$E_s(x) = \sum_{1 \le k \le N} S(x_k, a_k)$$

where $a_k$ is the degree of solvent accessibility of site $k$ and $S(x_k, a_k)$ is the statistical potential for observing amino acid $x_k$ at such a degree of solvent accessibility (Robinson et al., 2003). We can now define $E(x, y)$ for proteins in a similar form as for RNA:

$$E(x, y) = e^{(E_p(x) - E_p(y))p + (E_s(x) - E_s(y))s}$$

where $p$ and $s$ are, like $z$ in the case of RNA, parameters that control how much the difference in sequence energies affect the rate of substitution for pairwise potentials and solubility, respectively. Note again that when $p = s = 0$, $E(x, y) = 1$ for all $x, y$, reducing the model to one of independence among sites.

## Simulation Procedure

There are a number of ways to simulate data at a single position under an independent-sites model. Some of these are not applicable for simulating data that are context-dependent. We will discuss a few of these methods and their applicability. The first method (Fig. 1.2$a$) begins by drawing the nucleotide at the root node of the tree from the stationary distribution $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$. If we specify the rate matrix $\mathbf{Q}$ and a branch length $t$ we can calculate the transition probability matrix $\mathbf{P}(t) = \{p_{ij}(t)\} = e^{\mathbf{Q}t}$. This provides, for all $i$ and $j$, the probability that after a branch length of $t$ the descendent node is in state $j$ given that our ancestral node was at state $i$. Each branch will have its own transition probability matrix, since branch lengths may differ. We can then work our way up the tree starting from the root, choosing states at each node until we reach the tips. The usage of matrix exponentiation to calculate transition probabilities is attractive because it considers all the possible paths, or character histories, from $i$ to $j$ in time $t$. However the matrix exponentiation becomes

intractable when the rate matrix is large. This is the case with the dependent-sites model we have described, where the rate matrix $\mathbf{R}$ is $4^N \times 4^N$, and for any reasonable sequence length $N$ the matrix is quite large indeed.

Instead of using a transition probability matrix to consider all the possible paths from state $i$ to $j$ over time $t$ simultaneously we could instead simulate a single character history (Fig. 1.2*b*). One of the properties of the continuous time Markov chain is that if the process is in state $i$, the waiting time until we leave state $i$ is an exponentially distributed random variable with rate $q_{ii} = -\sum_{j \neq i} q_{ij}$. This means we find our root node state $i$ from the stationary distribution as before, but now draw an exponential random variable with rate $-q_{ii}$. If this time is less than $t$ we observe a change from $i$ to some other state $j$. The particular state $j$ is drawn with probability $p_{ij} = q_{ij}/-q_{ii}$. This procedure is repeated until the sum of the drawn waiting times exceeds the length of the branch $t$, at which point the state of the process is the state at the descendent node. This character history simulation is performed iteratively up the tree for all branches until we have our states at the tip nodes. This method of drawing character histories has the benefit that it can be used under the dependent-sites model we have described. This is done by using the full sequence as the unit of evolution and replacing the site rate matrix $\mathbf{Q}$ with the sequence rate matrix $\mathbf{R}$, and then drawing a sequence history along the branch.

Both of these methods have assumed that we could draw the state at the root of the tree directly from the stationary distribution. This is not trivial under the dependent-sites model, as the state space of all possible sequences is quite large ($4^N$ possible sequences) when compared with independence (4 possible nucleotides). However, the intuitive meaning of the process being at stationarity at the root is that the process has been underway for a long time before reaching the root of the tree, and we can simulate this directly (Fig. 1.2*c*). Under independence, if we pick any state $i$ as an ancestral state and then simulate its evolution along an exceedingly long branch before reaching the root, then the probability that we observe a particular state $j$ at the root is the same as having drawn directly from the stationary distribution. This method can be used for the dependent-sites model described as well, and is the method employed for all simulations in this study. We begin with an arbitrary sequence, not necessarily one that would likely be sampled from the true dependent stationary distribution. We then evolve this sequence along a very long root branch under the model of dependence as described above, allowing the sequence to evolve into one that would be sampled from the true dependent stationary distribution. The intuition should be clear: we need a sequence that corresponds to a fixed structure, so we choose a random sequence and allow it to evolve into one that corresponds to the structure (directional selection). This yields a sequence at the root of the tree that corresponds to the structure, that can be used as a starting point for the simulation of the tree itself under continued structural constraint (stabilizing selection).

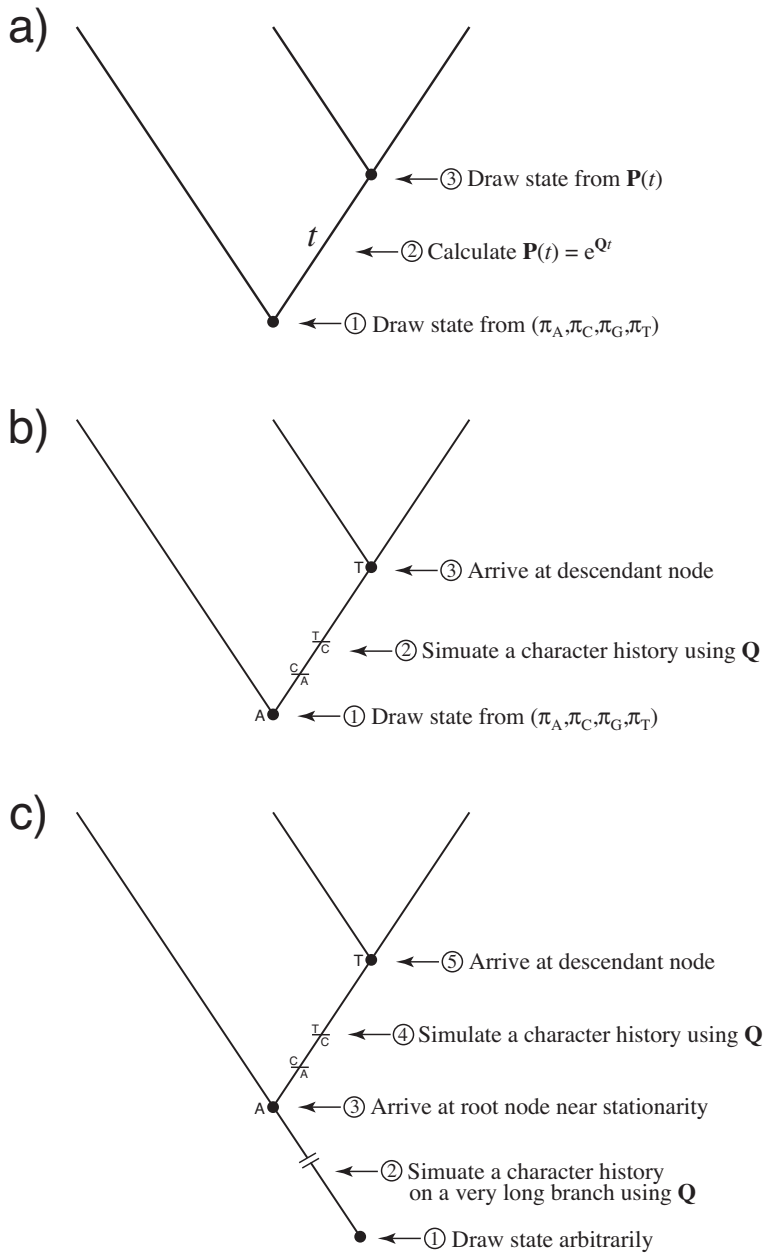



Figure 1.2: Three methods for simulating data under independence. (*a*) Using matrix exponentiation is intractable for dependent data (*b*) Simulating a character history can be done with context-dependency for an entire sequence, but drawing from the stationary distribution at the root node is still problematic. (*c*) Evolve into stationarity by simulating a very long character history before reaching the root, then continuing up the tree as in *b*.

### Structures Examined

In this study, we examine the effect of dependence introduced via structural constrain in both RNA and proteins. For RNA, we will focus on two structures: the *Bombyx mori* R2 element reverse-transcriptase 3' UTR (R2), a 300 nucleotide structure previously examined by Mathews et al. (1997) and the eukaryotic 5S rRNA structure (119 nucleotides) examined by Yu and Thorne (2006). Each simulated parameter set using these structures include 400 and 1000 replicates, respectively. For proteins, we will also use two structures: mammalian myoglobin (*Physester catodon*; PDB code 1MBD; 459 nucleotides) and 6-hydroxymethyl-7-8-dihydroxypterin pyrophosphokinase (*Escherichia coli*; PDB code 1HKA; 474 nucleotides), both examined by Rodrigue et al. (2005). Simulations using these protein structures consist of 1000 and 500 replicates, respectively.

### Energy at Stationarity

Since we have described the energy of a sequence as measuring how well a sequence fits a structure, we can visually inspect this process of approaching and sampling from the stationary distribution of sequences under the selective constraint by monitoring the energies of the sequences sampled. Figure 1.3*a* shows the energies of a sequence, initially sampled at random, evolving continuously under independence. As expected, the sequences sampled have similarly high energies since the vast majority of the $4^N$ possible sequences will not naturally fit the structure well. Contrast this with Figure 1.3*b*, which shows the energies of a sequence, similarly sampled at random originally, but evolving under the model of dependence. The sequences sampled converge to an area of the sequence space with much lower energies and remains there indefinitely. This indicates that the selective constraints of the structure limit the sequences that can be sampled to those that fit the structure reasonably well. That the chain fails to leave this area of the sequence space is indication that we are in fact sampling sequences from the stationary distribution. In this manner we can empirically determine the branch length necessary to be sampling from the stationary distribution with high probability prior to the simulation of sequences along the trees.

## 1.3 Results and Discussion

### Rate Variation Among Sites

We expect that the constraints imposed by structure will affect among-site rate variation; at stationarity, a site that is tightly constrained will experience a low rate of substitution relative to unconstrained sites. To confirm this, we simulated the evolution of a sequence at stationarity for varying levels of dependence and observed the number of changes occurring at different sites in the sequence. The results are shown in Figure 1.4. Under independence, RNA stem and loop positions observed similar rates of substitution (Fig. 1.4*a*), whereas under dependence ($z = 0.01$) the rate of substitution at stem position decreased and at

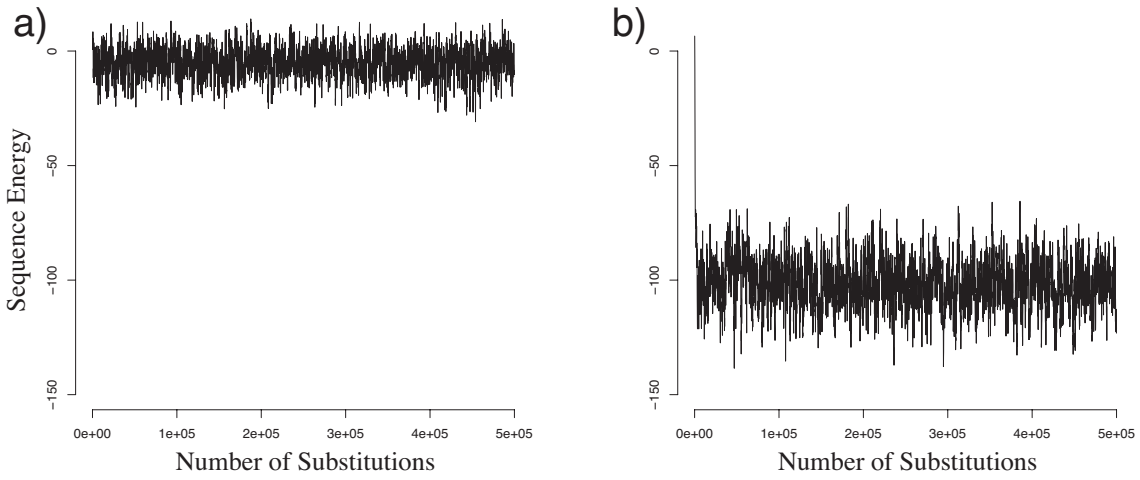



Figure 1.3: Energies sampled every 100 substitutions from a continuously-evolving sequence. ($a$) Independence among sites. Energies sampled are similar to that initially sampled at random. ($b$) Dependence due to structural constraint. Low energies indicate that sequences sampled are those that fit the structure. The sequence evolves from a randomly-sampled starting state of high energy to sample those states of low energy that correspond to the structure.

loop positions increased (Fig. 1.4$b$). Further increasing the level of dependence did not seem to affect the change in substitution rate (data not shown). For proteins, the substitution process at a particular site can depend upon a number of other sites determined by the site's location in the folded protein. While under independence the rate of substitution observed was similar regardless of number of contacted other sites (Fig. 1.4$c$), under dependence we observed a clear negative correlation between the number of sites upon which a particular site is dependent and the rate of change that site experienced (Fig. 1.4$d$). Similar results were obtained for all RNA and protein structures examined.

It is worth noting that the RNA model induces a higher rate of substitution among loop sites than stem sites. This is quite different from what is observed in alignments of certain RNAs, in which loop regions are often highly conserved. This difference is in part because, while this model accounts for the dependencies introduced by the maintenance of the structure of the molecule, the model does not explicitly consider its function. If loop regions are conserved due to functional constraint of binding to another molecule, the dependence of these sites on their binding site is not captured by our model, which looks only at the structure of the single RNA. Clearly both types of constraint are biologically relevant, and it would be straightforward to imagine expanding the model beyond a single sequence to consider two RNAs (or proteins) that interact, introducing dependencies both within and between the structures. It should also be noted that while our model does not capture stabilizing selection on RNA loop regions, neither do the independence-assuming

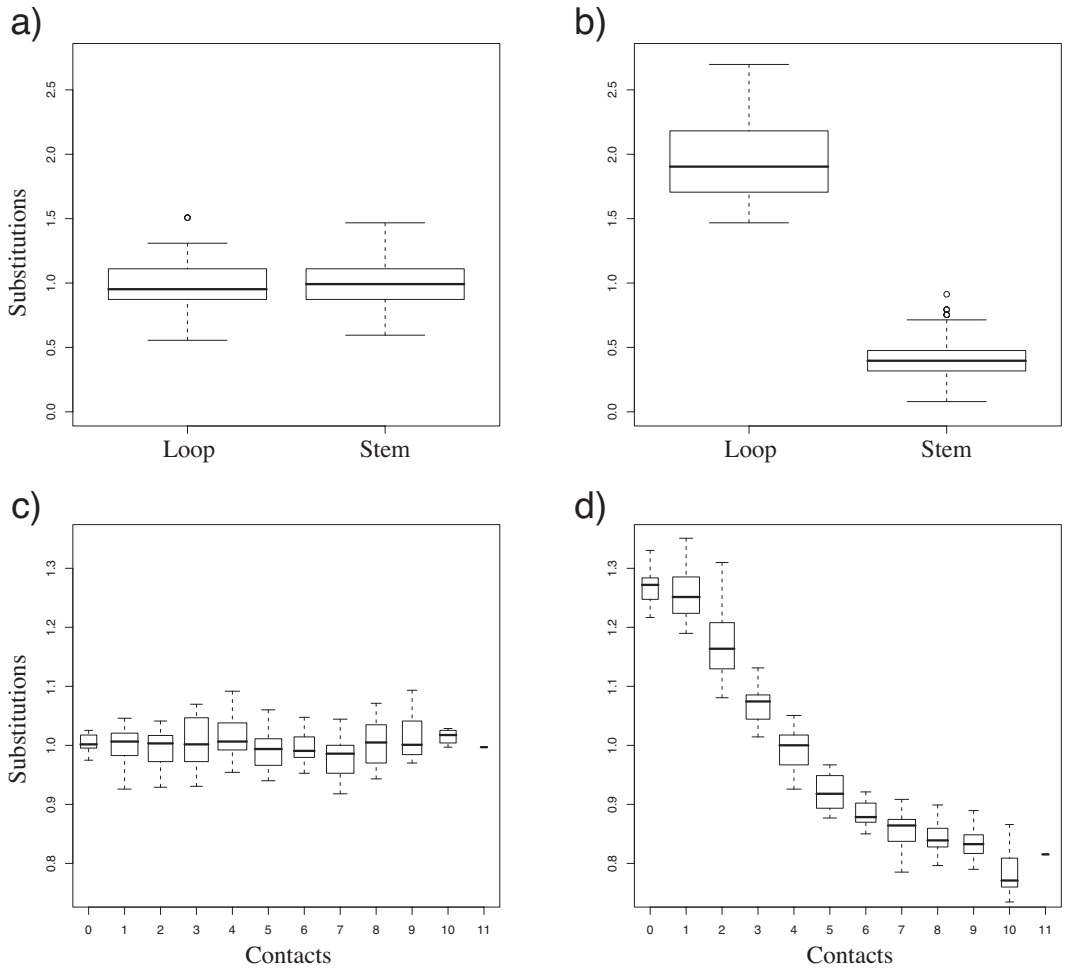



Figure 1.4: Number of substitutions at sites of varying constraint. Under independence RNA stem and loop sites experience similar rates of substitutions ($a$), but under dependence stem sites observe fewer and loop sites observe more substitutions ($b$). For proteins, under independence all amino acid sites experience similar rates of substitution ($c$), whereas under dependence the rate of substitution is inversely proportional to the number of other sites with which the given site is in contact ($d$).

models typically used for phylogenetic inference.

## Effect of Dependence on RNA

We simulated sequences on the four-taxon tree using the predicted RNA structure of the *Bombyx mori* R2 element reverse-transcriptase 3' UTR (300 nucleotides) previously examined by Mathews et al. (1997). We did this for a constant tree size ($V = 1.75$) and a range of branch length proportions at varying levels of dependence, and then estimated the topology from the data assuming independence. Figure 1.5*a* shows the accuracy of maximum likelihood (ML) at estimating the true topology for these simulated data sets (400 replicates). The shorter the internal branch, the more difficult is the estimation problem. As expected, ML performs well for all tree shapes on data simulated with no dependence among sites. But as the level of dependence among sites increases the accuracy of ML decreases, particularly when the internal branch is short. Perhaps most striking is the decrease in accuracy resulting from even small levels of dependence in the data ($z = 0.1$), with accuracy falling to nearly 50% when the internal branch is short.

These simulations also provide a sense for just how short the internal branch must be before ML will begin to see a decrease in accuracy resulting from dependent evolution among sites. While it might be encouraging if ML had difficulty only when the internal branch was quite short, this is not the case. Appreciable decreases in accuracy are observed over a wide range of internal branch lengths, indicating that the effects of dependent evolution on phylogenetic inference are not restricted to extreme topological cases.

It is important to note that while our structural model does induce rate variation among sites, this is at least partially accounted for in the GTR+$\Gamma$ model used for analysis. This means that observed decreases in accuracy are more likely to result from differences resulting from the context-dependent nature of the substitution process induced by the model of structural constraint.

The decreased accuracy resulting from dependence in the data is not a particular property of maximum likelihood however. Figures 1.5*b*, *c* show the analysis of the same data using neighbor-joining (with ML distances) and parsimony, respectively. While the baseline expectations of how well the methods will perform when the data are independent differ, the trend is the same for all methods that assume independence: the effect of dependence is to reduce the accuracy of the methods, particularly when the problem is difficult, as is the case when branches differ markedly in length. We might note that neighbor-joining appears to do as well as maximum likelihood in many cases, and in some cases seems to perform better. It would be tempting to attempt to draw broader conclusions from these simulations about the relative performance of these methods, but it must be remembered that we show here only a small portion of the possible parameter space of topologies, branch lengths, model parameters, and have only shown a four-taxon case using a single structure. We refrain from drawing any such conclusions, and instead focus on the observation that all of these methods seem to suffer by failing to account for the dependence.

To examine whether these results were specific to the structure examined or more general,

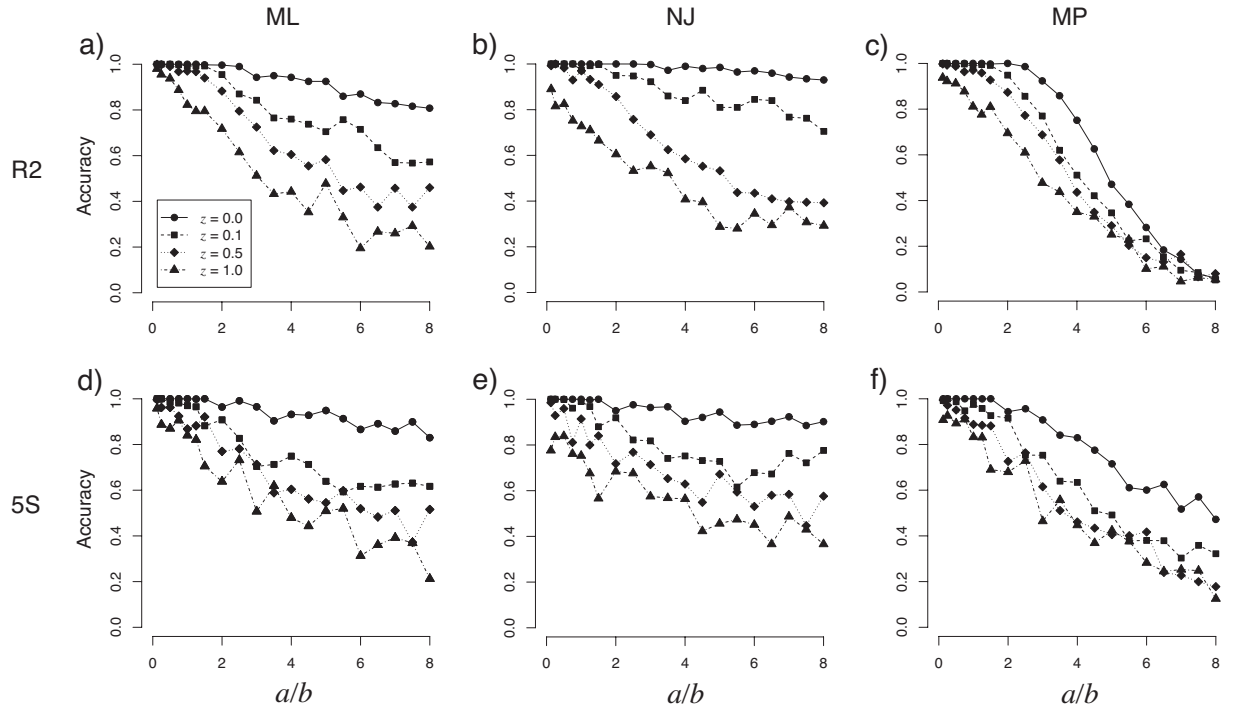



Figure 1.5: The accuracy of independence-assuming phylogenetic methods to infer the correct topology using RNA sequences constrained by structure simulated on a tree of total length $V = 1.75$. As the level of dependence in the data ($z$) increases, the methods are increasingly unable to infer the correct topology. This is especially true as the branch length ratio ($a/b$) becomes large and the problem becomes difficult. Structures: *Bombyx mori* R2 element reverse transcriptase 3' UTR (R2) [300 nucleotides, 400 replicates] and 5S rRNA (5S) [119 nucleotides, 1000 replicates]. Methods: Maximum Likelihood GTR+$\Gamma$ (ML), Neighbor-Joining using ML distances (NJ), Parsimony (MP).

we simulated data using the eukaryotic 5S rRNA (119 nucleotides) as the reference structure. We did so on a slightly shorter tree length ($V = 1.0$) over the same range of branch length proportions and levels of dependence (1000 replicates). The analysis of these simulated sequence sets (Fig. 1.5*d*-*f*) are qualitatively consistent with the previous results: methods that assume independence experience a reduction in accuracy over a wide range of branch length proportions as the level of dependence increases. This suggests that these decreases in accuracy are not specific to a single structure, but are a more general property of the effect of dependent evolution in RNA.

The performance of phylogenetic methods assuming independence is also affected by the overall length of the underlying tree as well as its topology. Figure 1.6 shows the effect on accuracy of ML estimation using simulated R2 element RNA sequences over a range of branch length proportions on trees of total length 0.25, 1.0, and 1.75. The effect of a fixed

level of dependence ($z = 0.5$) is to reduce accuracy relative to independence ($z = 0.0$) as shown before, but the effect is greater when the overall tree length is greater. On a larger tree (Fig. 1.6*c*) reductions in accuracy are observed at small branch length proportions, whereas on a small tree (Fig. 1.6*a*) the branch length proportion must be larger before reductions in accuracy are observed. This demonstrates how tree length and topology may interact to cause difficulties in estimation on dependence-containing data; dependence seems to have the greatest effect when the tree is very large and the internal branch is short. Results for neighbor-joining and parsimony were qualitatively similar and for brevity we will largely focus the remainder of the four-taxon case discussion on results for maximum likelihood, which are representative of trends observed using all methods examined.

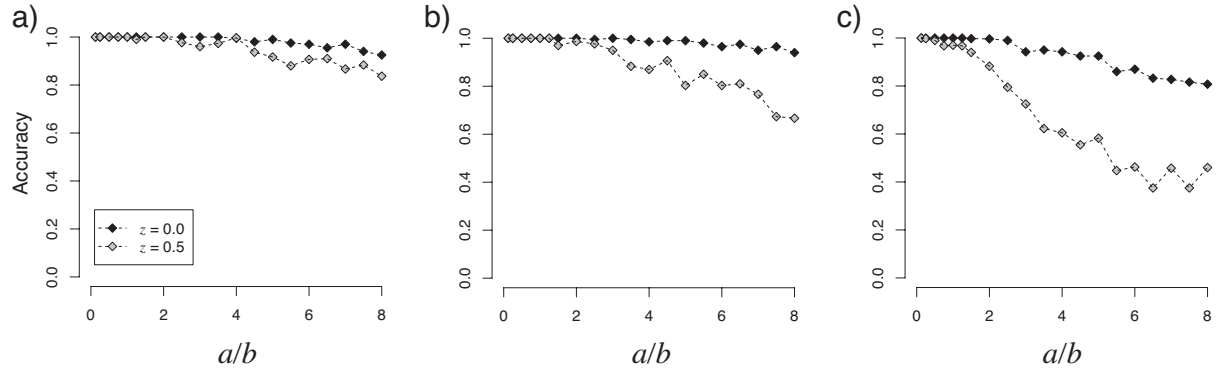



Figure 1.6: The total tree length ($V$) effects the accuracy of Maximum Likelihood on simulated RNA sequences constrained by structure (R2). Dependence in the data (bottom curves) reduces the accuracy relative to independence (top curves), and this effect is more pronounced when the underlying tree is larger. (*a*) $V = 0.25$ (*b*) $V = 1.0$ (*c*) $V = 1.75$. Qualitatively similar results were obtained for other independence-assuming methods and levels of dependence.

## Effect of Dependence on Proteins

For proteins the dependence involves two components: pairwise interactions and solubility constraints. To explore how each of these affect inference we used the reference structure of mammalian myoglobin (*Physeter catodon*; PDB code: 1MBD; 459 nucleotides), previously studied by Rodrigue et al. (2005), to simulate data on a tree topology with $a/b = 5$ and a total length $V$ of 1.3 (Fig. 1.7*a*) or 2.08 (Fig. 1.7*b*) across a wide range of dependence parameter values (1000 replicates). The larger tree shows the same trend as RNA: increased levels of dependence result in decreased accuracy. However the effect seems to be less severe, particularly when the level of dependence is small. Furthermore the dependence due to pairwise interactions has a much greater effect than dependence due to solubility constraints.

Importantly, there is very little effect whatsoever observed when the tree length is small until levels of dependence become quite large indeed, even when the topology itself poses a challenging problem.

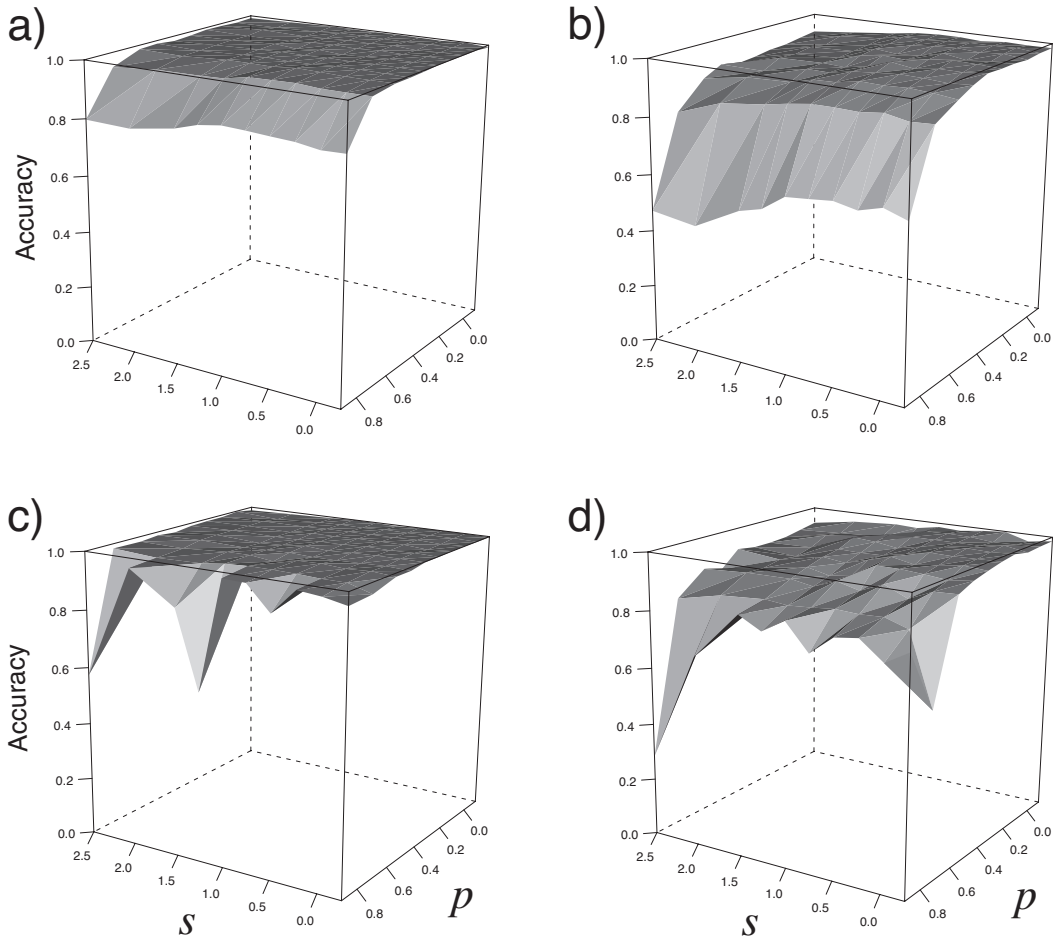



Figure 1.7: Accuracy of phylogenetic inference using Maximum Likelihood using sequences generated under varying levels of dependence due to protein structure constraints: solubility ($s$) and pairwise interactions ($p$). Accuracy is reduced when dependence is strong and tree length is large. All panels represent the same tree topology ($a/b = 5$). Structures: mammalian myoglobin (MYO) [459 nucleotides, 1000 replicates] and 6-hydroxymethyl-7-8-dihydroxypterin pyrophosphokinase (PKA) [474 nucleotides, 500 replicates]. ($a$) MYO, $V = 1.3$ ($b$) MYO, $V = 2.08$ ($c$) PKA, $V = 1.3$ ($d$) PKA, $V = 2.08$.

We then repeated these simulations using the reference structure of 6-hydroxymethyl-7-8-dihydroxypterin pyrophosphokinase (*Escherichia coli*; PDB code: 1HKA; 474 nucleotides), also examined previously by Rodrigue et al. (2005). While the structures of these two proteins

are quite different, the results using the two structures are quite similar (Fig. 1.7$c, d$). There is some additional variance due to fewer replicates (500) but the trend is the same. This suggests that dependence among sites in proteins may have similar effects on phylogenetic inference regardless of the precise nature of the structure.

It is encouraging to see that for proteins, unlike RNA, small levels of dependence in the data do not seem to have a strong effect on the accuracy of phylogenetic methods. Estimates of the level of dependence in actual data will be considered below, but another consideration is whether or not the protein model, which reduces protein structure to two parameters of solubility and pairwise interactions, can adequately account for the complexity of actual protein structures. Vendruscolo and Domany (1998, 2000) and Park et al. (2000) have argued that there are limits to the utility of pairwise interaction potentials and hydrophobicity constraints in protein structure prediction. It is likely that the structural fitness of a sequence would be more accurately represented by the actual Gibbs free energy of the sequence, but at present this approach is computationally demanding. While the simplified approach adopted here is well-justified, the conclusions drawn for proteins may not be the final word.

## Effective Sequence Length

How phylogenetic methods behave when the data are neutral and independent may be used as a reference for describing how phylogenetic methods perform when ideal conditions are not met. We may consider the *effective* sequence length ($L_e$) as the length of independent neutral sequence that behaves in the same manner (in terms of phylogenetic accuracy) as our dependence-containing sequences. This is similar in spirit to the concept of an effective population size in population genetics. Because we expect dependence to introduce correlated substitutions, we expect the effective sequence length to be smaller than the actual sequence length (Huelsenbeck and Nielsen, 1999). How much smaller is of interest, and will depend on several factors including the actual sequence length, the nature of the structural constraints, the relative importance of the dependence, and the topology and length of the underlying tree.

Figure 1.8 quantifies the effective sequence length for one case examined. Each panel represents a different underlying tree topology ($a/b$) of the same overall tree length ($V = 1.75$). For each topology we first simulated under independence sequences of different lengths and assessed the phylogenetic accuracy obtained by using these sequences. Shown in Figure 1.8 as the curves, these indicate the expected accuracy when using sequences of $n$ independent neutral sites. For each topology we then simulated RNA sequences of length 300 nucleotides (using the R2 reverse transcriptase structure) under dependence ($z = 0.1$) and assessed the accuracy using these dependent sequences, indicated by the horizontal lines. Where these observed (dependent) accuracies intersect our expected (independent) curve we can project to the $x$-axis to estimate the effective sequence length for these dependence-containing data. The presence of dependence in the data results in a large decrease in effective sequence length, particularly for topologies in which the internal branch is relatively short.

One could argue that we might have easily predicted the effective length for RNA by

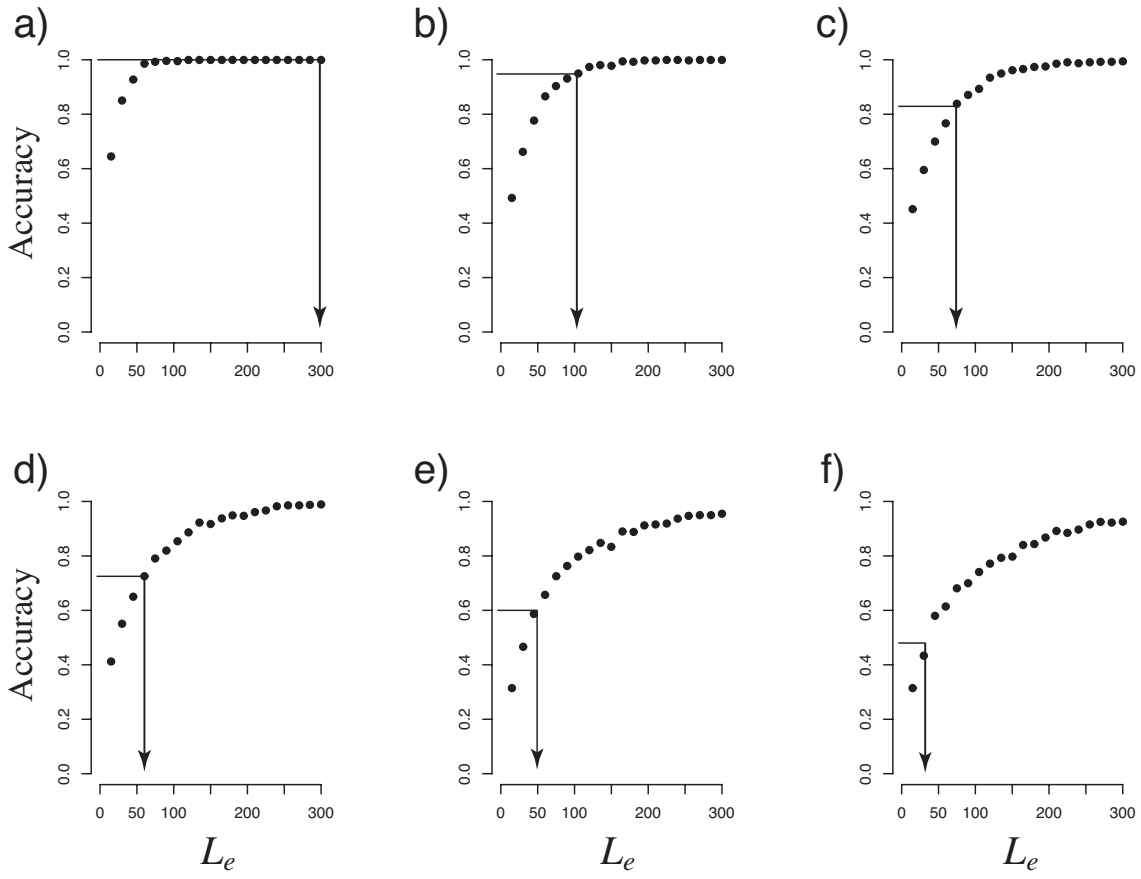



Figure 1.8: The effective sequence length ($L_e$) as a means of quantifying the phylogenetic information content of a sequence that contains dependence. All panels represent a fixed tree length ($V = 1.75$) and level of dependence ($z = 0.1$). ($a$) $a/b = 0.5$. ($b$) $a/b = 2$. ($c$) $a/b = 3$. ($d$) $a/b = 4$. ($e$) $a/b = 6$. ($f$) $a/b = 8$. The plotted curves indicate the accuracy of Maximum Likelihood on these trees using independent data of varying lengths, or the expected accuracy if the data were independent. The accuracy of Maximum Likelihood on the simulated RNA sequences (R2, actual length = 300 nucleotides) on each topology is shown by the horizontal lines. Where these horizontal lines cross the curve, drop to the $x$-axis to estimate the effective sequence length: the length of independent, neutral sequence that displays the same amount of error in estimation that the actual dependence-containing sequence displays.

simply considering paired sites to be as informative as a single unpaired site. In the structure used for the simulated RNA sequences there were 168 stem and 132 loop positions, which by this method would predict an effective sequence length of 216 nucleotides. Alternatively, if all stem positions were considered to be invariable, the effective sequence length would be predicted to be 132 nucleotides. However, what we observe is that the dependence in our data leads to much lower accuracies, and subsequently much lower effective sequence lengths than both of these expectations, observing effective sequence lengths of less than 100 nucleotides. This implies that models simply accounting for covariation in the data are not accounting for all aspects of structural constraint and that these structural constraints lead to greater information loss.

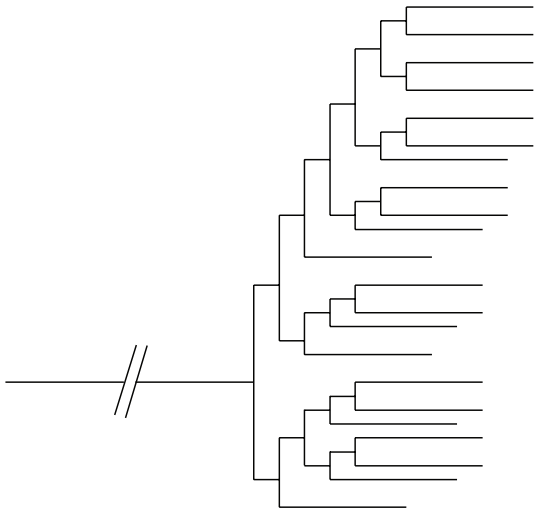
While we suggest that the concept of effective sequence length is useful for thinking about the effect of dependence on data, and is particularly useful for assessing these effects in our simulations, determining the effective sequence length requires knowledge about the true tree and importance of the structural constraints. The practicing systematist would therefore need to make some very strong assumptions in order to use the concept of effective sequence length to explicitly guide analysis.

## Larger Datasets

In order to understand how dependence among sites might affect phylogenetic inference we have focused on the four-taxon case using a single sequence/structure. This allowed us to thoroughly explore the relevant parameter space and gain some intuition for when we might expect error. However, using only four taxa or such a limited amount of sequence data is hardly something done in practice. It would therefore be useful to understand how the effects we have observed extend when the methods are presented with more taxa or more sequence data.

To address the question of how the methods perform on trees larger than four taxa, we simulated sequences on a 22-taxon tree using the used the R2 element RNA structure. In this tree (Fig. S1) all terminal branches are of the same length (0.05 expected substitutions per site), and are five times longer than internal branches (0.01 expected substitutions per site). In some sense this makes for a relatively easy estimation problem: unlike the four-taxon case all terminal branches are of equal length, and the overall tree length is quite small. We simulated 500 RNA datasets on this tree for each of a range of levels of dependence and analyzed these datasets using the same methods used in the four-taxon case. We calculated the Robinson-Foulds metric (Robinson and Foulds, 1981) to compare the estimated tree to the true tree, and the results are shown in Figure 1.10. As expected, dependence in the data increases the amount of topological estimation error, in spite of the estimation problem not being an incredibly difficult one. Notably, even small amounts of dependence are sufficient to cause appreciable decreases in accuracy. We expect that on trees of greater length or containing variance in branch lengths might present more challenging problems and therefore be more sensitive to the effects of dependence. While these simulations are hardly a thorough exploration of the space of possible trees larger than four taxa, they give

Figure 1.9: The 22-taxon tree used for simulation. All terminal branches are of length 0.05 substitutions/site, and internal branches of length 0.01 substitutions/site. The short overall tree length renders this a relatively simple case study of 22 taxa.



a sense for how the problems observed might scale with the number of taxa.

Addressing the question of how very long dependence-containing sequences affect the analyses is not as straightforward. This is because one of the limitations of conditioning on an actual, fixed structure is that the sequences are constrained to a fixed length. To test this question we concatenated our R2 element datasets to create three very long ($\geq 30,000$ nucleotides) sets of sequences. We similarly concatenated 5S sequences to create 2 sets of sequences ($\geq 45,000$ nucleotides). We opted to use the same structure repeatedly to ensure that the same kind of dependence is introduced, as there is no guarantee that different structures will not contain conflicting signal. The results are consistent with what was observed on shorter sequences (Fig. S2). When the dependence is large ($z \geq 0.5$), ML fails to estimate the correct topology when the problem is difficult. When the dependence is small ($z = 0.1$) ML is able to recover the true tree most of the time. Curiously, the neighbor-joining algorithm (using GTR+$\Gamma$ distances) performs very well for all levels of dependence on all trees. Parsimony behaves qualitatively similarly to the results on shorter sequences (Fig. 1.5$c, f$), While these limited number of replicates are hardly conclusive, they give a sense for how these independence-assuming methods might handle a great deal of dependence.

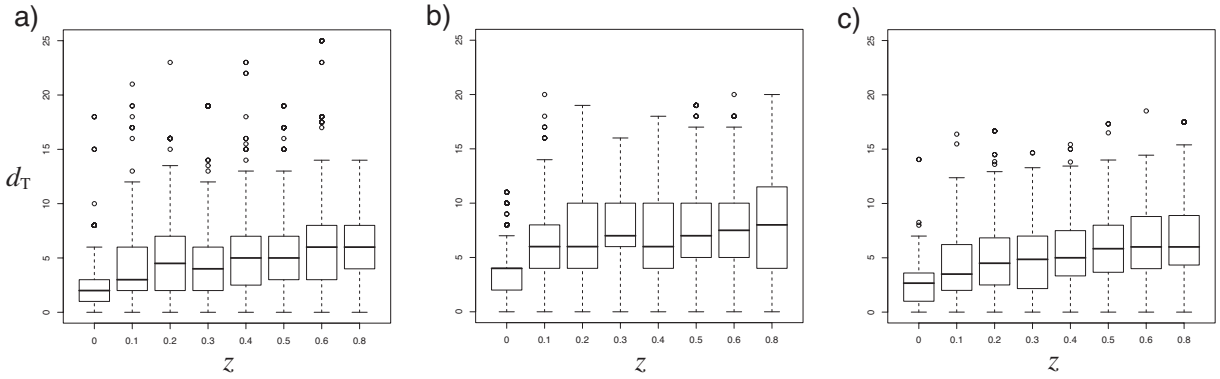



Figure 1.10: Accuracy of independence-assuming phylogenetic methods for 22-taxon simulations of RNA constrained by structure (R2; 500 replicates). The Robinson-Founds distance metric compares the estimated tree to the true tree for datasets under varying levels of dependence ($z$). For all methods, small amounts of dependence introduce error in tree estimation. ($a$) Maximum Likelihood ($b$) Neighbor-Joining ($c$) Parsimony.

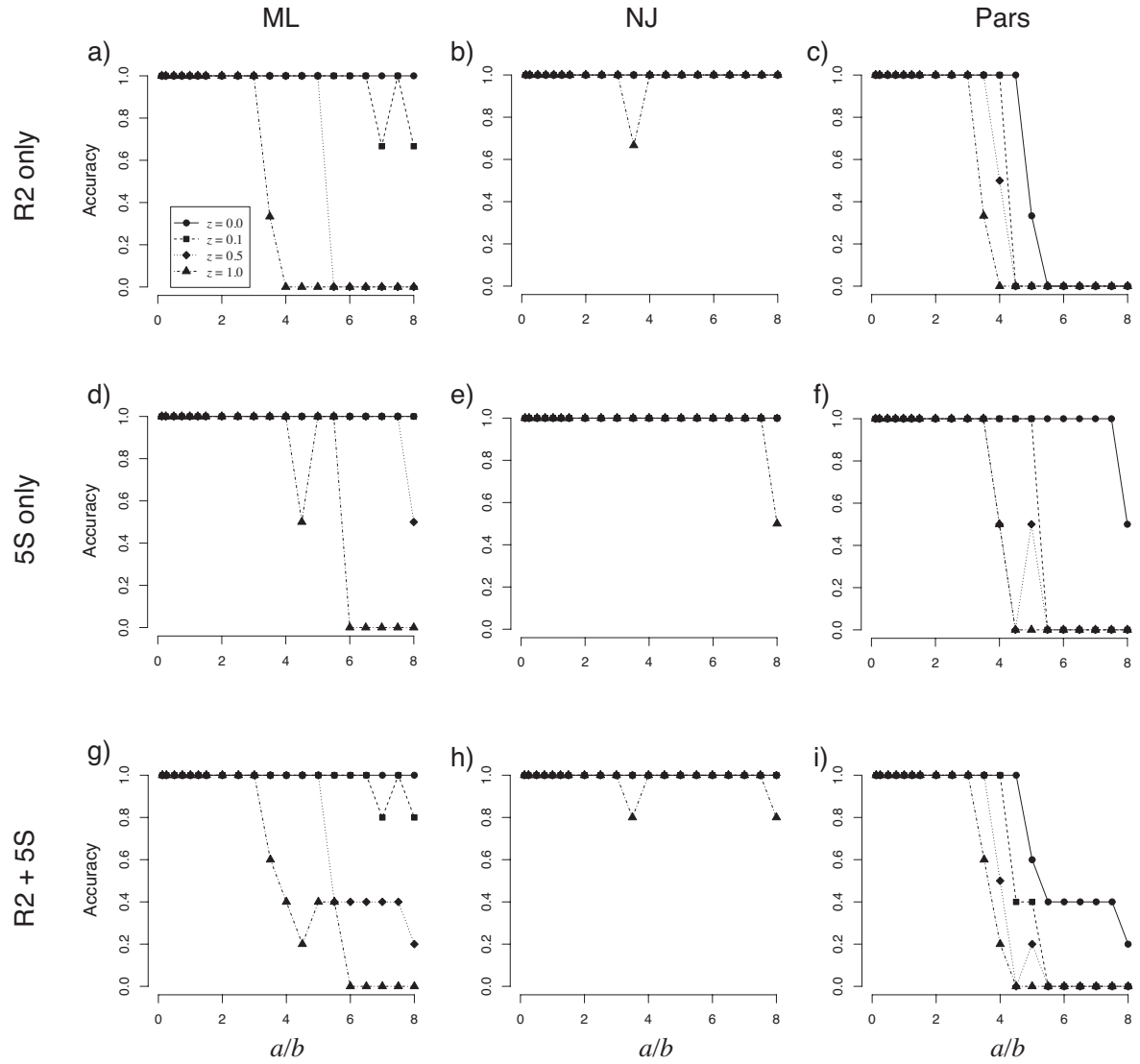
## Estimates of Dependence

Our simulations have shown that failure to account for dependence among sites, such as dependence due to structural constraints, can greatly impair inference of the underlying tree topology. We have shown this for a wide range of levels of dependence, but it would be useful to have a sense for what might be reasonable levels of dependence to expect in actual data. Yu and Thorne (2006) estimated the level of dependence due to secondary structure for a set of eight 5S rRNA sequences to be 0.3661. In our simulations we observe a significant impact on accuracy at lower levels of dependence than this ($z = 0.1$, see Fig. 1.5). This implies that failure to account for secondary structure of RNA may often lead to inaccurate inference of the true topology.

However, it is important to note that these kinds of models allow two methods of specifying the importance of structural constraint. One is an explicit level of dependence, as specified by the tuning parameters discussed here ($z$ for RNA, $p$ and $s$ for proteins). Another form of constraint is more implicit, namely how much flexibility is allowed in the structure. Here we have presented a model in which the (implicit) requirements of the structure are strict, but the (explicit) level of dependence has been varied. Yu and Thorne (2006), however, allowed more internal flexibility in the structures they examined. This implies that the explicit level of dependence in a model such as what we have presented might be lower than what Yu and Thorne (2006) presented because the implicit constraint is greater. How much lower is a reasonable question, and will be important in determining the level of decreased phylogenetic accuracy to be expected as a result.

For proteins the story is also complicated. While it is clear that there is dependence due to secondary structure in proteins (Thorne et al., 1996; Goldman et al., 1998) estimates of

Figure 1.11: Simulated RNA sequences were concatenated to form a few very long sequences, which were then analyzed using independence-assuming methods. R2: 3 long sequences ($\geq 30,000$ nucleotides). 5S: 2 long sequences ($\geq 45,000$ nucleotides). R2+5S: pooled analysis of all 5 long sequences. $V = 1.75$ for all panels. Maximum Likelihood and Parsimony results behave as with short sequences: they fail when the dependence is strong and the tree is difficult. Neighbor-Joining performs well in most cases, unlike with shorter sequences.



the level of dependence vary considerably. The model we have described and the model under which these estimates were obtained utilized similar levels of implicit flexibility, so we focus on the estimates themselves. Rodrigue et al. (2005) used a model that involved the same pair potentials we employ, but did not utilize solubility constraints. They estimated levels of dependence due to pairwise interactions to be in the range of $0.36 - 0.70$. Robinson et al. (2003) used a model that included both pair potentials and solubility, and obtained estimates of pairwise dependence an order of magnitude less than Rodrigue *et al.* $(0.028 - 0.038)$ while also estimating the dependence due to solubility $(0.88 - 0.95)$. The large difference in pairwise interaction estimates could be due to differences in the modeling of the pairwise interactions, or because the Rodrigue et al. (2005) model lacked solubility constraints. Choi et al. (2007) used the Robinson et al. (2003) model to estimate pairwise and solubility dependence for a wide range of proteins (Choi et al., 2007, Figure 1), which not surprisingly agree with the Robinson et al. (2003) estimates. The difference between the Robinson *et al.* / Choi *et al.* estimates and the Rodrigue *et al.* estimates is an important one. As we have shown, pairwise interactions of the level Robinson *et al.* describe have little effect on our ability to estimate the true topology in spite of our assumptions of independence. If however pairwise dependence is of the level Rodrigue *et al.* describe the impact on phylogenetic estimation is quite large.

## Use of Energy as Fitness

The use of the energy of a sequence on a particular structure is but one possible surrogate for the fitness of a sequence, and may have its limitations. It is possible, for example, that a given sequence might be able to fold well into many possible structures; that while a given sequence might have a low energy on the structure of interest, it might have an even lower energy on an alternate structure. This implies that this sequence would in reality spend more time folded in the alternative structure than the one of interest. In this case we might argue that the sequence energy itself is not a good proxy for the fitness of the sequence. A better surrogate for fitness in this case might be the probability that a sequence will fold into the desired structure. However, this would involve considering the energy of a sequence on all its possible structures, and as we are allowing the sequence itself to change this becomes computationally prohibitive, particularly as the sequence length increases.

## Additional Model Limitations

The model we have presented is one in which dependence among sites results from the existence of a structure that must be maintained in order to perform some function. One limitation to this is that we do not allow the structure itself to evolve along the tree. This might be reasonable for short phylogenetic distances, but the fact remains that even closely-related sequences vary widely in their structural homology across taxa. Accounting for variance in the structural constraints across the tree will be a challenge for future research.

Another way in which the kind of model we have described might be developed is to allow for more than one substitution at a time. Huelsenbeck and Nielsen (1999) developed a compound Poisson model that allows for this, and it might be a natural pairing with the type of model described here; evaluating the energy/fitness of a sequence two substitutions away is a straightforward extension. Allowing more than one substitution at a time might be particularly useful when the intrinsic constraints are very strong, enabling sequences to cross fitness valleys more easily.

## 1.4 Conclusions

We have shown that failure to account for dependence among sites due to secondary and tertiary structure can lead to inaccurate estimation of the underlying tree topology. This is particularly true when the dependence is strong, as may be the case with RNA, when the internal branch is relatively short, and when the overall tree length is large. These findings have direct implications for anyone interested in phylogenetic estimation or analyses dependent thereupon. We have also shown the effect is stronger than might have been expected under simpler models of dependence, such as considering paired RNA sites as one. This indicates that there is room for improvement in phylogenetic methods by accounting for the nature of the dependencies in the data. We have introduced the concept of an effective sequence length as an intuitive means of quantifying the effects of dependence, and have presented a general method of simulating data on phylogenetic trees under complex models of evolution.

While in this paper we have focused on RNA and protein structures to introduce the dependencies among sites, the findings here may extend to the general case in which there may be dependence among characters. Morphological characters, for example, may contain large amounts of dependence, although it may be much more difficult to model the particular nature thereof. But our findings that the presence of dependence in the data, if unaccounted for, may lead to error in phylogenetic estimation should hold regardless of how well we understand the nature of the dependence itself. This suggests that in cases where we may be unable to model the dependence, being able to simply detect the presence of dependence in the data might be valuable.

# Chapter 2

# The Dynamics of the Compensatory Substitution Process Within a Population

## 2.1 Introduction

Complex interactions among loci play an important role in many evolutionary processes. They have arisen in the context of adaptation as Sewall Wright's shifting balance theory (Wright, 1931, 1932), in speciation as Dobzhansky-Muller incompatibilities (Dobzhansky, 1936; Muller, 1939), and in the intramolecular evolution of proteins and RNAs whose structures are important for function. An understanding of the the dynamics of such dependent evolutionary processes is therefore applicable for a wide range of fundamental questions in evolutionary biology.

Often these problems are cast in the framework of a fitness landscape, with peaks corresponding to genotypes of high fitness and valleys to genotypes of low fitness. When the landscape is very flat and all genotypes of similar fitness different loci may appear to evolve independently. But when the landscape is more rugged and certain combinations of alleles are deleterious, the means and rate at which an entire population can cross from one peak to another become the central issues.

From a combination of alleles that are of high fitness, a mutation at either locus will result in a suboptimal combination of alleles and consequently a haplotype of low fitness. However, a second mutation can restore the combination of alleles to one of high fitness. An entire population moving from one fitness peak to another in this manner is called compensatory substitution.

Theoretical population geneticists have used simple but powerful models, typically using two loci and two alleles at each locus, to discern the rate at which such shifts can occur, as a function of the population size, the mutation rate, the the strength of selection against intermediates, and the frequency of recombination between loci. With only four

allelic combinations, the frequency of these haplotypes in a population becomes a difficult three-dimensional diffusion problem. The approach has uniformly been to simplify to a two-dimensional diffusion problem in order to obtain analytical results. Simulations were frequently used to verify results and explore problems for which analytic solutions could not be found.

Kimura (1985) was the first to show that compensatory substitutions could occur, using the simplest possible model of compensatory evolution with two linked loci and two alleles at each locus. Subsequent efforts have extended Kimura's results to models of greater complexity by allowing for: different fitnesses for the two "peak" haplotypes (Iizuka and Takefu, 1996; Michalakis and Slatkin, 1996), different fitnesses for the two deleterious intermediate haplotypes (Iizuka and Takefu, 1996; Stephan, 1996; Michalakis and Slatkin, 1996; Innan and Stephan, 2001), different mutation rates (Iizuka and Takefu, 1996; Stephan, 1996; Innan and Stephan, 2001), reversible mutations (Higgs, 1998; Innan and Stephan, 2001), and expanding to a two locus four allele model (Higgs, 1998). These models are shown in Figure 2.1.

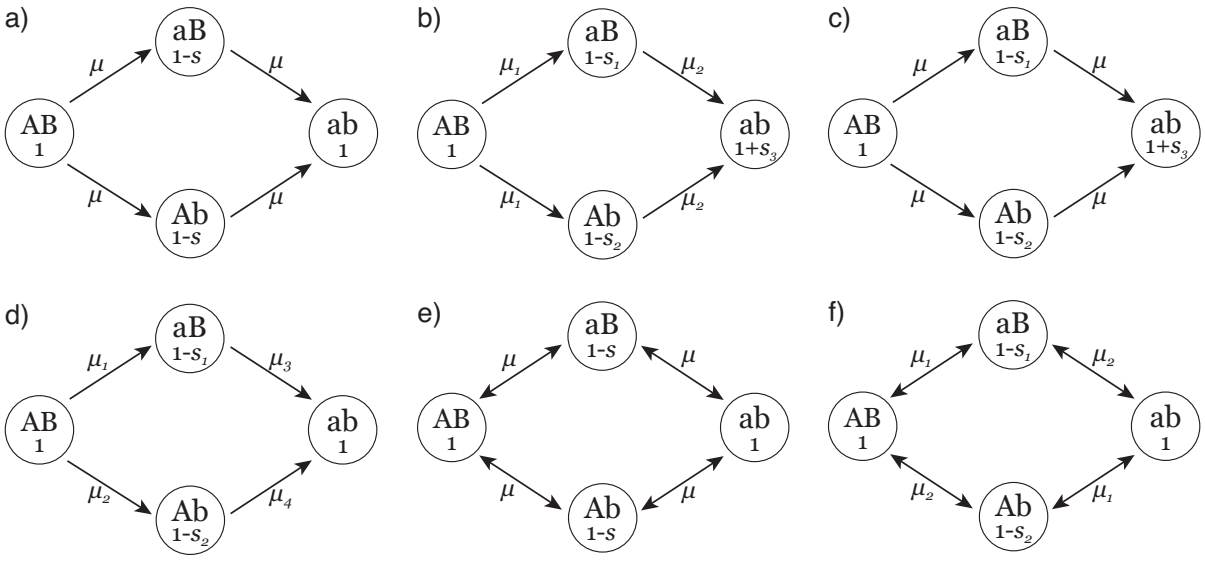



Figure 2.1: Mutation-selection models used previously for study of the compensatory substitution process in a population. (a) Kimura (1985) (b) Iizuka and Takefu (1996) (c) Michalakis and Slatkin (1996) (d) Stephan (1996) (e) Higgs (1998) (f) Innan and Stephan (2001)

Each of these studies made simplifying assumptions about the strength of selection and mutation, and in order to compare their results it is useful to know in what ranges of parameter space their results may be applicable. Relative statements to the effect of "when $\mu \ll s$" are fine for derivation purposes, but are not especially helpful when trying to compare approaches. The figures from these papers and the simulations used by the authors are perhaps better indicators of the range of parameter space in which these authors consider

their approaches reasonable. Figure 2.2 summarizes the approximate ranges of parameter space explored by each of these papers, based on their simulations and figures. This focuses on mutation and selection, both scaled by population size, although most of these papers explored recombination as well.

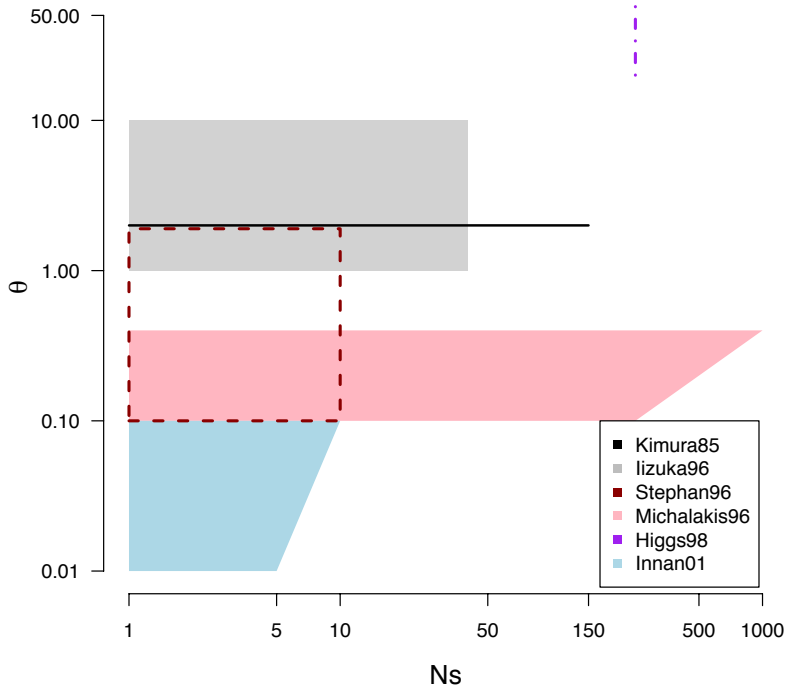



Figure 2.2: The parameter ranges explored by previous work on compensatory evolution.

The majority of this theoretical work assumed that selection was very strong, . But there are many cases in which selection is not expected to be strong. Rousset et al. (1991) observed a high frequency of deleterious intermediate states among homologous stem pairs of RNA between species, arguing this was evidence that selection on RNA intermediates could not be as strong as previously supposed. Innan and Stephan (2001) extended the theoretical work on compensatory evolution to the case of weak selection.

There are two pathways to compensatory substitution. One is observed when selection is very strong and the deleterious intermediate cannot become fixed in the population. In this case compensatory substitution can only occur if the second mutation arises before the first mutation is lost due to selection, and the new high-fitness double mutant may then drift to fixation. This kind of two-at-a-time substitution will be called a "Type 2" event (Fig. 2.3b). However, when selection is very weak there is some chance (particularly if the population

size is small) that the deleterious intermediate can become fixed before a second mutation arises and goes to fixation. This will be referred to as a "Type 1" event (Fig. 2.3a). When selection is of intermediate strength both of these substitution pathways are possible (Fig. 2.3c).

The dynamics of the compensatory substitution process for both Type 1 and Type 2 events is the focus of this work. Using simulation, the relationship between selection, mutation, recombination, and the probability of a deleterious intermediate fixing in the population before compensatory substitution is explored. This work also shows that the expected time to compensatory substitution depends upon the fixation pathway. These results yield insights into how fitness valleys can be crossed when selection is more moderate.

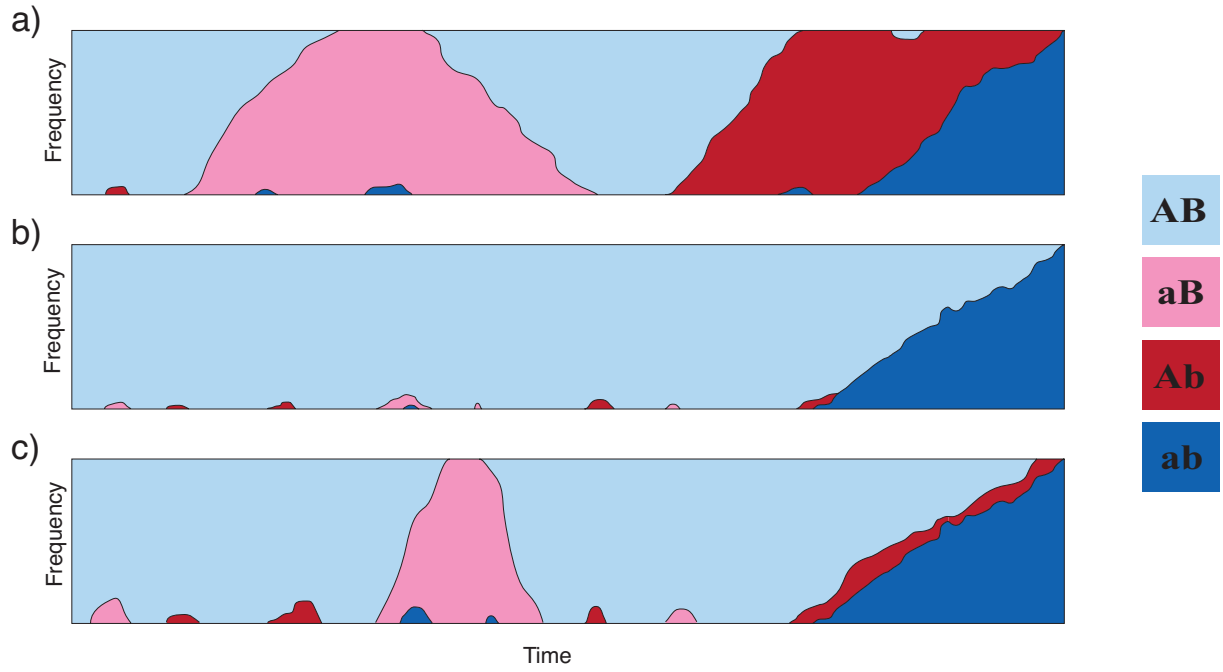



Figure 2.3: Illustrations of compensatory substitution pathways. (a) Type 1. Under weak selection the deleterious intermediate can readily fix in the population. (b) Type 2. Under strong selection the deleterious intermediates cannot rise above low frequency, and compensatory substitution must await the second mutation arising while the first is still polymorphic. (c) Under intermediate selection both Type 1 and Type 2 events are possible. Shown is a Type 2 event because while the initial state is lost before the deleterious intermediate during the final path to fixation of the double mutant, the intermediate does not fix.

## 2.2 Methods

The mutation-selection model used for simulation is the two-locus, two-allele model of Higgs (1998) as shown in Figure 2.1e. This is the simplest possible model of compensatory evolution that allows reversible mutation. In a population of $2N$ haploid individuals, reversible mutations occur at rate $\mu$ per locus per copy per generation.

The forward evolution of the population is simulated using Wright-Fisher sampling. In each generation mutations may be introduced, then recombinations may occur, and finally haplotypes are resampled with replacement with probability proportional to their fitness, before continuing to the next generation. The number of mutations occurring in the population is drawn as a Poisson($\mu$) random variable, where $\mu$ is the mutation rate. The haplotypes to be mutated are selected uniformly without replacement, and a mutation introduced on each haplotype at the first or second locus with equal probability. The number of recombination events in each generation is drawn as a Poisson($r$) random variable, where $r$ is the recombination rate. For each recombination event, two haplotypes are selected uniformly without replacement, recombined, and returned to the population.

The haplotype frequencies in the next generation are determined by sampling with replacement from the current population. This is done by drawing a multinomial random variable from the current allele frequencies, weighted by their fitnesses and normalized. Let $\boldsymbol{h} = (h_0, h_1, h_2, h_3)$ be the frequencies of the haplotypes $AB, aB, ab, Ab$ respectively, such that $\sum_i h_i = 1$. Let $\boldsymbol{w} = (1, 1-s, 1, 1-s)$ be the corresponding haplotype fitnesses, and $\boldsymbol{p} = (p_0, p_1, p_2, p_3)$ be the vector of sampling probabilities. The sampling probability $p_i$ of haplotype $i$ is

$$p_i = \frac{h_i w_i}{\sum_j h_j w_j}.$$

The allele frequencies in the next generation are then $\boldsymbol{h'} \sim \text{Multinomial}(2N, \boldsymbol{p})$. This multinomial random variable is itself drawn as a series of conditional binomial random variables (Davis, 1993). If $\mathcal{S}$ is the set of haplotypes sampled already in this generation, then $h'_i \sim \text{Binomial}(2N - K_\mathcal{S}, q_i)$ where $K_\mathcal{S} = \sum_{k \in \mathcal{S}} h'_k$ is the sum of the counts of the haplotypes sampled thus far and $q_i = p_i / \sum_{j \notin \mathcal{S}} p_j$ is the sampling probability of the current haplotype renormalized by the other haplotypes that are yet unsampled.

After each round of mutation-recombination-resampling the composition of the population is evaluated to determine the allele frequencies. Whenever a haplotype becomes fixed in the population that is different from the last fixed haplotype, the generation and haplotype are recorded. Each simulation begins with the population fixed for the $AB$ haplotype and ends when the population becomes fixed for the $ab$ haplotype. If the last fixed haplotype before the simulation is completed was $AB$, it is a Type 2 event (Fig. 2.3b). If the last fixed haplotype before completion was a deleterious intermediate ($aB$ or $Ab$), it is a Type 1 event (Fig. 2.3a).

Type 2 events encompass two scenarios: either the deleterious intermediate haplotype is lost from the population before the $AB$ haplotype, leaving only the two high-fitness haplo-

Table 2.1: Simulation Parameters

| Parameter | Symbol | Value |
|---|---|---|
| Mutation Rate | $\theta = 4N\mu$ / locus / generation | (0.001, 0.01, 0.1, 1.0) |
| Recombination Rate | $Nr$ / generation | (0, 5) |
| Selection Coefficient | $Ns$ | (0, 0.1, 0.2, ..., 3.0) |
| Population Size | $2N$ [haploid] | 200 |

types, or the deleterious intermediate persists until after the $AB$ haplotype is lost. While the latter case may result in one of the two new alleles becoming fixed before the other, as long as the intermediate haplotype does not fix it remains a Type 2 event (Fig. 2.3c).

Using this Wright-Fisher simulation strategy, the compensatory substitution process was examined for combinations of a range of mutation rates, recombination rates, and selection coefficients (Table 2.1). The simulation values chosen are similar to those of Innan and Stephan (2001), and when scaled by the population size reflect reasonable estimates for a variety of natural populations.

Five hundred simulation replicates were performed for each combination of parameters when $\theta \geq 0.01$. When $\theta = 0.001$ only 100 replicates were performed in the absence of recombination, and up to 100 replicates were performed in the presence of recombination. This was due to the very long computation times required under these conditions. Each simulation was analyzed to determine the overall time to compensatory substitutions (in generations), the type of fixation pathway observed (Type 1 or Type 2), and the number of and timing of any intermediate fixed states observed during the process.

## 2.3 Results and Discussion

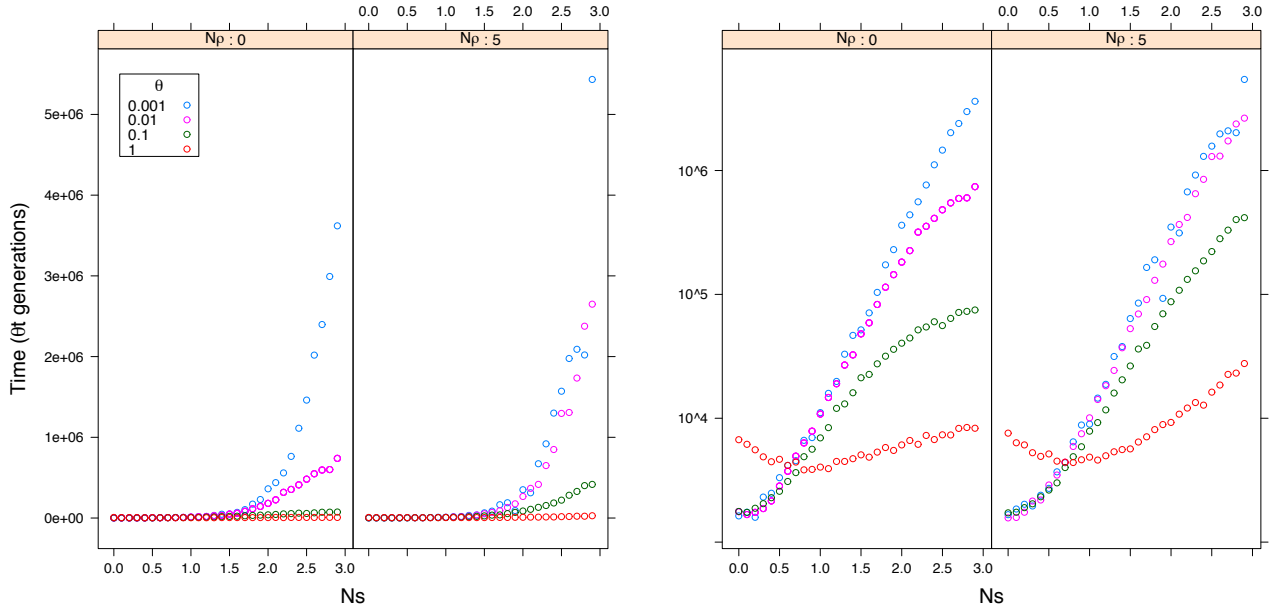
### Expected time to compensatory substitution

While the primary focus of this work is to elucidate differences between compensatory substitution pathways, the marginal results (regardless of path) should be consistent with prior analyses. Figure 2.4 shows the relationship between mean fixation time, selection, and recombination for different mutation rates ($\theta = 4N\mu$). The time to fixation is shown in units of $4N\mu$ generations.

As selection against the deleterious intermediates increases the expected time generally increases, and the combination of moderate to large selection coefficients and small $\theta$ results in the expected time to compensatory substitution being quite long. As expected, recombination also increases the time to fixation, although only noticeably so when selection is not too small. When the mutation rate is moderately large ($\theta = 1$) the expected time to compensatory substitution is a convex function with a minimum greater than 0, a pattern

noted by others as well (Kimura, 1985; Iizuka and Takefu, 1996). These results are highly consistent with those of Innan and Stephan (2001), who also explored relatively low mutation rates and selection coefficients.

These results indicate, perhaps counterintuitively, that mutation rate may play an even larger role than does selection in determining time to compensatory substitution, at least for relatively weak levels of selection. Humans for example, with an estimated mutation rate on the order of $10^{-8}$ (Xue et al., 2009) and an effective population size of only $10^4$ (Tenesa et al., 2007), might expect small amounts of selection to lead to long compensatory substitution times. Other organisms with higher mutation rates or larger population sizes might be able to compensate more quickly across a broader range of selection.




(a) (b)

Figure 2.4: Mean time to compensatory substitution. (a) Selection against intermediates greatly increases the expected time to fixation when *theta* is small. Recombination increases the time required, particularly when selection is strong. (b) Log transform reveals that when $\theta$ is large the time to compensatory substitution is a convex function.

## Two pathways to compensatory substitution

Most previous theoretical work on compensatory substitution has assumed that selection was very strong, and consequently that deleterious intermediate haplotypes could not become fixed in the population. But there are many cases in which selection might be expected to be weaker to moderate, or that population sizes are small enough to mitigate the strength of

selection. It is therefore of interest to understand the dynamics of the possible compensatory substitution pathways.

The probability that compensatory substitutions occur without the deleterious intermediate first becoming fixed (Type 2) is expected to be a function of selection, mutation, and recombination, all scaled by population size. The probability of a compensatory substitution being of Type 2 can be estimated by the proportion of simulation runs in which this is observed (Fig. 2.5). The proportion of Type 2 events is very small when selection is weak or absent and $\theta$ is small. As expected, this proportion increases with selection. Interestingly, when $\theta = 1$ the proportion of Type 2 substitutions is nearly 40% even in the absence of selection against the intermediate. This is the same parameter space in which the expected time to fixation was convex (Fig. 2.4), and is the level of polymorphism used by Kimura (1985) for the initial work on compensatory substitution.

Figure 2.5 also shows that Type 2 compensatory substitution events are very rare when the mutation rate is relatively low, meaning that the vast majority of compensatory substitution events involve the fixing of the deleterious intermediate haplotype. Type 2 events require having a second mutation enter the population while the first is still polymorphic in order to generate the *ab* haplotype. If the mutation rate is very low, then this may take longer than waiting for the unlikely fixation of a deleterious intermediate. If so, the expectation is that the overall time to fixation should be very large, and this is indeed what was observed (Fig. 2.4).

Recombination reduces the proportion of Type 2 substitutions observed, and this is fairly intuitive. : Consider when there is only one copy of each of the $a$ and $b$ alleles in the population and there is to be a recombination event. If the two alleles are on the same haplotype then recombination will break them apart if that haplotype is selected to be recombined. If the two alleles are on different haplotypes, then recombination can create a high-fitness *ab* haplotype, but only if *both* of the haplotypes bearing these alleles are selected. As a result recombination serves to break up *ab* haplotypes far more frequently than it assembles new ones, decreasing the probability of Type 2 events. Interestingly however, the effect of recombination is not observed when the mutation rate is high ($\theta = 1$), where Type 2 events dominate over a wide range of selective strength.

## A simple Markov model of compensatory substitution

A useful way to conceptualize the process of compensatory substitution is strictly in terms of the fixation events. Ignoring polymorphism for the moment, the transitions between fixed states of the population can be represented by a discrete-time three-state Markov chain (Fig. 2.6). It is important to stress that these state changes represent fixation events in the population; the gain and subsequent loss of an allele are represented in this model by simply not leaving the current state. The amount if time required for each fixation event to occur is not considered here, merely the state changes themselves.

Because simulations always begin as fixed for the $AB$ haplotype and end when the population becomes fixed for the *ab* haplotype, the fixed state *ab* is considered an absorbing

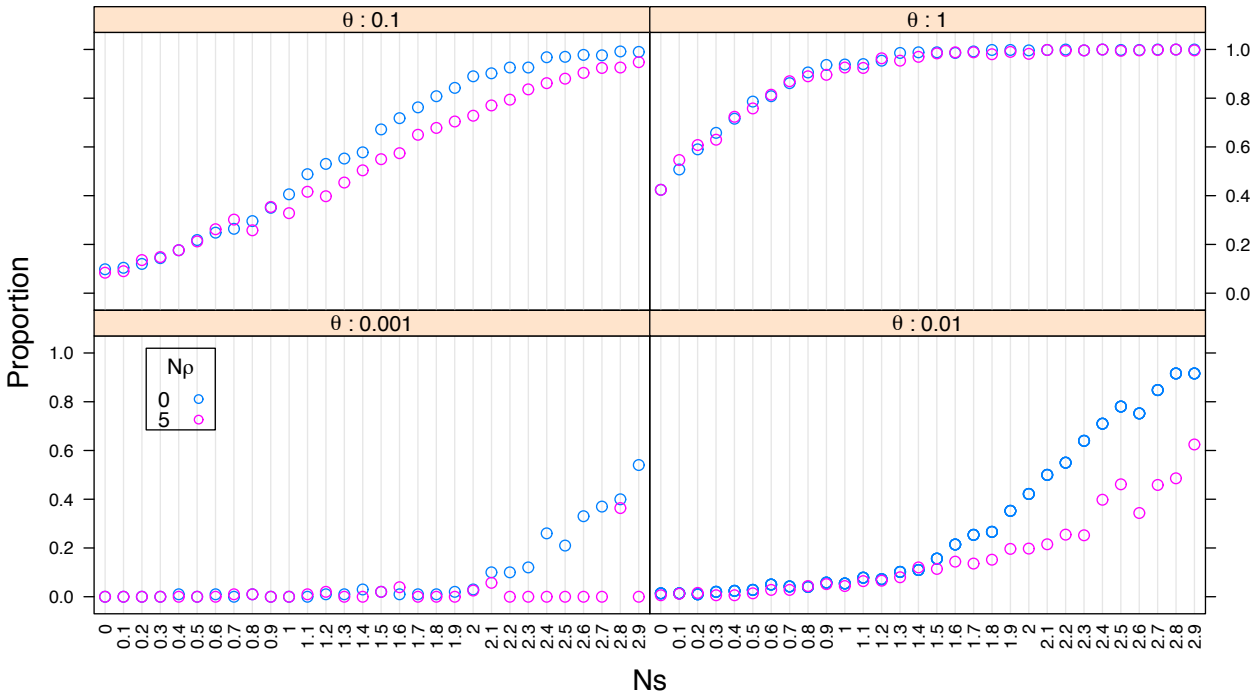



Figure 2.5: The proportion of Type 2 Compensatory Substitutions. As selection increases, the proportion of Type 2 events increases. Recombination decreases this proportion, as do lower mutation rates.

state. The deleterious haplotypes $aB$ and $Ab$, both of relative fitness $1 - s$ are considered together as a single state representing the population being fixed for one of the deleterious haplotypes. Because both the $AB$ and $ab$ haplotypes have fitness 1 and because mutation is symmetric, the probability of moving from state $aB/Ab$ to either $AB$ or $ab$ is $1/2$.

In this Markov model there is some probability $p$ of moving from state $AB$ directly to state $ab$. When $p = 0$ the compensatory substitution can only occur via the fixation of the intermediate, and when $p = 1$ the intermediate can never become fixed. However, this probability $p$ is not the same as the proportion of Type 2 substitution events shown in figure 2.5, but is a lower bound on that proportion. This is because it is possible to move first to the deleterious intermediate, return to the original state, then proceed to the final state directly without passing through the intermediate. This situation would be considered a Type 2 substitution event because the final substitution path did not include the deleterious intermediate becoming fixed.

The connection between this model and the proportion of figure 2.5 is as follows. Given that the population is in state $AB$, the probability that it moves directly to $ab$ without

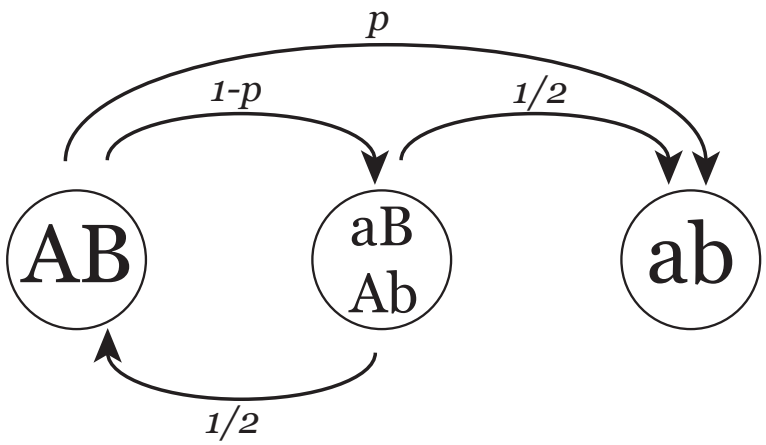



Figure 2.6: A discrete-time three-state Markov chain representing novel fixation events in the population for the different haplotypes during simulation. Haplotypes $aB$ and $Ab$ are grouped together as one state, representing the class of deleterious intermediates with the same fitness. The probability of moving from one state to another, conditional on making a move, are shown.

returning to $AB$ is $p$. The probability that it moves from $AB$ to $ab$ via the deleterious intermediate, without returning to $AB$ is $(1-p)/2$. Thus the relative probability that it moves from $AB$ to $ab$ directly before moving to $ab$ via the deleterious intermediate is

$$\frac{p}{p+(1-p)/2} = \frac{2p}{1+p}, \tag{2.1}$$

and this is the proportion of compensatory substitutions that are of Type 2. The population could of course have returned to $AB$ from the deleterious intermediate as well. But every time it does, it returns to the initial state, and the probabilities of reaching $ab$ via the two alternate paths from the initial state have just been described.

## Reversions to the ancestral state

The population can in fact leave the original state $AB$ for the intermediate state $aB/Ab$ and return again to $AB$ a number of times before eventually becoming fixed for the $ab$ haplotype, either via the intermediate or directly. Each time in state $AB$ the probability of "failure" (leaving then returning to $AB$) is $(1-p)/2$, and the probability of "success" (reaching $ab$ by any means) is $(1+p)/2$. The number of returns to $AB$ is therefore a Geometric$((1+p)/2)$ random variable on the space {0,1,2,...} and has an expected value of $(1-p)/(1+p)$. It should be noted the number of reversions to $AB$ does not depend on whether the final fixation occurs via the deleterious intermediate or not.

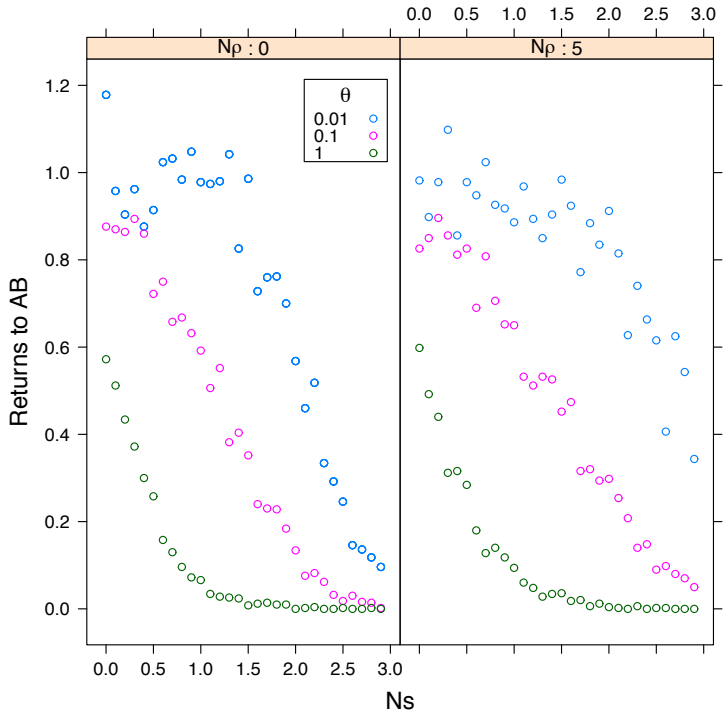



Figure 2.7: Average number of times returning to state $AB$ after having been fixed for a deleterious intermediate state. The process will return to state $AB$ some number of times before the $ab$ haplotype fixes, with or without the deleterious intermediate fixing en route. This number is a geometric random variable with success parameter a function of the relative probability of Type 2 substitutions.

Figure 2.7 shows the average number of times in simulation that the population fixes a deleterious intermediate state and then returns to state $AB$ before fixing in state $ab$ via any pathway. The number of returns to the initial state is higher when the mutation is low. This is because low mutation rates lead to lower probability of forming $ab$ haplotypes and therefore lowering the probability of fixing the $ab$ state sooner. Recombination's effect can be seen when selection is moderately large and mutation is low, preventing $ab$ haplotypes from being formed and providing more opportunities for reversion to the initial state. Increasing selection against intermediates reduces the number of reversions by removing deleterious alleles from the population; by preventing deleterious mutants from fixing in the population it also prevents reversions (as well as Type 1 fixation events).

Together with Figure 2.5 the dynamics of the substitution process may be understood. When selection is strong, the deleterious intermediates cannot fix. In the Markov model, this means that $p = 1$, and implies that the number of times returning to $AB$ should approach zero (Fig. 2.7) and that the proportion of Type 2 substitutions should approach 1 (Fig. 2.5).

In fact, increasing selection or mutation has the same general effect: an increase in $p$ and consequently an increase in the proportion of Type 2 substitutions and a decrease in the number of returns to state $AB$.

The number of returns to $AB$ can also be used to estimate the probability $p$ of direct double substitution. Let $x$ be the number of returns to $AB$. Since $x$ is a Geometric random variable with mean $\frac{1-p}{1+p}$, then the maximum likelihood estimate for the direct double substitution probability is

$$\hat{p} = \frac{1-x}{1+x}. \tag{2.2}$$

Estimates of this direct double-substitution probability are shown in Figure 2.8. These estimates offer insight into the fine-scale dynamics of how compensatory substitutions occur, as a function of selection, mutation, and recombination. More importantly, by focusing on the direct double-substitution probability this Markov model (Fig. 2.6 allows the reversions to be captured in the model. This is important for parameterizing substitution models used for phylogenetic inference, where the rates can be estimated and such reversions can be allowed. Some phylogenetic models of the compensatory substitution process (Tillier, 1994; Tillier and Collins, 1995) allow for double substitutions, but the single and double substitution rates are unrelated. The kind of approach outlined here could assist the development of new substitution models for phylogenetic inference that are more firmly grounded in population genetic principles.

## Relative time of the two pathways

An important question is whether the expected amount of time to compensatory substitution differs for Type 1 and Type 2 events. It is not obvious that these two events should take the same amount of time, or if differences between them should remain constant across a wide range of selection, mutation, and recombination. The times to fixation shown in Figure 2.4 do not consider whether events were of Type 1 or 2, and also include the time required for any reversions to state $AB$ before fixation (Fig. 2.7). These returns to $AB$ increase the overall time to fixation and reduce the power to detect any differences between the time required for Type 1 and Type 2 events. In order to compare the time required for these alternative fixation pathways, only the final path from $AB$ to $ab$ should be considered.

Recalling the Markov model of Figure 2.6, the quantity of interest is the number of generations from the last time the population enters state $AB$ until the time it enters state $ab$, and this quantity was calculated for each simulation run. If the expected time for Type 1 and Type 2 pathways was genuinely the same, then the difference between the mean times required for each path is expected to be zero. This difference in means, normalized by the variance, is shown in Figure 2.9. If there is truly no distinction between the two paths, then this normalized difference should be distributed $\sim$ Normal(0,1), and strong deviations from this reveal differences in the expected time for the paths.

This analysis reveals that under weak selection Type 1 events require a comparatively longer amount of time to occur, and this is particularly true when the mutation rate is high.

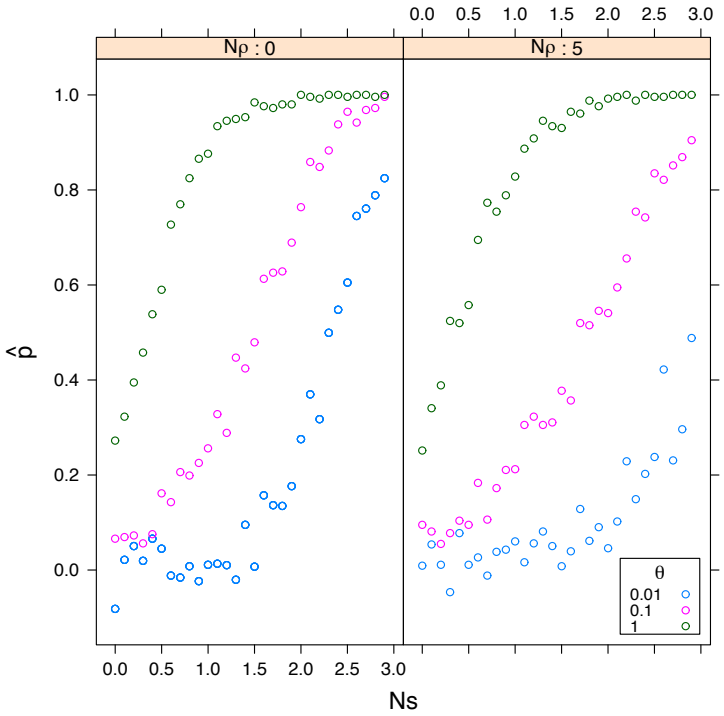



Figure 2.8: Estimates of the direct double-substitution probability $p$ from the observed number of reversions to the initial state. This probability serves as a lower bound on the probability of Type 2 substitution evnts, and could prove useful in the development of substitution models of compensatory substitution.

When selection is weak, Type 1 events are far more probable because the fixation barrier is low. But Type 1 events require two mutations and two fixation times, whereas Type 2 events require only one fixation time for the two mutations. So in the comparatively rare event that two mutations occur close in time, then they will fix relatively quickly. Increasing the mutation rate increases the probability of observing two such mutations in a close window and shortens the waiting time until it occurs, contributing to the increased difference between Type 1 and Type 2 events when the mutation rate is high.

The more subtle observation is that when selection is stronger there is little difference between the two paths observed, and this appears true across all mutation rates examined. With increased selection the probability of Type 2 events increases, but even though Type 1 events are less likely to occur under strong selection, when they do occur they take the same amount of time as Type 2 events. It is possible that if selection were yet stronger that the overall trend would continue and the time required for Type 1 events might be *less* than Type 2 events, but Type 1 events are so improbable at high levels of selection that they are simply never observed.

Recombination had little effect on either the differences observed under weak selection or the lack of differences observed at stronger selection. This suggests that the effect recombination has on the expected time to fixation occurs during the reversions to the initial state rather than the final compensatory substitution pathway itself.

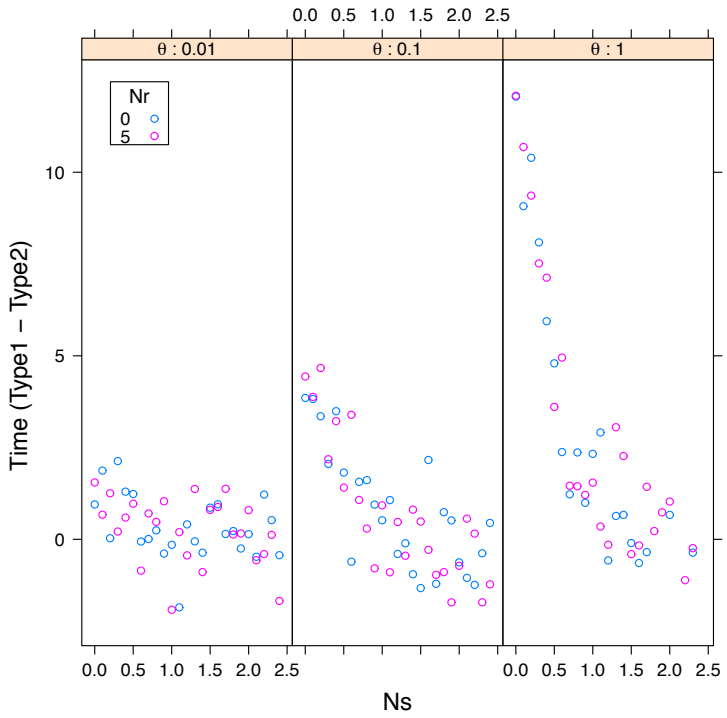



Figure 2.9: The normalized difference between Type 1 and Type 2 substitution events in the time required to enter state $ab$ since the final time we entered state $AB$. When selection is weak, Type 1 events take longer than do Type 2 events, particularly when $\theta$ is large. Recombination does not seem to affect this difference.

# Chapter 3

# A Phylogenetic Model of Compensatory Evolution Accounting for Population Genetic Dynamics

## 3.1 Introduction

RNA plays a central role in molecular biology, acting as an information carrier (mRNA), translator (tRNA and rRNA), as well as regulatory molecule (miRNA). Because it is largely conserved across the tree of life, rRNA is of central importance for use in phylogenetic inference. Metagenomic studies have revealed much about microbial biodiversity, often identified exclusively by differences in rRNA.

There is a strong relationship between the important functions performed by RNA and the physical structures formed as a result of the sequence folding upon itself; maintaining the correct structure is essential for preserving function. Consequently, the evolution of individual positions in the sequence depends on other sites as well, as it is interactions among sites that produce the unique helical structure of RNA.

The primary interactions required for RNA structure formation are Watson-Crick pairings between sites not immediately adjacent in the primary sequence. A mutation at one of these positions will disrupt the bond and negatively affect the structure. However, the bond can be restored by a second "compensatory" mutation at the other position. The process of compensatory substitution has been well-studied in a population genetic context (Kimura, 1985; Iizuka and Takefu, 1996; Michalakis and Slatkin, 1996), often in the specific case of RNA (Stephan, 1996; Higgs, 1998; Innan and Stephan, 2001). The signature of compensatory evolution can often be observed in alignments of distantly-related species, and served as the basis for early, accurate predictions of the secondary structure itself (Woese and Pace, 1993; Gutell, 1996).

Phylogenetic substitution models often assume that sites evolve independently of all others, which is clearly not valid for RNA. To account for these interactions, substitution

models specifically for RNA expanded the unit of evolution from the single nucleotide to the pair of nucleotides (doublet), but did not allow the possibility for more than one substitution at a time (Schöniger and von Haeseler, 1994; Muse, 1995; Rzhetsky, 1995). It was Tillier (1994) and subsequently Tillier and Collins (1995, 1998) who first allowed the possibility of direct transition from one Watson-Crick pair to another (which requires changes at two sites) in a single substitution event. Savill et al (2000) provide a thorough description of the large number of models developed and compare the fit of the various models, finding that models that allow such double substitutions fit empirical data much better than do models that do not allow double substitutions.

However, these are purely *descriptive* models, with parameters designed to account for patterns in the data. The goal here is to present an *interpretive* model for the evolution of RNA, in which parameters have meaning rooted in the process of evolution rather than simply fitting the pattern of the data. The model presented here not only allows the possibility of double substitutions, but is parameterized in such a way as to better connect to the population genetic processes that underlie compensatory substitution. The model employs a single parameter for the relative rates of double and single substitutions that has direct interpretability with respect to the strength of natural selection acting against non-canonical stem pairs. The model is implemented in a fully Bayesian statistical inference framework and is demonstrated using a dataset of eukaryotic 5S rRNA.

## 3.2 Model and Implementation

### Paired Substitution Model

The evolution of paired sites is modeled as a continuous-time Markov chain, where the state space is the sixteen possible two-nucleotide combinations, or doublets. This model therefore considers substitution events between different doublets. Let $x$ and $y$ be doublets, and adopt the convention that the first nucleotide of each doublet is the 5' and the second nucleotide the 3' position in the sequence. Define the exchangeability matrix between the four different nucleotides as

$$\mathbf{S} = \{s_{ij}\} = \begin{pmatrix} - & \alpha & \beta & \gamma \\ \alpha & - & \delta & \epsilon \\ \beta & \delta & - & \eta \\ \gamma & \epsilon & \eta & - \end{pmatrix} \tag{3.1}$$

and the stationary frequencies of the sixteen possible doublet states

$$\boldsymbol{\pi} = (\pi_{AA}, \pi_{AC}, \ldots, \pi_{GT}, \pi_{TT}).$$

Now letting $\mathcal{W} = \{AT, CG, GC, TA\}$ be the set of Watson-Crick pairs, the instantaneous rate matrix to describe changes from $x$ to $y$ is defined as

$$\mathbf{Q} = \{q_{x,y}\} = \begin{cases} u\pi_y s_{x_1 y_1} & \text{if } x_1 \neq y_1 \text{ and } x_2 = y_2 \\ u\pi_y s_{x_2 y_2} & \text{if } x_1 = y_1 \text{ and } x_2 \neq y_2 \\ u\pi_y s_{x_1 y_1} s_{x_2 y_2} d & \text{if } x_1 \neq y_1, x_2 \neq y_2, \text{ and } x, y \in \mathcal{W} \\ 0 & \text{if } x_1 \neq y_1, x_2 \neq y_2, \text{ and } x, y \notin \mathcal{W} \\ -\sum_{y \neq x} q_{x,y} & \text{if } x = y \end{cases} \tag{3.2}$$

where $u$ is a rate-scaling factor to ensure that branch lengths are interpretable in terms of average number of substitutions per site, and $d$ regulates the relative proportion of double to single substitutions. Gamma-distributed rate variation among pairs of sites is also incorporated, with shape parameter $\alpha$ (Yang, 1993). The set of parameters specific to the doublet substitution model are $\boldsymbol{\theta}_p = (\boldsymbol{\pi}, \mathbf{S}, d, \alpha)$.

Note that while the focus of this paper is the evolution of RNA sequences, evolution is actually occurring at the level of the DNA in which the RNA is encoded. Consequently, throughout this paper thymine (T) is referenced rather than uracil (U).

### Double Substitution Rate

The parameter $d$, representing the relative rate of double to single substitutions between doublets, is the primary parameter of interest in this model, and the intuition for this parameterization is as follows. Assume, as most phylogenetic models do, free recombination among sites and consider two independent sites $A$ and $B$. Mutations that are destined to go to fixation (substitutions) arise at each site at rates $\lambda_A$ and $\lambda_B$, respectively. While rare, it is possible for mutations destined to become fixed to arise at the same time at each site, and these joint events will occur at rate $\lambda_A \lambda_B$. Thus the rate of double substitution, under neutrality, should be roughly the product of the two individual site rates. The parameter $d$ allows this rate of double substitutions to increase above that allowed by the purely neutral process.

The double substitution parameter $d$ is similar in spirit to the nonsynonymous/synonymous rate ratio parameter $\omega$ of codon substitution models (Nielsen and Yang, 1998). When $d = 0$ then only single substitutions are allowed and the model collapses to the single-site model specified by the exchangeability matrix (GTR). When $d = 1$ the rate of double substitutions is equal to that expected under neutrality, namely that double substitutions arise at a rate proportional to the product of the individual single substitution rates. When $d > 1$ the rate of double substitutions is enriched relative to the expectation under neutrality.

It has been shown that for paired sites evolving within a population, the proportion of compensatory substitution events that occur two-at-a-time increases with the strength of selection against the deleterious intermediate (Chapter 2). This means that a value of $d > 1$

can be interpreted as evidence for negative selection against intermediates. Furthermore, it implies that the larger the estimated value of $d$, the stronger the negative selection has been.

An alternative that must be considered is that there has been positive selection for different doublets, an interpretation closer to that of $\omega$. Positive selection for new doublets, without negative selection against intermediates, would cause an increase in the overall rates of both single and double substitutions, but should not necessarily affect their proportion. This is because simply obtaining a new doublet does not require that those substitutions be made together. Thus, this interpretation can be discarded.

The combined effects of negative selection against deleterious intermediates and positive selection for new doublets should lead to very large estimates of $d$. Assuming negative selection is already sizable however, it may be difficult to distinguish the presence of positive selection from an additional increase in negative selection against intermediates. Again the signature may in the absolute rate of single substitution; positive selection should act to increase this rate in spite of selection against deleterious intermediates, whereas increased negative selection should reduce single substitution rates further.

## Unpaired Substitution Model

As with paired sites, evolution at unpaired sites is modeled as a continuous-time Markov chain, but on the state space of the four possible nucleotides. A general time-reversible (GTR) model is assumed (Tavaré, 1986) with gamma-distributed rate variation among sites (Yang, 1993). The unpaired process has a vector of the four nucleotide stationary frequencies $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$, and an exchangeability matrix $\mathbf{S}$ of the same form as that described in equation 3.1. The instantaneous rate matrix to describe changes from nucleotide $i$ to $j$ can be defined as

$$\mathbf{Q} = \{q_{i,j}\} = \begin{cases} u\pi_j s_{ij} & \text{if } i \neq j \\ -\sum_{j \neq i} q_{i,j} & \text{if } i = j \end{cases} \tag{3.3}$$

where $u$ is again a scaling factor such the branch lengths have meaning in terms of the expected number of substitutions per site. The parameters specific to the unpaired model are collectively $\boldsymbol{\theta}_u = (\boldsymbol{\pi}, \mathbf{S}, \alpha)$, where $\alpha$ is the shape parameter of the gamma distributed rate variation for unpaired sites.

## Implementation

### Alignment

An $n \times m$ alignment of RNA sequences $\mathbf{D}$ is assumed, where $n$ is the number of taxa and $m$ is the number homologous sequence positions in each taxa. This alignment is treated as observed data rather than as a random variable. The model further assumes a known structure of the RNA that defines the pairing of RNA stem positions. Using this structure, the alignment is partitioned into two subsets: one for paired ($\mathbf{D}_p$) and one for unpaired ($\mathbf{D}_u$)

positions. The structure is assumed to be shared among all species and to remain constant over the course of evolution.

### Phylogeny

All species are assumed to be related via an unrooted bifurcating phylogenetic tree. Let $\tau$ represent the tree topology, and let $\boldsymbol{v} = \{v_1, v_2, \ldots, v_{2n-3}\}$ represent the vector of $2n-3$ independent branch lengths of the phylogeny. In some cases the topology and branch lengths will be assumed to be known, and in others they will be considered as random variables.

### Shared Parameters

It would be possible to analyze the paired and unpaired partitions completely independently, or for these partitions to share certain parameters. In my implementation the two alignment partitions always share the same topology and branch lengths. It would be possible for other parameters to be shared as well, but while it seems reasonable to assume that sequences share the same species tree, it is a stronger assumption to make that the exchangeability parameters themselves should be the same between paired and unpaired partitions. Consequently, these parameters are not shared in my current implementation. It could be done sensibly through inclusion of an additional scaling factor on the length of the tree were introduced to allow for proportionally differential rates between partitions.

### Statistical Model

Define a character as either an unpaired site or a paired doublet. The likelihood of the alignment for a particular character is the probability of that data given all of the parameters of the model. To do so this requires the transition probability matrices for each branch of the tree. For both the paired and unpaired models, the transition probability matrix between states after some branch length $v$ is obtained in standard fashion by matrix exponentiation

$$\mathbf{P}(v) = e^{\mathbf{Q}v}. \tag{3.4}$$

Using these transition matrices, the likelihood for a particular character can be calculated from the tips to the root of the tree using the sum-product algorithm, also known as the pruning algorithm (Felsenstein, 1981), to calculate the likelihood at each node conditional on everything above it.

If all characters are assumed to be independent of all others, then the likelihood for each data partition can be calculated as the product of the individual character likelihoods,

$$P(\mathbf{D}_p \mid \boldsymbol{\theta}_p, \tau, \boldsymbol{v}) = \prod_{c=1}^{m_p} P(\mathbf{D}_p^{(c)} \mid \boldsymbol{\theta}_p, \tau, \boldsymbol{v}), \tag{3.5}$$

for the paired-sites partition, and

$$P(\mathbf{D}_u \mid \boldsymbol{\theta}_u, \tau, \boldsymbol{v}) = \prod_{c=1}^{m_u} P(\mathbf{D}_u^{(c)} \mid \boldsymbol{\theta}_u, \tau, \boldsymbol{v}), \tag{3.6}$$

for the unpaired sites partition, where $m_p$ and $m_u$ are the number of paired and unpaired characters in each partition. Because all characters, and consequently all partitions on characters, have been assumed to be independent, the likelihood of the full alignment is the product of the individual partition likelihoods given in equations 3.5 and 3.6.

A fully Bayesian framework for parameter estimation is adopted, with each parameter considered a random variable with an associated prior and posterior probability density. Letting $\boldsymbol{\theta} = (\boldsymbol{\theta}_p, \boldsymbol{\theta}_u, \tau, \boldsymbol{v})$ be all of the parameters of the model, the joint posterior probability of the model parameters is

$$P(\boldsymbol{\theta} \mid \mathbf{D}) = \frac{P(\mathbf{D} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})}{P(\mathbf{D})}, \tag{3.7}$$

where $P(\boldsymbol{\theta})$ is the set of prior probabilities of the model parameters and $P(\mathbf{D})$ is the marginal probability of the data.

### Prior Distributions

The prior densities chosen for model parameters are shown in Table 3.1. The Dirichlet distribution is the conjugate prior to the multinomial distribution and a sensible prior for stationary frequencies and exchangeability parameters, and a flat Dirichlet distribution is relatively uninformative. For tree topology, either no prior knowledge of the topology is assumed, in which case a uniform prior on all trees is assumed, or it is assumed that the topology is known, which corresponds to a point density prior on the tree of interest. Branch lengths are drawn from independent, identically-distributed exponential distributions, and likewise are the shape parameters of the gamma-distributed rate variation, with exponential rate parameters of 40 and 2, respectively. In general, the prior densities described thus far are standard, relatively uninformative ones.

In advance of any analyses, the Cauchy(0,1) distribution on the interval (0, $\infty$), also known as a half-Cauchy, was selected as a prior density for the parameter $d$ controlling the relative rate of double to single substitutions. The Cauchy is a weakly informative prior density, and is a reasonable choice in this situation. It's large density close to zero befits our assumption that the rate should be small (neutral) unless the data show otherwise. However, because the possibility exists that predominantly double substitutions occur in the presence of large amounts of selection against intermediates, the parameter $d$ must in principle be able to take very large values. The heavy-tailed Cauchy will allow for this very naturally. A second, related justification for this choice is that the half-Cauchy can be derived as the ratio of two independent exponential random variables. This has been used previously as a prior on nonsynonomous-synonymous rate ratios and transition-transversion rate ratios (Huelsenbeck et al., 2006). While the double and single substitution rates are not truly independent processes, the parameter $d$ is certainly a ratio of rates and is similar in spirit.

Table 3.1: Prior distributions on model parameters

| Partition | Parameter | Symbol | Prior |
|---|---|---|---|
| Shared | Tree Topology (when not fixed) | $\tau$ | Uniform over all topologies |
| | Branch Lengths | $v$ | Each branch independent Exponential(2) |
| Paired | Doublet Frequencies | $\pi_p$ | Dirichlet(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1) |
| | Exchangeability | $S_p$ | Dirichlet(1,1,1,1,1,1) |
| | Double Substitution Parameter | $d$ | Cauchy(0,1) |
| | Rate Variation Shape | $\alpha_p$ | Exponential(2) |
| Unpaired | Base Frequencies | $\pi_u$ | Dirichlet(1,1,1,1) |
| | Exchangeability | $S_u$ | Dirichlet(1,1,1,1,1,1) |
| | Rate Variation Shape | $\alpha_u$ | Exponential(2) |

### MCMC

Markov chain Monte Carlo (MCMC) was used to sample the joint posterior distribution of all model parameters (Metropolis et al., 1953; Hastings, 1970). A Markov chain whose state space is the set of all model parameters, and whose stationary distribution is the joint posterior distribution of those parameters, was constructed. Periodic sampling of this Markov chain at stationarity therefore yields a sample from the posterior distribution, and the frequency with which the chain visits a particular parameter configuration is proportional to the joint posterior probability of the parameter states.

The chain moves among different possible parameter values. In each cycle a parameter is selected and a modified value of that parameter proposed. The ratios of new to old value prior probability, likelihood, and proposal probability are calculated. If the product of these $c \geq 1$, then the new parameter value is always accepted. If $c < 1$, it is accepted stochastically with probability $c$ (Metropolis et al., 1953; Hastings, 1970). A C++ computer program was implemented to perform this MCMC sampling on the model described.

Standard proposal mechanisms are used for each parameter. For stationary frequencies ($\boldsymbol{\pi}$) and exchangeability parameters ($\mathbf{S}$), new configurations are chosen as Dirichlet random variables with weights from the current state. New values of branch lengths ($\boldsymbol{v}$), double substitution parameter ($d$), and rate variation shape parameter ($\alpha$) are proposed by multiplying the current value by the factor $e^{z(\eta-\frac{1}{2})}$, where $\eta$ is a Uniform$(0,1)$ random variable and $z$ is a tuning parameter. Changes to the topology are using nearest-neighbor interchange or subtree pruning and re-grafting with equal probability.

Chains were run for 4 million update cycles in analyses in which the topology is a random variable, and for 2 million cycles in analyses in which the topology is fixed. The first 10% of the samples from runs were discarded as burn-in, and multiple chains were run for analyses using empirical data. The program TRACER (Rambaut and Drummond, 2007) was used for assessment of MCMC convergence.

## 3.3 Materials and Methods

### Empirical Data

An alignment of eukaryotic 5S rRNA sequences (113 species: 45 plants and 68 fungi) was downloaded from the 5S Ribosomal Database (Szymanski et al., 2002). The 5S rRNA structure is relatively conserved across the tree of life, and sequences are available for a wide variety of organisms, making it a good choice for assessing the adequacy of the model. The general eukaryotic 5S structure used for analysis is shown in Figure 1a of Szymanski et al. (2002), and consists of 37 pairs of sites and the remaining sites are unpaired. Although 5S rRNA sequences are typically around 119 nucleotides in length, the alignment includes gap characters for insertions present in some taxa, increasing the total alignment length to 140 nucleotides. Since the focus of this work is not alignment itself, all analyses performed here condition upon the alignment as if it were observed rather than estimated.

### Tree Topology

The primary goal of this work is the inference of the parameters of the evolutionary process, particularly the relative double substitution rate, not topological inference. Furthermore the 5S rRNA described, containing 113 taxa and only 103 characters (paired and unpaired), is insufficient for topological inference. For inference using this data the topology will be considered a random variable in order to obtain marginal estimates of the parameters of interest. For evaluation of the model however it will sometimes be convenient to condition on a particular topology. This is to ensure that the same parameter space is being explored, given finite computation time. It is important to note that this tree need not be the "true" tree in any sense, nor need it be an especially good one, because it will only be used for internal model controls.

To obtain a tree upon which to condition, the 5S rRNA dataset was analyzed using maximum likelihood in the program PAUP* 4.0b10 (Swofford, 2003). For this analysis base pairings were ignored and all sites assumed to be independent. The model of nucleotide substitution assumed was the HKY85 model (Hasegawa et al., 1985), with gamma-distributed rate variation (Yang, 1994). The gamma distribution was discretized into four categories and the shape parameter was assumed to be 0.5. Stationary frequencies were estimated using the empirical base frequencies and the transition-transversion rate ratio was assumed to be 2. A heuristic search was performed for 115 hours to optimize the topology and branch lengths, using neighbor-joining to generate a starting tree Saitou and Nei (1987) and using tree bisection and reconnection as a swap mechanism. One of the 4174 high-likelihood trees found was selected at random, is shown in Figure 3.1, and will be referred to subsequently as "the ML tree."

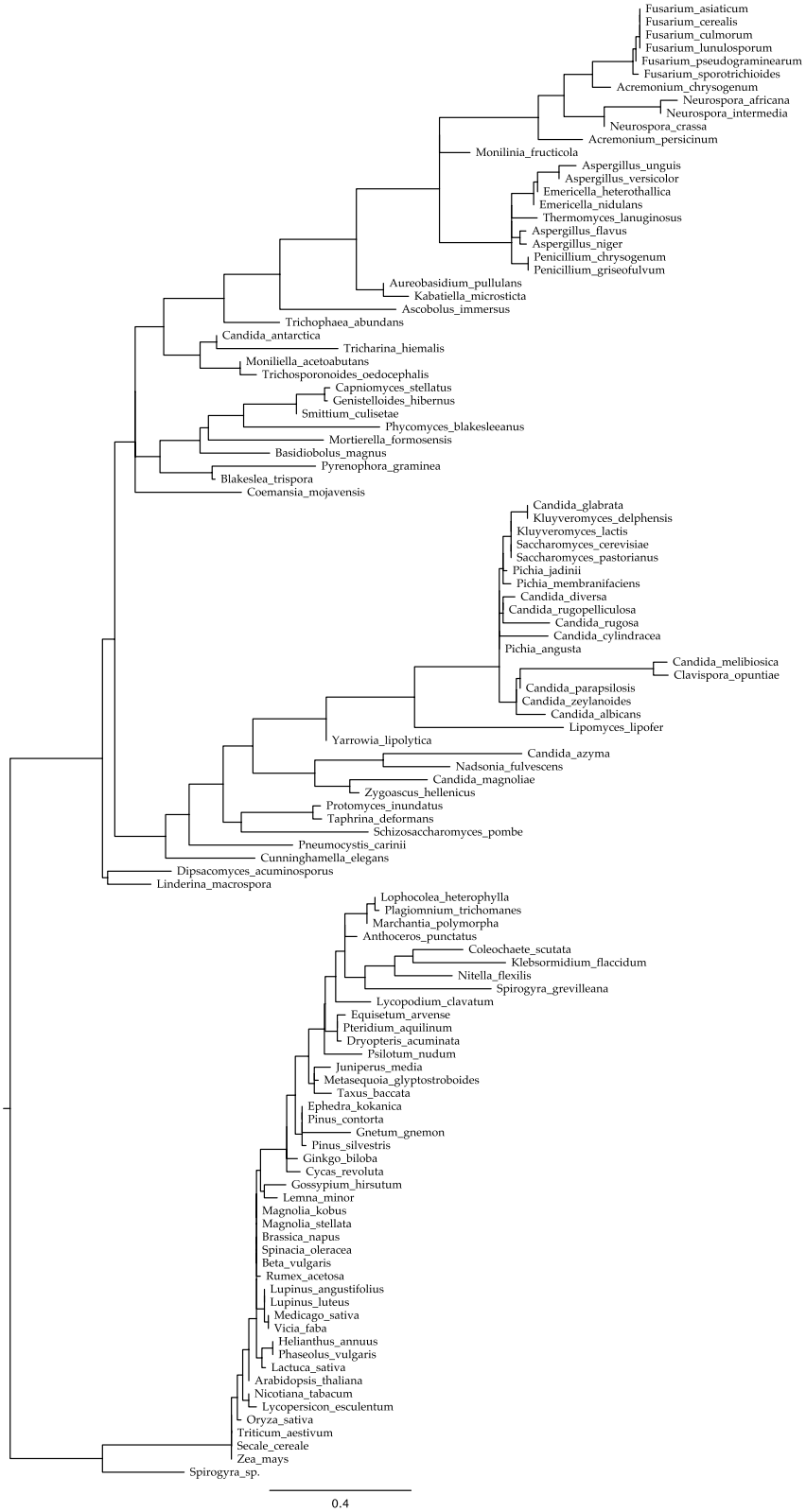



Figure 3.1: A maximum likelihood tree used when conditioning on a fixed topology.

Table 3.2: Alignment Simulation Parameters

| Parameter | Value |
|---|---|
| $\boldsymbol{\pi_p}$ | (0.015, 0.025, 0.011, 0.15, 0.0175, 0.019, 0.175, 0.014, 0.014, 0.22, 0.01, 0.06, 0.14, 0.04, 0.065, 0.0245 ) |
| $\boldsymbol{\pi_u}$ | (0.31, 0.24, 0.23, 0.22) |
| $\mathbf{S_p} = \mathbf{S_u}$ | (0.1, 0.2, 0.1, 0.1, 0.2, 0.1) |
| $\alpha_p = \alpha_u$ | 1.0 |
| $d$ | {0, 1, 2, 3, 4, 5, 10, 20} |

## Simulated Data Generation

A computer program was implemented for the generation of simulated datasets under the paired and unpaired substitution models described. Simulation of alignments is done in the following manner. The rate matrices for the paired and unpaired models are specified by equations 3.2 and 3.3, respectively. A Gamma(0.5) distribution was discretized into four categories with equal probability. Each character draws randomly from these four categories a scalar by which the branch lengths are multiplied, introducing rate variation among sites. The transition probability matrix for each branch is calculated using equation 3.4 and the adjusted branch lengths for that character. Internal and terminal node states are then determined via a preorder tree traversal, in which the state of the current node is drawn stochastically given the state of the ancestral node and the transition probability matrix for that branch.

Simulated alignments of the same dimensions and pairing structure as the empirical 5S rRNA dataset were generated using the ML tree and branch lengths (Fig. 3.1). Table 3.2 summarizes the values of substitution model parameters used for simulations. These parameter values are close to the maximum likelihood values inferred using the empirical data, and may be considered as reasonable for RNA datasets. Table 3.2 also shows the range of values of the double substitution rate parameter used to generate simulated datasets with varying levels of double substitutions.

## Bayes Factor Calculation

Model comparison was accomplished by the calculation of Bayes factors (Kass and Raftery, 1995). The comparisons of interest are the special cases when double substitutions are disallowed ($d = 0$) and when double substitutions occur only at a neutral rate ($d = 1$). Because the model will only be compared against models nested within it, Bayes factors are estimated using the Savage-Dickey ratio (Verdinelli and Wasserman, 1995; Suchard et al.,

2001). If model $M_0$ is nested within model $M_1$, then the Bayes factor in favor of $M_1$ is estimated by the ratio of the prior and posterior densities evaluated at the constrained value,

$$B_{10} = \frac{P(d = d_0 \mid M_1)}{P(d = d_0 \mid \mathbf{D}, M_1)}. \tag{3.8}$$

Approximating the posterior distribution of $d$ was done by first noting that its form resembled that of a Gamma distribution, then finding the values of the shape and rate parameters of the Gamma distribution that maximized the likelihood of the posterior samples of $d$. It should be noted that the Gamma distribution with shape parameter not equal to 1 has zero density when evaluated at 0. The density evaluated at $d = 1 \times 10^{-10}$ is used to approximate the density at $d = 0$.

## 3.4 Results

### Analysis of Simulated Data

Assessment of the accuracy with which the method can infer the relative rate of double substitutions requires first knowing the true relative rate of double substitutions. While this is not possible using actual data, it is using simulation. To that end a computer program was implemented for the simulation of sequences on a phylogenetic tree.

Simulated alignments were generated of the same dimensions and pairing structure as the empirical 5S rRNA dataset using the ML tree (Fig. 3.1). A range of values of the double substitution rate parameter $d$ were used, and the other model parameters were fixed at values shown in table 3.2. Seven datasets were created for each value of the double substitution rate. The values chosen are close to the maximum likelihood values inferred using the empirical data, and may be considered as reasonable for RNA datasets.

For each simulated alignment the values of all model parameters were estimated using the inference method described. The true ML topology was assumed as given because here the interest is whether or not the model can estimate the parameters of the molecular evolutionary process. The distribution of posterior means of the double substitution rate $d$ inferred from these simulated datasets is shown in Figure 3.2. The method is able to well-estimate the true value of the double substitution rate and can detect subtle increases above neutrality, even using the very limited data of the 5S RNA structure (only 37 paired characters). Consequently a high degree of confidence can be had in the method's ability to estimate double substitution rates elevated above neutral, and that truly neutral rates can be estimated as such.

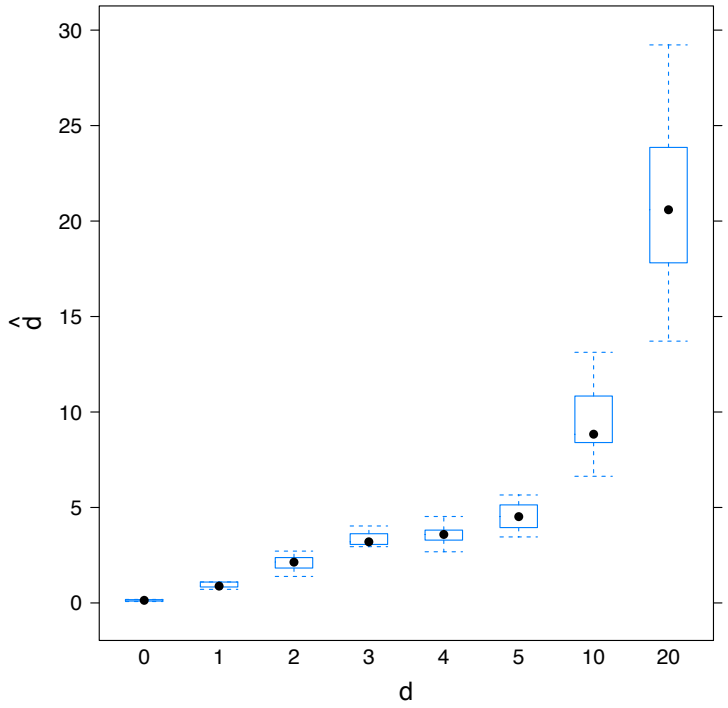



Figure 3.2: Estimated values of the double substitution parameter $d$ on simulated 5S RNA datasets. Shown is the distribution of the posterior means from several simulated datasets for each value of the true parameter.

## Analysis of 5S rRNA Data

### Parameter Estimates

The 5S rRNA dataset was analyzed in two ways: by allowing the topology to be a random variable and by fixing the topology to that of the ML tree. The estimates of the double substitution rate parameter $d$ are shown in Figure 3.3. The posterior mean of $d$ when the tree is free is 9.724, and the 95% highest posterior density interval is (5.261, 14.956). When the topology is constrained to that of the ML tree the posterior mean is 7.591 with 95% highest-posterior density interval (4.110, 11.763). While these two estimates are noticeably different, they are both significantly higher than expected by chance, indicating strong support for the frequent occurrence of double substitutions in this dataset and consequently the presence of strong selection acting against deleterious intermediates.

Certain model parameters such as stationary frequencies and rate variation parameters are typically insensitive to the topology used. This does not appear to be the case for the double substitution rate parameter, but this might not be altogether unexpected. The ML tree was obtained by maximizing the likelihood under a single-substitution-only model

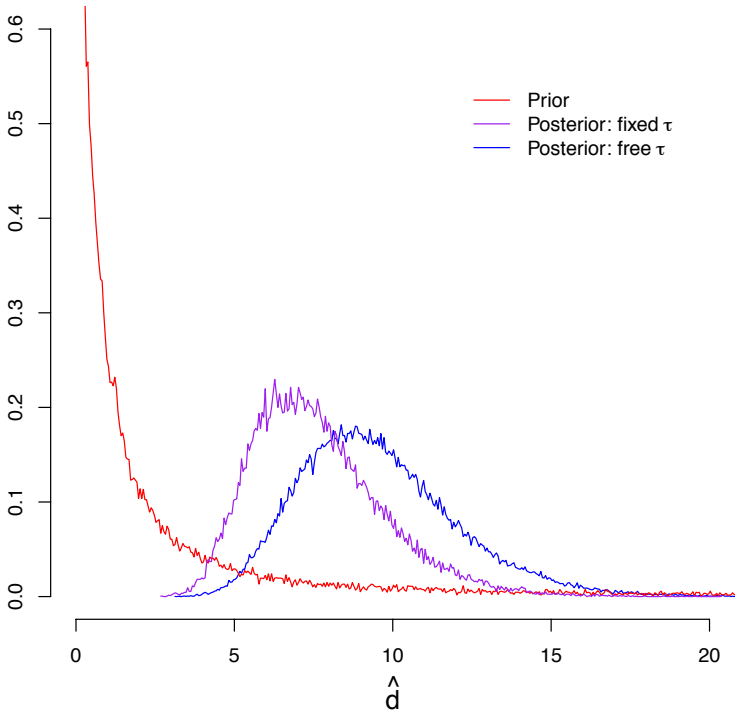



Figure 3.3: Posterior estimates of the double substitution parameter $d$ on the 5S rRNA dataset. A strong signature of negative selection against deleterious intermediate pairings is observed. The double substitution rate is underestimated when a topology unlikely to be sampled in the posterior is conditioned upon.

with independent sites. It is plausible that such a tree might rarely be sampled in the posterior distribution of trees under the model allowing double substitutions. To test whether the dependence of estimates of $d$ on the topology is particular to the ML tree or more general, several trees were sampled at random from the posterior distribution of trees. These topologies were then conditioned upon and all other model parameters re-estimated. The estimates of $d$ obtained were highly similar to those obtained when also estimating the topology (data not shown). Thus the variance in parameter estimates of $d$ might only be noticeable when the tree used is one unlikely to be sampled in the posterior. Furthermore, such differences in estimates were also observed for the exchangeability parameters of the unpaired loop partition $\mathbf{S_u}$, indicating that the tree topology's effect is on rates in general and not specific to the double substitution parameter. It is therefore recommended that the tree be considered a random variable in analyses, or at least that the tree upon which the analyses are conditioned be obtained using additional datasets.

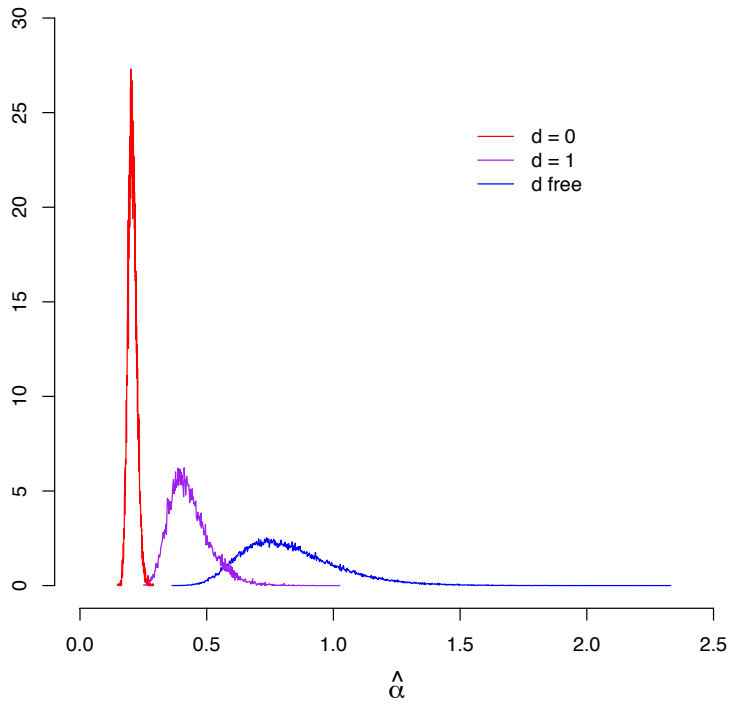



Figure 3.4: Rate variation among sites decreases when double substitutions are permitted. Show are the posterior distributions of the $\alpha$ shape parameter of the gamma-distributed rate variation under different constraints on the double substitution rate. Small values of $\alpha$ correspond to large amounts of rate variation among sites. $\alpha$ increases (rate variation decreases) when the rate of double substitution is unconstrained compared to when double substitutions occur at a neutral rate or are disallowed.

Estimates of the double substitution rate are also positively correlated with estimates of the rate variation shape parameter $\alpha$ for the paired sites, and consequently negatively correlated with the amount of rate variation itself (Fig. 3.4). When the model is constrained to $d = 0$, disallowing double substitutions, estimates of the $\alpha$ shape parameter are small, implying there is a great deal of rate variation among sites. When $d = 1$, the neutral case, the estimates increase, and when $d$ is unconstrained and takes on larger values, $\alpha$ increases yet again, implying a reduction in the rate variation among sites. While accounting for rate variation among sites is essential in studies of molecular evolution (see Yang, 1996), approaches to doing have largely been descriptive. In this case, at least some of the rate variation among sites observed has been implicitly accounted for by employing a more realistic substitution model.

### Model Comparison

There are two comparisons of interest for this model that allows double substitutions. One is to a model that disallows such substitutions ($d = 0$). The other is to a model that allows double substitutions, but only at a rate expected if the evolution of doublets is neutral with respect to natural selection ($d = 1$). Both of these represent special cases of the model in which the double substitution rate can vary.

Bayes factors (Kass and Raftery, 1995) comparing the general model to each of the two specific cases were computed using the Savage-Dickey ratio (Verdinelli and Wasserman, 1995; Suchard et al., 2001). The posterior distribution of $d$ was fit to a Gamma(15.25, 1.5995) distribution, and this density, the sampled posterior, and analytical priors are shown in Figure 3.5. Bayes factors were then calculated using equation 3.8 by taking the ratio of the prior and the posterior densities, both evaluated at the fixed point of interest. Intuitively, if the prior density is much higher than the posterior evaluated at $d_0$, then the data have moved the posterior away from the constrained value. Inspection of Figure 3.5 reveals that the prior density is far greater than the posterior at both $d = 0$ and $d = 1$. Twice the natural log of the Bayes factors comparing the general model against the cases in which $d = 0$ and $d = 1$ are 692.7 and 38.3 respectively. Given that a value of $2\log B$ greater than 10 is considered decisive evidence (Jeffreys, 1961), there is overwhelming evidence that 5S rRNA are better described by a model that allows double substitutions, and allows them at a greater than neutral rate.

### Prior Sensitivity Analysis

While the choice of a Cauchy(0,1) distribution as a prior on the double substitution rate parameter was well-motivated, as described above, it is essential to demonstrate that the prior choice does not exert undue influence on the nature of the posterior distribution. The 5S rRNA dataset was analyzed using a series of different prior distributions on the double substitution rate. Parameters were estimated on the fixed ML topology in order to better constrain the parameter space. The marginal posterior distribution of the double substitution rate for each choice of prior distribution is shown in Figure 3.6, with each compared to sampling under the prior itself. An overly informative prior with little density in the plausible ranges of the double substitution rate (Fig. 3.6d) can be seen to markedly shift the posterior estimates, and such strongly informative priors on unreasonable ranges should clearly be avoided. However, prior distributions with density in the appropriate ranges all seem to perform similarly, even on the short 5S rRNA data, indicating reasonable robustness to prior specification.

## Permutation Analysis

It is important to know that the model will not infer a rate of double substitution when one is not present. In the case of RNA, the dataset itself can be used to test the model's sensitivity

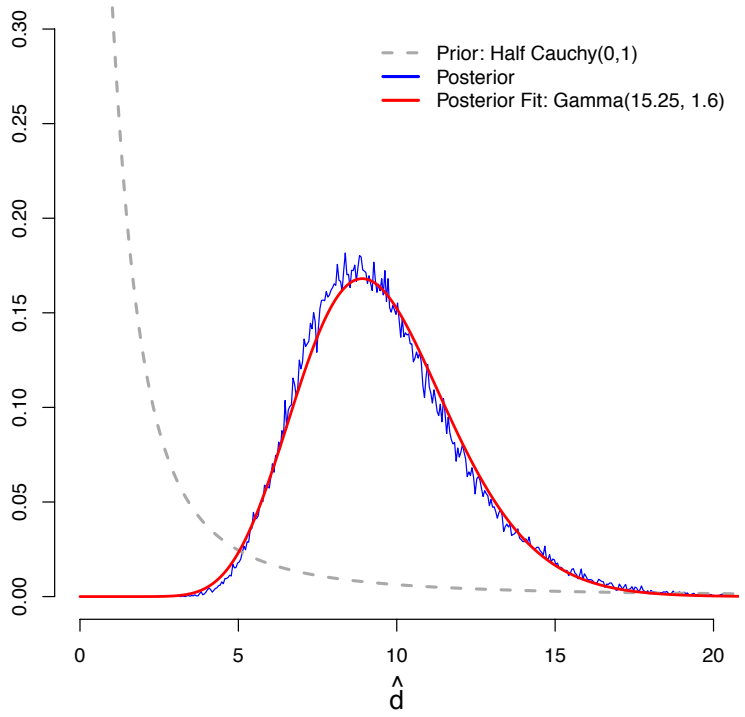



Figure 3.5: Savage-Dickey approach to calculating Bayes factors for nested models. The sample from the posterior of the parameter $d$ is fit to a gamma distribution. The analytical prior density and fitted posterior density are evaluated at $d = 0$ and $d = 1$ to calculate the Bayes factors comparing the general model to each of these two special cases.

to false positives. This is because the inference of a double substitution rate depends on the paired structure of the RNA. By altering the pairing assignments of the columns of the alignment, new datasets can be created that contain only the original columns of the alignment and preserve the overall nucleotide frequencies, but should display less of a pattern indicative of double substitution. Permuted datasets are expected to have an inferred value of the double substitution rate markedly less than that of the true alignment.

Alignment permutations can be obtained either by shuffling the original columns with respect to the fixed structure, or by redefining the structure itself. Ten permuted datasets were created by sampling without replacement pairs of sites from the set of individual sites that are paired in the original structure. In this manner no columns are used more than once and unpaired sites remain unpaired, allowing them to serve as an additional control. These permuted datasets were analyzed assuming the same fixed tree as before (Fig. 3.1), and the posterior distributions of the double substitution rate parameter are summarized in

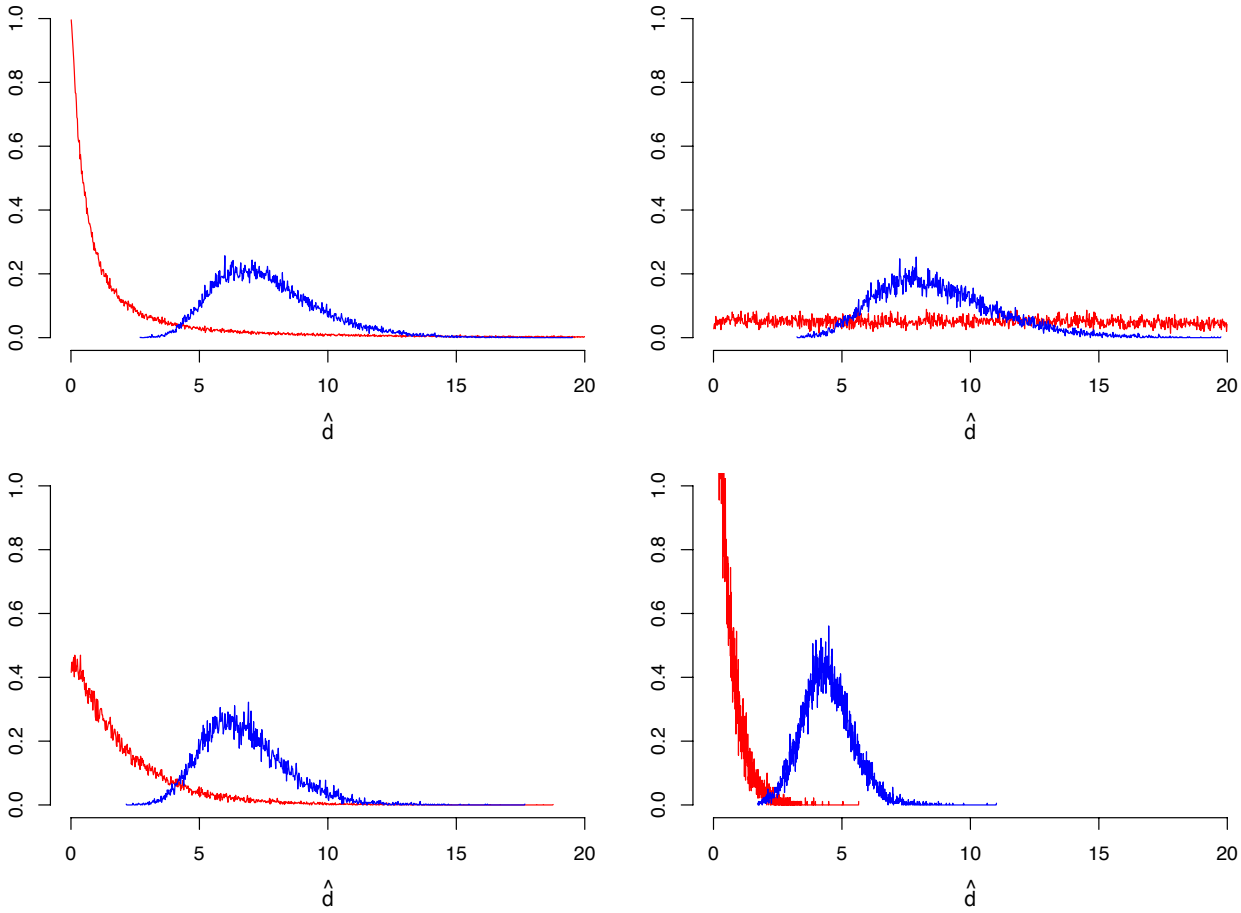



Figure 3.6: Alternative prior distribution choices for the double substitution rate parameter. The prior distributions are shown in red, and the posterior distributions in blue. ($a$) Cauchy(0, 1) ($b$) Uniform(0, 20) ($c$) Exponential(0.5) ($d$) Exponential(2)

Figure 3.7. As expected, the posterior distributions inferred for most permuted datasets fall somewhere between 0 and 1, and are uniformly much lower than the true data.

Random permutations of a highly structured data such as RNA stems can produce datasets that retain many characteristics of the original, leading to an inferred rate greater than neutral but still less than the original. This is observed in the fourth permuted alignment (Fig. 3.7), which contains an elevated level of Watson-Crick pairings relative to all other permutations and to that expected by chance (data not shown). These permutation tests are all in concert with expectations of the method, and with the simulation results described above provide ample evidence that the method is capable of distinguishing genuine rates of double substitutions and can generally avoid false positives.

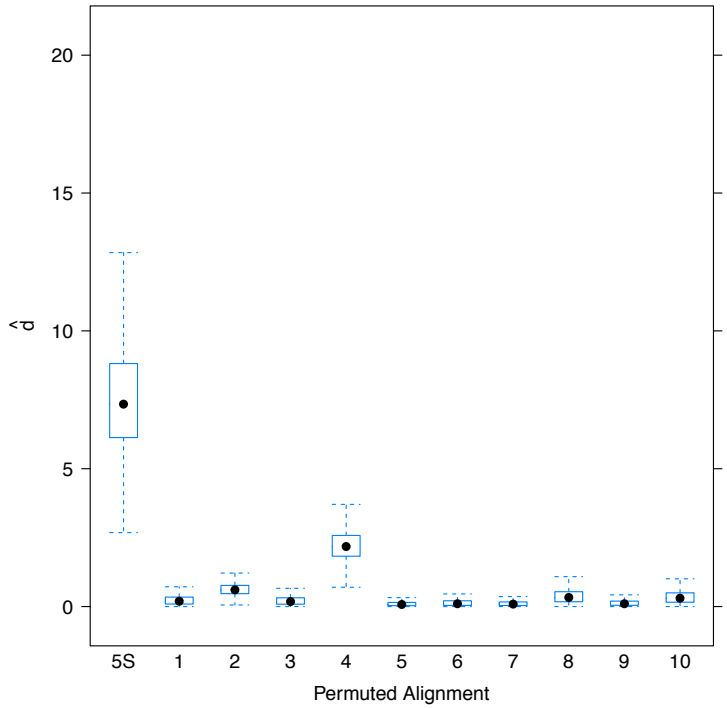



Figure 3.7: Posterior distributions of the double substitution rate using the 5S rRNA alignment and ten permutations of that alignment. Permuted alignments, in which the pairings are misspecified, provide a negative control and produce markedly lower estimates of the double substitution rate.

## 3.5 Discussion

The unique properties of rRNA, such as the high degree of structural conservation paired with a relatively low degree of sequence conservation, make them well-suited for studies of how natural selection affects the evolution of primary sequence. These analyses can only be done however with the appropriate statistical tools. Nielsen and Yang (1998) showed that positive selection could be detected in protein-coding DNA using evolutionary substitution models, leading to numerous studies detecting the effects of selection acting on protein-coding genes. Such tools for the direct detection of natural selection have been lacking for RNA, and the model presented here seeks to provide researchers with such a tool with which to investigate the role of natural selection acting on various kinds of RNAs.

Because of the characteristic helical structure of RNA, the epistatic interactions between paired sites are of primary importance in constraining their evolution. Theoretical population geneticists, interested in the properties of such epistatic interactions, have studied the evolution of compensatory evolution specifically in the case of RNA (Stephan, 1996; Higgs,

1998; Innan and Stephan, 2001). The insights into the dynamics of the evolutionary process provided by these studies can serve as the basis for an evolutionary model that accounts for interactions in a meaningful way.

The model presented here has focused on compensatory substitutions, and the relative occurrence of double and single substitutions between stem pairs. Other models for the evolution of RNA that allow double substitution have previously been developed (Tillier, 1994; Tillier and Collins, 1995, 1998), and these models have been shown to better fit RNA datasets than do models that disallow double substitutions (Savill et al., 2001). But the model presented here involves several key differences, making it useful not only for the improved estimation of phylogeny, but for better understanding the role natural selection has played in the evolutionary history of the species. Additionally it is a step towards bridging the gap between the population genetic processes that generate sequence change and the patterns observed from interspecific data.

Unlike previous models (Tillier, 1994; Tillier and Collins, 1995), here the rate of double substitution is a function of two quantities: the rates of single substitutions for the bases changing and the relative rate parameter $d$. It is important that these two be separated in order to connect the model to the population genetic processes. The single substitution rate from state $i$ to state $j$ can be interpreted as the rate at which, in a population fixed for type $i$, mutations of type $j$ arise that are destined to go to fixation. This means these rates should be a function of the mutation rate and the effects of natural selection. If natural selection acts against new mutations by keeping them at low frequency or by removing them altogether, then these rates will be lower. Simultaneous double substitutions require first having two mutations in order to form a compensatory haplotype that can the rise to fixation via genetic drift. Therefore, if natural selection acts strongly against single substitutions it will also affect the total rate of double substitutions, since the waiting time until two mutations are present is increased. Since the total double substitution rate depends on the single substitution rates, it will be affected accordingly.

As natural selection acts against deleterious intermediates the total rate of both single and double substitutions is expected to decrease, but the *relative* rate of double substitution is expected to increase. When selection is very strong there is little chance of a single deleterious mutation going to fixation, but it may remain in the population long enough to be combined with a second compensatory mutation, which then may fix neutrally. So the proportion of compensatory substitutions that occur via this method, or the relative rate, will increase as selection against intermediates increases (Chapter 2). This is the role of the double substitution rate parameter $d$, allowing the relative rate of double substitutions to increase independently of the decreases observed from the single substitution rates.

The relative rate $d$ takes on direct meaning, since it is expected to increase only in the presence of natural selection against deleterious intermediates. Under neutrality the relative rate should be no more than 1, and will increase as selection is applied. With such a neutral expectation, the presence of natural selection acting on stem sites can be tested directly. The idea of comparing rates of double and single substitutions is not unique to this work; Tillier and Collins (1998) had compared estimated rates of double to single substitutions

under their model as well, but the parameterization of the model presented here allows such comparisons implicitly and gives them specific meaning with respect to what is known about the population genetics of compensatory substitution.

When applied to the 5S rRNA dataset, not surprisingly there is a clear signature of natural selection acting against intermediates that might disrupt the structure. The ability to use the model itself to draw conclusions about the process of compensatory evolution is appealing, particularly in a Bayesian context, because it allows uncertainty to be incorporated. For example, Meer et al. (2010) examined the fitness landscape of tRNA compensatory substitutions, but did so by conditioning on a fixed topology and estimating ancestral states via maximum likelihood, using only pairs of sites with unambiguous patterns. The sensitivity of estimates about compensatory substitution to topology have been demonstrated (Fig. 3.3), and ancestral states could be integrated over rather than conditioned upon. Both of these, as well as the utilization of all of the data (of varying ambiguity) can be incorporated into a Bayesian analysis such as that presented here.

## Model Limitations and Extensions

The inferences of the model must be conditioned on the various assumptions made. The assumption that all sequences must share the same structure (or that homology is preserved) is a fairly reasonable one, but the assumption that the structure remain constant across the entire tree is a stronger assumption. If the structure is not constant across the tree then the strength of natural selection inferred should be an underestimate, as evidenced by the permutation analyses. The model has also assumed that the only interactions among sites are between the paired sites. This is certainly not the case; stacking interactions between adjacent positions are also important (Walter et al., 1994). Accounting for such neighbor interactions would add a great deal of complexity to the model, requiring a model of the entire sequence as the evolutionary unit to encompass all of the overlapping interactions (Robinson et al., 2003; Yu and Thorne, 2006).

Another potential objection to the model presented is that it considers all non-canonical pairings to be equally deleterious. Pairings between guanine and thymine (uracil) are more stable than other non-canonical pairings, and can at times appear to have been conserved over interspecific timescales (Rousset et al., 1991). The stationary frequencies of G-T doublets are typically inferred to be higher than those of other non-canonical doublets, and that was observed in these analyses as well. These elevated stationary frequencies result in increased rates of compensatory evolution via single substitution through a G-T intermediate. And since the double substitution rates depends upon the single rates, the additional stability of the G-T pairs is indirectly incorporated into the double substitution rates as well. Consequently, the elevated stability of the G-T pairings are implicitly, if never explicitly, considered by the model presented.

The assumption of linkage equilibrium among all sites is highly unreasonable, since many RNA are relatively short. Linkage, even among neutral sites, has a profound effect on the pattern of substitutions, creating a process that is no longer Poisson and an overdispersed

pattern of fixation events (Watterson, 1982). However, nearly all phylogenetic substitution models make this assumption, often using sequences that are similarly short; linkage is not a problem specific to this particular model, rather a challenge for the field at large. That being said, population genetic theory has predicted the importance of recombination for compensatory evolution (Kimura, 1985; Stephan, 1996) and empirical studies have shown a relationship between the distance between two paired sites in the sequence and compensatory fitness interactions (Stephan and Kirby, 1993; Piskol and Stephan, 2008). Consequently, there may be a great deal to gain for considering linkage explicitly for the specific case of RNA.

The strength of natural selection may not be uniform across all pairs of sites within a particular structure, as assumed here. Certain stem positions, such as those near the ends of stems, have been shown to be more conserved than others (Piskol and Stephan, 2008) and have likely experienced stronger selection. A similar problem was addressed by Huelsenbeck et al. (2006), who implemented the model of Nielsen and Yang (1998) for the detection of positive selection among codons under a Dirichlet process prior, allowing the detection of positive selection at individual codon positions. Implementing the model described here under a similar nonparametric clustering method would be a natural extension of the current work and would allow such site-specific detection of selection strength at different stem positions.

# Bibliography

M. Anisimova and C. Kosiol. Investigating protein-coding sequence evolution with probabilistic substitution models. *Molecular Biology and Evolution*, 26:255–271, 2009.

U. Bastolla, J. Farwer, E. W. Knapp, and M. Vendruscolo. How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins*, 44:79–96, 2001.

T. A. Castoe, A. P. Jason de Konig, H.-M. Kim, W. Gu, B. P. Noonan, G. Naylor, Z. J. Jiang, C. L. Parkinson, and D. D. Pollock. Evidence for an ancient adaptive episode of convergent molecular eolution. *PNAS*, 106:8986–8991, 2009.

S. C. Choi, A. Hobolth, D. M. Robinson, H. Kishino, and J. L. Thorne. Quantifying the impact of protein tertiary structure on molecular evolution. *Molecular Biology and Evolution*, 24:1769–1782, 2007.

C. S. Davis. The computer generation of multinomial random variates. *Computational Statistics & Data Analysis*, 16:205–217, 1993.

T. Dobzhansky. Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila* pseudoobscura hybrids. *Genetics*, 21:113–135, 1936.

J. Felsenstein. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–411, 1978.

J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

J. Felsenstein and G. A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13:93–104, 1996.

N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11:725–736, 1994.

N. Goldman, J. Thorne, and D. Jones. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149:445–458, 1998.

R. R. Gutell. Comparative sequence analysis and the structure of 16S and 23S rRNA. In A. E. Dahlberg and R. A. Zimmermann, editors, *Ribosomal RNA: Structure, Evolution, Processing and Function in Protein Biosynthesis*, pages 111–128. CRC Press, 1996.

M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

P. G. Higgs. Compensatory neutral mutations and the evolution of RNA. *Genetica*, 102/103: 91–101, 1998.

J. P. Huelsenbeck and D. M. Hillis. Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, 42:247–264, 1993.

J. P. Huelsenbeck and R. Nielsen. Effect of non-independent substitution on phylogenetic accuracy. *Systematic Biology*, 48:317–328, 1999.

J. P. Huelsenbeck, S. Jain, S. W. D. Frost, and S. L. Kosakovsky Pond. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proceedings of the National Academy of Science, U.S.A.*, 103:6263–6268, 2006.

M. Iizuka and M. Takefu. Average time until fixation of mutants with compensatory fitness interaction. *Genes Genet. Syst.*, 71:167–173, 1996.

H. Innan and W. Stephan. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. *Genetics*, 159:389–399, 2001.

H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, 1961.

D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

D.T. Jones. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, 287:797–815, 1999.

T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–123. Academic Press, 1969.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

M. Kimura. The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics*, 64:7–19, 1985.

C. Kleinman, N. Rodrigue, C. Bonnard, H. Philippe, and N. Lartillot. A maximum likelihood framework for protein design. *BMCBI*, 7(326), 2006.

D. H. Mathews, A.R. Banerjee, D.D. Luan, T.H. Eickbush, and D.H. Turner. Secondary structure model of the rna recognized by the reverse transcriptase from the r2 retrotransposable element. *RNA*, 3:1–16, 1997.

D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288:911–940, 1999.

D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure. *PNAS*, 101:7287–7292, 2004.

M. V. Meer, A. S. Kondrashov, Y. Artzy-Randrup, and F. A. Kondrashov. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature*, 464:279–282, 2010.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

Y. Michalakis and M. Slatkin. Interaction of selection and recombination in the fixation of negative-epistatic genes. *Genet. Res.*, 67:257–269, 1996.

H. J. Muller. Reversibility in evolution considered from the standpoint of genetics. *Bio. Rev. Camb. Philos. Soc.*, 14:261–280, 1939.

S. V. Muse. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics*, 139:1429–1439, 1995.

S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome. *Molecular Biology and Evolution*, 11:715–724, 1994.

R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148:929–93, 1998.

K. Park, M. Vendruscolo, and E. Domany. Toward and energy function for the contact map representation of proteins. *Proteins: Structure, Function, and Genetics*, 40:237–248, 2000.

R. Piskol and W. Stephan. Analyzing the evolution of RNA secondary structures in vertebrate introns using Kimura's model of compensatory fitness interactions. *Molecular Biology and Evolution*, 25:2483–2492, 2008.

A. Rambaut and A. J. Drummond. Tracer v1.4. http://beast.bio.ed.ac.uk/Tracer, 2007.

D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*, 20:1692–1704, 2003.

N. Rodrigue, N. Lartillot, D. Bryant, and H. Philippe. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347:207–217, 2005.

F. Rousset, M. Pelandakis, and M. Solignac. Evolution of compensatory substitutions through g-u intermediate state in *Drosophila* rRNA. *PNAS*, 88:10032–10036, 1991.

A. Rzhetsky. Estimating substitution rates in ribosomal RNA genes. *Genetics*, 141:771–783, 1995.

N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

N. J. Savill, D. C. Hoyle, and P. G. Higgs. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics*, 157:399–411, 2001.

M. Schöniger and A. von Haeseler. A stochastic model and the evolution of autocorrelated dna sequences. *Molecular Phylogenetics and Evolution*, 3:240–247, 1994.

W. Stephan. The rate of compensatory evolution. *Genetics*, 144:419–426, 1996.

W. Stephan and D. A. Kirby. RNA folding in *Drosophila* shows a distance effect for compensatory fitness interactions. *Genetics*, 135:97–103, 1993.

M. A. Suchard, R. E. Weiss, and J. S. Sinsheimer. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution*, 18:1001–1013, 2001.

D. L. Swofford. *PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods*. Sinauer Associates, Inc., Sunderland, Massachusetts, 1998.

D. L. Swofford. *PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods. Version 4*. Sinauer Associates, Inc., Sunderland, Massachusetts, 2003.

M. Szymanski, M.Z. Barciszewska, V.A. Erdmann, and J. Barciszewski. 5S ribosomal RNA database. *Nucleic Acids Research*, 30:176–178, 2002.

S. Tavaré. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures in Mathematics in the Life Sciences*, 17:57–86, 1986.

A. Tenesa, P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and P. M. Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, 17:520–526, 2007.

J. Thorne, N. Goldman, and D. Jones. Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13:666–673, 1996.

E. R. M. Tillier. Maximum likelihood with multiparameter models of substitution. *Journal of Molecular Evolution*, 39:409–417, 1994.

E. R. M. Tillier and R. A. Collins. Neighbor joining and maximum likelihood with rna sequences: addressing the interdependence of sites. *Molecular Biology and Evolution*, 12: 7–15, 1995.

E. R. M. Tillier and R. A. Collins. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics*, 148:1993–2002, 1998.

M. Vendruscolo and E. Domany. Pairwise contact potentials are unsuitable for protein folding. *Journal of Chemical Physics*, 109:11101–11108, 1998.

M. Vendruscolo and E. Domany. Protein folding using contact maps. *Vitamins and Hormones*, 58:171–212, 2000.

I. Verdinelli and L. Wasserman. Computing Bayes factors using a generalization of the SavageDickey density ratio. *Journal of the American Statistical Association*, 90:614–618, 1995.

A. E. Walter, D. H. Turner, J. Kim, Lyttle M. H., Müller P., Mathews D.H., and Zuker M. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *PNAS*, 91:9218–9222, 1994.

G. A. Watterson. Substitution times for mutant nucleotides. *Journal of Applied Probability*, 19:59–70, 1982.

C. R. Woese and N. R. Pace. Probing RNA structure, function, and history by comparative analysis. In R. F. Gesteland and J. F. Atkins, editors, *The RNA World*, pages 91–117. Cold Spring Harbor Laboratory Press, 1993.

S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*, 1:356–366, 1932.

Y. Xue, Q. Wang, Q. Long, B. L. Ng, H. Swerdlow, J. Burton, C. Skuce, R. Taylor, Z. Abdellah, Y. Zhao, Asan, D. G. MacArthur, M. A. Quail, N. P. Carter, H. Yang, and C. Tyler-Smith. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology*, 19:1453–1457, 2009.

Z. Yang. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10:1396–1401, 1993.

Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39:306–314, 1994.

Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139: 993–1005, 1995.

Z. Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*, 11:367–372, 1996.

J. Yu and J. L. Thorne. Dependence among sites in RNA evolution. *Molecular Biology and Evolution*, 23:1525–1537, 2006.