

Abstract

MANNINO, FRANK VINCENT. Site-to-Site Rate Variation in Protein Coding Genes. (Under the direction of Spencer V. Muse)

The ability to realistically model gene evolution improved dramatically with the rejection of the assumption that rates are constant across sites. Rate heterogeneity models allow for better estimates of parameters and site specific inferences such as the detection of positive selection. Recently developed models of codon evolution allow for both synonymous and nonsynonymous rates to vary independently according to discretized gamma distributions. I applied this model to mitochondrial genomes and concluded that synonymous rate variation is present in many genes, and is of appreciable magnitude relative to the amount of nonsynonymous heterogeneity. I then extending this model to allow for the two rates to vary according to a dependent bivariate distribution, permitting tests for the significance of correlation of rates within a gene. I present here the algorithm to discretize this bivariate distribution and the application of the model to many real data sets. Significant correlation between synonymous and nonsynonymous rates exists in roughly half of the data sets that I examined, and the correlation is typically positive. These data sets range over a wide group of taxa and genes, implying that the trend of correlation is general. Finally, I performed a thorough investigation of the statistical properties of using discretized gamma distributions to model rate variation, looking at the bias and variance in parameter estimates. These discretized distributions are common in modeling heterogeneity, but have weaknesses that must be well understood before making inferences.

SITE-TO-SITE RATE VARIATION IN PROTEIN CODING GENES

by

Frank Vincent Mannino

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

BIOINFORMATICS

Raleigh

2006

APPROVED BY:

Chair of Advisory Committee

Dedication

To my wife Andrea, whose support made this possible.

Biography

Frank Mannino was born on May 10, 1980 to Vincent and Grace Mannino in Patchogue, NY. He lived his entire childhood in Farmingville, NY on Long Island. Frank was active in school with orchestra and tennis, and earned the rank of Eagle Scout in June of 1998. Frank graduated from Sachem High School in 1998 as salutatorian of his class. He continued on to the University of Pennsylvania. While there, Frank was a founding father of the Mu colony of Phi Sigma Kappa. During the summers, he interned at Brookhaven National Laboratory which provided some of the building blocks for his research and computing skills. Frank received his Bachelor's degree in Mathematical Biology in the spring of 2001 with a double minor in mathematics and psychology.

Frank then received a National Science Foundation Training Grant from North Carolina State University in 2001 to pursue his PhD in Bioinformatics. Frank has had a well rounded graduate experience. He has attended bioinformatics retreats, national conferences, was president of the Bioinformatics Graduate Student Association and has participated in the University's intramural sports program and played in a community tennis ladder. Frank also had the privilege to study twice at the University of Tokyo with Hirohisa Kishino. Frank's research has been focused on molecular evolution under the direction of Spencer Muse and he will graduate in 2006. During his graduate program, Frank married his high school sweetheart Andrea in North Carolina.

Acknowledgements

I'd like to start by first thanking my committee, Spencer Muse, Jeff Thorne, Bruce Weir and Bill Atchley, for their help along the way. Specifically I'd like to thank Spencer for his patience and guidance with my research. Many people have helped my progress towards my PhD, whether it was scientific advice or someone to go out and have a beer with. While it would be impossible to list everyone, the following people were most important. Sergei Kosakovsky Pond helped with many programming problems. Hirohisa Kishino opened his vast knowledge to me and showed great hospitality during my trips to Tokyo. Errol Strain was always available for advice on bioinformatics or my pool game. David Aylor was always willing to help with any problem I encountered or just discuss ACC sports. Andrea Johnson helped tremendously with countless programming tips and trivia nights. Doug Robinson provided me with useful job search advice, a golf partner, and most important, lots of his beer.

I'd like to thank the people associated with Bioinformatics Research Center for always doing their best to make everything run smoothly, specifically Juliebeth Briseno, Alex Rogers, and Debbie Hibbard. They shared knowledge and food, both of which were essential for my research. To the people responsible for the Genomic Sciences graduate program, Barbara Sherry, Bruce Weir, and Stephanie Curtis, I wish to express my thanks for the opportunities I have been afforded in my pursuit of a PhD.

I'd like to thank my entire family, specifically my parents, Grace and Vincent. Their support for the last 26 years has allowed me to be in the position I am today. And lastly, I'd like to thank my wife Andrea. No matter how frustrated I got, she

was always there to support me and I could not have done this without her.

Contents

LIST OF TABLES	x
LIST OF FIGURES	xi
1 INTRODUCTION	1
Introduction	2
Dissertation structure	5
Data	6
Models	7
Markov chains	7
Nucleotide Models	9
RNA models	11
Amino Acid Models	12
Codon Models	13
Likelihood Function	15
Rate Variation	16
How to model rate variation	16
Why rate variation should be included	22
Discretizing a Continuous Distribution	23

Gamma model extensions	25
Synonymous rate variation	26
Statistical Inference	28
Model Selection	29
Site specific rates	30
Positive Selection	31
References	33

2 EXTENSIVE SITE-TO-SITE VARIABILITY OF SYNONYMOUS SUBSTITUTION RATES IN MITOCHONDRIAL GENOMES	48
Abstract	49
Introduction	49
Materials and Methods	51
Sequences, Alignments, and Trees	51
Statistical methods	52
Results and Discussion	55
Concatenated vs. Separated Genes	55
Synonymous rate variation is ubiquitous, and of non-negligible mag- nitude	55
Positive Selection	60
Implications	63
Acknowledgments	64
References	65
Appendix	71
Further Results	79
Nucleotide Model	79

Mean vs Median Discretization	80
Effect of nucleotide frequencies in model	80
3 A NEW MODEL OF SITE-TO-SITE RATE VARIATION WITHIN GENES TO ESTIMATE THE CORRELATION OF SYNONY- MOUS AND NONSYNONYMOUS RATES	85
Introduction	86
Methods	87
Correlation of rates under independent model	87
Bivariate Model	88
Bivariate Discretization	91
Maximum likelihood Estimation	92
Testing the Correlation Hypothesis	93
Estimators of ρ	93
Materials	94
Results and Discussion	94
Correlation of synonymous and nonsynonymous rates	94
Branch Lengths	95
Site-specific rate estimates	97
Implications	98
References	99
Appendix	101
Bivariate discretization	101
4 STATISTICAL PROPERTIES OF MODELING EVOLUTION- ARY RATE HETEROGENEITY WITH DISCRETIZED GAMMA	

DISTRIBUTIONS	119
Introduction	120
Theory	120
Variation limits	120
Number of classes	123
Normalized median discretization	124
Effect of discretization method	127
Simulation study	127
Results and Discussion	130
Fit of models based on number of rate classes	130
Estimates of Coefficients of Variation	131
Estimates of K	132
Estimates of shape parameter	134
Effect of tree shape and number of taxa	134
Conclusions	135
References	139
5 CONCLUSION	141
Final thoughts	142

List of Tables

2.1	Data sets with significant synonymous rate variation.	57
2.2	Significance of synonymous rate variation.	58
2.3	Clades.	71
4.1	CV estimates for simulations under tree 1.	137
4.2	CV estimates for simulations under tree 2.	138

List of Figures

1.1	Sample phylogenetic tree.	7
1.2	Rate variation across sites.	20
1.3	Shapes of a gamma distribution.	21
1.4	Discretization of a gamma distribution.	25
2.1	Box plot of coefficients of variation for clades.	59
2.2	Comparison of coefficients of variation of synonymous and nonsynonymous rates.	60
2.3	Box plot of coefficients of variation for genes.	61
2.4	Carnivores and Insects phylogenetic trees.	74
2.5	Mollusks and Nematodes phylogenetic trees.	75
2.6	Neognathae and Palaeognathae phylogenetic trees.	76
2.7	Percomorpha and Platyhelminthes phylogenetic trees.	77
2.8	Primates and Rodents phylogenetic trees.	78
2.9	Box plot of coefficients of variation for clades.	82
2.10	Comparison of coefficients of variation of synonymous and nonsynonymous rates.	83
2.11	Box plot of coefficients of variation for genes.	84
3.1	Correlation of synonymous and nonsynonymous rates.	96

3.2	Correlation differences based on number of classes.	97
3.3	Box differences in bivariate distribution.	103
3.4	Algorithm to integrate interior box.	108
3.5	Algorithm to integrate edge box.	110
3.6	Algorithm to integrate corner box.	111
3.7	Interpretation of rate classes.	115
3.8	Relative computational time for discretization.	116
4.1	Effect of the number of classes on rates.	124
4.2	Variances under continuous and discretized gamma distributions.	126
4.3	Simulation tree shape 1.	128
4.4	Simulation tree shape 2.	129
4.5	Log likelihood values based on classes.	131
4.6	CV estimates under small shape parameter.	133
4.7	Estimates of K	134

Chapter 1

INTRODUCTION

Introduction

Investigating the mechanisms by which genes evolve over time constitutes a major area of research in molecular evolution. Understanding this unobservable history has presented many challenges to scientists. The basic elements of molecular evolution are the DNA and the protein coding genes, although evolution is studied at many stages between the DNA and the eventual gene function, including nucleotide sequences, amino acid sequences, RNA secondary structure and protein structures. Because we cannot observe evolution for most cases, statistical models are necessary to understand past occurrences and make future predictions.

Over the past 40 years, models of molecular evolution have become increasingly more complex, and this trend is likely to continue. More complex generally implies models with more statistical parameters but can also refer to the computational efforts necessary to implement a model. We are left to ask several questions regarding this observation. Why are we interested in more complex models other than being able to conclude that a new model fits a data set better than a simpler version? What has allowed us the ability to create and apply these complex models? And lastly, are more complex models necessarily better and what are the tradeoffs involved in model selection?

Our interest in more complex models stems from the basic motivation for studying molecular evolution, to ask questions about biological phenomena. The application of statistical models to sequence evolution allows us to make inferences about the underlying biological processes and secondary aspects such as phylogenetic trees (e.g., HUELSENBECK and RONQUIST, 2001), divergence times (e.g., KISHINO *et al.*, 2001) or positive selection (e.g., YANG *et al.*, 2000). Without a rigorous statistical framework, we cannot assess any significance to the results, and

furthermore, we cannot perform hypothesis tests. Usually, more biologically realistic models account for finer evolutionary details making them more complex, and this realism will allow us to make better inferences. Many studies have shown that using too simplistic a model can cause biases in the estimates of parameters such as tree branch lengths (e.g., HOLMQUIST and PEARL, 1980), transition/transversion ratio (e.g., WAKELEY, 1994) and predictions of sites under positive selection (e.g., KOSAKOVSKY POND and MUSE, 2005).

The ability to develop more complex models stems from both the growth in computing power and the vast influx of new data. As models become more realistic, it is not surprising that the computational costs increase. Take, for example, the difference between considering all nucleotide sites evolving independently and considering all three nucleotides of a codon co-evolving. There are 64 (4^3) possible codons as opposed to 4 possible nucleotides. Even without invoking any particular model, intuition tells us that computational time should be cubed for such a jump in complexity. Luckily for researchers, computing power over the past few decades and through today has been doubling roughly every two years, a phenomenon known as Moore's Law. The other advancement allowing more complex models is the enormous amount of data being generated. Models with more parameters naturally require more data to obtain reliable estimates. Prior to the recent genomics boom, obtaining sequences was far more expensive and time consuming. With advancements in high-throughput sequencing, entire genomes are being sequenced for many varied species. These advances are clearly intertwined as larger data sets allow for the estimation of more parameters with reasonable accuracy, and increased computing power allows for the estimation of more parameters in a reasonable amount of time.

The large number of molecular evolutionary models often leave us wondering

which to use for a given problem. On one extreme we have the most simplistic and unrealistic models that offer the advantage of speed, because fewer parameters must be estimated. On the other end of the spectrum we have the most biologically realistic models that are computationally expensive. While one could argue that we should always fit the most realistic model our computing power will allow, overparameterizing increases the variance in the estimates. Therefore for a fixed data set, there is a limit to how complex we can get. Unfortunately, there will never be a clear way to determine which model to use *a priori*, as the best option depends on both the questions of the study and the nature of the particular data set. Several methods have been proposed to choose the best model from some subset of possible models, ranging from an *a priori* fixed set of models (e.g., POSADA and CRANDALL, 2001) to searching the entire set of reversible stochastic models (MUSE, 1999, 2000; HUELSENBECK *et al.*, 2004; KOSAKOVSKY POND *et al.*, 2006b,a). While discussing potential pitfalls of using one model over another is common in the literature (e.g., YANG *et al.*, 1994), we must remember that any model so far proposed is vastly more simplistic than the true forces acting upon the evolutionary history of genomes. Therefore, choosing the model to use should depend on how to best answer the questions of a research project within a reasonable time frame. However, we should acknowledge that sometimes the questions of interest involve precisely these mechanisms of evolution rather than secondary inferences such as divergence time estimation. For these cases, we use the model that fits the data set best based on some model comparison method (e.g., AKAIKE, 1974). This will tend to lead to more complex models, whereas if the model is only a tool for other evolutionary inferences, more complex models may offer little improvement, even if they fit the data better (e.g., YANG, 1997).

Dissertation structure

- Chapter 1
 - Examination of the background of molecular evolution.
 - Presentation of the statistical framework used in later analyses.
 - Discussion of existing models of evolution.
 - Discussion of methods of modeling site-to-site rate variation.

- Chapter 2
 - Analysis of mitochondrial genes under model of synonymous rate variation.
 - Discussion of implications of abundance and magnitude of synonymous rate heterogeneity.
 - Accepted in Genetics.

- Chapter 3
 - Presentation of a new bivariate model of rate heterogeneity.
 - Discussion of statistical properties of model and test for significance of correlation.
 - Detailed description of discretization algorithm for bivariate distribution.
 - Analysis of data under new model demonstrating significant amounts of correlation.

- Chapter 4
 - Examination of statistical properties of discretized gamma distributions.
 - Simulation study to look at bias and variance in estimates of heterogeneity

Data

The gene data in molecular evolution consist of homologous sequences from different species, represented by a multiple sequence alignment which can be obtained through several methods (e.g., THOMPSON *et al.*, 1994). This alignment will be assumed fixed and correct throughout the analyses. The sequence can be comprised of nucleotide bases, codons, RNA base pairs, or amino acids. Consider a basic example of nucleotides from three species (Sp1, Sp2 and Sp3). An alignment of length S can be denoted by D , with $D_s, s = 1, 2, \dots, S$, representing each site (or column) in the alignment

Sp1	A G G T A T C . . .
Sp2	A T G T A T C . . .
Sp3	A G C T A T A . . .

When dealing with a data set of greater than two species we must consider that the sequences are not independently related to each other, but rather they are connected through their evolutionary history. Therefore any calculations require that this history be taken into account to gain maximum information from the data and avoid taxon sampling bias. This history is represented by the phylogenetic tree, denoted as T . A sample tree for the above data set is seen in Figure 1.1, where Sp4 and Sp5 are the unobservable ancestral sequences.

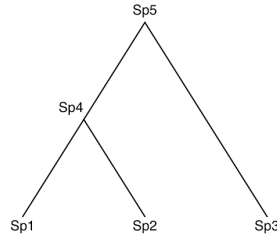


Figure 1.1: **Sample phylogenetic tree.** Here we have a sample tree for three species, Sp1, Sp2 and Sp3. The internal nodes are Sp4 and Sp5. These represent the unknown ancestral species.

The above tree is rooted, meaning the location of the ancestor to all extant species is known, as opposed to an unrooted tree. For some studies, the main interest lies in estimating this tree, while others simply use a fixed (and assumed correct) tree. The latter is done for all analyses in this work and this assumption is generally acceptable, provided that the fixed tree is reasonably similar to the true tree (e.g., YANG *et al.*, 1995, 2000).

Models

Markov chains

The evolution of sequences is often described by a stochastic process known as a Markov chain, consisting of discrete states and probabilities of changes between states over time. Units of time can comprise a discrete or continuous scale. For gene evolution a continuous time Markov chain is used. The basic premise of such a model is that the dependence of the state of a random variable at a future time depends on the state at the current time and that given the current state, future states are independent of any previous states. This structure allows us to calculate probabilities of changes given the amount of evolutionary time and a model.

For biological problems, the states for a continuous time Markov chain are the genomic units that are being studied (e.g., nucleotides, codons, amino acids). Calculating the probability of an entire sequence of length S evolving into another sequence would require the set of character states to be 4^S in length for nucleotides. Even for incredibly short sequences it is computationally impractical to consider such a large number of states and we therefore make the assumption that each site in the data set evolves independently of every other.

The probabilities of change between states necessary to calculate the likelihood function are typically expressed as a probability transition matrix, P . This transition matrix depends on the amount of time, t . The matrix of instantaneous rates, defined as $Q = \frac{dP(t)}{dt}|_{t=0}$, comprise our parameters of interest. These rates allow modeling at the level that evolution is occurring, substitutions, and permit us to use a single model for all tree branches. The individual rates of the rate matrix Q are expressed as q_{ij} for changes from state i to state j . The diagonals of such a matrix are defined so that the rows sum to zero ($q_{ii} = -\sum_{j \neq i} q_{ij}$). A matrix of transition probabilities can be calculated by exponentiating the product of the rate matrix and time ($P = e^{Qt}$), which can be performed in numerous ways (MOLER and VAN LOAN, 1978) but is often done using eigenvalue/eigenvector decomposition or a Taylor series approximation, $P = \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!}$. Because the instantaneous rates in the rate matrix are multiplied by the branch lengths the values are confounded and in practice, one of the rates is set to equal 1. The remaining substitution rates estimated will be ratios.

The stationary distribution of the Markov chain is referred to as the equilibrium frequency and denoted by π_i for nucleotide i . The vector of such frequencies, $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$, satisfy the equation $\pi Q = \pi$. For Markov chain applications to sequence evolution the frequencies of the states are assumed to have reached

stationarity. A useful property of some stochastic models is time reversibility, defined as $\pi_i q_{ij} = \pi_j q_{ji}$. The benefit of such a condition is the ability to calculate likelihoods by proceeding forward or backward through time, which makes traversing a phylogenetic tree easier and allows the use of unrooted trees. In general, the rate matrix is scaled such that $\sum_i \pi_i q_{ii} = -1$, so that the branch lengths will be in terms of the expected number of substitutions per site.

Nucleotide Models

The most basic unit of gene evolution is a nucleotide. The states in such a model are the four nucleotide bases: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The first stochastic model proposed of this kind was the Jukes-Cantor (JC) model (JUKES and CANTOR, 1969) which allows one instantaneous rate of change, $q_{ij} = \mu$, for all possible nucleotide changes and equal nucleotide equilibrium frequencies, $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$. The rate matrix Q can be represented as

$$\mathbf{JC} \quad \begin{matrix} & A & C & G & T \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \left(\begin{array}{cccc} * & 0.25\mu & 0.25\mu & 0.25\mu \\ 0.25\mu & * & 0.25\mu & 0.25\mu \\ 0.25\mu & 0.25\mu & * & 0.25\mu \\ 0.25\mu & 0.25\mu & 0.25\mu & * \end{array} \right) \end{matrix}.$$

The * represents the element along the diagonal necessary for the rows to sum to 0. Given the evolutionary time, t , we can solve for the probabilities of change from nucleotide i to nucleotide j as

$$P_{ij}(t) = \begin{cases} 0.25 + 0.75e^{-4\mu t}, & j = i \\ 0.25 - 0.25e^{-4\mu t}, & j \neq i. \end{cases}$$

Again we see that μ and t are confounded as we can only estimate the product μt . The transition probabilities were originally used to calculate distances between pairs of sequences. The development of methods to calculate likelihoods (FELSENSTEIN, 1973; NEYMAN, 1971) and the description of the pruning algorithm (FELSENSTEIN, 1981) for efficient evaluation of a likelihood over a tree allowed for the analysis of multiple sequences rather than only pairwise sequences. These advances, combined with the influx of large amounts of new sequence data, led to several newer models being proposed relaxing certain constraints of the JC model and increasing the biological realism.

The Kimura 2 parameter (K2P) model (KIMURA, 1980) focused on the differences between purines (A and G) and pyrimidines (C and T), allowing separate rates for transitions ($A \Leftrightarrow G, C \Leftrightarrow T$) and for transversions ($A \Leftrightarrow C, A \Leftrightarrow T, C \Leftrightarrow G$ and $G \Leftrightarrow T$). The transition/transversion ratio, frequently denoted as κ , is believed to be significantly greater than 1, a fact supported by empirical studies (e.g., BROWN *et al.*, 1982). Therefore the K2P model typically fits data sets better than the JC model. The rate matrix for the K2P model is expressed as

$$\mathbf{K2P} \begin{matrix} & A & C & G & T \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} * & 0.25 & 0.25\kappa & 0.25 \\ 0.25 & * & 0.25 & 0.25\kappa \\ 0.25\kappa & 0.25 & * & 0.25 \\ 0.25 & 0.25\kappa & 0.25 & * \end{pmatrix} \end{matrix}.$$

Further models introduced more realistic features. Because all nucleotides do not occur equally in genes, we have little reason to believe that the equilibrium frequencies are all 0.25, and the Felsenstein 81 (F81) model (FELSENSTEIN, 1981) corrected this by allowing nucleotide equilibrium frequencies to differ. The Hasegawa,

Kishino and Yano (HKY) model (HASEGAWA *et al.*, 1985) incorporated both improved features of the K2P and F81 models. The most parameter rich model that is still time reversible, allows differing nucleotide frequencies and a unique rate for each of the 6 possible nucleotide changes and is known as the General Time Reversible (REV or GTR) model (TAVARÉ, 1986; LANAVE *et al.*, 1984). The rate matrices for these three models are:

$$\mathbf{F81} \begin{matrix} & A & C & G & T \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} * & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & * & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & * & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & * \end{pmatrix} \end{matrix},$$

$$\mathbf{HKY} \begin{matrix} & A & C & G & T \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} * & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & * & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & * & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & * \end{pmatrix} \end{matrix},$$

$$\mathbf{REV} \begin{matrix} & A & C & G & T \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} * & R_{AC}\pi_C & R_{AG}\pi_G & R_{AT}\pi_T \\ R_{AC}\pi_A & * & R_{CG}\pi_G & R_{CT}\pi_T \\ R_{AG}\pi_A & R_{CG}\pi_C & * & R_{GT}\pi_T \\ R_{AT}\pi_A & R_{CT}\pi_C & R_{GT}\pi_G & * \end{pmatrix} \end{matrix}.$$

RNA models

Basic nucleotide models treat each site as acting independently of every other site, which clearly is contrary to biological knowledge. For example, we can consider the constraints imposed by nucleotide base pairing in RNA secondary structure. This can often be a useful tool in molecular evolution due to the wide range of species from which RNA data is available (e.g., HIGGS, 2000). The general structure of such a model is a 16×16 rate matrix consisting of the nucleotide pairs (AA, AC, ..., TT), with instantaneous rates between the dinucleotides. In some cases, this rate matrix is reduced to 6×6 , allowing only those base pairs that are favorable (TILLIER, 1994) or to 7×7 , allowing favorable base pairs and grouping all mismatches into one group (HIGGS, 2000; TILLIER and COLLINS, 1998). These models also vary in whether they allow double mutations (SCHÖNIGER and VON HAESLER, 1994; SAVILL *et al.*, 2001; TILLIER and COLLINS, 1998) or disallow double mutations (MUSE, 1995; RZHETSKY, 1995). More recent work has assumed no RNA model structure other than reversibility and searched the model space for the best fitting models using a genetic algorithm (KOSAKOVSKY POND *et al.*, 2006b). The general consensus of the results obtained by using RNA models is that selective forces acting upon RNA structure are clearly present and models allowing double mutations seem to fit better than those without.

Amino Acid Models

Early amino acid models were not defined mechanistically through a rate matrix, but rather derived empirically from real data sets (DAYHOFF *et al.*, 1978; HENIKOFF and HENIKOFF, 1992; JONES *et al.*, 1992). While these models may perform well for the protein families from which they were created, we have little

reason to believe that all proteins follow the same evolutionary patterns given the strong amounts of functional constraints. However, to create a 20×20 mechanistically defined rate matrix was not initially possible due to computational restrictions. Maximum likelihood estimates for proteins were initially based on very simplistic models, assuming an equal rate of change between all amino acids which is comparable to a JC nucleotide model (KISHINO *et al.*, 1990). Some researchers have gone as far as fitting a general reversible model (ADACHI and HASEGAWA, 1996; DIMMIC *et al.*, 2002), which is difficult with 189 transition parameters to estimate. Such mechanistic models provide improvement over empirical models (YANG *et al.*, 1998) and allow for greater flexibility in model composition and inferences. The most advanced models attempt to incorporate physical and chemical properties of amino acids (e.g., hydrophobicity, size) and/or secondary and tertiary structure into the models when such structure is known (e.g., THORNE *et al.*, 1996; GOLDMAN *et al.*, 1996, 1998; PARISI and ECHAVE, 2001; ROBINSON *et al.*, 2003; KOSHI and GOLDSTEIN, 1998). Other methods try to infer model structure without prior input (KOSAKOVSKY POND *et al.*, 2006a). However, without the strong support for any specific mechanistic model, the use of empirically derived models continues, although often analysis of structure is taken into account after fitting the model (e.g., SIMON *et al.*, 2002; GAUCHER *et al.*, 2001; BLOUIN *et al.*, 2003).

Codon Models

While nucleotide models are simple and realistic for certain cases, when working with protein coding genes, the effect of codon structure causes the assumption that all nucleotides evolve under the same constraints to break down quickly. Independently, two groups proposed models that account for this codon structure

(GOLDMAN and YANG, 1994; MUSE and GAUT, 1994). Unlike nucleotide models, the states in the Markov chain are not single bases, but groups of three bases, increasing the number of states to 64 (4^3). For practical purposes we can ignore stop codons, which leaves us with 61 states for the universal genetic code. Because our interest is with instantaneous rates, an assumption made in these models causes rates between codons differing by 2 or 3 nucleotides to be set to 0. This assumption was made for both convenience and for codon models to be a more natural extension of nucleotide models, though recent work has argued against this approach (WHELAN and GOLDMAN, 2004). Codon models take advantage of information from the genetic code and differentiate between nonsynonymous nucleotide changes that alter the amino acid and synonymous nucleotide changes that do not. The instantaneous rates for these substitutions are β and α , respectively. Because the selective pressures will act on these changes differently we model them separately. The instantaneous rates of change from codon i to codon j for the MG94 (MUSE and GAUT, 1994) and GY94 (GOLDMAN and YANG, 1994) models are

$$\begin{aligned}
 \text{MG94} \quad q_{ij} &= \begin{cases} \alpha\pi_{n_j}, & i \rightarrow j \text{ is a synonymous one-step change} \\ & \text{(e.g., CGA} \rightarrow \text{CGG)} \\ \beta\pi_{n_j}, & i \rightarrow j \text{ is a nonsynonymous one-step change} \\ & \text{(e.g., CGA} \rightarrow \text{GGA)} \\ 0, & \text{otherwise (e.g., CGA} \rightarrow \text{AGG)} \end{cases} \\
 \text{GY94} \quad q_{ij} &= \begin{cases} \kappa\pi_j, & i \rightarrow j \text{ is a one-step synonymous transition} \\ & \text{(e.g., CGA} \rightarrow \text{CGG)} \\ \kappa\omega\pi_j, & i \rightarrow j \text{ is a one-step nonsynonymous transition} \\ & \text{(e.g., CGA} \rightarrow \text{CAA)} \\ \pi_j, & i \rightarrow j \text{ is a one-step synonymous transversion} \\ & \text{(e.g., CGA} \rightarrow \text{CGT)} \\ \omega\pi_j, & i \rightarrow j \text{ is a one-step nonsynonymous transversion} \\ & \text{(e.g., CGA} \rightarrow \text{CCA)} \\ 0, & \text{otherwise (e.g., CGA} \rightarrow \text{AGG)} \end{cases}
 \end{aligned}$$

Under the above notation, π_{n_j} is the equilibrium frequency of the nucleotide in codon j that has changed from codon i , π_j is the equilibrium frequency of codon j , and ω represents the ratio of nonsynonymous to synonymous rates. A slight variant of the GY94 model accounts for differences in properties of amino acids formed from codons i and j , but is rarely used.

The GY94 model incorporates rate differences between transitions and transversions similarly to the HKY nucleotide model. However, any nucleotide model structure can be “crossed” with either codon model. For example, the MG94 model can allow for separate parameters for all six possible nucleotide changes like the REV nucleotide model. Notation for such a model is written as MG94 \times REV.

Although adding a great deal of biologically accurate features, codon models also add a lot of computational cost, as a 4×4 rate matrix becomes a 61×61 rate matrix (or other, depending on the genetic code). Summing over ancestral sequences and exponentiating the rate matrix becomes more difficult.

Likelihood Function

Studies considered in the work will focus on likelihood based estimations of parameters, with the likelihood function represented by

$$L(\theta|D) = P(D|\theta),$$

where θ represents all parameters in the model, including branch lengths. Based on our previous alignment, we now calculate the probability of our data given a tree and a model as $P(D|\theta, \text{Model}, T) = P(S1, S2, S3|\theta, \text{Model}, T)$. Because all likelihood calculations rely on a given tree and model, these are omitted from future formulas. If we assume independence, the likelihood for the evolution of the

entire sequence will reduce to the product of the likelihoods for each site,

$$L(\theta|D) = \prod_{s=1}^S L_s(\theta|D) = \prod_{s=1}^S P(D_s|\theta).$$

However, without knowledge of the ancestral nodes we cannot evaluate this probability. Therefore, we must sum over all possible sequences at these internal nodes, which can be a large number of summations if the number of taxa is large. Because of the Markov chain memoryless property we know that the probabilities of states at a given node depend solely on the nodes that are direct descendants. Because all internal nodes in a bifurcating tree depend upon only two other nodes the calculations can be simplified a great deal using the pruning algorithm (FELSENSTEIN, 1973, 1981). An example based on the tree in Figure 1.1 is

$$P(D|\theta) = \sum_{Sp5} \sum_{Sp4} \pi_{Sp5} P(Sp5 \rightarrow Sp3) P(Sp5 \rightarrow Sp4) P(Sp4 \rightarrow Sp1) P(Sp4 \rightarrow Sp2).$$

While the independence assumption appears overly simplistic, for many years it was necessary for computational purposes. Some models have partially relaxed this assumption by allowing evolution dependent on neighboring sites (e.g., MUSE and GAUT, 1994; GOLDMAN and YANG, 1994; YANG, 1995; FELSENSTEIN and CHURCHILL, 1996; HWANG and GREEN, 2004).

Rate Variation

How to model rate variation

All models discussed above make a key assumption, often for computational tractability, that all sequence sites evolve at an equal rate. However, the underlying biology tells us that different sites evolve at different rates, due to factors such as natural genomic variation like hotspots or selection acting on sites. As computing power

grew, the ability to allow sites to have differing rates became a possibility and eventually the standard. Incorporating variation in models also improves estimates of other parameters and allows us to ask new scientific questions.

Allowing each site to have a distinct rate, while biologically the most realistic, is clearly not feasible due to the enormous number of parameters to be estimated, although this has been attempted in some studies (MEYER and VON HAESELER, 2003; NIELSEN, 1997; KELLY and RICE, 1996). FELSENSTEIN (2001) points out that a model with a rate for each site would lead to the number of parameters increasing at the same pace as the increase in sequence length, causing the likelihood methods to lose their properties of consistency. Without a very large number of taxa, data sets would lack the statistical power to estimate everything, and even in a best case scenario, only some site specific parameters will be estimated well. We must therefore make certain concessions to have workable models, usually accomplished by assuming that rates follow some statistical distribution.

The most basic method to modeling rate variation would be to allow each site to belong to one of two classes, with a slow or a fast rate. If we knew the class for each site, then the likelihood calculations would be straightforward. However, this information is never known, so instead we must sum over both possible classes and the likelihood function becomes

$$L(\theta|D) = \prod_{s=1}^S \sum_{m=1}^2 P(D_s|\theta_m)p_m,$$

where θ_m represents the set of parameters for site m (most of θ_m will be consistent across sites) and p_m is the probability of rate class m . This rate variation doubles the computational time necessary for likelihood calculation.

This discrete distribution of rate variation consists of three free parameters, the fast rate, the slow rate, and the probability of one of the rate classes. In some

cases the rate classes are assumed to be equally probable, $P(\text{fast rate class}) = P(\text{slow rate class})$. Some of the earliest uses of rate variation attempted to model invariant sites (FITCH and MARGOLIASH, 1967; FITCH and MARKOWITZ, 1970), which simply assumes that the rate for the slow class is 0.

We can extend this discrete distribution to include M rate classes instead of two. Now we must estimate M rates and $M - 1$ corresponding probabilities. The likelihood function similarly sums over all classes as

$$L(\theta|D) = \prod_{s=1}^S \sum_{m=1}^M P(D_s|\theta_m)p_m.$$

The two drawbacks of such a discrete distribution are that there are only a finite number of possible rates and that the number of parameters to be estimated increases as we increase the number of classes. Alternatively, we can consider using a continuous distribution, which has the most desirable properties for characterizing such variation because there is little reason to believe that rates of evolution can be summarized better by a fixed number of discrete rates. In addition continuous distributions are generally described by very few parameters, usually one or two.

Under a continuous distribution we do not know the true rate for any site, and must integrate over the distribution. If we use a rate α defined by some continuous distribution $f(\alpha)$ then our likelihood function becomes

$$L(\theta|D) = \prod_{s=1}^S \int_0^{\infty} P(D_s|\theta, \alpha)f(\alpha)d\alpha.$$

There is no closed form for the integral, and the only way to calculate the likelihood would be through numerical integration techniques. The computational burden incurred in evaluating the likelihood function and the need to numerically integrate for each sequence site, makes the use of continuous distributions impractical for data sets of any reasonable size.

One of the most commonly used distributions has been a gamma, first described for rate variation by UZZELL and CORBIN (1971). If we let some rate α vary according to a gamma distribution with parameters μ_α and λ , the probability density function of α is defined as

$$\alpha \sim \text{Gamma}(\mu_\alpha, \lambda) = \frac{1}{\Gamma(\mu_\alpha)} \lambda^{\mu_\alpha} \alpha^{\mu_\alpha-1} e^{-\alpha\lambda}.$$

The rate for each site, α_s , is then a random variable drawn from the above distribution (see Figure 1.2). If the rates of evolution were constant across sites, the number of substitutions would be Poisson distributed, while if the rates of evolution were not constant, but rather followed a gamma distribution, then the distribution of substitutions would be a negative binomial. Many researchers demonstrated that the Poisson distribution did not fit data sets well and that the negative binomial fit much better (e.g., UZZELL and CORBIN, 1971; HOLMQUIST and PEARL, 1980; HOLMQUIST *et al.*, 1983). While this does not let us conclude that the true rates are drawn from a gamma distribution, it does provide evidence that a gamma distribution summarizes the underlying rates well.

When allowing rates to vary, the expected value of the distribution of variation must equal 1 because otherwise the rate variation is nonidentifiable and the values estimated for parameters will change (see Chapter 4 for more details). Very often, this bias will affect rates such as ω or the branch lengths, which are used later in analyses for inferences including divergence times. For the gamma distribution, this requires that we use the single parameter version where $\lambda = \mu_\alpha$, leading to a density function of

$$\alpha \sim \text{Gamma}(\mu_\alpha, \mu_\alpha) = \frac{1}{\Gamma(\mu_\alpha)} \mu_\alpha^{\mu_\alpha} \alpha^{\mu_\alpha-1} e^{-\alpha\mu_\alpha}.$$

This version is considered for all analyses in the work. One convenient property of the gamma distribution is its ability to take multiple shapes, namely when $\mu_\alpha < 1$

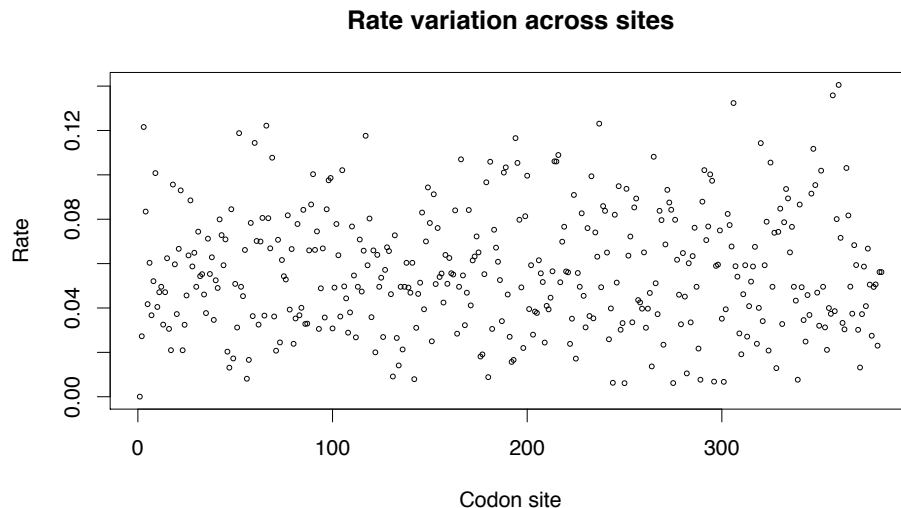


Figure 1.2: **Rate variation across sites.** This is an example of rate heterogeneity within a gene. The x-axis is the site in the gene and the y-axis is represent the rate of evolution for that site. The gene in this example is the primates CYTB data set used in Chapter 2, showing the estimated nonsynonymous rate.

the distribution is L-shaped, when $\mu_\alpha = 1$ the distribution is exponential and when $\mu_\alpha > 1$ the distribution is unimodal (see Figure 1.3). Small values of μ_α lead to large amounts of rate variation, while large values of μ_α correspond to very little rate variation with $\mu_\alpha = \infty$ reducing to the case with no variation at all.

Models estimating genetic distances using a gamma distribution were some of the first to incorporate site to site variation (UZZELL and CORBIN, 1971; HOLMQUIST *et al.*, 1983; LI *et al.*, 1990; JIN and NEI, 1990; NEI and GOJOBORI, 1986; LEE *et al.*, 1995; TAMURA and NEI, 1993; NEI *et al.*, 1976; GOLDING, 1983; TAKAHATA, 1991; LARSON and WILSON, 1989; HOLMQUIST and PEARL, 1980). In these cases the shape parameter was either taken to be a fixed (not estimated) value from previous studies or estimated using the negative binomial distribution (BLISS and FISHER, 1953) of substitutions through a method of mo-

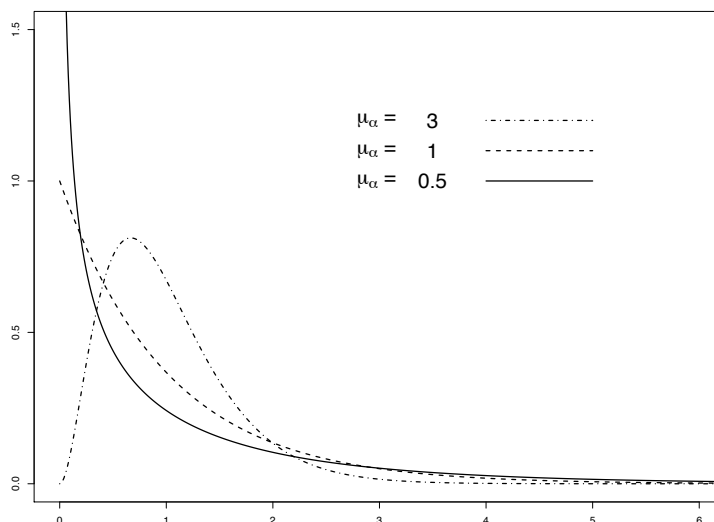


Figure 1.3: **Shapes of a gamma distribution.** Here we present three different shapes that a gamma distribution can take, depending on the shape parameter, μ_α . These distributions are the single parameter versions of the gamma, with an expected value of 1. The flexibility of the gamma distribution makes it advantageous for use in the modeling of rate variation.

ments or maximum likelihood estimator based on the parsimonious reconstruction of substitutions. However, these kinds of estimators were shown to overestimate μ_α , especially when the true value is small (SULLIVAN *et al.*, 1995; TAKEZAKI and GOJOBORI, 1999; YANG and KUMAR, 1996). YANG (1993) extended the use of a continuous gamma distribution to likelihood methods, but as we have seen, the cost of using a continuous distribution with the likelihood function is too great.

Other models developed to account for rate variation over sites include other known statistical distributions like a lognormal distribution (OLSEN, 1987; GUOY and LI, 1989; GOLDING, 1983), an inverse Gaussian (WADDELL *et al.*, 1997), beta and normal (YANG *et al.*, 2000), or general discrete distributions (YANG *et al.*, 2000; KOSAKOVSKY POND and MUSE, 2005). IRWIN *et al.* (1991) used a

sliding window analysis of rate estimation along a gene. MOROZOV *et al.* (2000) used Fourier series and wavelet transformation functions to model rate variation. FELSENSTEIN (2001) argues that there is little difference in using one over the other as the maximum likelihood estimates of the distributions are all similar. Nonetheless, the use of gamma distributions has become the standard method for modeling rate variation in molecular evolutionary analyses (e.g., KATZ *et al.*, 2004; HUELSENBECK and BOLIBACK, 2005).

Why rate variation should be included

A key feature to consider when implementing models is the magnitude of the effect of each assumption on the potential inferences. Assuming homogenous rates can introduce bias and/or increase variance in parameter estimation. Many simulation studies have been done testing for biases in parameter estimation under a suboptimal model, which is usually taken to mean a simpler model (one that is nested in the true model), because the true evolutionary history is more complex than any statistical model. Specifically, excluding rate variation when variation is truly present has been shown to cause underestimates in parameters such as a transition/transversion ratio (WAKELEY, 1994) and the branch lengths (TATENO *et al.*, 1994; YANG *et al.*, 1994; HUELSENBECK and NIELSEN, 1999; LEITNER *et al.*, 1997; HASEGAWA *et al.*, 1993; VAN DE PEER *et al.*, 1993; PALUMBI, 1989; UPHOLT, 1977). This bias gets more extreme as the amount of variation increases (WAKELEY, 1994). Choosing too simplistic a model of substitution (e.g., JC instead of HKY) will also cause underestimates in branch lengths, but of a less severe nature than ignoring variation (YANG *et al.*, 1994). For hypothesis testing, our concern lies in either an increased Type I or Type II error. ZHANG (1999)

showed that a likelihood ratio test can be misleading under incorrect models. For phylogenies, using models without rate variation can give too much support to certain taxa not being monophyletic (SULLIVAN and SWOFFORD, 1997) and can lead to biases in methods of tree reconstruction (GAUT and LEWIS, 1995; KUHNER and FELSENSTEIN, 1994; TAKEZAKI and GOJOBORI, 1999; VAN DE PEER *et al.*, 1993; SULLIVAN and SWOFFORD, 2001; JIN and NEI, 1990). All of this evidence gives sound statistical support to the notion that site-to-site variation needs to be modeled well.

Discretizing a Continuous Distribution

As a compromise between discrete and continuous distributions YANG (1994) created a new model of rate variation, allowing for a discretized approximation of continuous distributions, focusing on the gamma, although the methodology can be applied to any distribution (see YANG *et al.*, 2000).

Discretized distributions offer many of the advantages of continuous distributions, but without the computational obstacle. To discretize a continuous distribution we first divide up the distribution into intervals to represent each rate class, and then determine the rate within each interval. The intervals can all contain an equal proportion of the distribution or varying proportions that must then be estimated as separate parameters. For M rate classes, let the left boundaries be defined as L_1, \dots, L_M , and right boundaries as R_1, \dots, R_M , such that $L_1 = 0$, $R_M = \infty$, and $R_m = L_{m+1}$ for $m = 1, \dots, M - 1$. The interval boundaries are defined to follow

$$P(\text{rate class } m) = p_m = \int_{L_m}^{R_m} f(\alpha|\mu_\alpha) d\alpha.$$

For equiprobable rates classes, $p_m = \frac{1}{M}$ and for nonequiprobable rate classes, p_m are free parameters. However, due to the constraint of $\sum_{m=1}^M p_m = 1$, we need

to estimate only $M - 1$ proportions. Once we have intervals, the next step is calculating a rate, α_m to represent each interval, which can be taken as the mean or median value over that range. These values are found by solving for α_m in the following equations,

$$\begin{aligned}\alpha_m &= \frac{1}{p_m} \times \int_{L_m}^{R_m} \alpha f(\alpha|\mu_\alpha) d\alpha && \text{MEAN} \\ \frac{p_m}{2} &= \int_{L_m}^{\alpha_m} f(\alpha|\mu_\alpha) d\alpha && \text{MEDIAN}\end{aligned}$$

The relative merits of each method will be discussed in detail in Chapter 4. The mean rate can be calculated using the incomplete gamma ratio (YANG, 1994), while the median rate can be found using the same method used to find the interval boundaries. A sample discretization can be seen in Figure 1.4 for $\mu_\alpha = 2.0$ and 4 rate classes. Because we now have a discrete distribution rather than a continuous one, the integral in the likelihood function can be replaced with a summation,

$$L(\theta|D_s) = \prod_{s=1}^S \sum_{m=1}^M P(D_s|\theta_m) p_m.$$

The expected value of the discretized gamma distribution will have an expected value of 1 if we use mean rates, but not median rates. This is the case because the mean rates are calculated using expected values and the median rates are not. Therefore, median rates must be normalized to ensure proper interpretation of all other parameters in the model (see Chapter 4 for more details).

Discretizing the rates of the gamma distribution into equal classes is the common practice when including rate variation and will be used in this research. However, there are several potential pitfalls of equiprobable rate classes that will be described in detail in Chapter 4. Other methods of discretizing include use of Laguerre Quadratures to determine not only the rates but the probabilities of each rate class (FELSENSTEIN, 2001) and a hierarchical approach in which the intervals of the discretization are determined using a beta distribution (KOSAKOVSKY

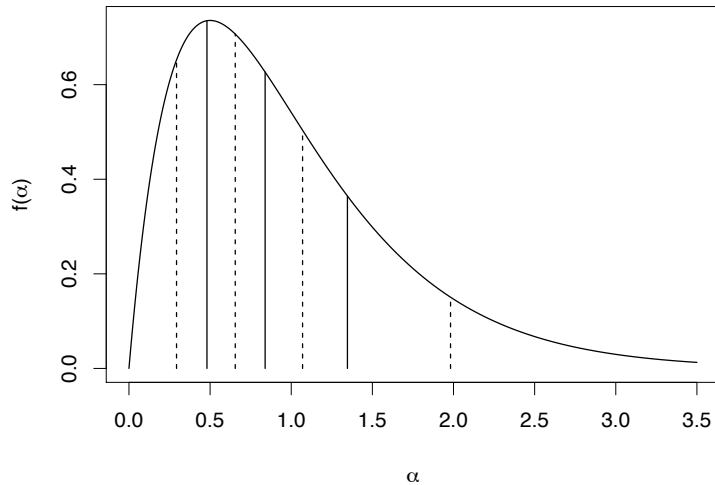


Figure 1.4: **Discretization of a gamma distribution.** Here we present a sample discretization. The solid lines divide the distribution into intervals and the dashed lines represent the estimated rate within each interval using the mean value. The shape parameter for this gamma is 2.0.

POND and FROST, 2005). Both of these approaches appear to fit data sets better and give better approximations of the continuous gamma distribution, without the need to estimate many other parameters.

Gamma model extensions

Many extensions to the gamma distribution have been made to improve such models. Going back to the early ideas that some sequence sites did not vary at all, GU *et al.* (1995) proposed a model combining the discrete gamma model with an additional rate class of invariant sites, adding only a single parameter, the proportion of invariant sites. YANG *et al.* (2000) proposed a series of models based on discrete distributions and discretized continuous distributions including a mixture of 2 gammas, a mixture of gamma and beta, and mixtures with truncated normal

distributions.

Synonymous rate variation

One problem with the commonly employed models dealing with rate variation over coding genes is the assumption that the variation is found solely on the nonsynonymous/synonymous ratio (NIELSEN and YANG, 1998; YANG *et al.*, 2000). The effect of this parameterization is to allow nonsynonymous rates to vary, while keeping synonymous rates constant. For example, following the M5 model of YANG *et al.* (2000), (but with notation consistent with this work), $\beta_s \sim \omega \times \text{Gamma}(\mu_\beta, \mu_\beta)$ and $\alpha_s = 1$. Because synonymous changes are silent at the protein level the assumption is that these changes must be neutral. More recent studies infer that synonymous rates are not constant, showing variation through a sliding window analysis (HURST and PÁL, 2001; ALVAREZ-VALIN *et al.*, 1998, 2000; SMITH and HURST, 1998; INA, 1996), or demonstrating evidence for selection on synonymous sites (AKASHI, 1995, 1999; EYRE-WALKER, 1999; SMITH and HURST, 1999; AKASHI and EYRE-WALKER, 1998; COMERON and KREITMAN, 1998; COMERON *et al.*, 1999; DUAN *et al.*, 2003; PAGANI *et al.*, 2005). Clearly this mounting evidence should not be ignored, and modeling synonymous rate variation could be potentially useful. Shifting the gamma distribution to both the synonymous and nonsynonymous rates would fix this problem, but would force an equal amount of variation on both rates, which seems unlikely, as the selective forces that cause a good proportion of nonsynonymous rate variation would have a much smaller effect on synonymous rates. In addition, the sites with higher nonsynonymous rates would be forced to also have high synonymous rates as constrained by that parameterization, removing the possibility of testing for site specific positive selection.

Instead, KOSAKOVSKY POND and MUSE (2005) presented a class of rate variation models where both synonymous and nonsynonymous rates are free to vary independently. The focus here will be on a Dual Gamma distribution model. Building off of the parameterization of the MG94 model, the variation can be applied to the synonymous (α) and nonsynonymous (β) rates for a site s as

$$\alpha_s \sim \text{Gamma}(\mu_\alpha, \mu_\alpha),$$

$$\beta_s \sim \omega \times \text{Gamma}(\mu_\beta, \mu_\beta).$$

Because the distributions are assumed independent the density function of the joint distribution is simply the product of the individual densities,

$$f(\alpha, \beta | \mu_\alpha, \mu_\beta, \omega) = f(\alpha | \mu_\alpha) f(\beta | \mu_\beta, \omega).$$

Again, due to the confounding of the mutation rate with time, the expected value of the synonymous rate is set to 1, which is a more natural extension of the models of nonsynonymous rate variation, where α is fixed as 1. In addition, KOSAKOVSKY POND and MUSE (2005) present other models that allow for synonymous rates to vary, such as a general discrete bivariate distribution.

Calculating the likelihood under this model is nearly identical to using a model without synonymous rate variation, except that we are dealing with rate variation in two dimensions,

$$L(\theta | D) = \prod_{s=1}^S \int_0^\infty \int_0^\infty P(D_s | \alpha, \beta, \theta) f(\alpha, \beta | \mu_\alpha, \mu_\beta, \omega) d\alpha d\beta.$$

As previously mentioned, dealing with such integrals is not computationally feasible and we again resort to discretization methods (YANG, 1994). Because the rates are assumed independent, the discretization process involves the same

methods, performed separately on each distribution. If we use M synonymous and N nonsynonymous rate classes ($M \times N$), then the likelihood becomes

$$L(\theta|D) = \prod_{s=1}^S \sum_{m=1}^M \sum_{n=1}^N P(D_s|\theta_{mn})h(\alpha_m, \beta_n|\mu_\alpha, \mu_\beta, \omega).$$

For the data sets tested by KOSAKOVSKY POND and MUSE (2005), they found significant synonymous rate variation in a wide range of genes from diverse taxa.

Statistical Inference

Estimating the parameters using maximum likelihood is difficult, because no closed form expression exists. Instead, we must use standard numerical optimization techniques to find estimates. This optimization causes the computational burden that makes such studies difficult. We can estimate the equilibrium frequencies as free parameters or from the counts in each data set. The more common approach is to use the observed values, as this reduces the number of parameters in the model and usually provides a good estimate of the maximum likelihood values. All models used in analyses for this research will use observed frequencies.

For codon models, the codon and nucleotide equilibrium frequencies can be calculated in two ways, assuming equal frequencies at all three codon positions

$$\pi_{ijk} = \frac{\pi_i \pi_j \pi_k}{1 - \sum \text{stops}},$$

or by allowing different frequencies

$$\pi_{ijk} = \frac{\pi_i^{(1)} \pi_j^{(2)} \pi_k^{(3)}}{1 - \sum \text{stops}},$$

where $\sum \text{stops} = \sum_{\text{stop codons } b} \pi_b$ and π_i^l is the frequency of π_i in the l^{th} codon position. The latter method tends to fit data sets better.

Model Selection

For any given data set, we may wish to determine whether or not the model with synonymous rate variation fits significantly better than a model with no synonymous rate variation. The former is our alternative hypothesis, while the latter is the null hypothesis. Let us again note that the alternative hypothesis reduces to the null if we constrain the shape parameter $\mu_\alpha = \infty$, and therefore these models are nested. The likelihood ratio test (LRT) statistic is the same as in the traditional case, namely $\Lambda = -2(L_O - L_A)$, where L_O and L_A are the maximum likelihood values under the null and alternative hypotheses, respectively. We must take notice of the fact that the constraint lies on the boundary of the range of the shape parameter. Therefore, the test statistic is not asymptotically a standard χ^2 , but rather a mixed χ^2 (or $\bar{\chi}^2$) (SELF and LIANG, 1987; SIEGEL, 1979; OTA *et al.*, 2000; WHELAN and GOLDMAN, 1999). The number of degrees of freedom is the difference in number of free parameters between the models. If the shape parameter at the boundary is the only constraint, then our asymptotic distribution for the test statistic is $\bar{\chi}_1^2$, which is a mix of 50% χ_0^2 (point mass at 0) and 50% χ_1^2 . If the models also differ by v non-boundary constraints then we should consider a $\bar{\chi}_{v+1}^2$ distribution (50% χ_v^2 and 50% χ_{v+1}^2). When multiple parameters are on the boundary of the parameter space, the asymptotic distribution becomes more complex, but can be computed (OTA *et al.*, 2000).

When comparing models that are not nested, the likelihood ratio test is not an appropriate method and we must use an alternative. Two commonly used methods are the Akaike Information Criteria (AIC) (AKAIKE, 1974) and the Bayesian Information Criteria (BIC) (SCHWARTZ, 1978). These are defined as

$$AIC = -2 \times \hat{L} + 2 \times \text{Number of parameters},$$

$$BIC = -2 \times \hat{L} + \log(n) \times \text{Number of parameters},$$

where \hat{L} is the maximum likelihood value, and n is the sample size of the data, which for our applications is taken as the number of sites in the data set. To compare two or more models using either of these tests, simply calculate the value for each model and select the lowest score. Sometimes model selection involves comparing many models, some of which are nested and some of which are not nested, in which case we use a combination of LRT and AIC tests. An example of this is the model selection methods of POSADA and CRANDALL (2001), which are implemented in the program Modeltest (POSADA and CRANDALL, 1998). This program hierarchically chooses the most optimal model from a collection of nucleotide models with varying structure and rate variation components.

Site specific rates

The natural interest after incorporating rate variation is to estimate the rates for each site, thereby allowing us to see which sites are evolving quickest and slowest and permitting us to discuss things like selective pressures acting on sites, or mutational hotspots in certain gene segments. The simplest approach is to find the rate class that fits a particular site the best. This can be done by calculating the marginal likelihood of the data at site s given class m as

$$l_m^{(s)} = P(D_s | \text{class} = m, \hat{\theta}),$$

where $\hat{\theta}$ represents the maximum likelihood values for all other parameters in the model. By maximizing $l_m^{(s)}$ over all m we find the class with the highest likelihood. However, this gives us no idea of how much better the data fits in the best class relative to the others, especially because it does not consider the probabilities for

each rate class. We can use the marginal likelihood to calculate the empirical Bayes posterior probability of a site belonging to a class as

$$r_m^{(s)} = \frac{l_m^{(s)} p_m}{\sum_{u=1}^M l_u^{(s)} p_u}.$$

This gives us a level of confidence in the class assignments (NIELSEN and YANG, 1998). Assigning the rate for a site from a single class can be misleading because for some sites all classes will have relatively equal support. Alternatively, we can use the posterior probabilities to calculate a weighted average rate for site s , $\alpha^{(s)}$, over all classes as

$$\alpha^{(s)} = \sum_{m=1}^M r_m^{(s)} \alpha_m.$$

These rate estimates have the advantage of taking into account all of the information of the distribution. Extensions of the above methods to the Dual Gamma model of synonymous and nonsynonymous rate variation are straightforward.

Positive Selection

One useful tool of models with rate variation is the ability to look for sites that are under positive selection. Positive selection is typically defined as the case where the nonsynonymous rate is greater than the synonymous rate ($\omega > 1$). However, if rate variation is ignored, a single ω is considered for all codons and to estimate positive selection we would need the mean value of ω to be greater than 1 over the entire data set, which will rarely occur. Positive selection, if present, would usually act upon a small subsets of sites, while most codons would be under neutral or purifying selection. Adding rate variation allows different codons to have different values of ω , and improving the power of detecting positive selection (NIELSEN and YANG, 1998).

However, in order to detect this positive selection, there needs to be at least one rate class in which $\omega > 1$. Whether this happens depends on the exact model used and number of rate classes implemented. Testing for the overall presence of some sites under positive selection can be done by testing the null hypothesis of a model that does not allow any classes of $\omega > 1$ to the alternative model that does allow positive selection. Confidence can be placed on the positive selection of each codon by summing the posterior probabilities of all classes for which $\omega > 1$ (NIELSEN and YANG, 1998; YANG *et al.*, 2000).

Under the Dual Gamma model a rate class mn is considered to be under positive selection if $\frac{\beta_n}{\alpha_m} > 1$. Again, there is no guarantee that any classes will meet this criterion. KOSAKOVSKY POND and MUSE (2005) demonstrated that the Dual Gamma model increased the probability of finding such a class and therefore increased the power to detect codons under positive selection.

References

- ADACHI, J., and M. HASEGAWA, 1996 Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42**: 459–468.
- AKAIKE, H., 1974 A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* **19**: 716–723.
- AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at “Silent” sites in drosophila DNA. *Genetics* **139**: 1067–1076.
- AKASHI, H., 1999 Within and between species DNA sequence variation and the ‘footprint’ of natural selection. *Gene* **238**: 39–51.
- AKASHI, H., and A. EYRE-WALKER, 1998 Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8**: 688–693.
- ALVAREZ-VALIN, F., K. JABBARI and G. BERNARDI, 1998 Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. *J. Mol. Evol.* **46**: 37–44.
- ALVAREZ-VALIN, F., K. JABBARI and G. BERNARDI, 2000 Nonrandom spatial distribution of synonymous substitutions in the GP63 gene from leishmania. *Genetics* **155**: 1683–1692.

- BLISS, C. I., and R. A. FISHER, 1953 Fitting the negative binomial distribution to biological data. *Biometrics* **9**: 176–200.
- BLOUIN, C., Y. BOUCHER and A. J. ROGER, 2003 Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. *Nucl. Acids Research* **31**: 790–797.
- BROWN, W. M., E. M. PRAGER, A. WANG and A. C. WILSON, 1982 Mitochondria DNA sequences of primates: the tempo and mode of evolution. *J. Mol. Evol.* **18**: 225–239.
- COMERON, J. M., and M. KREITMAN, 1998 The correlation between synonymous and nonsynonymous substitutions in drosophila: mutation, selection or relaxed constraints. *Genetics* **150**: 767–775.
- COMERON, J. M., M. KREITMAN and M. AGUADÉ, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in drosophila. *Genetics* **151**: 239–249.
- DAYHOFF, M. O., R. M. SCHWARTZ and B. C. ORCUTT, 1978 A model of evolutionary changes in proteins, pp. 345–352 in *Atlas of protein sequence and structure*. vol. 5, Suppl. 3 National Biomedical Research Foundation, Washington, D. C.
- DIMMIC, M. W., J. S. REST, D. P. MINDELL and R. A. GOLDSTEIN, 2002 rtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J. Mol. Evol.* **55**: 65–73.
- DUAN, J., M. S. WAINWRIGHT, J. M. COMERON, N. SAITOU, A. R. SANDERS *et al.*, 2003 Synonymous mutations in the human *dopamine receptor* D2 (DRD2)

- affect mRNA stability and synthesis of the receptor. *Human Molecular Genetics* **12**: 205–216.
- EYRE-WALKER, A., 1999 Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* **152**: 675–683.
- FELSENSTEIN, J., 1973 Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data of discrete characters. *Syst. Zool.* **22**: 240–249.
- FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**: 368–376.
- FELSENSTEIN, J., 2001 Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J. Mol. Evol.* **53**: 447–455.
- FELSENSTEIN, J., and G. A. CHURCHILL, 1996 A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**: 93–104.
- FITCH, W. M., and E. MARGOLIASH, 1967 Construction of phylogenetic trees. *Science* **155**: 279–284.
- FITCH, W. M., and E. MARKOWITZ, 1970 An improved method for determining codon variability in a gene and its applications to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4**: 579–593.
- GAUCHER, E. A., M. M. MIYAMOTO and S. A. BENNER, 2001 Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. *Proc. Natl. Acad. Sci. USA* **98**: 548–552.

- GAUT, B. S., and P. O. LEWIS, 1995 Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.* **12**: 152–162.
- GOLDING, G. B., 1983 Estimates of DNA and protein sequence divergence: an examination of some assumptions. *Mol. Biol. Evol.* **1**: 125–142.
- GOLDMAN, N., J. L. THORNE and D. T. JONES, 1996 Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.* **263**: 196–208.
- GOLDMAN, N., J. L. THORNE and D. T. JONES, 1998 Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**: 445–458.
- GOLDMAN, N., and Z. YANG, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**: 725–736.
- GU, X., Y. FU and W. LI, 1995 Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.* **12**: 546–557.
- GUOY, M., and W. LI, 1989 Phylogenetic analysis based on rRNA sequences supports the archaeobacterial rather than the eocyte tree. *Nature* **339**: 145–147.
- HASEGAWA, M., A. DIRIENZO, T. D. KOCHER and A. C. WILSON, 1993 Toward a more accurate time-scale for the human mitochondrial DNA tree. *J. Mol. Evol.* **37**: 347–354.
- HASEGAWA, M., H. KISHINO and T. YANO, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**: 160–174.

- HENIKOFF, S., and J. G. HENIKOFF, 1992 Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**: 10915–10919.
- HIGGS, P. G., 2000 RNA secondary structure: physical and computational aspects. *Quart. Rev. Biophys.* **33**: 199–253.
- HOLMQUIST, R., M. GOODMAN, T. CONROY and J. CZELUSNIAK, 1983 The spatial distribution of fixed mutations within genes coding for proteins. *J. Mol. Evol.* **19**: 437–448.
- HOLMQUIST, R., and D. PEARL, 1980 Theoretical foundations for quantitative paleogenetics. Part III: The molecular divergence of nucleic acids and proteins for the case of genetic events of unequal probability. *J. Mol. Evol.* **16**: 211–267.
- HUELSENBECK, J. P., and J. P. BOLLBACK, 2005 Empirical and hierarchical bayesian estimation of ancestral states. *Syst. Biol.* **50**: 351–366.
- HUELSENBECK, J. P., B. LARGET and M. E. ALFARO, 2004 Bayesian phylogenetic model selection using reversible jump markov chain monte carlo. *Mol. Biol. Evol.* **21**: 1123–1133.
- HUELSENBECK, J. P., and R. NIELSEN, 1999 Variation in the pattern of nucleotide substitution across sites. *J. Mol. Evol.* **48**: 86–93.
- HUELSENBECK, J. P., and F. RONQUIST, 2001 MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.
- HURST, L. D., and C. PÁL, 2001 Evidence for purifying selection acting on silent sites in BRCA1. *Trends in Genetics* **17**: 62–65.

- HWANG, D. G., and P. GREEN, 2004 Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. USA* **101**: 13994–14001.
- INA, Y., 1996 Pattern of synonymous and nonsynonymous substitutions: An indicator of mechanisms of molecular evolution. *J. Genet.* **75**: 91–115.
- IRWIN, D. M., T. D. KOCHER and A. C. WILSON, 1991 Evolution of the cytochrome-b gene of mammals. *J. Mol. Evol.* **32**: 128–144.
- JIN, L., and M. NEI, 1990 Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* **7**: 82–102.
- JONES, D. T., W. R. TAYLOR and J. M. THORNTON, 1992 The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**: 275–282.
- JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. M. MUNRO. Academic Press, New York.
- KATZ, L. A., J. G. BORNSTEIN, E. LASEK-NESELQUIST and S. V. MUSE, 2004 Dramatic diversity of ciliate histone H4 genes revealed by comparisons of patterns of substitutions and paralog divergences among eukaryotes. *Mol. Biol. Evol.* **21**: 555–562.
- KELLY, C., and J. RICE, 1996 Modeling nucleotide evolution: A heterogeneous rate analysis. *Math. Biosci.* **133**: 85–109.

- KIMURA, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.. *J. Mol. Evol.* **16**: 111–120.
- KISHINO, H., T. MIYATA and M. HASEGAWA, 1990 Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31**: 151–160.
- KISHINO, H., J. L. THORNE and W. J. BRUNO, 2001 Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**: 352–361.
- KOSAKOVSKY POND, S. L., and S. D. W. FROST, 2005 A simple hierarchical approach to modeling distributions of substitution rates. *Mol. Biol. Evol.* **22**: 223–234.
- KOSAKOVSKY POND, S. L., F. V. MANNINO, M. B. GRAVENOR, S. V. MUSE and S. D. W. FROST, 2006a Modeling protein sequence evolution through a genetic algorithm approach. *In preparation* .
- KOSAKOVSKY POND, S. L., F. V. MANNINO, M. B. GRAVENOR, S. V. MUSE and S. D. W. FROST, 2006b Modeling RNA sequence evolution subject to secondary structure constraints: A genetic algorithm approach. *In preparation* .
- KOSAKOVSKY POND, S. L., and S. V. MUSE, 2005 Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* **22**: 2375–2385.
- KOSHI, J. M., and R. A. GOLDSTEIN, 1998 Models of natural mutations including site heterogeneity. *Proteins* **32**: 289–295.

- KUHNER, M. K., and J. FELSENSTEIN, 1994 A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**: 459–468.
- LANAVE, C., G. PREPARATA, C. SACCONI and G. SERIO, 1984 A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20**: 86–93.
- LARSON, A., and A. C. WILSON, 1989 Patterns of ribosomal RNA evolution in salamanders. *Mol. Biol. Evol.* **6**: 131–154.
- LEE, Y., T. OTA and V. D. VACQUIER, 1995 Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* **12**: 231–238.
- LEITNER, T., S. KUMAR and J. ALBERT, 1997 Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.* **71**: 4761–4770.
- LI, W., M. GOUY, P. M. SHARP, C. O’HUGIN and Y. YANG, 1990 Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. *Proc. Natl. Acad. Sci. USA* **87**: 6703–6707.
- MEYER, S., and A. VON HAESELER, 2003 Identifying site-specific substitution rates. *Mol. Biol. Evol.* **20**: 182–189.
- MOLER, C., and C. VAN LOAN, 1978 Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev.* **20**: 801–836.
- MOROZOV, P., T. SITNIKOVA, G. CHURCHILL, F. J. AYALA and A. RZHETSKY, 2000 A new method for characterizing replacement rate variation in molecular

- sequences: Application of the faurier and wavelet models to drosophila and mammalian proteins. *Genetics* **154**: 381–395.
- MUSE, S. V., 1995 Evolutionary analysis of DNA sequences subject to constraints on secondary structure. *Genetics* **139**: 1429–1439.
- MUSE, S. V., 1999 Modeling the molecular evolution of HIV sequences, pp. 122–152 in *The Evolution of HIV*, edited by K. A. CRANDALL. The Johns Hopkins University Press, Baltimore, MD.
- MUSE, S. V., 2000 Examining rates and patterns of nucleotide substitution in plants. *Plant Mol. Biol.* **42**: 25–43.
- MUSE, S. V., and B. S. GAUT, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome. *Mol. Biol. Evol.* **11**: 715–724.
- NEI, M., R. CHAKRABORTY and P. A. FUERST, 1976 Infinite allele model with varying mutation rate. *Proc. Natl. Acad. Sci. USA* **73**: 4164–4168.
- NEI, M., and T. GOJOBORI, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**: 418–426.
- NEYMAN, J., 1971 Molecular studies of evolution: a source of novel statistical problems, pp. 1–27 in *Statistical decision theory and related topics*, edited by S. S. GUPTA and J. YACKEL. Academic Press, New York, NY.
- NIELSEN, R., 1997 Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. *Syst. Biol.* **46**: 346–353.

- NIELSEN, R., and Z. YANG, 1998 Likelihood models for detecting positive selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929–936.
- OLSEN, G. J., 1987 Earliest phylogenetic branchings - comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harb. Symp. Quant. Biol.* **52**: 825–837.
- OTA, R., P. J. WADDELL, M. HASEGAWA, H. SHIMODAIRA and H. KISHINO, 2000 Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol. Biol. Evol.* **17**: 798–803.
- PAGANI, F., M. RAPONI and F. E. BARALLE, 2005 Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl. Acad. Sci. USA* **102**: 6368–6372.
- PALUMBI, S. R., 1989 Rates of molecular evolution and the fraction of nucleotide positions free to vary. *J. Mol. Evol.* **29**: 180–187.
- PARISI, G., and J. ECHAVE, 2001 Structural constraints and emergence of sequence patterns in protein evolution. *Mol. Biol. Evol.* **18**: 750–756.
- POSADA, D., and K. A. CRANDALL, 1998 Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.
- POSADA, D., and K. A. CRANDALL, 2001 Selecting the best fit model of nucleotide substitution. *Syst. Biol.* **50**: 580–601.
- ROBINSON, D. M., D. T. JONES, H. KISHINO, N. GOLDMAN and J. L. THORNE, 2003 Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.* **20**: 1692–1704.

- RZHETSKY, A., 1995 Estimating substitution rates in ribosomal RNA genes. *Genetics* **141**: 771–783.
- SAVILL, N. J., D. C. HOYLE and P. G. HIGGS, 2001 RNA sequence evolution with secondary structure constraints: Comparison of substitution rate models using maximum-likelihood methods. *Genetics* **157**: 399–411.
- SCHÖNIGER, M., and A. VON HAESLER, 1994 A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylo. Evol.* **3**: 240–247.
- SCHWARTZ, G., 1978 Estimating the dimension of a model. *Ann. Stat.* **6**: 461–464.
- SELF, S. G., and K.-Y. LIANG, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Stat. Assoc.* **82**: 605–610.
- SIEGEL, A. F., 1979 The noncentral chi-squared distribution with zero degrees of freedom and testing for uncertainty. *Biometrika* **66**: 381–386.
- SIMON, A. L., E. A. STONE and A. SIDOW, 2002 Inference of functional regions in proteins by qualification of evolutionary constraints. *Proc. Natl. Acad. Sci. USA* **99**: 2912–2917.
- SMITH, N. G. C., and L. D. HURST, 1998 Molecular evolution of an imprinted gene: Repeatability of patterns of evolution within the mammalian insulin-like growth factor type II receptor. *Genetics* **150**: 823–833.
- SMITH, N. G. C., and L. D. HURST, 1999 The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153**: 1395–1402.

- SULLIVAN, J., K. E. HOLSINGER and C. SIMON, 1995 Among site rate variation and phylogenetic analysis of 12S rRNA in Sigmodontine rodents. *Mol. Biol. Evol.* **12**: 988–1001.
- SULLIVAN, J., and D. L. SWOFFORD, 1997 Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J. Mammal. Evol.* **4**: 77–86.
- SULLIVAN, J., and D. L. SWOFFORD, 2001 Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated?. *Syst. Biol.* **50**: 723–729.
- TAKAHATA, N., 1991 Overdispersed molecular clock at the major histocompatibility complex. *Proc. R. Soc. Lond. B* **243**: 13–18.
- TAKEZAKI, N., and T. GOJOBORI, 1999 Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. *Mol. Biol. Evol.* **16**: 590–601.
- TAMURA, K., and M. NEI, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**: 512–526.
- TATENO, Y., N. TAKEZAKI and M. NEI, 1994 Relative efficiencies of the maximum likelihood, neighbor joining and maximum parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* **11**: 261–277.

- TAVARÉ, S., 1986 Some probabilistic and statistical problems in the analysis of DNA sequences, pp. 57–86 in *Lectures on Mathematics in the Life Sciences*, edited by R. M. MIURA. Amer. Math. Soc., Providence, R.I.
- THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Research* **22**: 4673–4680.
- THORNE, J. L., N. GOLDMAN and D. T. JONES, 1996 Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13**: 666–673.
- TILLIER, E. R. M., 1994 Maximum likelihood with multi-parameter models of substitution. *J. Mol. Evol.* **39**: 409–417.
- TILLIER, E. R. M., and R. A. COLLINS, 1998 High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics* **148**: 1993–2002.
- UPHOLT, W. B., 1977 Evolution of DNA sequence divergence from comparison of restriction endonuclease digests. *Nucl. Acids Research* **4**: 1257–1265.
- UZZELL, T., and K. W. CORBIN, 1971 Fitting discrete probability distribution to evolutionary events. *Science* **172**: 1089–1096.
- VAN DE PEER, Y., J. NEEFS, P. DE RIJK and R. DE WACHTER, 1993 Reconstructing evolution from eukaryotic small ribosomal subunit RNA sequence: Calibration of the molecular clock. *J. Mol. Evol.* **37**: 221–232.

- WADDELL, P. J., D. PENNY and T. MOORE, 1997 Hadamard conjugations and modeling sequence evolution with unequal rates across sites. *Mol. Phylo. Evol.* **8**: 33–50.
- WAKELEY, J., 1994 Substitution rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* **11**: 436–442.
- WHELAN, S., and N. GOLDMAN, 1999 Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **16**: 1292–1299.
- WHELAN, S., and N. GOLDMAN, 2004 Estimating the frequency of events that cause multiple nucleotide changes. *Genetics* **167**: 2027–2043.
- YANG, Z., 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**: 1396–1401.
- YANG, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**: 105–111.
- YANG, Z., 1995 A space-time process model for the evolution of DNA sequences. *Genetics* **139**: 993–1005.
- YANG, Z., 1997 How often do wrong models produce better phylogenies?. *Mol. Biol. Evol.* **14**: 105–108.
- YANG, Z., N. GOLDMAN and A. FRIDAY, 1994 Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**: 316–324.

- YANG, Z., N. GOLDMAN and A. FRIDAY, 1995 Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* **44**: 384–399.
- YANG, Z., and S. KUMAR, 1996 Appropriate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.* **13**: 650–659.
- YANG, Z., R. NIELSEN, N. GOLDMAN and A. PEDERSEN, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431–449.
- YANG, Z., R. NIELSEN and M. HASEGAWA, 1998 Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**: 1600–1611.
- ZHANG, J., 1999 Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol. Biol. Evol.* **16**: 868–875.

Chapter 2

EXTENSIVE SITE-TO-SITE VARIABILITY OF SYNONYMOUS SUBSTITUTION RATES IN MITOCHONDRIAL GENOMES

Mannino FV and SV Muse

Abstract

We investigate the presence of site-to-site variability of synonymous substitution rates using a recently developed model of codon evolution. In an analysis of all 13 protein coding genes from the completed mitochondrial genomes of a total of 111 taxa sampled from ten animal clades, we find extensive variability of synonymous rates in the majority of genes (77% of the analyzed data sets showed statistically significant levels). Furthermore, the magnitude of synonymous rate variation as measured by the coefficient of variation is comparable to that of nonsynonymous rates, including three data sets where there is more synonymous rate heterogeneity than nonsynonymous heterogeneity. These findings raise interesting questions about underlying mechanisms of selection at silent sites, and they have important implications for analyses— especially studies of adaptive evolution— that are based on the premise that silent rates are equal across all sites in a gene.

The main body of this chapter was prepared and submitted as a manuscript for publication in *Genetics*. The section titled “Further results” consists of extra analysis not including in the manuscript. The plots at the end of this chapter are identical to the other plots in the manuscript, but use a different estimator of the coefficient of variation (see Chapter 4 for justification).

Introduction

The identification and quantification of sources of nucleotide substitution rate heterogeneity are central to molecular studies of evolution. Researchers need to understand how genes and genomes evolve at the molecular level, and accurate estimation of substitution rates is a prerequisite for such investigations. All statistical

methods for studying the molecular evolutionary process make assumptions regarding the underlying process of sequence change, whether implicit or explicit, that allow inferences to be made with acceptable levels of computational expense and with enough statistical power to make inferences. When studying the evolution of protein coding genes, one such assumption has traditionally been that the rate of synonymous substitution is constant from position to position within the gene being analyzed (e.g., GOLDMAN and YANG, 1994; MUSE and GAUT, 1994; NIELSEN and YANG, 1998); this assumption is sometimes extended to include constancy across multiple genes when concatenated genes are analyzed as a single unit (e.g. YANG *et al.*, 2000).

Mitochondrial genomes are often used in evolutionary studies (e.g. CAO *et al.*, 1994; ZARDOYA and MEYER, 1996; PARHAM *et al.*, 2006; ARIS-BROSOU and YANG, 2003) because of useful properties of the organelle's DNA, including lack of introns, lack of recombination (OLIVIO *et al.*, 1983), and maternal inheritance (GILES *et al.*, 1980). The relatively small size (~ 16 kb) and limited number of protein coding genes (12-13) in animals makes sequencing simple and cheap. Finally, and perhaps most importantly, the evolutionary rate of mitochondrial genomes is quite high in many lineages (BROWN *et al.*, 1979), making mitochondrial genes ideal tools for studies of relatively recent evolutionary events. The widespread use of these organellar genomes is demonstrated in the presence of over 700 completely sequenced mitochondrial genomes in GenBank, a number that continues to grow rapidly.

In this work we survey patterns of synonymous substitution rates using a collection of all protein coding genes from ten clades, encompassing a wide range of animals. We find strong evidence that synonymous rates are not constant across sites, neither within nor between genes, and that the magnitude of this vari-

ability is comparable to the well-documented variability of nonsynonymous rates. Our results are consistent with recent work examining synonymous rate variation (KOSAKOVSKY POND and MUSE, 2005; HURST and PÁL, 2001; KOSAKOVSKY POND and FROST, 2005b).

Materials and Methods

Sequences, Alignments, and Trees

We compiled data sets for ten distinct clades using the complete mitochondrial genomes available from GenBank at NCBI. The clades and numbers of taxa sampled were carnivores (10), insects (15), mollusks (8), nematodes (7), neognathae (12), palaeognathae (11), percomorpha (19), platyhelminthes (9), primates (13), and rodents (7). Taxa were selected to provide clade sizes sufficient for maintaining statistical power (SULLIVAN *et al.*, 1999), to limit computational costs, and to avoid having clades with so much sequence divergence that accurate inference was impossible because of saturation effects. The complete set of taxa and GenBank accession numbers can be found in the Table 2.3 in the Appendix.

The typical metazoan mitochondrial genome contains 13 protein coding genes: NADH dehydrogenase subunits 1-6 and 4L (NADH1 - NADH6,NADH4L), cytochrome c oxidase subunits I, II, and III (COX1-COX3), ATP synthase F0 subunits 6 and 8 (ATP6,ATP8), and cytochrome b (CYTB). These genomes also typically include 2 rRNA genes and 22 tRNA genes. In three clades (mollusks, nematodes, and platyhelminthes) the ATP8 gene is not present in one or more species; ATP8 was excluded from the analyses of those three clades. The 10 clades \times 13 genes/genome $-$ 3 ATP8 genes yielded a total of 127 gene \times clade combina-

tions in this study. We will refer to each of these gene×clade pairs as a *data set*. For many genomes there are regions of overlapping genes ranging in size from 1 to 46 nucleotides. ATP8 and ATP6 overlapped in all taxa that contained ATP8. In addition ATP6/COX3 frequently overlapped as well as NADH4/NADH4L and NADH5/NADH6. We removed all codons containing any overlapping nucleotides from the data sets, because those sites are likely to experience evolutionary constraints that are quite different from “typical” positions.

The protein sequences were aligned using ClustalW (THOMPSON *et al.*, 1994) then adjusted manually. DNA alignments were obtained by reverse translation. Phylogenetic trees for each clade were found using neighbor-joining (SAITOU and NEI, 1987) as implemented in the MEGA2 program (KUMAR *et al.*, 1993). The trees were built using the concatenation of all 13 protein sequences. Pairwise distances were estimated using the Poisson correction model. This procedure leads to a single topology for each of our ten clades (see Figures 2.4-2.8 in Appendix). Our trees were identical, or nearly identical, to published phylogenies (CAMPOS *et al.*, 1998; REYES *et al.*, 2004; BAXTER *et al.*, 1998; MIYA *et al.*, 2001; DYKE and VAN TUINEN, 2004; VAN TUINEN *et al.*, 2000; WHEELER *et al.*, 2001), but detailed phylogenetic studies of all taxa were not available for every clade. We acknowledge the fact that these trees are not likely to all be exactly correct. However, inferences about patterns of substitution rates are robust to slight flaws in the estimated tree (YANG *et al.*, 2000)

Statistical methods

Constructing the likelihood function (FELSENSTEIN, 1981) for codon-based alignments requires codon-based evolutionary models such as those of GOLDMAN and

YANG (1994) or MUSE and GAUT (1994). Standard “discretized gamma” models (YANG, 1993, 1994) allow for heterogeneity of nonsynonymous rates (NIELSEN and YANG, 1998); however, these models share the assumption that synonymous substitution rates are identical across all sites in the sequence. Recently, KOSAKOVSKY POND and MUSE (2005) introduced a new class of codon models, an extension of the MUSE and GAUT (1994) model that allows both synonymous and nonsynonymous rates to vary across sites according to independent gamma distributions. Using their notation, codon s has associated with it a nonsynonymous rate β_s and a synonymous rate α_s . The bivariate distribution of these rates is governed by parameters μ_β , ω (nonsynonymous rates), and μ_α (synonymous rates) as follows:

$$\alpha_s \sim \text{Gamma}(\mu_\alpha, \mu_\alpha)$$

$$\beta_s \sim \omega \times \text{Gamma}(\mu_\beta, \mu_\beta)$$

The synonymous distribution has mean one and variance $\frac{1}{\mu_\alpha}$. The nonsynonymous distribution has mean ω and variance $\frac{\omega^2}{\mu_\beta}$. This model allows us to estimate the variability of both synonymous and nonsynonymous rates across sites, along with their coefficients of variation (CV = standard deviation / mean). The means of the discretized approximations will be identical to those of their continuous counterparts, while the discretized variances will be slightly less than those of the continuous distributions (see KOSAKOVSKY POND and MUSE (2005) for complete details).

KOSAKOVSKY POND and MUSE (2005) describe a variety of models and suggest the use of the Akaike Information Criterion (AIC, (AKAIKE, 1974)) in conjunction with likelihood ratio tests to select from among those models. Following their guidelines and notation, the model used in the current work is MG94×REV Dual

Gamma 4×4 . This model incorporates gamma-based rate heterogeneity of both synonymous and nonsynonymous rates, accounts for nucleotide substitution biases using the general time reversible (REV) model (e.g. TAVARÉ, 1986), and discretizes each of the gamma distributions into four categories using the median values over distribution intervals following the ideas in YANG (1994). Choosing the number of discrete classes to use for the gamma distribution is a matter of compromise between computational time and a closer approximation to the true continuous distribution. Separate sets of base frequencies were used at each of the three codon positions to account for the extensive biased composition observed in our mtDNA data sets. Using this model, data sets were analyzed using the HyPhy program of KOSAKOVSKY POND *et al.* (2005). Maximum likelihood estimates of all model parameters were computed, and the presence of synonymous rate variation was identified using likelihood ratio tests. In this setting, the null hypothesis of no synonymous rate heterogeneity is reached at the boundary-like condition of $\mu_\alpha = \infty$. Because of this fact, a slight modification to the traditional likelihood ratio test must be applied. The standard likelihood ratio test statistic, $-2(L_O - L_A)$, is still used, but instead of comparing it to a χ_1^2 distribution, it is compared to a mixed χ^2 (or $\bar{\chi}^2$) distribution that consists of 50% χ_0^2 and 50% χ_1^2 (see SIEGEL, 1979; SELF and LIANG, 1987). Studies have shown that when testing for the presence or absence of rate heterogeneity, the asymptotic distribution of this test statistic is described much better by the mixture distribution than by the nonmixture (OTA *et al.*, 2000; WHELAN and GOLDMAN, 1999).

For completeness, the data sets were also analyzed under mean discretization and the MG94×HKY (HASEGAWA *et al.*, 1985) model of nucleotide substitution. These results are omitted since the model variations have almost no effect on the parameter estimates and test results presented in this study. Almost all data

sets fit better under the REV models, as determined by a likelihood ratio test. Model selection using AIC also tended to favor median discretization. The mean discretization model leads to slightly more extreme rates and therefore slightly more variation than does the median model. Thus, our choice to present results based on median discretization is conservative in some sense. Results of tests for the presence of synonymous rate variation differed in only two of our 127 data sets, and estimates of the CVs were similar in all cases.

Results and Discussion

Concatenated vs. Separated Genes

Multi-gene data sets have frequently been analyzed by concatenating the genes into a single alignment (e.g. YANG, 1995; PUPKO *et al.*, 2002) We first tested for each clade the straw man hypothesis that all 13 genes were evolving with equal rates by comparing the fit of the concatenated collection of all 13 genes to a model with separate parameters for each gene. This test was rejected in each of the ten cases with p-values less than 10^{-15} , and rejects the notion that all genes are evolving at the same rates and with equal amounts of variation. More practically, it justifies the separate analysis of each gene \times clade data set.

Synonymous rate variation is ubiquitous, and of non-negligible magnitude

For each data set we tested the null hypothesis of no site-to-site synonymous rate heterogeneity, $H_0 : \mu_\alpha = \infty$, using the likelihood ratio test described above. This hypothesis was rejected at the 0.05 significance level in 98 of the 127 data sets (Ta-

ble 2.1), providing strong evidence that variability of synonymous substitution rates is the norm in animal mitochondrial genomes, not the exception. The results in Table 2.2 reinforce our claim that this result is robust to the precise model chosen for the analyses.

In Figure 2.1 we present box plots that summarize estimates of the coefficients of variation for synonymous and nonsynonymous substitution rates in our ten clades. Each individual box plot shows the distribution of the CV estimates across the 13 genes in each data set. Several observations are immediate. First, and most important, the top panel shows clearly that the CV for synonymous rates is typically nonzero. Only one of the clades (mollusks) have any substantial number of genes estimated to have little or no synonymous rate heterogeneity. Second, we see that (as expected) the same is true for nonsynonymous rates. Also in agreement with expectations, the variability of nonsynonymous rates is less uniform across clades than is that of synonymous rates. Finally, the magnitude of the difference between heterogeneity in the two types of rates is surprisingly low. While it is true that nonsynonymous rates tend to have more variability (as reflected in their higher CV estimates), the CV estimates for synonymous rates are usually 30-50 percent of those for nonsynonymous rates in these clades. In other words, these data suggest that nonsynonymous rates tend to be only 2-3 times as variable as synonymous rates. Keeping in mind that standard statistical methods assume that there is absolutely no variability in synonymous rates, this result is striking.

Table 2.1: **Data sets with significant synonymous rate variation.** Each data set, or clade \times gene pair, contained significant synonymous rate variation at the 0.05 level if marked by a *. NA indicates a data set that was missing from this study.

Clade	ATP6	ATP8	NADH1	NADH2	NADH3	NADH4	NADH4L	NADH5	NADH6	COX1	COX2	COX3	CYTB
carnivores	*	*	*	*	*		*	*		*	*		
insects	*		*	*	*	*	*	*		*	*	*	*
mollusks		NA			*	*					*		
nematodes	*	NA	*	*	*	*		*		*	*	*	
neognathae	*	*	*	*	*	*	*	*	*	*	*	*	*
palaeognathae		*	*	*		*	*	*	*	*	*	*	*
percomorpha	*	*	*	*	*	*	*	*	*	*	*	*	*
platyhelminthes		NA	*	*		*	*		*	*			*
primates	*	*	*		*		*	*		*	*	*	*
rodents	*	*	*	*	*		*	*	*	*	*	*	*

Table 2.2: **Significance of synonymous rate variation.** Counts of the 127 data sets with and without significant synonymous rate variation for our primary model (REV Median) and 3 alternative models. The last line shows the number of data sets that have identical results regardless of the model chosen for the analyses.

Model	Significant Synonymous Variation	
	Yes	No
REV Median	98	29
HKY Median	93	34
REV Mean	96	31
HKY Mean	94	33
Identical results for all models	88	24

In Figure 2.2 we present a different view of these estimates, creating a scatter plot of synonymous (X axis) versus nonsynonymous (Y axis) CVs for each data set. For example, the primates NADH4L data set has a synonymous CV of 0.57 and a nonsynonymous CV of 1.18, indicating about half as much synonymous variation as nonsynonymous. The plane was divided into three sectors based on the relationship between the synonymous and nonsynonymous CV estimates. Not surprisingly, most genes were estimated to have nonsynonymous CVs greater than their synonymous CVs (the two left sectors). However, there were three data sets out of the 127 where the synonymous CV was actually greater than the nonsynonymous CV. The middle sector delineates data sets where the nonsynonymous CV was estimated to be no more than twice the synonymous CV. It is quite remarkable that roughly half the data sets fall into this range, reinforcing the notion that synonymous substitution rates vary across sites with appreciable magnitude.

The mollusks were unique in the finding that only three of 12 genes (ATP8 is missing) rejected the hypothesis no synonymous variation. Similarly, the nema-

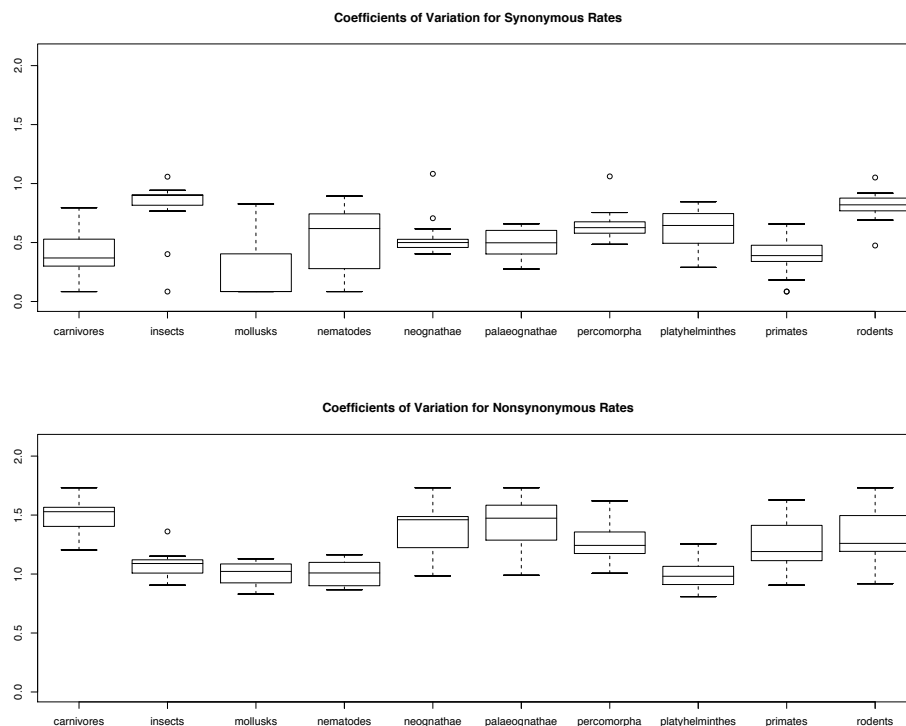


Figure 2.1: **Box plot of coefficients of variation for clades.** Box plot of the estimated synonymous and nonsynonymous coefficients of variation for all genes within each of the 10 clades. Each box indicates the median and first and third quartiles of the distribution of CV estimates for each gene in the corresponding clade.

todes only had significant synonymous variation in 7 of 12 genes. One possible reason might be that with only eight taxa (seven in nematodes), there was little power. However, the rodents had seven taxa and rejected in 12 of 13 genes, demonstrating that ample power can be found in data sets with low numbers of taxa.

As seen in the box plots in Figure 2.3, the distribution of the synonymous coefficients of variation seems relatively similar across genes, while those for nonsynonymous rates appear more variable. The more consistent amount of synonymous variation along a genome suggests that the forces causing the variation might be

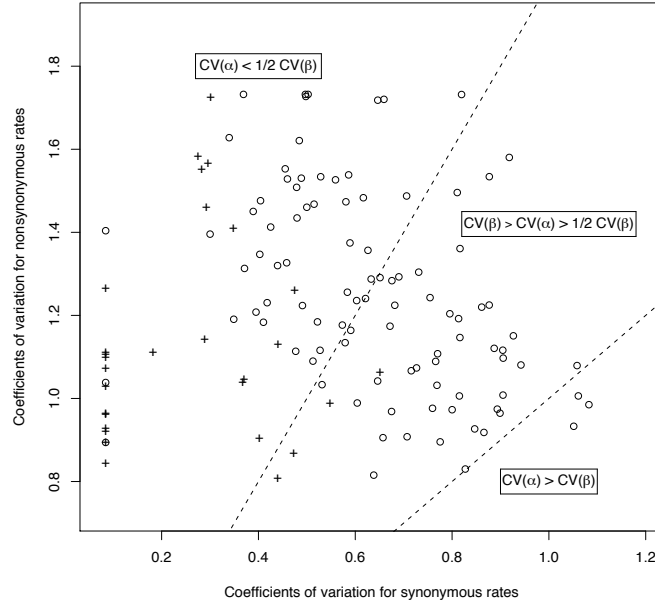


Figure 2.2: **Comparison of coefficients of variation of synonymous and nonsynonymous rates.** This scatter plot illustrates the relationship between estimated CVs for synonymous and nonsynonymous rates for each data set, highlighting the comparable magnitude of the two types of variability. Each plotted circle represents a gene with significant synonymous rate variation, while each plus is a gene without significant variation. The dashed lines separate the graph into regions of synonymous rate coefficient of variation less than half that of nonsynonymous rate ($CV(\alpha) < \frac{1}{2}CV(\beta)$), less than the nonsynonymous rate but greater than one half that rate, and greater than the nonsynonymous rate. These regions contain 73, 51, and 3 data sets, respectively.

more genome-specific than gene-specific; the higher variability of nonsynonymous rates is anticipated because of functional constraints that vary from gene to gene.

Positive Selection

Although not our central thesis, an obvious question to address with this type of data is whether sites are under positive selection. Widely used statistical methods allow one to calculate a posterior probability for positive selection at each individ-

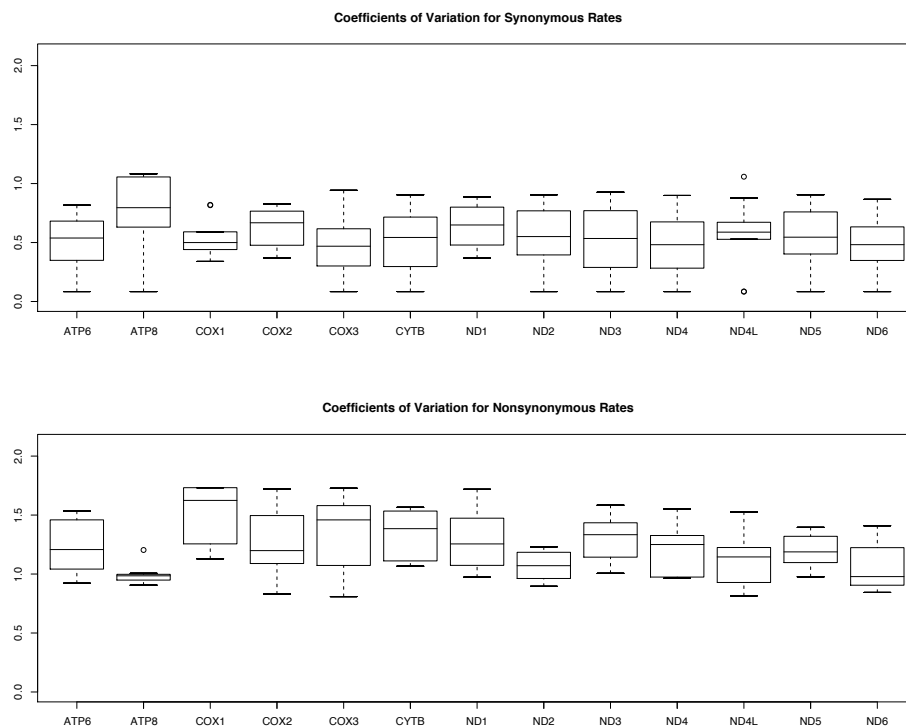


Figure 2.3: **Box plot of coefficients of variation for genes.** Box plot of the synonymous and nonsynonymous coefficients of variation for all clades within each of the 13 genes.

ual codon in a data set (NIELSEN and YANG, 1998). In order to identify such sites it is necessarily the case that the estimated nonsynonymous rate is greater than the synonymous rate for one or more of the discretized rate classes. In this study only 14 data sets contained classes in which the nonsynonymous rate exceeded the synonymous rate. In three cases, two of the $4 \times 4 = 16$ classes suggested the presence of positive selection, while the remaining 11 data sets produced only one such class. However, the posterior probabilities of a site belonging to a class under positive selection was generally low, with the largest values around 0.75. Taken in full, there is only weak evidence for the presence of positive selection in these mitochondrial genomes.

It is interesting, however, to compare these results with those from methods in which synonymous substitution rates are assumed to be equal across sites, as is typical in current methods for detecting selection. We simplified our model by allowing only nonsynonymous rates to be variable across sites and applied the positive selection procedure of NIELSEN and YANG (1998). Using the notation of KOSAKOVSKY POND and MUSE (2005) this simplified model is denoted MG94 \times REV *Nonsynonymous* Gamma 4, and it is very similar to the M5 model of YANG *et al.* (2000). These analyses led to no genes with a class having an estimated nonsynonymous rate greater than the synonymous rate. When the number of classes was increased to 16 (to match the 16 classes used above in the bivariate model and to keep computational expense roughly equal), only three data sets produced classes indicative of positively selected sites. While neither approach (i.e., either with or without variable synonymous rates) implies extensive positive selection, the difference between the two analyses is consistent with either (i) the presence of enhanced power when accounting for synonymous rate heterogeneity, or (ii) inflated Type I error rates when incorporating synonymous rate heterogeneity. We have seen no evidence for the latter in prior work (KOSAKOVSKY POND and FROST, 2005a; KOSAKOVSKY POND and MUSE, 2005), nor is there any good reason to expect these procedures to have poor sampling properties. However, it is quite possible that some sites have low mutation rates, leading to low synonymous rates at those positions. If this is the case, traditional methods will still compare the “average” synonymous rate across sites to the nonsynonymous rate inferred for that site, leading to a lack of positive selection signal. Looking at the genes with potential positive selection, we find two interesting observations. First, six of the seven clades with ATP8 contain a class with positive selection for that gene. Note in Figure 2.3 that this gene has the highest amount of synonymous variation of the

studied loci, an observation supporting the possibility of the positive selection signal being masked in studies ignoring synonymous rate heterogeneity. Second, the insect clade had five genes with potential positive selection. Similar to the ATP8 results, the insect clade had the highest levels of synonymous rate heterogeneity of the studied groups (Figure 2.1).

Implications

The most significant finding of this study is the rejection of the common assumption that the synonymous substitution rate is fairly constant across sites within individual genes. We demonstrate that in mitochondrial genomes from taxonomically diverse organisms these rates are substantially variable at most genes. The presence of this variability is both puzzling and potentially troublesome. It leads to the obvious biological question regarding the mechanism(s) underlying the variability. Codon bias might produce such a pattern, however, the degree of variability within single genes makes this explanation somewhat tenuous. More likely candidates seem to be the presence of mutational hotspots (GALTIER *et al.*, 2005), or weak purifying selection at silent sites to avoid features such as transcription factor binding sites or secondary structures (TUPLIN *et al.*, 2002). The more problematic implications of the finding are those related to potential statistical artifacts that arise when synonymous rate heterogeneity is ignored. While we present some evidence for effects on positive selection studies (and additional evidence is provided by KOSAKOVSKY POND and MUSE (2005)), any analysis that hinges on estimation of synonymous substitution rates might be potentially at risk.

Acknowledgments

This work has been supported by the National Science Foundation grants to SVM, and by predoctoral fellowships to FVM from the National Institutes of Environmental Health Sciences and the National Science Foundation. We thank Sergei Kosakovsky Pond and Errol Strain for helpful suggestions.

References

- AKAIKE, H., 1974 A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* **19**: 716–723.
- ARIS-BROUSO, S., and Z. YANG, 2003 Bayesian model of episodic evolution support a late precambrian explosive diversification of the metazoa. *Mol. Biol. Evol.* **20**: 1947–1954.
- BAXTER, M. L., P. D. LEY, J. R. GAREY, L. X. LIU, P. SCHELDEMAN *et al.*, 1998 A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**: 71–75.
- BROWN, W. M., M. GEORGE JR. and A. C. WILSON, 1979 Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **76**: 1967–1971.
- CAMPOS, A., M. P. CUMMINGS, J. L. REYES and J. P. LACLETTE, 1998 Phylogenetic relationships of platyhelminthes based on 18S ribosomal gene sequences. *Mol. Phylo. Evol.* **10**: 1–10.
- CAO, Y., J. ADACHI, A. JANKE, S. PÄÄBO and M. HASEGAWA, 1994 Phylogenetic relationships among Eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J. Mol. Evol.* **35**: 519–527.

- DYKE, G. J., and M. VAN TUINEN, 2004 The evolutionary radiation of modern birds (Neornithes): reconciling molecules, morphology and the fossil record. *Zool. J. Linn. Soc.* **141**: 153–177.
- FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**: 368–376.
- GALTIER, N., D. ENARD, Y. RADONDY, E. BAZIN and K. BELKHIR, 2005 Mutation hot spots in mammalian mitochondrial DNA. *Genome Research* **In Press**: In Press.
- GILES, R. E., H. BLANC, H. M. CANN and D. C. WALLACE, 1980 Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **77**: 6715–6719.
- GOLDMAN, N., and Z. YANG, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**: 725–736.
- HASEGAWA, M., H. KISHINO and T. YANO, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**: 160–174.
- HURST, L. D., and C. PÁL, 2001 Evidence for purifying selection acting on silent sites in BRCA1. *Trends in Genetics* **17**: 62–65.
- KOSAKOVSKY POND, S. L., and S. D. W. FROST, 2005a Not so different after all: comparison of various methods for detecting amino-acid sites under selection. *Mol. Biol. Evol.* **22**: 1208–1222.

- KOSAKOVSKY POND, S. L., and S. D. W. FROST, 2005b A simple hierarchical approach to modeling distributions of substitution rates. *Mol. Biol. Evol.* **22**: 223–234.
- KOSAKOVSKY POND, S. L., S. D. W. FROST and S. V. MUSE, 2005 HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**: 676–679.
- KOSAKOVSKY POND, S. L., and S. V. MUSE, 2005 Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* **22**: 2375–2385.
- KUMAR, S., 1995 Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics* **143**: 537–548.
- KUMAR, S., K. TAMURA and M. NEI, 1993 *MEGA: Molecular Evolutionary Genetics Analysis, version 1.01*. The Pennsylvania State University, University Park, PA.
- MIYA, M., A. KAWAGUCHI and M. NISHIDA, 2001 Mitogenomic exploration of higher Teleostean phylogenies: A case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial DNA sequences. *Mol. Biol. Evol.* **18**: 1993–2009.
- MUSE, S. V., and B. S. GAUT, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome. *Mol. Biol. Evol.* **11**: 715–724.
- NIELSEN, R., and Z. YANG, 1998 Likelihood models for detecting positive selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929–936.

- OLIVIO, P. D., M. J. VAN DE WALLE, P. J. LAIPIS and W. W. HAUSWIRTH, 1983 Nucleotide sequence evidence for rapid genotypic shifts in bovine mitochondrial dna D-loop. *Nature* **306**: 400–402.
- OTA, R., P. J. WADDELL, M. HASEGAWA, H. SHIMODAIRA and H. KISHINO, 2000 Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol. Biol. Evol.* **17**: 798–803.
- PARHAM, J. F., J. R. MACEY, T. J. PAPENFUSS, C. R. FELDMAN, O. TÜRKOZAN *et al.*, 2006 The phylogeny of mediterranean tortoises and their close relatives based on complete mitochondrial genome sequences from museum specimens. *Mol. Phylo. Evol.* **38**: 50–64.
- PUPKO, T., D. HUCHON, Y. CAO, N. OKADA and M. HASEGAWA, 2002 Combining multiple data sets in a likelihood analysis: Which models are the best?. *Mol. Biol. Evol.* **19**: 2294–2307.
- REYES, A., C. GISSI, F. CATZEFLIS, E. NEVO, G. PESOLE *et al.*, 2004 Congruent mammalian trees from mitochondrial and nuclear genes using bayesian methods. *Mol. Biol. Evol.* **21**: 397–403.
- SAITOU, N., and M. NEI, 1987 The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406–425.
- SELF, S. G., and K.-Y. LIANG, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Stat. Assoc.* **82**: 605–610.
- SIEGEL, A. F., 1979 The noncentral chi-squared distribution with zero degrees of freedom and testing for uncertainty. *Biometrika* **66**: 381–386.

- SULLIVAN, J., D. L. SWOFFORD and G. J. P. NAYLOR, 1999 The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol. Biol. Evol.* **16**: 1347–1356.
- TAVARÉ, S., 1986 Some probabilistic and statistical problems in the analysis of DNA sequences, pp. 57–86 in *Lectures on Mathematics in the Life Sciences*, edited by R. M. MIURA. Amer. Math. Soc., Providence, R.I.
- THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Research* **22**: 4673–4680.
- TUPLIN, A., J. WOOD, D. J. EVANS, A. H. PATEL and P. SIMMONS, 2002 Thermodynamic and phylogenetic prediction of RNA secondary structure in the coding region of hepatitis C virus. *RNA- A Publication of the RNA Society* **82**: 824–841.
- VAN TUINEN, M., C. G. SIBLEY and S. B. HEDGES, 2000 The early history of modern birds inferred from DNA sequences of nuclear and mitochondrial ribosomal genes. *Mol. Biol. Evol.* **17**: 451–457.
- WHEELER, W. C., M. WHITING, Q. D. WHEELER and J. M. CARPENTER, 2001 The phylogeny of the extant hexapod orders. *Cladistics* **17**: 113–169.
- WHELAN, S., and N. GOLDMAN, 1999 Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **16**: 1292–1299.

- YANG, Z., 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**: 1396–1401.
- YANG, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**: 105–111.
- YANG, Z., 1995 A space-time process model for the evolution of DNA sequences. *Genetics* **139**: 993–1005.
- YANG, Z., R. NIELSEN, N. GOLDMAN and A. PEDERSEN, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431–449.
- ZARDOYA, R., and A. MEYER, 1996 Phylogenetic relationships of mitochondrial protein-coding genes in resolving relationships among vertebrates. *Mol. Biol. Evol.* **13**: 933–942.

Appendix

Table 2.3: **Clades.** Species used in analyses

Clades Species	Common name	GenBank Accession Number
Carnivores		
<i>Arctocephalus forsteri</i>	New Zealand fur seal	NC_004023
<i>Canis familiaris</i>	dog	NC_002008
<i>Eumetopias jubatus</i>	Steller sea lion	NC_004030
<i>Felis catus</i>	cat	NC_001700
<i>Halichoerus grypus</i>	gray seal	NC_001602
<i>Odobenus rosmarus rosmarus</i>	Atlantic walrus	NC_004029
<i>Phoca vitulina</i>	harbor seal	NC_001325
<i>Ursus americanus</i>	American black bear	NC_003426
<i>Ursus arctos</i>	brown bear	NC_003427
<i>Ursus maritimus</i>	polar bear	NC_003428
Insects		
<i>Anopheles gambiae</i>	African malaria mosquito	NC_002084
<i>Anopheles quadrimaculatus A</i>	mosquito	NC_000875
<i>Apis mellifera ligustica</i>	common honey bee	NC_001566
<i>Bombyx mandarina</i>	wild silkworm	NC_003395
<i>Bombyx mori</i>	domestic silkworm	NC_002355
<i>Ceratitis capitata</i>	Mediterranean fruit fly	NC_000857
<i>Chrysomya putoria</i>	tropical African latrine blowfly	NC_002697
<i>Cochliomyia hominivorax</i>	primary screw-worm	NC_002660
<i>Crioceris duodecimpunctata</i>	spotted asparagus beetle	NC_003372
<i>Drosophila melanogaster</i>	fruit fly	NC_001709
<i>Drosophila yakuba</i>	fruit fly	NC_001322
<i>Ostrinia furnacalis</i>	Asian corn borer	NC_003368
<i>Ostrinia nubilalis</i>	European corn borer	NC_003367
<i>Pyrocoelia rufa</i>	firefly	NC_003970
<i>Tribolium castaneum</i>	red flour beetle	NC_003081
Mollusks * = Missing ATP8		
<i>Albinaria caerulea</i>	door snail	NC_001761
<i>Cepaea nemoralis</i>	banded wood snail	NC_001816
<i>Crassostrea gigas</i> *	Pacific oyster	NC_001276
<i>Katharina tunicata</i>	black chiton	NC_001636
<i>Loligo bleekeri</i>	Bleeker's squid	NC_002507
<i>Pupa strigosa</i>	opisthobranch gastropod	NC_002176
<i>Roboastra europaea</i>	sea slug	NC_004321
<i>Venerupis (Ruditapes) philippinarum</i> *	Japanese littleneck clam	NC_003354
Nematodes * = Missing ATP8		
<i>Ancylostoma duodenale</i> *	human hookworm	NC_003415
<i>Ascaris suum</i> *	pig roundworm	NC_001327
<i>Brugia malayi</i> *	agent of lymphatic filariasis	NC_004298
<i>Caenorhabditis elegans</i> *	roundworm	NC_001328
<i>Necator americanus</i> *	human hookworm	NC_003416
<i>Onchocerca volvulus</i> *	river blindness roundworm	NC_001861
<i>Trichinella spiralis</i>	trichinosis nematode	NC_002681

Table 2.3 continued

Clades Species	Common name	GenBank Accession Number
Neognathae		
<i>Arenaria interpres</i>	ruddy turnstone	NC_003712
<i>Aythya americana</i>	redhead	NC_000877
<i>Buteo buteo</i>	common buzzard	NC_003128
<i>Ciconia boyciana</i>	oriental stork	NC_002196
<i>Ciconia ciconia</i>	white stork	NC_002197
<i>Corvus frugilegus</i>	rook	NC_002069
<i>Coturnix japonica</i>	Japanese quail	NC_003408
<i>Falco peregrinus</i>	peregrine falcon	NC_000878
<i>Gallus gallus</i>	chicken	NC_001323
<i>Haematopus ater</i>	blackish oystercatcher	NC_003713
<i>Smithornis sharpei</i>	grey-headed broadbill	NC_000879
<i>Vidua chalybeata</i>	steelblue widowfinch	NC_000880
Palaeognathae		
<i>Anomalopteryx didiformis</i>	little bush moa	NC_002779
<i>Apteryx haastii</i>	great spotted kiwi	NC_002782
<i>Casuarius casuarius</i>	southern cassowary	NC_002778
<i>Dinornis giganteus</i>	giant moa	NC_002672
<i>Dromaius novaehollandiae</i>	emu	NC_002784
<i>Emeus crassus</i>	eastern moa	NC_002673
<i>Eudromia elegans</i>	elegant crested-tinamou	NC_002772
<i>Pterocnemia pennata</i>	Darwin's rhea	NC_002783
<i>Rhea americana</i>	greater rhea	NC_000846
<i>Struthio camelus</i>	ostrich	NC_002785
<i>Tinamus major</i>	great tinamou	NC_002781
Percomorpha		
<i>Antigonia capros</i>	deepbody boarfish	NC_003191
<i>Arctoscopus japonicus</i>	sailfin sandfish	NC_002812
<i>Cololabis saira</i>	Pacific saury	NC_003183
<i>Crenimugil crenilabis</i>	fringelip mullet	NC_003170
<i>Dactyloptena peterseni</i>	starry flying gurnard	NC_003194
<i>Elassoma evergladei</i>	Everglades pygmy sunfish	NC_003175
<i>Exocoetus volitans</i>	tropical two-wing flyingfish	NC_003184
<i>Gasterosteus aculeatus</i>	three spined stickleback	NC_003174
<i>Helicolenus hilgendorfi</i>	Rockfish	NC_003195
<i>Mastacembelus favus</i>	tire track eel	NC_003193
<i>Monopterus albus</i>	swamp eel	NC_003192
<i>Mugil cephalus</i>	flathead mullet	NC_003182
<i>Pagrus major</i>	red seabream	NC_003196
<i>Paralichthys olivaceus</i>	bastard halibut	NC_002386
<i>Platichthys bicoloratus</i>	stone flounder	NC_003176
<i>Rivulus marmoratus</i>	mangrove rivulus	NC_003290
<i>Stephanolepis cirrhifer</i>	thread-sail filefish	NC_003177
<i>Takifugu rubripes</i>	torafugu	NC_004299
<i>Trachurus japonicus</i>	Japanese jack mackerel	NC_002813

Table 2.3 continued

Clades Species	Common name	GenBank Accession Number
Platyhelminthes * = Missing ATP8		
<i>Echinococcus multilocularis</i> *	echinococcosis tapeworm	NC_000928
<i>Fasciola hepatica</i> *	liver fluke	NC_002546
<i>Hymenolepis diminuta</i> *	rat tapeworm	NC_002767
<i>Paragonimus westermani</i> *	lung fluke	NC_002354
<i>Schistosoma japonicum</i> *	oriental blood fluke	NC_002544
<i>Schistosoma mansoni</i> *	human blood fluke	NC_002545
<i>Schistosoma mekongi</i> *	blood fluke	NC_002529
<i>Taenia crassiceps</i> *	tapeworm	NC_002547
<i>Taenia solium</i> *	pork tapeworm	NC_004022
Primates		
<i>Cebus albifrons</i>	white-fronted capuchin	NC_002763
<i>Gorilla gorilla</i>	gorilla	NC_001645
<i>Homo sapiens</i>	human	NC_001807
<i>Hylobates lar</i>	common gibbon	NC_002082
<i>Lemur catta</i>	ring-tailed lemur	NC_004025
<i>Macaca sylvanus</i>	Barbary ape	NC_002764
<i>Nycticebus coucang</i>	slow loris	NC_002765
<i>Pan paniscus</i>	pygmy chimpanzee	NC_001644
<i>Pan troglodytes</i>	chimpanzee	NC_001643
<i>Papio hamadryas</i>	hamadryas baboon	NC_001992
<i>Pongo pygmaeus</i>	orangutan	NC_001646
<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	NC_002083
<i>Tarsius bancanus</i>	western tarsier	NC_002811
Rodents		
<i>Cavia porcellus</i>	domestic guinea pig	NC_000884
<i>Mus musculus</i>	house mouse	NC_005089
<i>Myoxus glis</i>	fat dormouse	NC_001892
<i>Rattus norvegicus</i>	Norway rat	NC_001665
<i>Sciurus vulgaris</i>	Eurasian red squirrel	NC_002369
<i>Thryonomys swinderianus</i>	greater cane rat	NC_002658
<i>Volemys kikuchii</i>	Taiwan vole	NC_003041

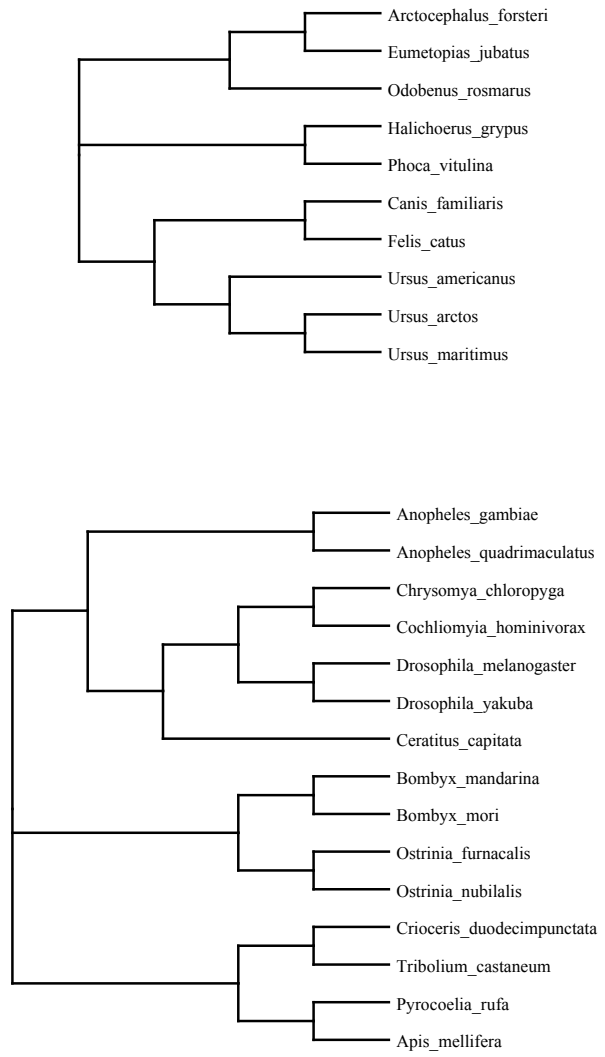


Figure 2.4: Carnivores and Insects phylogenetic trees.

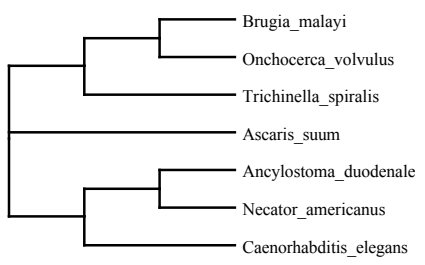
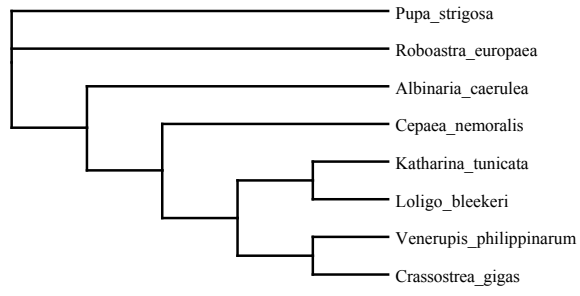


Figure 2.5: Mollusks and Nematodes phylogenetic trees.

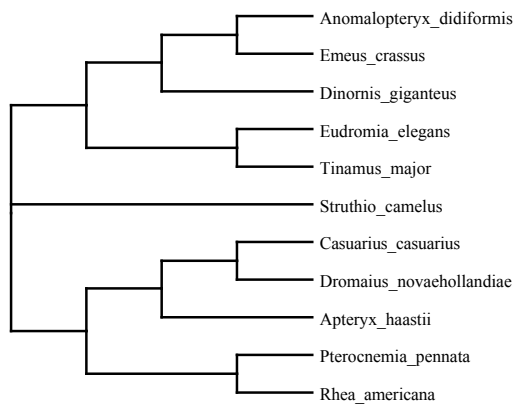
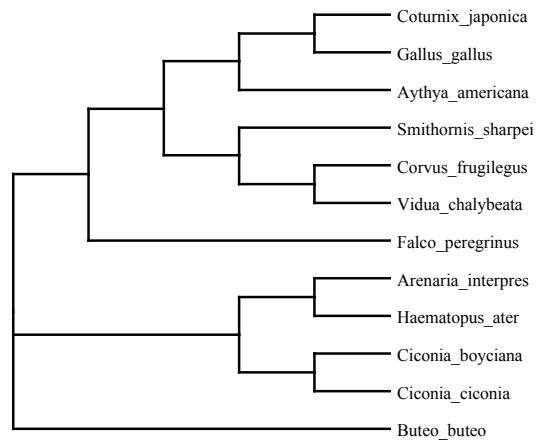


Figure 2.6: Neognathae and Palaeognathae phylogenetic trees.

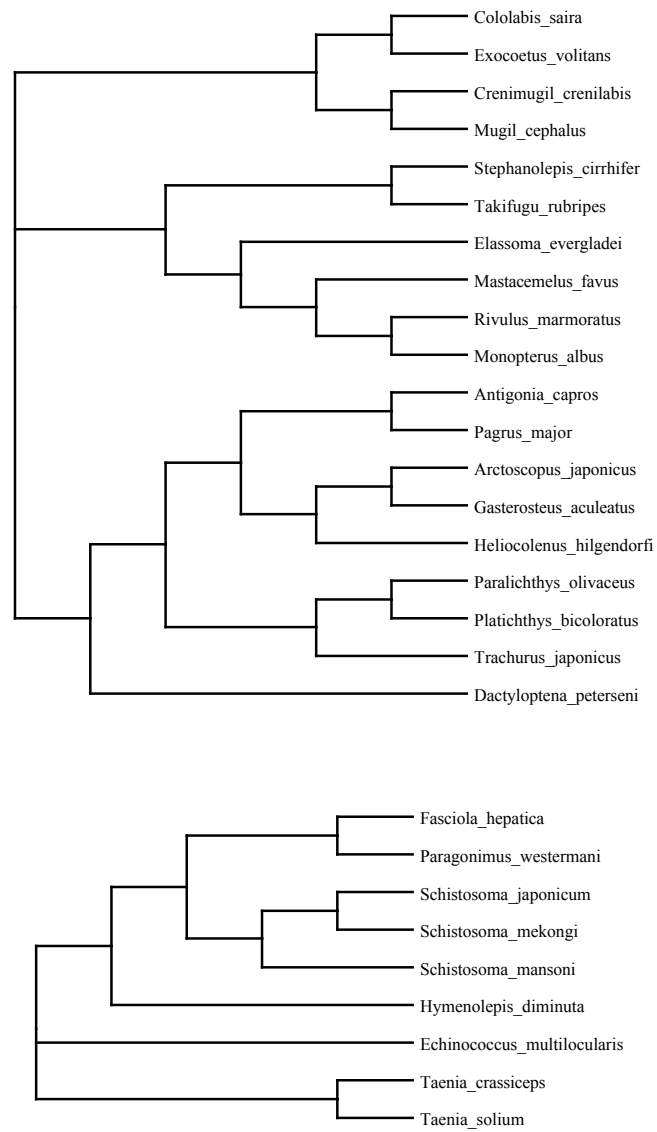


Figure 2.7: Percomorpha and Platyhelminthes phylogenetic trees.

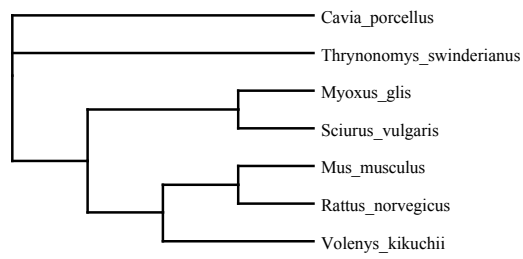
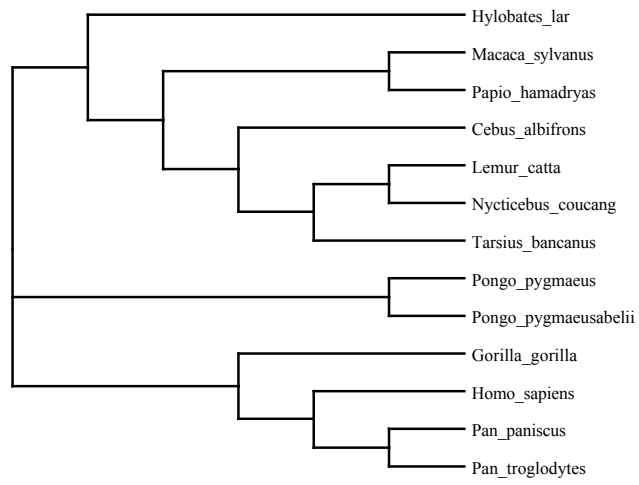


Figure 2.8: **Primates and Rodents phylogenetic trees.**

Further Results

Nucleotide Model

Mitochondrial genomes are known to contain strong differences in nucleotide rates (KUMAR, 1995) and for this reason we tested the codon model crossed with several nucleotide models. In every data set, the F81 model was rejected in favor of a more complex model. Because of the poor fit of the F81 model with mean discretization, we did not perform the analysis for this model using median values. Under the MG94 \times REV Dual Gamma Median parameterization with median rates, the HKY model was rejected 115/127 (90.6%) times in favor of the more complex GTR model. Similarly under all other model variations, the HKY model was rejected either 114 or 115 times, with 12 of the same genes favoring the HKY model regardless of discretization method or presence/absence of synonymous variation. This shows that nucleotide model selection is fairly independent of other model features

Another interesting behavior to check would be the effect on the estimation of synonymous variation when the incorrect model (HKY vs. GTR) is implemented. Under a median discretization, the coefficients of variation for α differed by more than 0.5 in three genes. In addition, there are 25 genes that differ by more than 0.1. However, the error does not seem to be biased in either direction. The estimation of the nonsynonymous rate coefficient of variation appears to have even less influence from the nucleotide model, with a correlation of over 0.99 of estimates between HKY and REV models. Only a single data set had $CV(\beta)$ estimates differing by more than 0.1 and no evidence of directional bias was found. The synonymous CV does differ slightly and thus implies the potential for the wrong

nucleotide model leading to poorer estimations of variation. However, using the wrong model does not appear favor either overestimating or underestimating the coefficients of variation, or conclusions about the significance of synonymous rate variation.

Mean vs. Median Discretization

When discretizing a continuous distribution, either the mean or median values over the intervals are used as the rate for that class. Both are valid methods with slightly different characteristics. The mean values over a range tend to be more extreme, especially in the highest class where the difference can be large if the tail of the distribution is long.

Because models with mean and median discretizations are not nested models, choosing the best one is accomplished using the AIC scores (AKAIKE, 1974). In total, the final model for a gene used the median discretization 84 times and mean discretization 43 times indicating that there might be a preference for median discretization likely due to the fact that a mean model has more extreme values and therefore more variance. For our inferences on synonymous rate variation, the effect of the discretization method was negligible.

Effect of nucleotide frequencies in model

We also looked at the effect of not accounting for positional nucleotide frequency bias (using a 1×4 rather than a 3×4 model). The frequency model used has no effect on the inferences of significance for 107 of 127 data sets, but for those 20 differing data sets, 5 failed to reject H_o in the 1×4 case, and 15 rejected H_o when it was not rejected in the 3×4 model. In other words, assuming equal

nucleotide frequencies for all 3 codon positions will lead to conclusions of more cases of synonymous variation than is truly present in the data. Digging deeper, the maximum likelihood values of μ_α tend to be underestimated, typically by small values around 0.1, but in four cases, the value was underestimated by greater than 90. This is causing an overestimate of the $CV(\alpha)$, and in turn giving too much confidence to the presence of synonymous rate variation. Conversely, values of μ_β and ω were generally overestimated, but never differed by more than 0.14. This presents evidence of the dangers of assuming nucleotide base equality between positions when drastically different values are present, as in these mtDNA data sets.

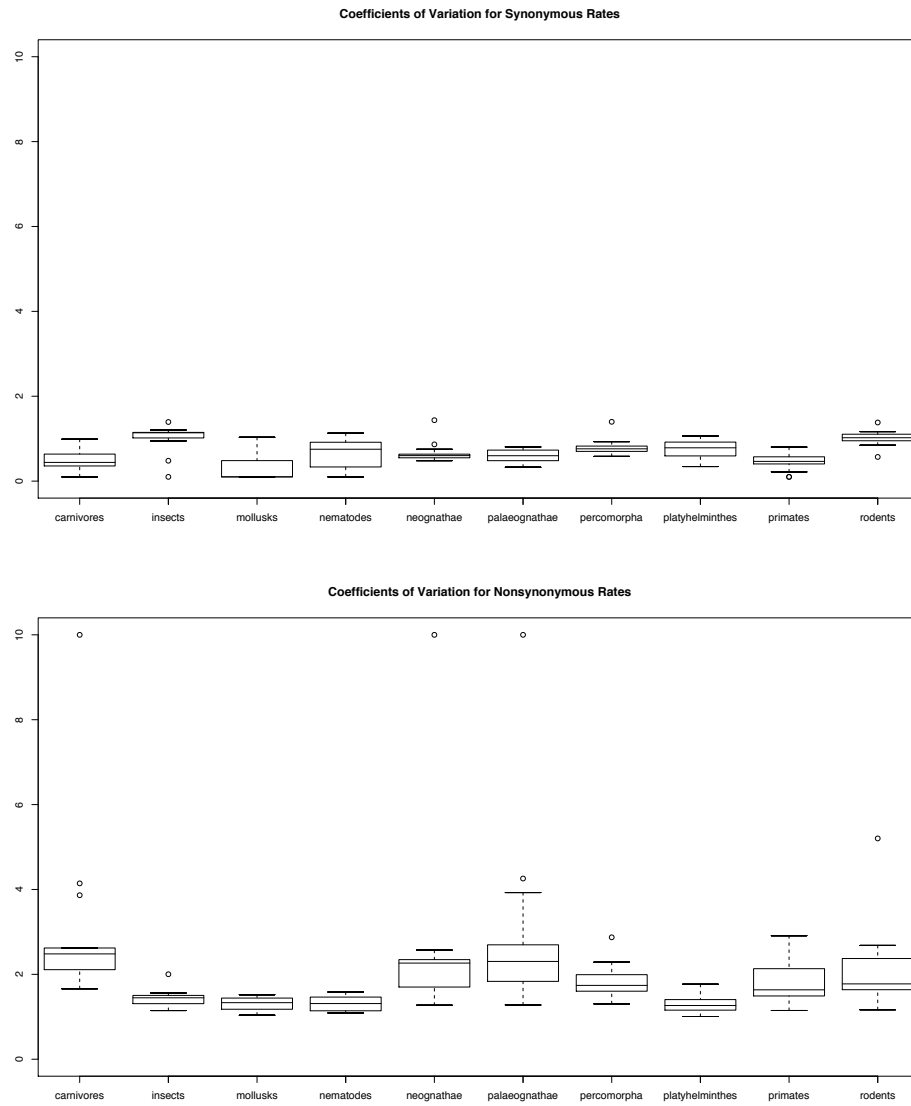


Figure 2.9: **Box plot of coefficients of variation for clades.** Box plot of the synonymous and nonsynonymous coefficients of variation, using the continuous estimator, for all genes within each of the 10 clades. Each box indicates the median and first and third quartiles of the distribution of CV estimates for each gene in the corresponding clade.

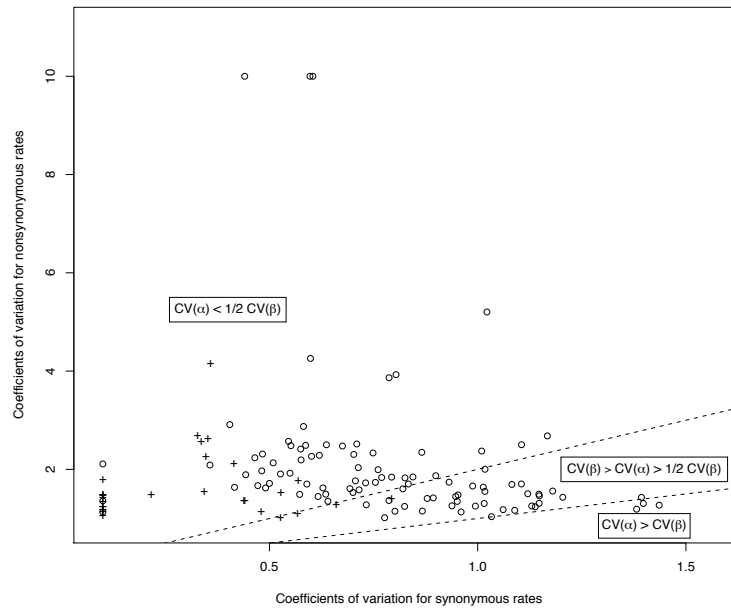


Figure 2.10: **Comparison of coefficients of variation of synonymous and nonsynonymous rates.** This scatter plot illustrates the relationship between the continuous CV estimates for synonymous and nonsynonymous rates for each data set, highlighting the comparable magnitude of the two types of variability. Each plotted circle represents a gene with significant synonymous rate variation, while each plus is a gene without significant variation. The dashed lines separate the graph into regions of synonymous rate coefficient of variation less than half that of nonsynonymous rate ($CV(\alpha) < \frac{1}{2}CV(\beta)$), less than the nonsynonymous rate but greater than one half that rate, and greater than the nonsynonymous rate. These regions contain 85, 39, and 3 data sets, respectively.

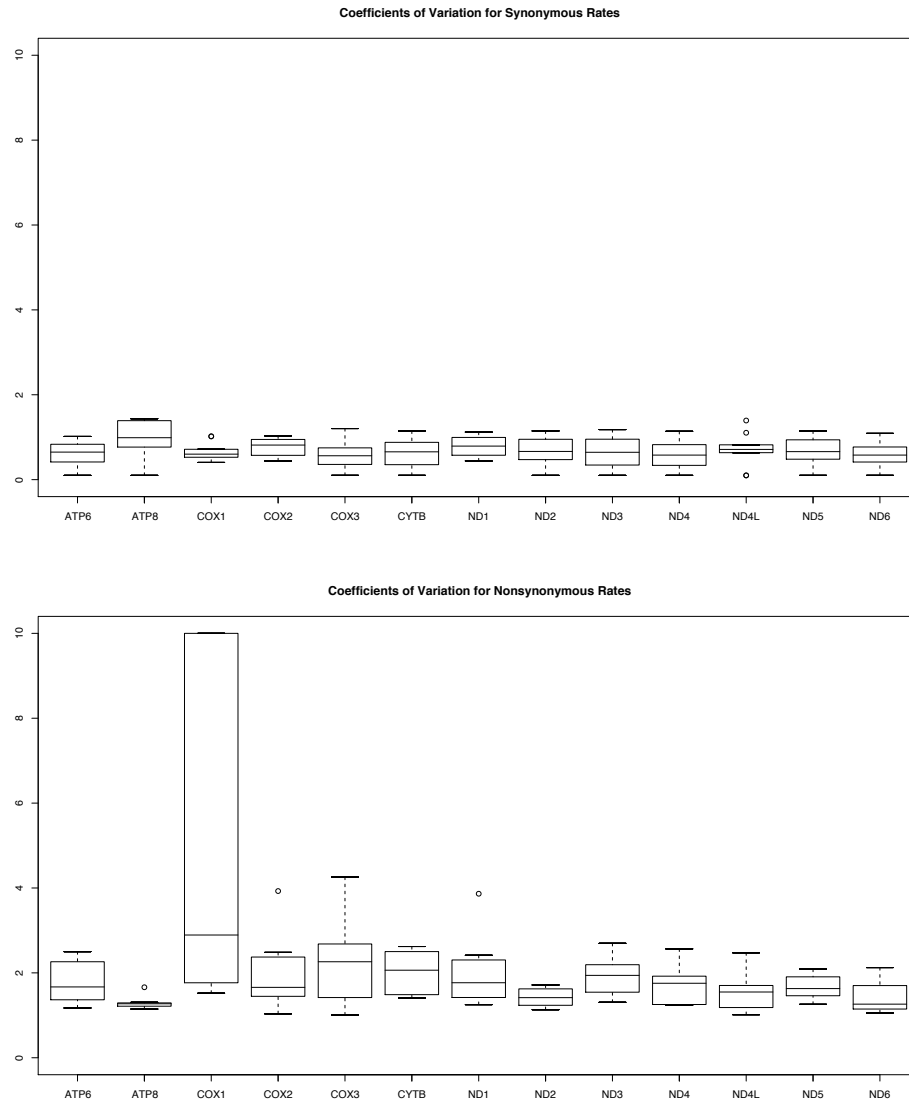


Figure 2.11: **Box plot of coefficients of variation for genes.** Box plot of the synonymous and nonsynonymous coefficients of variation, using the continuous estimator, for all clades within each of the 13 genes.

Chapter 3

A NEW MODEL OF SITE-TO-SITE RATE VARIATION WITHIN GENES TO ESTIMATE THE CORRELATION OF SYNONYMOUS AND NONSYNONYMOUS RATES

Mannino FV and SV Muse

Introduction

Evidence for site-to-site heterogeneity of both synonymous and nonsynonymous rates has recently been presented in a wide range of genes and species, but little is known about the relationship between the two types of rates. Previous studies estimated synonymous and nonsynonymous rates on a per gene basis and found support for a correlation of those rates (SMITH and HURST, 1999; WYCKOFF *et al.*, 2005; YOUNG and DEPAMPHILIS, 2005; MAKALOWSKI and BOGUSKI, 1998a,b; DUNN *et al.*, 2001; COMERON and KREITMAN, 1998; COMERON *et al.*, 1999; WOLFE and SHARP, 1993; OHTA and INA, 1995; MOUCHIROUD *et al.*, 1995). Other studies took a sliding window approach within a gene and found local regions of correlation (ALVAREZ-VALIN *et al.*, 1998, 2000), but these types of methods are often subject to influence of the chosen window size. The exact mechanisms leading to the correlation is under debate, but suggestions include correlated selective constraints, doublet/compensatory mutations, or mutational hot spots.

The statistical approaches are able to detect correlations between synonymous and nonsynonymous rates at a relatively crude level, related directly to the size of region used to estimate the rates. For example, if each gene is used to estimate one synonymous rate and one nonsynonymous rate, then only mechanisms creating a correlation between the average rates over the entire length of the gene will be detected. A sensitivity at the level of a gene is sufficient to detect, for instance, regional variations in mutation rates that impact both synonymous and nonsynonymous rates of nucleotide substitution. However, it would not be sufficient to detect correlations resulting from the increased rate of mutation tied to a specific mutagenic pentamer— the effect of such a mechanism would likely be lost in the averaging process.

Given that some mutation mechanisms act at the level of individual sites, while others act at the level of entire genomes, we propose a model of protein-coding sequence evolution that will allow us to investigate correlations of these synonymous and nonsynonymous rates within individual codons. Using this probabilistic model, we test for the presence of correlated rates in a collection of 73 data sets. Our results suggest that the rates are correlated in a significant proportion of genes, roughly half. This correlation tends to be positive, although we occasionally see negative correlation. The magnitude of these correlation estimates depends upon the method of estimation and also the number of discretized rate classes. The large number of significant data sets leads us to conclude that mechanisms affecting the synonymous and nonsynonymous rates are often acting in correlated manners.

Methods

Correlation of rates under independent model

Using the site specific rate estimates based on the empirical Bayes posterior probability, we can calculate the correlation between synonymous and nonsynonymous rates. We use the weighted rates over all classes for our estimates of the rates at each codon. When this correlation is computed based on the data sets analyzed in Chapter 2, we find that the average correlation is around 0.05, regardless of which specific synonymous rate variation model we use. However, while this approach could give us a feel for which data sets potentially have correlated rates, it lacks good statistical methodology to test for significance. In addition, because correlation is not incorporated in the models, we have no indication of what artifacts are contributing to this estimate.

Bivariate Model

Recent work (KOSAKOVSKY POND and MUSE, 2005) has expanded codon models to allow for synonymous rate variation by allowing both synonymous, α , and nonsynonymous, β , rates to vary according to a gamma distribution,

$$\alpha \sim \text{Gamma}(\mu_\alpha, \mu_\alpha),$$

$$\beta \sim \omega \times \text{Gamma}(\mu_\beta, \mu_\beta).$$

The density function for each rate can be written as

$$f(\alpha|\mu_\alpha) = \frac{1}{\Gamma(\mu_\alpha)} \mu_\alpha^{\mu_\alpha} \alpha^{\mu_\alpha-1} e^{-\alpha\mu_\alpha},$$
$$f(\beta|\mu_\beta, \omega) = \frac{1}{\Gamma(\mu_\beta)} \left(\frac{\mu_\beta}{\omega}\right)^{\mu_\beta} \beta^{\mu_\beta-1} e^{-\beta\frac{\mu_\beta}{\omega}}.$$

Because of the assumed independence between α and β , the joint distribution is simply the product of the marginal distributions,

$$f(\alpha, \beta|\mu_\alpha, \mu_\beta, \omega) = f(\alpha|\mu_\alpha)f(\beta|\mu_\beta, \omega).$$

The mean of α is one, and the mean of β is ω ; their variances are $\frac{1}{\mu_\alpha}$ and $\frac{\omega^2}{\mu_\beta}$, respectively. The general conclusions of these studies (see KOSAKOVSKY POND and MUSE (2005) and Chapter 2) has been that synonymous variation plays a significant role in the evolution of most genes, and studies ignoring this variation should be questioned.

Building off of the methodology of KOSAKOVSKY POND and MUSE (2005), we have extended this line of research by developing a bivariate distribution for α and β that allows for the presence of correlation between synonymous and nonsynonymous rates. We begin with the same basic approach to modeling rate variation as in the two-gamma independent model, but instead of having the mean of β equal

to ω , we set it equal to $\omega\alpha^d$. By relating the means of the two distributions, we create a bivariate model that can capture correlation between the rates. The model also reduces to the independent model (KOSAKOVSKY POND and MUSE, 2005) when $d = 0$, providing a straightforward way to formally test for the presence of correlation.

We begin by modeling synonymous variation in the same way as the independent model,

$$f(\alpha|\mu_\alpha) = \frac{1}{\Gamma(\mu_\alpha)} \mu_\alpha^{\mu_\alpha} \alpha^{\mu_\alpha-1} e^{-\alpha\mu_\alpha}.$$

The nonsynonymous rate under the new model is defined as a gamma random variable dependent on the synonymous rate,

$$\beta \sim \omega \times \alpha^d \times \text{Gamma}(\mu_\beta, \mu_\beta),$$

and the density function of β is now conditional on α ,

$$f(\beta|\alpha, \mu_\beta, \omega, d) = \frac{1}{\Gamma(\mu_\beta)} \left(\frac{\mu_\beta}{\omega\alpha^d}\right)^{\mu_\beta} \beta^{\mu_\beta-1} e^{-\beta\frac{\mu_\beta}{\omega\alpha^d}}.$$

In this parameterization, the amount of correlation between the rates depends upon $d \in (-\infty, +\infty)$. The correlation is negative when d is negative, and vice versa.

If we let η denote our vector of parameters, $(\mu_\alpha, \mu_\beta, \omega, d)$, then our bivariate density function becomes

$$\begin{aligned} f(\alpha, \beta|\eta) &= f(\beta|\alpha, \eta) f(\alpha|\eta) \\ &= \frac{1}{\Gamma(\mu_\alpha)} \mu_\alpha^{\mu_\alpha} \alpha^{\mu_\alpha-1} e^{-\alpha\mu_\alpha} \frac{1}{\Gamma(\mu_\beta)} \left(\frac{\mu_\beta}{\omega\alpha^d}\right)^{\mu_\beta} \beta^{\mu_\beta-1} e^{-\beta\frac{\mu_\beta}{\omega\alpha^d}} \end{aligned}$$

The marginal distribution of α is identical under the dependent and independent

models, and therefore the expected value and variance of α are the same:

$$\begin{aligned} E(\alpha) &= 1 \\ \text{Var}(\alpha) &= \frac{1}{\mu_\alpha}. \end{aligned}$$

We can also solve for expected value and variance of β ,

$$E(\beta) = \frac{\omega \Gamma(\mu_\alpha + d)}{\mu_\alpha^d \Gamma(\mu_\alpha)} \quad (\text{defined if } \mu_\alpha + d > 0),$$

$$\text{Var}(\beta) = \frac{\omega^2}{\mu_\alpha^{2d}} \left[\frac{(1 + \mu_\beta) \Gamma(2d + \mu_\alpha) \Gamma(\mu_\alpha) - \mu_\beta (\Gamma(\mu_\alpha + d))^2}{(\Gamma(\mu_\alpha))^2 \mu_\beta} \right] \quad (\text{defined if } \mu_\alpha + 2d > 0).$$

Our primary purpose in designing this particular parameterization was to enable examination of potential correlations between α and β . Our parameterization provides reasonably simple expressions for the covariance and correlation, ρ ,

$$\begin{aligned} \text{Cov}(\alpha, \beta) &= d \frac{\omega}{\mu_\alpha^{d+1}} \frac{\Gamma(\mu_\alpha + d)}{\Gamma(\mu_\alpha)} \quad (\text{defined if } \mu_\alpha + d > 1), \\ \rho(\alpha, \beta) &= \frac{d \sqrt{\frac{\mu_\beta}{\mu_\alpha}} \Gamma(\mu_\alpha + d)}{\sqrt{\Gamma(2d + \mu_\alpha) \Gamma(\mu_\alpha) (1 + \mu_\beta) - \mu_\beta (\Gamma(\mu_\alpha + d))^2}} \quad (\text{defined if } \mu_\alpha + 2d > 0). \end{aligned} \tag{3.1}$$

The above equations demonstrate the direct relationship between d and ρ . However, the shape parameters also affect the correlation, although in opposite manners. Larger values of μ_α lead to less variation in α and cause the synonymous rates to approach one, with the gamma distribution converging to a point mass around the mean. Because the correlation is based on α^d , as α values tend towards unity, this term will have less of an affect on β , reducing the correlation. Conversely, as μ_β increases, the gamma distribution portion of β will have values nearing one, reducing the influence of the gamma distribution on β , increasing the effect of α^d on β , and leading to more correlation. The ω parameter has no effect

on correlation because it is only a scaling parameter for β . Because of the parameter conditions necessary for ρ to be defined, we will consider only the parameter space where $\mu_\alpha + 2d > 0$.

The likelihood function is calculated in the same way as it was under the independent model,

$$L(\theta|D) = \prod_{s=1}^S \int_0^\infty \int_0^\infty P(D_s|\alpha, \beta, \theta) f(\alpha, \beta|\eta) d\alpha d\beta.$$

Bivariate Discretization

While the likelihood function takes the same form, its computational cost is still too high to evaluate exactly, and we must again resort to discretizing the bivariate distribution into a (small) number of categories. If we discretize the distribution into M synonymous classes and N nonsynonymous classes our likelihood function becomes

$$L(\theta|D) = \prod_{s=1}^S \sum_{m=1}^M \sum_{n=1}^N P(D_s|\theta_{mn}) f(\alpha_{mn}, \beta_{mn}|\eta),$$

where θ_{mn} represents α_{mn} , β_{mn} and all non-varying parameters. The calculation of this likelihood function is far more tractable than with a continuous distribution of variation, and it is practical for use with data sets of reasonable size. As an example, for a data set with 29 taxa and a sequence length of 192 codons, the calculation of the likelihood under a 4×4 discretization takes about 25 seconds on a Mac OS X 1 GHz G4.

When synonymous and nonsynonymous rates are modeled as independent variables, discretization of the continuous distribution reduces to the separate discretization of two univariate gamma distributions, a process that can be done quickly and with a high level of accuracy (YANG, 1994). However, the correlated bivariate case is more involved. Dividing the nonsynonymous distribution into

intervals and calculating representative rates for each of those classes requires integrating over the synonymous rates for each category. We therefore develop an iterative algorithm to perform this discretization (see Appendix for details). This algorithm divides up the bivariate distribution into roughly equal rate classes and calculates the mean values within each rate class using numerical integration.

Maximum likelihood Estimation

Estimation of the nucleotide and codon frequencies are done using the observed counts in the data set. Codon frequencies consider positional differences within a codon. All other parameters are estimated using maximum likelihood. Because there are no closed form expressions for such estimates, we use a modified Newton-Raphson method for numerical optimization. The first and second derivatives of the likelihood function are approximated using Richardson's extrapolation (as implemented in PRESS *et al.*, 1992). The algorithm iterates through each parameter separately initially under a low level of convergence. When no parameters change by more than a fixed amount, the convergence criterion is strengthened and the process is repeated.

For branch lengths the range of possible values includes 0, which is a boundary value. Any time a parameter approaches 0, the branch length is converted to a log scale for the optimization process to prevent problems of approaching a boundary and requiring a lower bound. All other parameters have extreme lower and upper bounds set simply to avoid potential, although very unlikely to ever be explored, parameter values that cause problems with the likelihood function.

Each data set is initially run under a basic codon model with no rate variation and with the appropriate nucleotide model to find quick but good initial starting

values for the branch lengths and nucleotide parameters. In addition, each data set is run under the Dual Gamma model of KOSAKOVSKY POND and MUSE (2005), which considers the synonymous and nonsynonymous rates as independent. Fitting this simpler model allows hypothesis testing of the significance of correlation between rates.

Testing the Correlation Hypothesis

Because the Dual Gamma model is a special case of the bivariate model, we can compare these models and test for the significance of correlation between the synonymous and nonsynonymous rates. Our null hypothesis, H_O , is that the rates are not correlated, while our alternative hypothesis, H_A , is that the rates are correlated. To test these hypotheses, we use the likelihood ratio test (LRT), where $\Lambda = -2(L_O - L_A)$ is asymptotically distributed as a χ_v^2 . The values L_O and L_A are the maximum likelihood values under each hypothesis and v is the number of degrees of freedom between the models. Testing for the significance of correlation relies on a single parameter, d , and we therefore compare the likelihood ratio test statistic with a χ_1^2 distribution.

Estimators of ρ

We consider two possible estimators for the amount of correlation, ρ , the discretized distribution correlation and the correlation from the continuous distribution based on the maximum likelihood estimates $\hat{\eta}$. The continuous correlation estimate appears to almost always be lower in absolute magnitude than the discretized estimate. We present both estimates here, although preliminary analysis combined with results from Chapter 4 lead us to prefer the use of the continuous estimate.

Materials

The data sets analyzed under the bivariate model were all previously used in other studies investigating synonymous rate variation. 65 data sets from Chapter 2 were reanalyzed here as well as 8 data sets used by KOSAKOVSKY POND and MUSE (2005) representing a wider range of genes and taxa. Only the REV model of nucleotide substitution was considered and all bivariate results are compared to the mean rather than the median results discussed in chapter 2, because that keeps consistency with the new methodology.

Results and Discussion

Correlation of synonymous and nonsynonymous rates

Using the LRT, we can test for the significance of the correlation of rates for each data set. For the data sets from KOSAKOVSKY POND and MUSE (2005), five out of eight had significant correlation (four positive and one negative). For those used in Chapter 2, 28 out of 65 data sets had significant correlation (23 positive and 5 negative). Finding correlation in roughly half of our data sets, leads us to believe that this phenomenon is reasonably common in evolution. We would expect more positive correlations because factors like selection on doublet/compensatory mutations or mutational hot spots would affect synonymous and nonsynonymous rates in the same manner. However, six data sets had significant negative correlation, leading us to question which mechanisms are causing this effect. While our methods cannot determine exactly why these rates are correlated, the significant presence of this correlation warrants future research.

To estimate the amount of correlation, we can use either the correlation of rates

in the discretized distribution or the continuous correlation (see Equation (3.1)) based on the maximum likelihood estimates of η . We can see these estimates plotted against each other for all data sets in Figure 3.1, distinguishing between those data sets with and without significant correlation. We see the immediate trend that the discretized estimates are always more extreme (farther away from 0) than the continuous estimates. Without performing a simulation study we cannot know which of these estimators are better. However, results from Chapter 4 support the use of continuous CV estimators over discretized, possibly indicating a general superiority of continuous estimators. Therefore, we choose to use the more conservative continuous ρ estimates in this work.

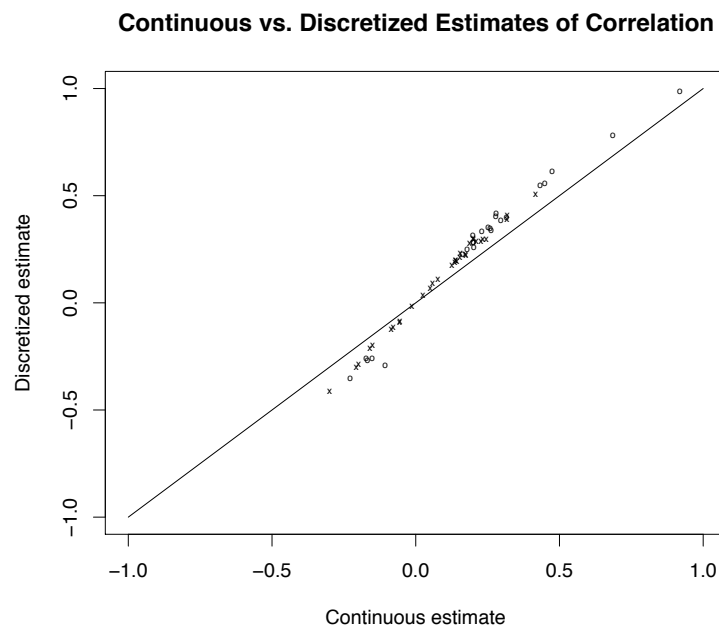


Figure 3.1: **Correlation of synonymous and nonsynonymous rates.** Here we present the continuous and discretized estimates of correlation for each data set, distinguishing between significant (o) and non-significant (x) correlation. The solid line shows that for positive correlations, the discretized estimate is greater, while for negative correlations the discretized estimate is smaller.

The number of classes appears to have a direct effect on the estimates of correlation. Increasing the number of rate classes will lower the absolute value of the estimate. For two data sets (primates NADH1 and primates NADH5), we fit our model with differing number of rate classes (3, 4, 5, 6 and 8) and the estimates of correlation can be seen in Figure 3.2 for both ρ estimators. Most likely, this trend indicates that estimates are biased under fewer rate classes. However, for these two data sets, the hypothesis test of correlation did not change with four or more rate classes. A possible cause of this bias is simply a poor fit of the discretized approximation to the true continuous distribution with a small number of classes. Because the discretization itself is an approximation and not an exact discretization formally proving such a poor fit would be difficult.

Branch Lengths

We compared the branch length estimates between the Dual Gamma model and bivariate model. We found a high correlation of estimates between the two models, all over 0.996. While there is no evidence of large errors in branch length estimation, we also wish to check for directional bias found when ignoring potential correlation. Of the 30 data sets with significant rate correlation, 14 tended to underestimate branch lengths, while 16 tended to overestimate. We conclude that ignoring correlation between the rates appears to have little influence on the branch lengths.

Site-specific rate estimates

Using the weighted rates estimates for each site, we can determine the effect of ignoring rate correlation. First off we can examine the correlations between rates

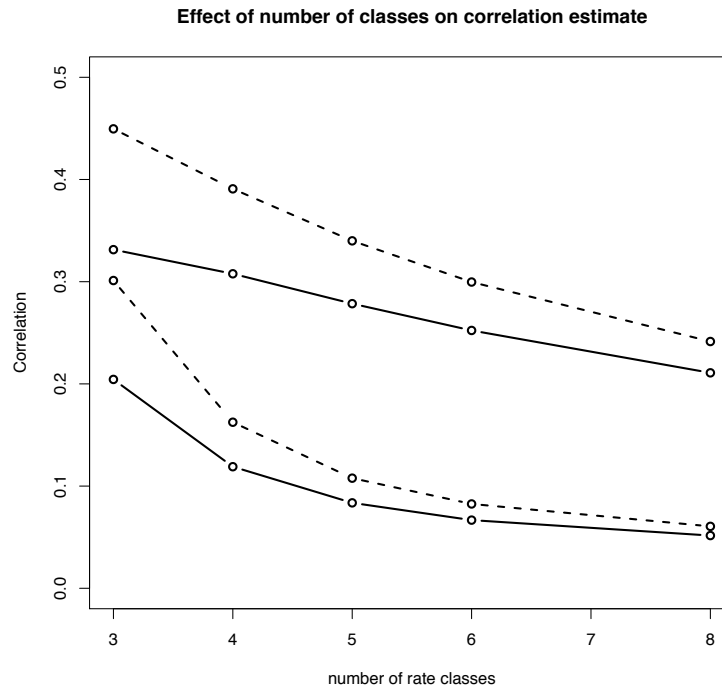


Figure 3.2: **Correlation differences based on number of classes.** Here we present the effect of the estimates of correlation by the number of rate classes. The dashed and solid lines represent the discrete and continuous estimates, respectively. The top two lines are for Primates NADH1 and the bottom two are for Primates NADH5. We see that increasing the number of classes leads to lower estimates of correlation and also lessens the difference between the two estimators.

under the independent and dependent models. The nonsynonymous rates do not differ greatly, while the synonymous rates were sometimes off by more. Looking at directional bias we see that the Dual Gamma model tends to overestimate both synonymous and nonsynonymous rates, with 50 and 45 data sets overestimating, respectively. Because our interests in modeling rate variation are for these precise types of estimates, potential differences in estimation pose questions about the practical effects of ignoring rate correlation.

Implications

Detecting correlation between synonymous and nonsynonymous rates helps provide clues about the mechanisms of gene evolution. While significant correlation has been detected between genes (e.g., SMITH and HURST, 1999; WOLFE and SHARP, 1993), no rigorous study has provided the ability to detect correlation within genes. Here we have described methods to test for this correlation, although we cannot distinguish the causes. Finding many genes with correlated synonymous and nonsynonymous rates leads us to question possible causes, including mutational hot spots, codon bias, doublet/compensatory mutations, or other selective pressures such as the conservation of binding or splice sites. The ability to measure this correlation will give us a clearer understanding of the processes of gene evolution and possibly lead to better inferences.

References

- ALVAREZ-VALIN, F., K. JABBARI and G. BERNARDI, 1998 Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. *J. Mol. Evol.* **46**: 37–44.
- ALVAREZ-VALIN, F., K. JABBARI and G. BERNARDI, 2000 Nonrandom spatial distribution of synonymous substitutions in the GP63 gene from leishmania. *Genetics* **155**: 1683–1692.
- BEST, D. J., and D. E. ROBERTS, 1975 The percentage points of the χ^2 distribution. *Appl. Statist.* **24**: 385–388.
- COMERON, J. M., and M. KREITMAN, 1998 The correlation between synonymous and nonsynonymous substitutions in drosophila: mutation, selection or relaxed constraints. *Genetics* **150**: 767–775.
- COMERON, J. M., M. KREITMAN and M. AGUADÉ, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in drosophila. *Genetics* **151**: 239–249.
- DUNN, K. A., J. P. BIELAWSKI and Z. YANG, 2001 Substitution rates in drosophila nuclear genes: Implications for translational selection. *Genetics* **157**: 295–305.

- KOSAKOVSKY POND, S. L., and S. V. MUSE, 2005 Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* **22**: 2375–2385.
- MAKALOWSKI, W., and M. S. BOGUSKI, 1998a Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. USA* **95**: 9407–9412.
- MAKALOWSKI, W., and M. S. BOGUSKI, 1998b Synonymous and nonsynonymous distances are correlated in mouse and rat genes. *J. Mol. Evol.* **47**: 119–121.
- MOUCHIROUD, D., C. GAUTIER and G. BERNARDI, 1995 Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. *J. Mol. Evol.* **40**: 107–113.
- OHTA, T., and Y. INA, 1995 Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. *J. Mol. Evol.* **41**: 717–720.
- PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY and W. T. VETTERLING, 1992 Richardson extrapolation and the Bullrsch-Stoer method, pp. 718–725 in *Numerical recipes in FORTRAN: The art of scientific computing*. Cambridge University Press, Cambridge, England.
- SMITH, N. G. C., and L. D. HURST, 1999 The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153**: 1395–1402.
- URROZ, G. E., 2005 Double integrals using Simpson's 1/9 rule Website http://www.engineering.usu.edu/cee/faculty/gurro/Software_Calculators/S%cilab_Docs/SCILAB_Notes&Functions.htm.

- WOLFE, K. H., and P. SHARP, 1993 Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* **37**: 441–456.
- WYCKOFF, G. J., C. M. MALCOM, E. J. VALLENDER and B. T. LAHN, 2005 A highly unexpected strong correlation between fixation probability or nonsynonymous mutations and mutation rate. *Trends in Genetics* **21**: 381–385.
- YANG, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**: 105–111.
- YOUNG, N. D., and C. W. DEPAMPHILIS, 2005 Rate variation in parasitic plants: correlated and uncorrelated patterns among plastid genes of different function. *BMC Evolutionary Biology* **5**.

Appendix

Bivariate discretization

We first divide our distribution into synonymous intervals of equal probability. Because the marginal distribution of α is a univariate gamma distribution, the boundaries can be found by using the cumulative distribution function. We label the left synonymous boundaries L_1, \dots, L_M , and right boundaries R_1, \dots, R_M , such that $L_1 = 0, R_M = \infty$, and $R_m = L_{m+1}$ for $m = 1, \dots, M - 1$. We can then find the boundary points such that

$$\frac{1}{M} = \int_0^\infty \int_{L_m}^{R_m} f(\alpha, \beta | \eta) d\alpha d\beta = \int_{L_m}^{R_m} f(\alpha | \eta) d\alpha$$

These boundary values are computed using methods described by YANG (1994). We first note the relationship of the gamma and chi-squared distributions, which

states that a $\text{Gamma}(\mu_\alpha, \frac{1}{2})$ distribution is equivalent to a $\chi^2_{2\mu_\alpha}$ distribution. Therefore we can calculate the gamma distribution points as

$$F_G(p|\mu_\alpha, \lambda) = \frac{F_{\chi^2}(p|2\mu_\alpha)}{2\lambda}. \quad (3.2)$$

The values of F_{χ^2} can be calculated using several algorithms (e.g., BEST and ROBERTS, 1975). These synonymous intervals are delimited by the solid vertical lines in Figure 3.3.

We then label the nonsynonymous boundaries for synonymous interval m as $B_{m,1}, \dots, B_{m,N}$ for the bottom, and $T_{m,1}, \dots, T_{m,N}$ for the top, such that $B_{m,1} = 0, T_{m,N} = \infty$, and $T_{m,n} = B_{m,n+1}$ for $n = 1, \dots, N - 1$. Because of the dependence the nonsynonymous distribution will be different for each synonymous interval, and therefore the boundaries will be different. Ideally we would want to then solve for the boundaries according to

$$\frac{1}{MN} = \int_{B_{m,n}}^{T_{m,n}} \int_{L_m}^{R_m} f(\alpha, \beta|\eta) d\alpha d\beta,$$

but this integral cannot be evaluated empirically, so for each interval we want to approximate the nonsynonymous distribution. We find the expected synonymous rate (α_m^*) over interval m in the same manner as with a univariate gamma discretization by calculating $\alpha_m^* = M \int_0^\infty \int_{L_m}^{R_m} \alpha f(\alpha, \beta|\eta) d\alpha d\beta = M \int_{L_m}^{R_m} \alpha f(\alpha|\eta) d\alpha$. The conditional density $f(\beta|\alpha_m^*, \eta)$ is then a univariate gamma distribution and its exact intervals can be determined using the relationship in Equation 3.2. These approximate nonsynonymous boundaries are the dashed lines in Figure 3.3. Each box in the figure now corresponds to one rate class. The region can be divided into three types of boxes. Internal boxes have four defined boundaries. Edge boxes have defined boundaries in one dimension, but a boundary of 0 or ∞ in the other. Corner boxes contain a 0 or ∞ boundary in both dimensions.

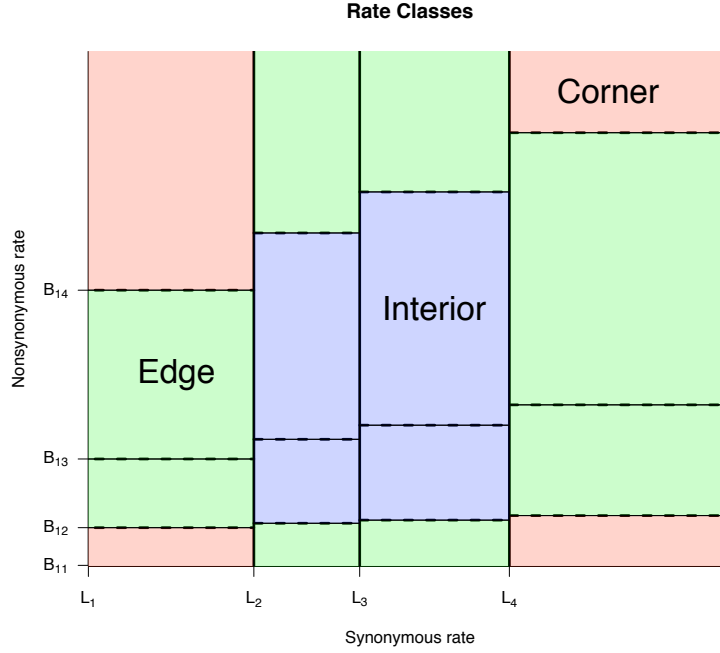


Figure 3.3: **Box differences in bivariate distribution.** This plot shows the division of bivariate distribution into boxes based on the synonymous intervals (solid lines) and nonsynonymous intervals (dashed lines). There are three types of boxes in the discretization, which are treated differently because box boundaries of 0 or ∞ cannot be used in the numerical integration. Interior boxes have four defined boundaries. Edge boxes have three defined boundaries and either 0 or ∞ for the fourth. Corner boxes have two defined boundaries and two boundaries of 0 or ∞ .

Because a temporary synonymous value was used instead of integrating over all synonymous values, the probabilities for all boxes are not equal to $\frac{1}{MN}$ as they would be if α and β were independent. Before obtaining the rates within each box we must calculate the probability of that box,

$$p_{ij} = \int_{B_{i,j}}^{T_{i,j}} \int_{L_i}^{R_i} f(\alpha, \beta | \eta) d\alpha d\beta.$$

After calculating each of these probabilities, the rates for each box can be found

as

$$\alpha_{ij} = \frac{1}{p_{ij}} \int_{B_{i,j}} \int_{L_i}^{R_i} \alpha f(\alpha, \beta | \eta) d\alpha d\beta$$
$$\beta_{ij} = \frac{1}{p_{ij}} \int_{B_{i,j}} \int_{L_i}^{R_i} \beta f(\alpha, \beta | \eta) d\alpha d\beta.$$

As previously mentioned, these integrals cannot be calculated analytically, and we must employ numerical integration techniques. These methods approximate the integral by breaking up the function into K pieces where the area (or volume in a two dimensional case) can be calculated exactly for a given geometrical shape. As we increase the number of these intervals, the approximation approaches the true value. The simplest of these methods would be right or left Riemann sums, which create rectangles (or cuboids) for which we can easily calculate the area. A modification of right and left sums is the trapezoidal rule, which forms trapezoids rather than rectangles and is in fact the same as computing the right and left sums and taking the average of those. The midpoint rule also creates rectangles but with the height equal to the middle value of the interval. While these methods are quick and easy to implement, many intervals are needed to get a decent approximation for complex functions that cannot easily be approximated by rectangles or trapezoids.

Simpson's rule approximates the integral by fitting quadratic polynomials over $2K$ intervals. Every two consecutive intervals consist of three points that are used to create the polynomial. For $\int_a^b f(x)$ and $2K$ intervals, we have $x_0 = a, x_1, \dots, x_{2n-1}, x_{2n} = b$ endpoints. The area under the quadratic formed by the first two intervals, marked by the points x_0, x_1 , and x_2 , is

$$\frac{b-a}{6K} [f(x_0) + 4f(x_1) + f(x_2)],$$

and the entire integral is

$$\begin{aligned}\int_a^b f(x) &\approx \frac{b-a}{6K} \sum_{k=0}^{K-1} [f(x_{2k}) + 4f(x_{2k+1}) + f(x_{2k+2})] \\ &= \frac{b-a}{6K} [f(x_0) + 2 \sum_{k=1}^{K-1} f(x_{2k}) + 4 \sum_{k=1}^{K-1} f(x_{2k+1}) + f(x_{2K})].\end{aligned}$$

The final value for the integral is the same as $\frac{1}{3}$ of the trapezoidal integration value plus $\frac{2}{3}$ of the midpoint integration value when both are calculated with K intervals. Despite the fact that the method is using quadratic functions, the approximation is exact for up to cubic polynomial functions. Although Simpson's method is more time consuming, the resulting error is much less than other methods and the computing time required is still better than other more complex methods. To obtain the same level of accuracy using a simpler method would often require a large increase in the number of intervals.

Based on the algorithm of URROZ (2005), we can numerically integrate a bivariate function using Simpson's method with $2K \times 2K$ intervals as

$$\int_a^b \int_c^d f(x, y) \approx \left(\frac{b-a}{3K}\right) \left(\frac{d-c}{3K}\right) \sum_{k=0}^{K-1} \sum_{l=0}^{K-1} S_{kl},$$

where

$$\begin{aligned}S_{kl} = &f(x_{2k,2l}) + 4f(x_{2k,2l+1}) + f(x_{2k,2l+2}) + 4f(x_{2k+1,2l}) + 16f(x_{2k+1,2l+1}) + \\ &4f(x_{2k+1,2l+2}) + f(x_{2k+2,2l}) + 4f(x_{2k+2,2l+1}) + f(x_{2k+2,2l+2}).\end{aligned}$$

It is important to note that the discretization of a bivariate distribution is not symmetrical. Our method discretized the synonymous dimension into intervals, then the nonsynonymous dimension. Different rates would be obtained if the process was done in reverse. Our main reason for choosing to discretize synonymous distribution first is that the marginal distribution for this rate is known and is a

standard gamma. Performing the discretization in reverse would be more difficult and would require approximations for all box boundaries rather than only the nonsynonymous boundaries.

Regions of the distribution

The best way to perform the integration is to break up each rate class box into smaller pieces that can be integrated very quickly and accurately. Any integral over a range can be broken up into the sum of integrals of the range broken into smaller pieces. Ideally, we would not want to integrate parts of the distribution that have a negligible effect on the final integral value. However, in practice we do not know which parts are negligible and we use our algorithm to traverse the space finding all non-negligible regions without spending too much time integrating the irrelevant regions. We must first determine the level of accuracy desired in the final box probabilities and rates, which we call the convergence factor (CF). We chose to use a CF value of 10^{-6} , because more accurate estimates did not greatly affect the calculated rates or the log-likelihood values. The goals of the algorithm are to find all regions of a box that contribute a weight that is within six decimal places of accuracy of the true value and also to estimate each numerical integration to within this level of accuracy.

Determining six decimal places of accuracy can be problematic because the true values of the integral, which are unknown, range over many orders of magnitude. To help decide when we have reached convergence, all functions are scaled to an approximate value based on the Dual Gamma model. This scaling allows regions of integration with drastically different resulting values to use the same CF value. If little or no correlation exists, the scaling factor is very accurate, and the weighted integral should be close to 1.0. But when strong correlation exists,

these estimates can be off by large amounts and therefore as we get a better idea of the final value, the scaling factor is continuously updated. When the integral has converged, the value is returned to the original scale. For example, consider a box with a nonsynonymous rate of 0.0016. We can quickly calculate what the rate would be for the independent model given the set of parameters η . If we suppose this rate estimate is 0.0018, then the function we will numerically integrate becomes $g(\alpha, \beta|\eta) = \frac{f(\alpha, \beta|\eta)}{0.0018}$.

Determining how to break up each box into pieces to integrate depends upon the type of box. As expected, internal boxes are the simplest to integrate because none of their values are on the boundary of the parameter space (0 or ∞). Splitting the region into pieces is beneficial because Simpson's integration works well on flatter functions and smaller ranges. If the boundaries of the box differ by a factor of more than 10, the region is split as many times as necessary so that no region is ever evaluated by the Simpson's algorithm that violates the factor of 10 rule. While the number of regions integrated increases, the number of intervals necessary for the algorithm to converge can be greatly reduced, and can actually be faster overall when the density of the function is steep. This rule applies to boundaries in both the synonymous and nonsynonymous dimensions. See Figure 3.4.

This same factor of 10 rule is applied to the edge boxes in the dimension that has two defined boundaries. In the other dimension we have issues that arise from the boundary of the box being at one of the endpoints of the range of the parameter. For the lower bound at 0, the function cannot be evaluated when the shape parameter is less than 1 and therefore cannot be a boundary for the Simpson's integration, because the function of the endpoint is one of the necessary values in the numerical integration. For the upper bound at ∞ we cannot use this value in the Simpson's function because the length of the intervals would be

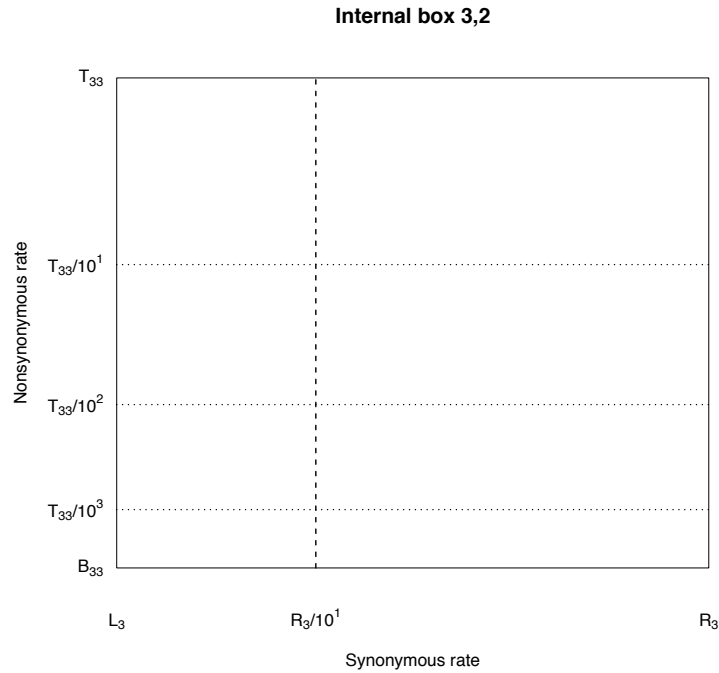


Figure 3.4: **Algorithm to integrate interior box.** This plot demonstrates the application of the factor of 10 rule to divide an interior box for faster numerical integration. All subboxes are numerically integrated separately.

infinite. Choosing an arbitrary bound to cut off the integration can be dangerous and lead to large underestimates of the integral. Setting an extremely large bound leads to wasted computational effort integrating regions of low density. To combat this problem, we are using the fact that an integral over $(0, a)$ can be divided into the infinite sum of integrals

$$\int_0^a f(x) = \sum_{i=0}^{\infty} \int_{\frac{a}{LF^{i+1}}}^{\frac{a}{LF^i}} f(x),$$

where LF is the lower bound factor. We use the value of 10 for LF, which was found to perform well in tests of accuracy and speed of integration. This creates subboxes within each edge box that do have defined boundaries. It is important to note that smaller LF values will lead to more accurate integral approximations,

but requires performing more Simpson function evaluations thereby increasing the time. Larger LF values will be faster only if the range of integration can be handled well by Simpson's method. Similarly, for the upper boundary we have an upper factor (UF) and can split the integral as

$$\int_a^\infty f(x) = \sum_{i=0}^{\infty} \int_{aUF^i}^{aUF^{i+1}} f(x),$$

with the best UF values found to be around 2. In the nonsynonymous direction, we can calculate the mode of the distribution given a fixed synonymous rate. Therefore, when the mode is in the current edge box, the infinite sum of integrals starts at the subbox with the mode and proceeds in both directions until the weights of the subboxes are below the CF. In the synonymous direction we cannot calculate the mode easily because no closed form expression exists for $f(\alpha|\beta)$. For this reason we simply start at the known boundary and use the same stopping criteria. In addition, we check to ensure that the function and the weights in the subboxes are decreasing and that therefore the mode has been reached. If the nonsynonymous dimension is divided according to the factor of 10 rule, the first sum over subboxes starts with the fixed boundary and the rest start at the subbox with the highest weight from the previous nonsynonymous interval and work in both directions. Occasionally the weight of the integral at the start of the infinite sum will be so low as to be treated as 0 by the computer. This does not necessarily cause the stopping criteria to be triggered if no weight has been found in the current summation. See Figure 3.5.

The corner box is broken up in a manner similar to the edge boxes, but in this case both parameters are approaching a boundary so we must iterate over both

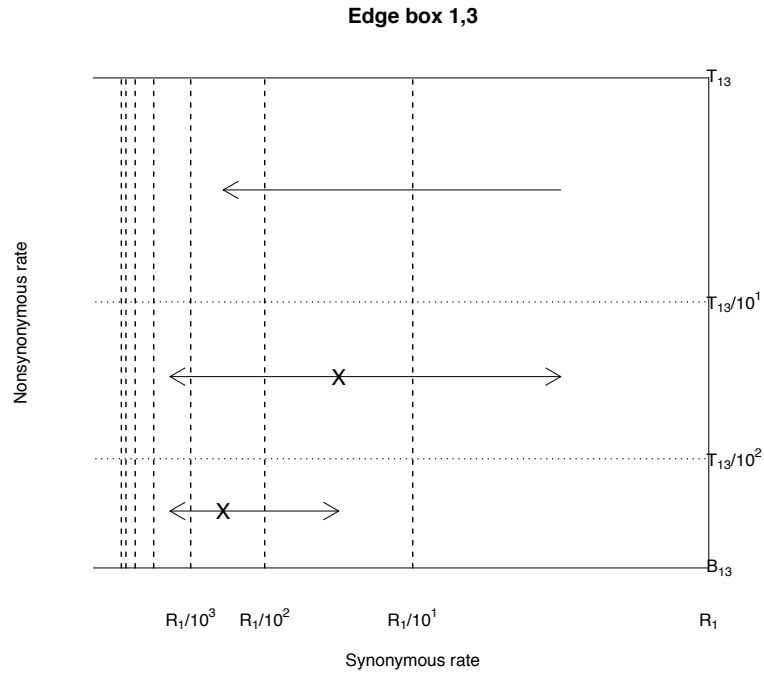


Figure 3.5: **Algorithm to integrate edge box.** This plot demonstrates the division of an edge box to determine which subboxes to numerically integrate. For this example, division of the box along the nonsynonymous axis follows the factor of 10 rule. Along the synonymous axis we start at the first subbox and proceed left until reaching our stopping criteria. For the remaining nonsynonymous intervals we start at the mode from the previous interval (denoted by X) and proceed in both directions until reaching the stopping criteria.

dimensions. For example, the bottom left box integral would be

$$\int_0^a \int_0^b f(x, y) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \int_{\frac{a}{LF^{i+1}}}^{\frac{a}{LF^i}} \int_{\frac{b}{LF^{j+1}}}^{\frac{b}{LF^j}} f(x, y).$$

Again using the fact that the nonsynonymous mode can be calculated, we fix the synonymous summation (i in above equation), then start at the mode and work in both directions exactly as done in the edge box, using the same stopping criteria. The algorithm then moves on to the next synonymous interval and repeats.

To stop in the synonymous direction some weight must have been reached and that weight must be decreasing and smaller than the CF. In addition, for an upper

boundary box, the function values themselves must be decreasing to ensure that the mode has been reached. See Figure 3.6.

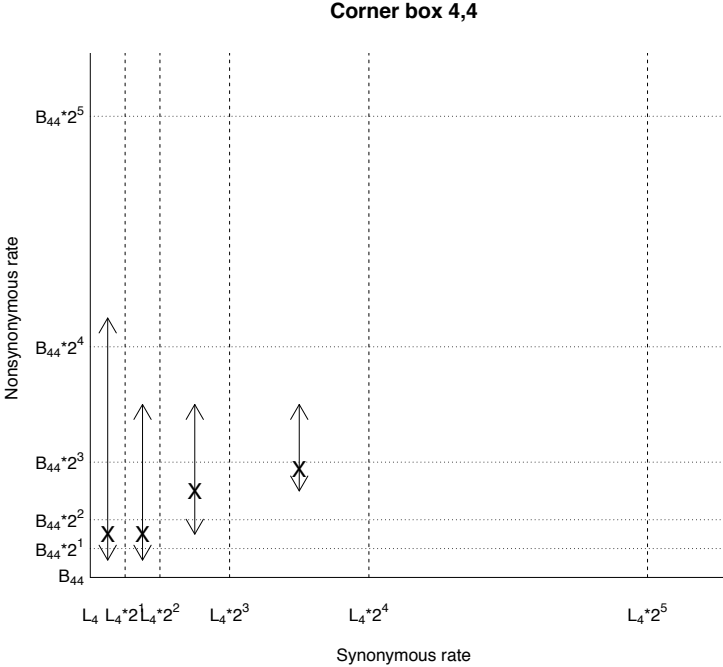


Figure 3.6: **Algorithm to integrate corner box.** This plot demonstrates the method of dividing up a corner box. For each synonymous interval a mode is calculated in the nonsynonymous dimension based on the median synonymous rate within the interval (denoted by X). Integration proceeds in both directions until reaching the stopping criteria. The algorithm then moves to the next synonymous interval, until reaching the synonymous dimension stopping criteria.

Once we have broken up a box into pieces, we must decide how many intervals to use in the Simpson’s integration. More intervals leads to a better integral approximation, but requires more time. For relatively flat regions of the function and smaller areas of integration, ten intervals are sufficient to achieve a certain level of accuracy. However in many cases the necessary number can be over one hundred. Because we have no clear way to know how many intervals we need for different integrals, all integrations start with $2n = 10$ intervals and double the number

until the integral has converged, as determined by a change in value less than CF. The starting point for this doubling of intervals was chosen large enough so that most integrations would only be done twice (ten and twenty intervals), but small enough so that the two integrations done are not tremendously time consuming. For example, starting with four intervals almost always requires performing the integration with eight and sixteen intervals, therefore starting with eight would yield exactly the same results in less time.

Other Algorithmic Details

Under certain parameter conditions, the nonsynonymous rates in the highest rate classes do not have a finite expected value and therefore we cannot assign rates. The situation occurs when $\mu_\alpha + d < 0$, which is the same condition required for the expected value of β to be defined in the continuous case. In other words, only when the correlation is negative and correlation parameter d is larger in magnitude than μ_α will this be a problem. Therefore, our parameters are limited so as not to violate this setting. This does not seem to cause an issue with finding the maximum likelihood estimates of parameters.

Dividing the synonymous distribution into intervals is exact, and therefore, we know that for a given synonymous interval m , $\sum_{n=1}^N p_{mn} = \frac{1}{M}$. If this equality does not hold for each interval, the result will be a discrete distribution whose total probability does not equal 1.0. Naturally small errors occur in the numerical integration and the resulting error is distributed equally among the probabilities of rate classes for that synonymous interval. In other words, after calculating each p_{mn} we update it to be

$$p_{mn} = \frac{p_{mn}}{\sum_{n=1}^N p_{mn}}.$$

It makes more sense to distribute this error separately for each synonymous interval rather than for the entire distribution because the error has a greater chance of being correctly attributed to the box causing the problem. Also, if we combine errors across all synonymous intervals, there is a possibility that much of the error will cancel out. A similar approach can be applied to the synonymous rates. Recall that when finding the nonsynonymous intervals we calculated the mean synonymous value over each synonymous interval (α_m^*). We also know that this temporary rate can be broken up into a linear combination of the rate for each class because the integrals can be divided up,

$$\int_0^\infty \int_{L_m}^{R_m} \alpha f(\alpha, \beta | \eta) d\alpha d\beta = \sum_{n=1}^N \int_{B_{mn}}^{T_{mn}} \int_{L_m}^{R_m} \alpha f(\alpha, \beta | \eta) d\alpha d\beta.$$

Therefore, the temporary rates calculated previously are by definition a combination of the rates for each box,

$$\sum_{n=1}^N \alpha_{mn} p_{mn} = \alpha_m^* p_m = \frac{\alpha_m^*}{M}.$$

Any error in the numerical integration values could then be distributed exactly as with the probabilities. For the nonsynonymous rates, a similar relationship holds if we calculate the corresponding temporary nonsynonymous means over an interval m ,

$$\sum_{n=1}^N \beta_{mn} p_{mn} = \beta_m^* p_m = \frac{\beta_m^*}{M},$$

$$\text{where } \beta_m^* = M \int_0^\infty \int_{L_m}^{R_m} \beta f(\alpha, \beta | \eta) d\alpha d\beta = M\omega \int_{L_m}^{R_m} \alpha^d f(\alpha | \eta) d\alpha.$$

However, we choose not to adjust the synonymous and nonsynonymous rates because the distribution is already proper and the effects of adjusting are minimal. However, these errors are still calculated to allow for a check against any major problems that Simpson's method has encountered.

For the independent bivariate model, the interpretation of a class is more meaningful than in the dependent model. For example, consider the rate class C_{mn} representing the m^{th} synonymous class and n^{th} nonsynonymous class. Under the independent model C_{mn} will have the same synonymous rate but a lower nonsynonymous rate than $C_{m(n+1)}$ and this holds for any m . For the bivariate case the relationship between rates depends heavily on the amount of correlation. In fact, its quite possible for $\beta_{3,2}$ to be greater than $\beta_{2,3}$ even though the second is in a higher nonsynonymous rate class, as is illustrated in Figure 3.7. Therefore, assigning rates α_{mn} and β_{mn} to a site based on a posterior probability has a less clear meaning, and instead we use the weighted average rates for each site s ,

$$\alpha^{(s)} = \sum_{m=1}^M \sum_{n=1}^N r_{mn}^{(s)} \alpha_{mn}$$

$$\beta^{(s)} = \sum_{m=1}^M \sum_{n=1}^N r_{mn}^{(s)} \beta_{mn}$$

where $r_{mn}^{(s)}$ is the posterior probability for each class. These weighted rates give a better understanding of the differences of rates between sites in a gene.

With a small shape parameter of a gamma distribution, problems often arise from significant density lying in ranges of values that can cause either computational underflow or overflow errors. To combat this potential problem, all function evaluations are performed on a log scale. In addition, when the boundaries of integration become too small to handle (i.e. the computer treats them as 0), we use our method of traversing an edge or corner box to store the boundaries of integration as the product of the box boundary and a factor of either the lower or upper factor ($\frac{a}{LF^i}$ or aUF^i). Because we are performing all function calls on a log scale, we can take advantage of properties of logs to split the product into a sum of logs and therefore allow much smaller values.

internals. Also, increasing the number of rate classes does not really change the regions of the distribution that the algorithm numerically integrates. Therefore, although increasing the number of classes has a direct and significant effect on computational time for evaluating the likelihood function, the effect on the process of discretization is minimal. This can be seen in Figure 3.8 which plots the time necessary to discretize different numbers of classes relative to the time required for three classes.

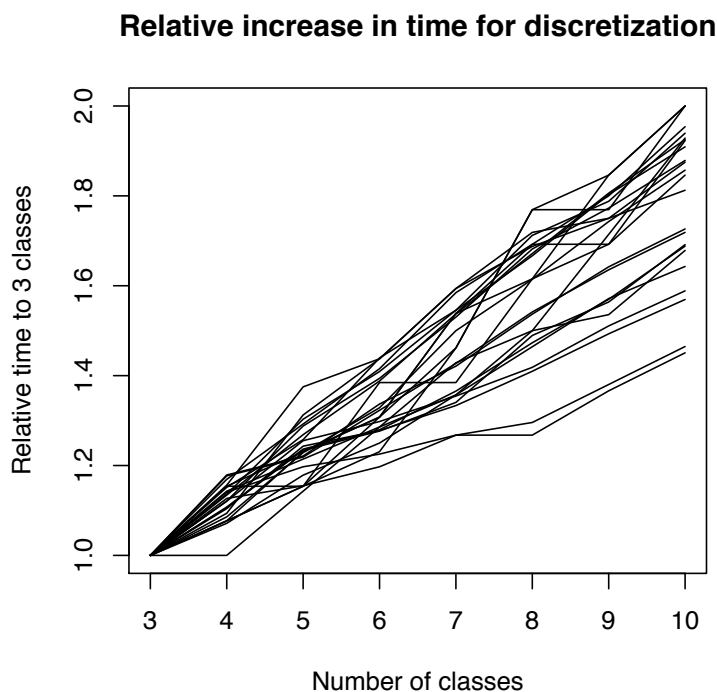


Figure 3.8: **Relative computational time for discretization.** Here we show the amount of computational time required to discretize a distribution relative to the time required for 3 rate classes. Each line represents a different set of values for η . We see that increasing the number of rate classes has a roughly linear affect on computation time.

Previous algorithmic attempts

Originally, the discretization was performed by using Simpson's method once for each box, with the upper and lower bounds (UB and LB) defined by some arbitrarily chosen value. For most cases, the upper bound led to a large area that was being integrated despite containing relatively little of the distribution. Conversely, the lower bound cut off too much probability, especially with a shape parameter less than 1. The number of intervals used in Simpson's method also varied greatly from the internal boxes to the external boxes.

To fix the upper and lower bound problems, we tried to make use of the Gamma-Poisson relationship, namely if $X \sim \text{Gamma}(\mu_\alpha, \lambda)$ and μ_α is an integer $P(X \leq x) = P(Y \geq \mu_\alpha)$, where $Y \sim \text{Poisson}(\frac{x}{\lambda})$. Also note that for the gamma distribution, as the shape parameter μ_α increases, the cumulative distribution function decreases. Therefore, even if our value of μ_α is not an integer, we can bound the probability by a gamma distribution with an integer value. Using a fixed error level (10^{-6} in this example) and for all possible integer values of α we can calculate the upper and lower bounds through the Poisson distribution. If LB and UB are our lower and upper bounds, respectively:

$$P(X \leq LB|\mu_\alpha, \lambda) \leq P(X \leq x|\text{floor}(\mu_\alpha), \lambda) = P(Y \geq \text{floor}(\mu_\alpha)) = 10^{-6}$$

$$\text{where } Y \sim \text{Poisson}\left(\frac{LB}{\lambda}\right)$$

$$P(X > UB|\mu_\alpha, \lambda) \leq P(X > x|\text{ceiling}(\mu_\alpha), \lambda) = P(Y < \text{ceiling}(\mu_\alpha)) = 10^{-6}$$

$$\text{where } Y \sim \text{Poisson}\left(\frac{UB}{\lambda}\right).$$

The problem with using this method is that if $\mu_\alpha < 1$, which is the case most often for typical data sets, the lower bound becomes 0, which is impossible to evaluate using Simpson's integration. Sometimes a useful upper bound was

found that reduced computation, however at times the upper bound was a large overestimate of the true error cut off point, leading to the same original problems. In addition, we do not know how the univariate bounds will carry over to our bivariate distribution.

Romberg Integration

As an alternative to Simpson's integration we tried Romberg integration which is an extension of the trapezoidal rule and makes use of the fact that once having calculated the function values under N intervals, those same points do not need to be recalculated for $2N$ intervals, only the new points in between, and saves considerable time. Several successive doublings are performed, and then the final result is extrapolated to the case where each interval has a width of 0, which corresponds to the true integral. This method was unfortunately too slow, as the number of doublings required became too great.

Chapter 4

STATISTICAL PROPERTIES OF MODELING EVOLUTIONARY RATE HETEROGENEITY WITH DISCRETIZED GAMMA DISTRIBUTIONS

Mannino FV and SV Muse

Introduction

With the realization that rates of evolution are not constant from site-to-site, models incorporating rate variation have become essential to molecular evolution. The use of statistical distributions such as a gamma to model rate variation is quite common. However, for likelihood based methods, continuous distributions add a level of computation that is too intensive for typical sized data sets. Because of this YANG (1994) proposed methodology to approximate a continuous distribution using a discretized version. Discretized distributions greatly reduce the computational burden of continuous distributions and have become the norm in molecular evolutionary literature. While many studies have looked at the statistical properties of ignoring rate variation (e.g., WAKELEY, 1994; YANG *et al.*, 1994; UPHOLT, 1977), to the best of our knowledge no thorough study has looked at the bias and variances in estimates when discretized gamma distributions are used to model rate variation. Here we examine these statistical properties, focusing on estimates of the coefficients of variation (CV) under various conditions. With the discretized distribution, the CV values and variances are bounded based on the number of rate classes. Because of this we perform a simulation study to analyze the estimates of CV under differing amount of variation, comparing the continuous and discretized estimators.

Theory

Variation limits

With a discrete approximation to a continuous distribution, users hope that the discretized version mimics the continuous properties, while providing computa-

tional tractability. When studying rate variation, one obvious area of interest is the amount of variation in the data set. However, here we must be careful because the amount of variation in the discrete approximation is bounded, based on the number of classes we choose to allow. For any discrete distribution with M random variables, $X_m > 0$, and all values equally probable:

$$\begin{aligned}
Var(X) &= \frac{1}{M} \sum_{m=1}^M (X_m - E(X))^2, \\
&= \left(\frac{1}{M} \sum_{m=1}^M X_m^2 \right) - (E(X))^2, \\
&\leq \frac{1}{M} \left(\sum_{m=1}^M X_m^2 + \sum_{i=1}^M \sum_{j=1, j \neq i}^M X_i X_j \right) - (E(X))^2, \\
&= \frac{1}{M} \left(\sum_{m=1}^M X_m \right)^2 - (E(X))^2, \\
&= (E(X))^2 (M - 1).
\end{aligned}$$

We see that the maximum variance is simply the mean squared times the number of classes minus one. For our discretized approximate distributions, the mean is always 1 and the bound becomes $M - 1$. The coefficient of variation ($CV = \frac{\text{Standard Deviation}}{\text{Mean}}$) is similarly bounded by

$$CV(X) \leq \sqrt{M - 1}. \quad (4.1)$$

In addition to the variance estimates, the highest rate class will also be bounded, because $\frac{\sum X_m}{M} = 1$ and all $X_m \geq 0$, the maximum value for X_m is M (Susko *et al.*, 2003). Because the number of rate classes used in analyses typically ranges from four to ten, we see that the largest amount of measurable CV ranges from $\sqrt{3}$ to 3 and the largest possible rate ranges from four to ten. When we allow non-equiprobable rate classes, this bound in Equation 4.1 does not apply, although some more complex bound may exist.

When using discretized gamma distributions, this limit on the amount of variation obviously presents issues with the interpretation of rates and estimates of variance and CV. The best course of action would seem to be simply to increase the number of classes to allow the model to capture more rate variation. However, this leads to a linear increase in computational cost. Also, the amount of variance in the discretized distribution will always be less than that of the continuous distribution with the same shape parameter, and the smaller the shape parameter, the greater this discrepancy. For example, if a discretized model estimates a shape parameter to be 2.5 with six rate classes the variance is 0.351, whereas the variance in the corresponding continuous distribution is 0.4, a difference of 0.05. However, with a shape parameter of 0.2 the variance under the discretized and continuous distributions become 2.91 and 5.0, respectively. The same properties are found if we consider coefficients of variation rather than variance. For this reason, we should not use variances under the discretized distributions as estimates of variances under a continuous distribution, unless the number of rate classes is quite large, which negates the computational savings of using the discretized distribution in the first place. One option is to consider the discrete distribution to be the model for which we mean to interpret the rate variation, and the continuous distribution as simply the tool by which we can obtain a discrete distribution based only on one parameter. Alternatively, we can fit the discretized distribution, but draw inferences only from estimates interpreted through the continuous distribution. In other words, if we choose to use discretized estimates of the variance or CV, we must not claim that these estimates correspond to the parameters in the continuous distribution.

Number of classes

Increasing the number of classes used in the discretization will lead to a discrete distribution that is more similar to the continuous distribution, and intuition dictates that more classes will lead to a better fit of the model to the data. However, this is not necessarily the case. The general tendency of increasing classes is to create a better fitting model, but it is quite possible - and not that uncommon - for a model with M classes to fit better than model with $M + i$ classes. Changing the number of rate classes adds new rates, but also shifts the placement of previous rates. For example, with a shape parameter of 1.3, the rates for four classes are 0.19, 0.55, 1.04, and 2.22. Discretizing into five classes gives us 0.16, 0.44, 0.76, 1.23, and 2.41. For data sets that fit well to the various rates under four classes, the fit under the five class model will be much poorer. Possibly no shape parameter under the five class model will lead to a collection of rates similar to those under the four class model. An example of this shift in rates between four and five classes can be seen in Figure 4.1 for a shape parameter of 1.3.

Choosing the number of discrete classes to use for the gamma distribution is a matter of compromise between computational time and a closer approximation to the continuous distribution. For analyses in Chapters 2 and 3 we used four classes for both synonymous and nonsynonymous distributions which has performed well in previous studies (Kosakovsky Pond and Muse, pers. comm.). Because the likelihood function must be summed over the number of classes, increasing from N to $N + 1$ classes will raise the computational time approximately by a factor of $\frac{(N+1)^2}{N^2}$.

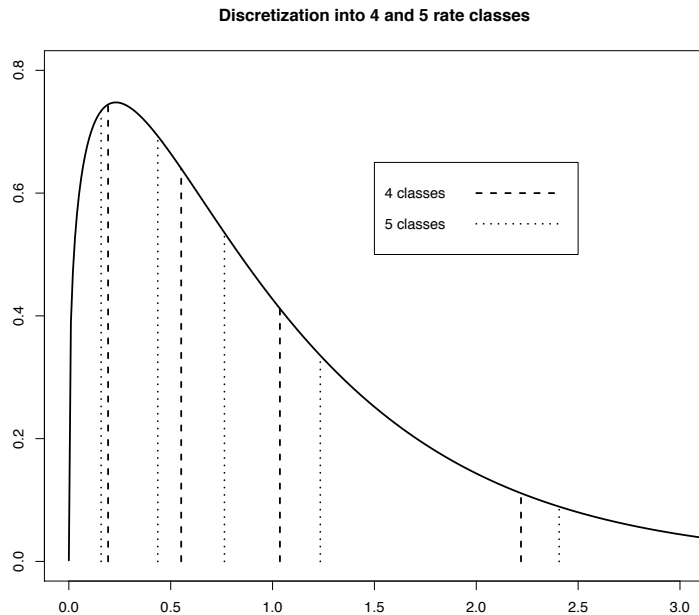


Figure 4.1: **Effect of the number of classes on rates.** Here we present a gamma distribution with a shape parameter of 1.3 and the discrete rates for both four (solid lines) and five (dashed lines) classes. This demonstrates that increasing the number of classes shifts all rates and explains why models with N and $N - 1$ discretized rates are not nested.

Normalized median discretization

When discretizing using the median values over the discrete intervals, it is necessary to normalize the rates to have a mean of 1. Failure to do this will not change the likelihood value of the fit of the model to the data set, but will cause errors in the other parameters estimated by changing the interpretation. Let us consider the model M5 of nonsynonymous rate variation from YANG *et al.* (2000) which can have nonsynonymous rates parameterized as

$$\beta = \omega \times \text{Gamma}(\mu_\beta, \mu_\beta).$$

The expected value of β is $E(\omega \times \text{Gamma}(\mu_\beta, \mu_\beta)) = \omega \times E(\text{Gamma}(\mu_\beta, \mu_\beta))$.

When the expected value of the discretized gamma distribution is 1, as is the case with a mean discretization, ω is the expected value of the nonsynonymous rates and $\hat{\omega} = \hat{E}(\beta)$. This is an easy to interpret and meaningful parameter. If the gamma distribution is discretized using median values and not normalized, the interpretation changes, and $\hat{\omega} = \frac{\hat{E}(\beta)}{E(\text{Gamma}(\hat{\mu}_\beta, \hat{\mu}_\beta))}$. Using $\hat{\omega}$ as an estimate of $\hat{E}(\beta)$ will therefore lead to an overestimation because the expected value of the discretized rates for an unnormalized median distribution will be less than 1. The severity of the bias depends upon both the number of classes in the discretization and the current value of the shape parameter. With more classes, the expected value gets closer to 1 and therefore the bias decreases. As the shape parameter decreases, the expected value drops, increasing the bias. For cases where multiple gamma distributions are used (e.g., KOSAKOVSKY POND and MUSE, 2005), a similar idea applies but the biases of parameter estimates will not be as simple to compute. While fitting an unnormalized median discretization model is not technically incorrect, care must be taken to adjust parameters afterwards before making any inferences.

As an example of this effect, consider the carnivores COX1 data set used in Chapter 2. Using a normalized median discretization gives a maximum likelihood estimate of $\hat{\omega} = 0.0088$, which is in line with the mean discretization estimate and the commonly accepted range for a nonsynonymous/synonymous ratio. If the discrete median is used but not normalized, the estimate becomes $\hat{\omega} = 378.847$. Although normalizing has no effect on the best log-likelihood value, clearly the effect on the parameters can be drastic.

Even more troubling, in many cases the purpose of including rate variation is to study the levels of variation in the data. In the unnormalized median case, when the shape parameter gets smaller than some value, which depends on the

number of classes, the variance will actually decrease which is contrary to the continuous gamma distribution for which the discrete rates are approximating. This undesired property expectedly disappears when the rates are normalized. Figure 4.2 shows the variances of an unnormalized and normalized distribution of rates under a median discretization, as well as under the mean discretization and continuous distribution for various shape parameters. We see both the problems of the unnormalized median and differences between continuous and discretized variances.

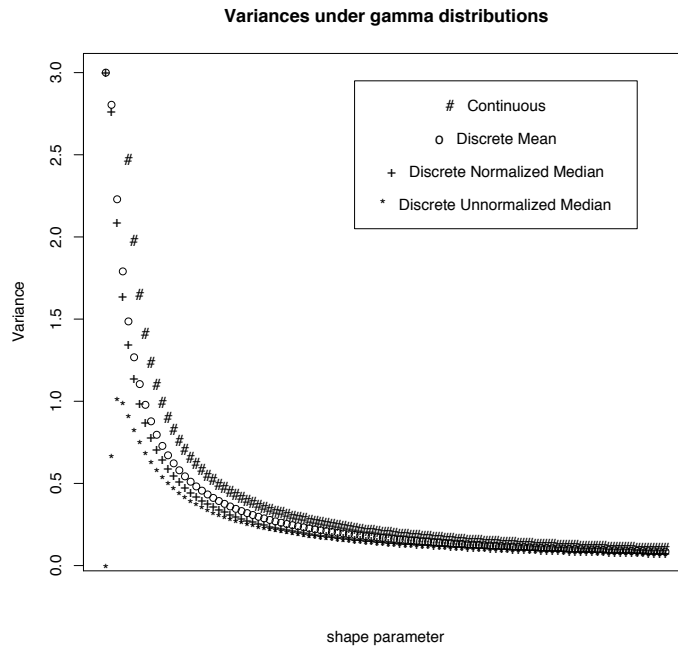


Figure 4.2: **Variances under continuous and discretized gamma distributions.** Here we have variances for various shape parameters under: 1) A continuous gamma distribution (#), 2) A discretized gamma using mean rates (o), 3) A discretized gamma using normalized median rates (+), and 4) A discretized gamma using unnormalized median rates (*). All discretized distributions use 4 rate classes. We see that variances are less for all discretized distributions than for the continuous distribution. In addition, little difference is seen between mean and normalized median, while unnormalized median clearly falls apart at small values of the shape parameter.

Effect of discretization method

Choosing to use a mean discretization versus a normalized median discretization has very little effect on the fit of the model and estimates of parameters. Mean rates tend to be more extreme and lead to slightly more variance in the distribution. However, evidence of a general superiority of one method over the other as determined by the fit of the models, which can be compared using an Akaike Information Criteria (AIC) (AKAIKE, 1974), has not been found.

Simulation study

We performed a simulation study to examine several facets of using a discrete gamma distribution. The core model in the simulation was a nucleotide rate matrix parameterized as

$$q_{ij} = \begin{cases} \pi_j, & i \rightarrow j \text{ is a transition} \\ \kappa\pi_j, & i \rightarrow j \text{ is a transversion} \end{cases}$$

where $\kappa \sim K \times \text{Gamma}(\mu_\kappa, \mu_\kappa)$. We chose a nucleotide model to reduce the computational constraints of codon models. The rate variation is placed solely on the transversion/transition ratio (κ) to mimic the most commonly used codon model with rate variation in which only the nonsynonymous/synonymous ratio varies. The variation was placed on the this ratio and not the inverse, because like the nonsynonymous/synonymous ratio, we expect smaller values to be more prevalent.

To ensure that the phylogenetic tree shape was not biasing our conclusions, we used two different trees with differing levels of divergence between clades. Both of these trees (see Figures 4.3 and 4.4) were taken as subsets of a tree constructed

using the placental mammal mitochondrial NADH5 gene. The branch lengths used for simulation were the values estimated from fitting the above nucleotide model with 4 rate classes to NADH5. To examine the effect of the number of species, we used trees of 20 and 40 taxa for each tree shape.

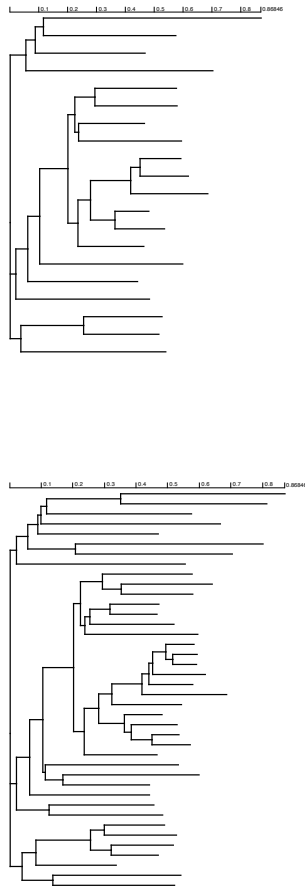


Figure 4.3: **Simulation tree shape 1.** Two of the trees used for simulation, shown with 20 and 40 taxa.

One hundred replicates were simulated for each combination of five shape parameters (0.01, 0.1, 0.5, 1.5, 4.0), three sequence lengths (500, 1500, 10000), and three methods of simulating variation (using a continuous distribution and using

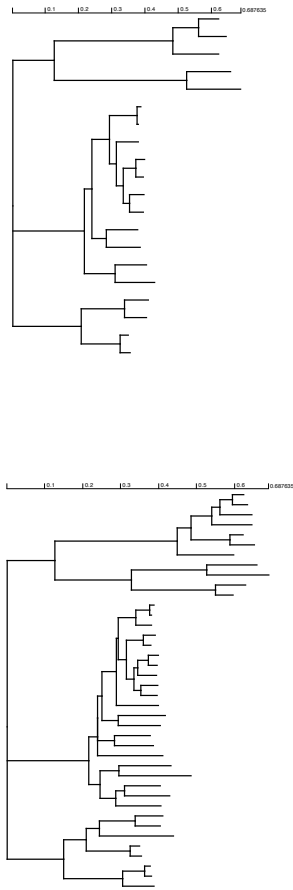


Figure 4.4: **Simulation tree shape 2.** Two of the trees used for simulation, shown with 20 and 40 taxa.

four and ten discretized rate classes). For all simulations we used equal base frequencies and a value of 0.5 for K . For each simulated data set, we found the maximum likelihood estimates under models with 4, 6, 8, 10, and 12 rate classes.

Results and Discussion

Fit of models based on number of rate classes

As previously mentioned, increasing the number of rate classes will provide a better approximation of the continuous version of the distribution, but not necessarily a better fit of the data set to the model. We use the simulation study results to examine how often a data set fits better with less classes. Here we consider only the data sets simulated under the continuous distributions because for those simulated under discretized distributions, the number of classes used in the simulation will affect the fit of the model. Of the 6000 simulated data sets, 5085 (84.75%) had log likelihood values that were strictly increasing with the number of classes. For comparison, the numbers when data was simulated under 4 and 10 discrete rate classes are 9.18% and 30.2%, respectively.

We can also examine the relative effect of each step of increasing the number of classes. For example, we may wish to ask whether going from 4 to 6 rate classes offers the same improvement as going from 10 to 12. In Figure 4.5 we see the difference in log likelihood values between models with different number of rate classes. Two things are immediately evident. First, longer sequences causes a greater increase in log likelihood values, and more often leads to an improved fit. For example, sequences of length 500 improved only 86.35% of the time when switching from 4 to 6 rate classes, while sequences of length 10000 improved 97.05% of the time. Secondly, increases in model fit occur less often as the number of classes increases. For sequences of length 1500, switching from 4 to 6 classes lead to an improvement 91.55% of the time, while switching from 10 to 12 classes lead to an improvement only 83.8% of the time.

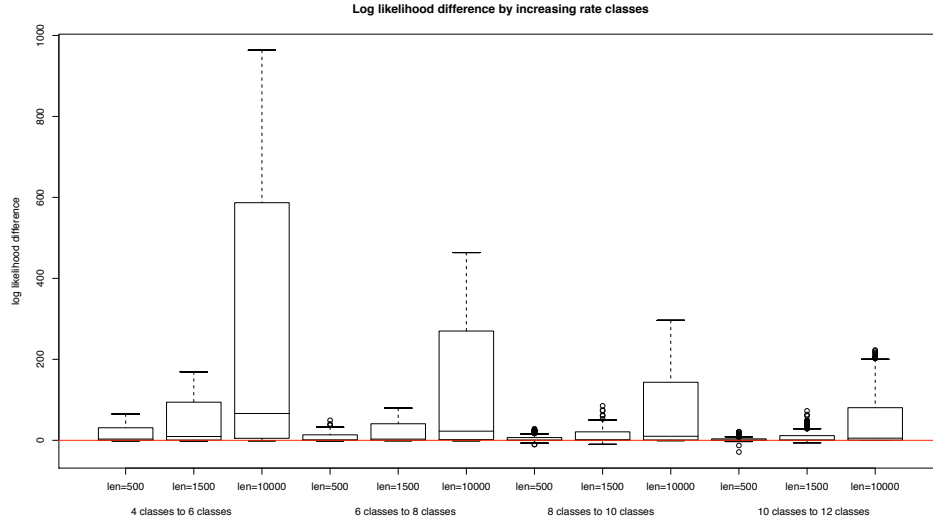


Figure 4.5: **Log likelihood values based on classes.** This plot shows the change in log likelihood values when increasing from N to $N + 2$ rate classes, for each of three sequence lengths. Each successive increase in rate classes has less of an effect on the likelihood.

These improvement of fit results confirm the trend that increasing the number of rate classes leads to a better fit of the model. However, the increase in log likelihood values decreases with more classes. In addition, longer sequences are more likely to improve with more rate classes.

Estimates of Coefficients of Variation

When fitting a discretized distribution of rate heterogeneity, we generally assume an underlying continuous distribution and approximate using a discretized version. Our interests here lie in estimating the amounts of variation by using the coefficients of variation. However, there are two possible ways to estimate the CV value, using the discrete distribution that was actually fit to the data set, or by using the CV of the continuous distribution defined by the estimate of the shape parameter. Since the goal is to best estimate the true CV value, we use the simulation

study to determine which approach works best. The means and variances for both estimates can be seen in Tables 4.1 and 4.2.

When the true shape parameter is small (0.01 or 0.1), the continuous CV estimate was less biased than the discrete estimate, although neither performed well (see Figure 4.6). This condition holds regardless of any other settings, but is more pronounced with a tree of 40 taxa. For these shape parameters, the discrete estimate always underestimates, because of the upper bound on the CV discussed earlier. The continuous CV can overestimate or underestimate the true values. Although the continuous CV estimates are better in these cases, the error is still quite large, even for very long sequences. Increasing the number of classes does help, but even with 12 classes, most estimates are off by greater than 3.

As the true shape parameter increases, the discrete CV estimates improve relative to the continuous, with both estimates being virtually equal when the shape parameter is 4.0. These simulation results give support to the notion that using the continuous CV estimate will provide estimates of the true CV at least as well as the discrete estimates, often giving much better estimates. However, this assumes that we have a continuous gamma distribution as the true underlying distribution of rates across sites. When we simulated under 4 discretized rate classes, the discrete CV estimate outperformed the continuous CV for all simulation settings. Likewise, under 10 simulated rate classes, the discrete CV generally provided better estimates, except when estimating under very few rates classes (4 or 6).

Estimates of K

Estimates of the expected value of κ , denoted by our parameter K , were greatly affected by the shape parameter and also by the number of classes used in esti-

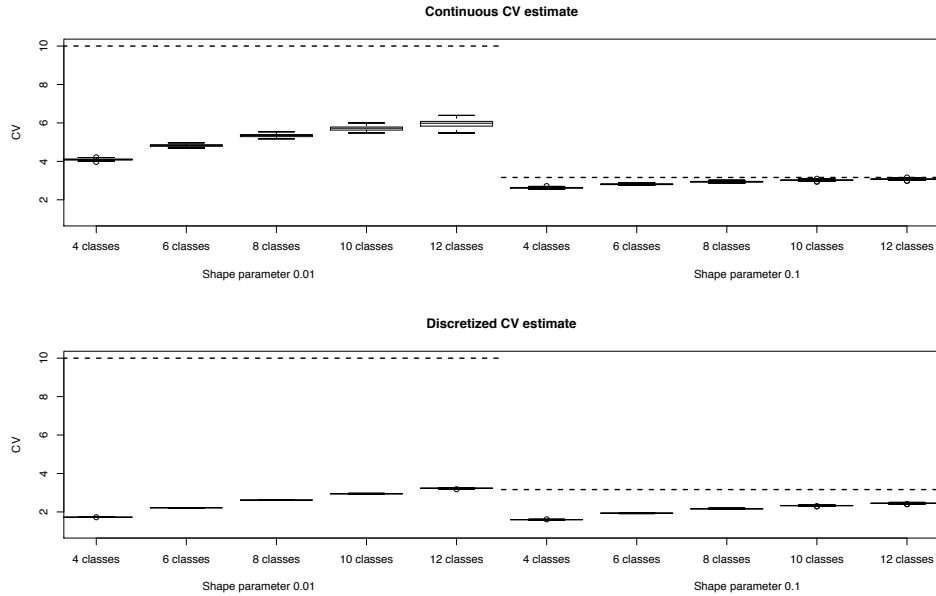


Figure 4.6: **CV estimates under small shape parameter.** This plot shows estimates of the coefficients of variation when the true shape parameter were small (0.01 and 0.1). The simulated data sets were analyzed under varying number of discretized classes. The top plot represents the continuous estimator and the bottom plot is the discretized estimator. The dashed lines are the true value of the CV in the simulation. The tree in the this plot is tree 2 with 40 taxa. The sequence length is 10000.

mation. For larger shape parameters (1.5 and 4.0), estimates were very unbiased, regardless of any settings. When the shape parameter was small (0.01 and 0.1), the positive bias in the estimate could be quite large. This bias decreases with more rate classes, but not with longer sequences. A plot of estimates of K under four rate classes can be seen in Figure 4.7. This parameter is meant to emulate the nonsynonymous/synonymous ratio, but we see here that estimates of ω could be greatly biased when the shape parameter is small.

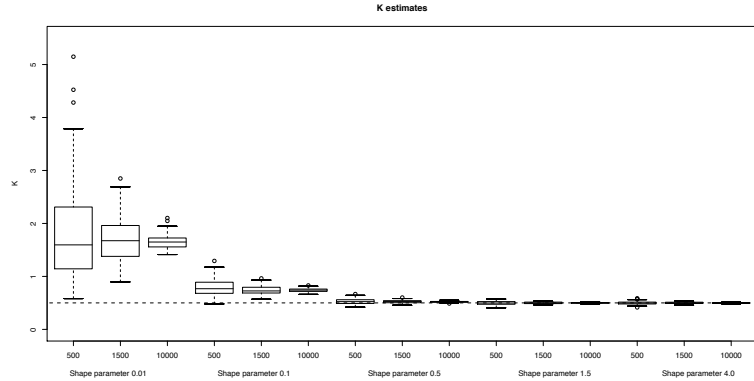


Figure 4.7: **Estimates of K.** Here we present a plot showing the estimates of K under tree 2, with 40 taxa, estimated under 4 rate classes. Each of the three sequence lengths are displayed in increasing order for each of the five shape parameters. The horizontal line represents the true simulated value. We see that for small shape parameters there is a bias in estimation which does not go away with longer sequences. For large shape parameters the estimates are very good.

Estimates of shape parameter

At larger values of the simulated shape parameter (0.5, 1.5, and 4.0), the estimate of this parameters was slightly negatively biased. This agrees with previously published results (e.g., EXCOFFIER and YANG, 1999). However, for small shape parameters we find the exact opposite result. Estimates for the shape parameter are overestimated. While this could be because we are estimating a parameter close to its range boundary, as we increase the number of rate classes this bias decreases.

Effect of tree shape and number of taxa

To examine the effect of phylogenetic tree shape, we considered a tree with short internal branches (Figure 4.3) and a tree with long internal branches (Figure 4.4). Continuous CV estimates were not greatly affected by either the tree shape or the

number of taxa. With more taxa, the variance in the estimates decreases and the variance is smaller under tree shape 1. However, there is little change in bias found between on the trees. Interestingly, when the shape parameter is small, trees with fewer taxa can perform better in estimating the CV, although this could potentially be an effect of the taxon sampling going from 40 to 20 tree tips.

Conclusions

One of the most interesting observations was that the length of the sequence seemed to have little effect on the bias of the CV estimates or K estimates implying that the asymptotic properties hold up well for even short sequences. Longer sequences naturally reduce the variance in the estimates, but any bias present in the short sequence is still present in the longest sequence. The number of rate classes used in the estimation on the other hand could have a large affect on the estimates, specifically implying that small shape parameter estimates with few classes cannot be trusted.

While we do not believe that the underlying distribution of rates across sites follows a discretized gamma distribution, we evaluated simulations of this kind to understand what our estimates would be like under these conditions. Expectedly, the estimates of CV that performed best under these simulations were the discrete CV. In addition, the log likelihood values were generally higher for the number of rate classes that matched the number of simulated rate classes. These results confirm that the simulation and estimation routines are performing as we expect.

To estimate the CV of a distribution of rate variation, we have two points of evidence in favor of the continuous estimate. First, the theoretical bounds on the discrete estimate make its use troublesome, as seen in Equation 4.1. Second,

the simulation results demonstrate that the continuous CV estimate outperforms the discrete estimate unless the shape parameter is large. Even for true values of CV that are within the bounds of the discrete estimator, the continuous CV is less biased. However, all estimates at low shape parameter should be taken with caution, as these can be greatly biased. Allowing for rate classes that are not equiprobable (e.g., KOSAKOVSKY POND and FROST, 2005) could possibly eliminate the problems with the discrete CV estimate.

Table 4.1: CV estimates for simulations under tree 1. Here we have the means continuous (cCV) and discretized (dCV) coefficient of variation estimates, with standard errors in parentheses. These results are for tree shape 1.

Tree	True CV	Sequence Length	Number of rate classes									
			4 classes		6 classes		8 classes		10 classes		12 classes	
			cCV	dCV	cCV	dCV	cCV	dCV	cCV	dCV	cCV	dCV
20 taxa	10	500	8.042 (9.189)	1.727 (0.004)	5.938 (4.618)	2.222 (0.008)	5.851 (2.52)	2.621 (0.013)	6.113 (0.467)	2.963 (0.019)	6.566 (0.576)	3.265 (0.028)
		1500	4.47 (2.731)	1.726 (0.002)	4.966 (0.176)	2.222 (0.005)	5.516 (0.217)	2.619 (0.008)	5.945 (0.262)	2.958 (0.014)	6.345 (0.382)	3.257 (0.02)
		10000	4.189 (0.044)	1.726 (0.001)	4.971 (0.063)	2.222 (0.002)	5.513 (0.078)	2.62 (0.003)	5.951 (0.089)	2.96 (0.004)	6.315 (0.103)	3.259 (0.006)
	3.162	500	2.689 (0.131)	1.609 (0.024)	2.862 (0.138)	1.948 (0.042)	2.973 (0.141)	2.174 (0.056)	3.042 (0.163)	2.334 (0.074)	3.077 (0.174)	2.448 (0.089)
		1500	2.676 (0.071)	1.608 (0.014)	2.856 (0.075)	1.948 (0.023)	2.966 (0.083)	2.173 (0.032)	3.034 (0.092)	2.332 (0.042)	3.073 (0.097)	2.448 (0.05)
		10000	2.664 (0.026)	1.606 (0.005)	2.84 (0.028)	1.944 (0.009)	2.947 (0.031)	2.166 (0.012)	3.014 (0.035)	2.324 (0.016)	3.054 (0.038)	2.439 (0.02)
	1.414	500	1.435 (0.071)	1.142 (0.041)	1.421 (0.073)	1.226 (0.052)	1.415 (0.074)	1.268 (0.057)	1.412 (0.075)	1.294 (0.061)	1.411 (0.075)	1.312 (0.063)
		1500	1.425 (0.044)	1.137 (0.026)	1.41 (0.045)	1.219 (0.032)	1.404 (0.045)	1.26 (0.035)	1.401 (0.046)	1.285 (0.037)	1.4 (0.046)	1.303 (0.039)
		10000	1.435 (0.014)	1.143 (0.008)	1.421 (0.015)	1.227 (0.011)	1.415 (0.015)	1.269 (0.012)	1.413 (0.015)	1.295 (0.012)	1.412 (0.015)	1.313 (0.013)
	0.816	500	0.851 (0.058)	0.745 (0.045)	0.832 (0.057)	0.766 (0.048)	0.825 (0.056)	0.777 (0.05)	0.821 (0.056)	0.784 (0.051)	0.819 (0.056)	0.789 (0.052)
		1500	0.85 (0.037)	0.744 (0.029)	0.831 (0.037)	0.766 (0.031)	0.824 (0.036)	0.777 (0.032)	0.82 (0.036)	0.784 (0.033)	0.818 (0.036)	0.788 (0.034)
		10000	0.848 (0.013)	0.743 (0.01)	0.829 (0.013)	0.764 (0.011)	0.822 (0.013)	0.775 (0.011)	0.819 (0.013)	0.782 (0.012)	0.817 (0.013)	0.787 (0.012)
0.5	500	0.517 (0.07)	0.469 (0.061)	0.502 (0.068)	0.474 (0.062)	0.497 (0.067)	0.477 (0.063)	0.494 (0.067)	0.479 (0.064)	0.492 (0.067)	0.48 (0.064)	
	1500	0.528 (0.035)	0.478 (0.03)	0.513 (0.034)	0.484 (0.031)	0.507 (0.034)	0.487 (0.032)	0.504 (0.034)	0.488 (0.032)	0.502 (0.034)	0.49 (0.033)	
	10000	0.531 (0.014)	0.482 (0.012)	0.516 (0.014)	0.487 (0.013)	0.51 (0.014)	0.49 (0.013)	0.507 (0.014)	0.492 (0.013)	0.506 (0.014)	0.493 (0.013)	
40 taxa	10	500	4.162 (0.219)	1.725 (0.004)	4.908 (0.294)	2.219 (0.008)	5.45 (0.412)	2.613 (0.02)	5.858 (0.549)	2.945 (0.039)	6.17 (0.634)	3.236 (0.052)
		1500	4.084 (0.112)	1.724 (0.002)	4.8 (0.152)	2.217 (0.005)	5.277 (0.256)	2.608 (0.017)	5.611 (0.393)	2.933 (0.032)	5.863 (0.481)	3.214 (0.046)
		10000	4.078 (0.04)	1.724 (0.001)	4.804 (0.052)	2.217 (0.002)	5.308 (0.074)	2.611 (0.004)	5.643 (0.11)	2.941 (0.008)	5.81 (0.222)	3.216 (0.022)
	3.162	500	2.497 (0.112)	1.568 (0.026)	2.708 (0.096)	1.899 (0.033)	2.874 (0.101)	2.136 (0.042)	2.988 (0.109)	2.311 (0.052)	3.06 (0.128)	2.44 (0.067)
		1500	2.465 (0.066)	1.561 (0.016)	2.676 (0.05)	1.889 (0.018)	2.843 (0.057)	2.123 (0.024)	2.952 (0.067)	2.294 (0.032)	3.021 (0.074)	2.421 (0.039)
		10000	2.457 (0.025)	1.56 (0.006)	2.658 (0.024)	1.883 (0.008)	2.819 (0.024)	2.113 (0.01)	2.932 (0.025)	2.285 (0.012)	3.008 (0.029)	2.414 (0.015)
	1.414	500	1.425 (0.055)	1.136 (0.032)	1.428 (0.059)	1.231 (0.042)	1.425 (0.06)	1.276 (0.046)	1.422 (0.06)	1.302 (0.048)	1.42 (0.06)	1.32 (0.05)
		1500	1.422 (0.035)	1.135 (0.021)	1.425 (0.039)	1.23 (0.027)	1.423 (0.04)	1.274 (0.031)	1.42 (0.04)	1.301 (0.032)	1.418 (0.04)	1.318 (0.034)
		10000	1.423 (0.013)	1.136 (0.008)	1.427 (0.014)	1.231 (0.01)	1.424 (0.014)	1.275 (0.011)	1.421 (0.014)	1.302 (0.011)	1.419 (0.014)	1.319 (0.012)
	0.816	500	0.846 (0.044)	0.741 (0.034)	0.829 (0.043)	0.764 (0.037)	0.822 (0.043)	0.775 (0.038)	0.819 (0.043)	0.782 (0.039)	0.817 (0.043)	0.786 (0.039)
		1500	0.848 (0.026)	0.742 (0.021)	0.831 (0.027)	0.766 (0.023)	0.824 (0.027)	0.777 (0.024)	0.82 (0.027)	0.783 (0.024)	0.818 (0.026)	0.788 (0.024)
		10000	0.853 (0.01)	0.747 (0.008)	0.836 (0.01)	0.77 (0.008)	0.829 (0.01)	0.782 (0.009)	0.826 (0.01)	0.788 (0.009)	0.824 (0.01)	0.793 (0.009)
0.5	500	0.528 (0.045)	0.478 (0.039)	0.514 (0.044)	0.485 (0.04)	0.508 (0.044)	0.488 (0.041)	0.506 (0.043)	0.49 (0.041)	0.504 (0.043)	0.492 (0.042)	
	1500	0.531 (0.024)	0.481 (0.02)	0.517 (0.023)	0.487 (0.021)	0.511 (0.023)	0.491 (0.021)	0.508 (0.022)	0.493 (0.021)	0.506 (0.022)	0.494 (0.021)	
	10000	0.531 (0.009)	0.482 (0.008)	0.517 (0.009)	0.488 (0.008)	0.512 (0.009)	0.492 (0.008)	0.509 (0.009)	0.494 (0.008)	0.507 (0.009)	0.495 (0.008)	

Table 4.2: CV estimates for simulations under tree 2. Here we have the means continuous (cCV) and discretized (dCV) coefficient of variation estimates, with standard errors in parentheses. These results are for tree shape 2.

Tree	True CV	Sequence Length	Number of rate classes									
			4 classes		6 classes		8 classes		10 classes		12 classes	
			cCV	dCV	cCV	dCV	cCV	dCV	cCV	dCV	cCV	dCV
20 taxa	10	500	13.883 (12.177)	1.729 (0.003)	10.464 (9.962)	2.228 (0.006)	8.834 (8.079)	2.629 (0.011)	8.139 (6.557)	2.971 (0.018)	7.844 (5.368)	3.273 (0.024)
		1500	11.737 (11.4)	1.728 (0.003)	6.132 (4.754)	2.225 (0.005)	5.653 (0.182)	2.624 (0.006)	6.119 (0.215)	2.966 (0.01)	6.504 (0.244)	3.268 (0.014)
		10000	6.888 (7.872)	1.727 (0.002)	5.035 (0.066)	2.224 (0.002)	5.623 (0.083)	2.624 (0.003)	6.068 (0.1)	2.965 (0.004)	6.429 (0.115)	3.265 (0.006)
	3.162	500	4.152 (5.616)	1.659 (0.033)	3.028 (0.178)	1.995 (0.048)	3.093 (0.197)	2.218 (0.072)	3.133 (0.219)	2.373 (0.096)	3.154 (0.233)	2.485 (0.116)
		1500	2.978 (0.148)	1.655 (0.018)	2.999 (0.097)	1.989 (0.026)	3.052 (0.105)	2.206 (0.039)	3.084 (0.112)	2.355 (0.05)	3.104 (0.118)	2.463 (0.06)
		10000	2.95 (0.054)	1.653 (0.007)	3 (0.04)	1.991 (0.011)	3.06 (0.044)	2.21 (0.016)	3.096 (0.047)	2.361 (0.021)	3.116 (0.05)	2.471 (0.026)
	1.414	500	1.522 (0.119)	1.19 (0.066)	1.474 (0.114)	1.262 (0.079)	1.455 (0.112)	1.298 (0.086)	1.446 (0.112)	1.321 (0.09)	1.441 (0.112)	1.336 (0.094)
		1500	1.497 (0.063)	1.177 (0.035)	1.451 (0.061)	1.247 (0.042)	1.434 (0.06)	1.283 (0.046)	1.426 (0.06)	1.305 (0.048)	1.421 (0.06)	1.321 (0.05)
		10000	1.498 (0.029)	1.179 (0.016)	1.452 (0.028)	1.248 (0.02)	1.435 (0.028)	1.284 (0.021)	1.426 (0.028)	1.306 (0.022)	1.422 (0.027)	1.321 (0.023)
	0.816	500	0.857 (0.114)	0.747 (0.089)	0.83 (0.111)	0.763 (0.095)	0.819 (0.11)	0.772 (0.098)	0.814 (0.11)	0.777 (0.1)	0.811 (0.11)	0.78 (0.102)
		1500	0.869 (0.058)	0.759 (0.045)	0.843 (0.056)	0.776 (0.048)	0.833 (0.056)	0.785 (0.05)	0.827 (0.056)	0.79 (0.051)	0.824 (0.056)	0.794 (0.052)
		10000	0.872 (0.026)	0.761 (0.02)	0.845 (0.025)	0.778 (0.021)	0.835 (0.024)	0.787 (0.022)	0.829 (0.024)	0.792 (0.022)	0.826 (0.024)	0.796 (0.022)
0.5	500	0.526 (0.139)	0.475 (0.121)	0.508 (0.135)	0.478 (0.123)	0.501 (0.133)	0.48 (0.125)	0.498 (0.132)	0.482 (0.126)	0.495 (0.131)	0.483 (0.126)	
	1500	0.518 (0.086)	0.47 (0.075)	0.502 (0.083)	0.473 (0.076)	0.495 (0.082)	0.475 (0.077)	0.491 (0.082)	0.477 (0.078)	0.489 (0.081)	0.477 (0.078)	
	10000	0.534 (0.031)	0.484 (0.027)	0.517 (0.03)	0.488 (0.027)	0.51 (0.029)	0.49 (0.028)	0.507 (0.029)	0.491 (0.028)	0.504 (0.029)	0.492 (0.028)	
40 taxa	10	500	4.491 (2.665)	1.725 (0.004)	5.234 (2.542)	2.221 (0.009)	5.781 (2.595)	2.617 (0.017)	5.946 (0.557)	2.952 (0.029)	6.269 (0.557)	3.247 (0.04)
		1500	4.093 (0.116)	1.724 (0.002)	4.832 (0.153)	2.218 (0.005)	5.342 (0.196)	2.612 (0.01)	5.711 (0.25)	2.944 (0.018)	5.964 (0.332)	3.228 (0.032)
		10000	4.093 (0.043)	1.724 (0.001)	4.821 (0.062)	2.218 (0.002)	5.338 (0.079)	2.613 (0.004)	5.715 (0.107)	2.946 (0.007)	5.962 (0.166)	3.231 (0.015)
	3.162	500	2.64 (0.114)	1.6 (0.023)	2.846 (0.134)	1.943 (0.042)	2.981 (0.152)	2.177 (0.06)	3.067 (0.172)	2.345 (0.078)	3.118 (0.189)	2.468 (0.095)
		1500	2.638 (0.062)	1.6 (0.013)	2.821 (0.062)	1.937 (0.02)	2.94 (0.064)	2.163 (0.026)	3.024 (0.074)	2.328 (0.035)	3.08 (0.086)	2.451 (0.044)
		10000	2.625 (0.028)	1.598 (0.006)	2.813 (0.03)	1.935 (0.01)	2.938 (0.03)	2.163 (0.012)	3.022 (0.033)	2.328 (0.015)	3.076 (0.036)	2.45 (0.019)
	1.414	500	1.464 (0.091)	1.158 (0.052)	1.443 (0.095)	1.241 (0.066)	1.432 (0.095)	1.281 (0.072)	1.426 (0.094)	1.305 (0.076)	1.422 (0.094)	1.321 (0.078)
		1500	1.467 (0.041)	1.161 (0.023)	1.446 (0.043)	1.244 (0.03)	1.436 (0.044)	1.284 (0.034)	1.43 (0.044)	1.309 (0.036)	1.426 (0.045)	1.325 (0.038)
		10000	1.467 (0.017)	1.161 (0.01)	1.447 (0.018)	1.245 (0.013)	1.438 (0.019)	1.286 (0.014)	1.432 (0.019)	1.311 (0.015)	1.429 (0.019)	1.327 (0.016)
	0.816	500	0.878 (0.07)	0.765 (0.054)	0.855 (0.069)	0.785 (0.059)	0.845 (0.069)	0.796 (0.061)	0.841 (0.069)	0.802 (0.062)	0.838 (0.069)	0.806 (0.063)
		1500	0.868 (0.042)	0.758 (0.032)	0.845 (0.042)	0.778 (0.035)	0.835 (0.042)	0.787 (0.037)	0.831 (0.042)	0.793 (0.038)	0.828 (0.042)	0.797 (0.039)
		10000	0.864 (0.014)	0.755 (0.011)	0.841 (0.013)	0.774 (0.011)	0.832 (0.013)	0.784 (0.012)	0.827 (0.013)	0.79 (0.012)	0.824 (0.013)	0.794 (0.012)
0.5	500	0.535 (0.082)	0.484 (0.071)	0.519 (0.08)	0.489 (0.073)	0.512 (0.079)	0.491 (0.074)	0.509 (0.078)	0.493 (0.075)	0.507 (0.078)	0.494 (0.075)	
	1500	0.537 (0.042)	0.486 (0.036)	0.52 (0.04)	0.491 (0.037)	0.514 (0.04)	0.493 (0.037)	0.51 (0.039)	0.494 (0.037)	0.508 (0.039)	0.495 (0.038)	
	10000	0.533 (0.016)	0.483 (0.014)	0.516 (0.016)	0.487 (0.014)	0.51 (0.015)	0.49 (0.014)	0.506 (0.015)	0.491 (0.015)	0.504 (0.015)	0.492 (0.015)	

References

- AKAIKE, H., 1974 A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* **19**: 716–723.
- EXCOFFIER, L., and Z. YANG, 1999 Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Mol. Biol. Evol.* **16**: 1357–1368.
- KOSAKOVSKY POND, S. L., and S. D. W. FROST, 2005 A simple hierarchical approach to modeling distributions of substitution rates. *Mol. Biol. Evol.* **22**: 223–234.
- KOSAKOVSKY POND, S. L., and S. V. MUSE, 2005 Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* **22**: 2375–2385.
- SUSKO, E., C. FIELD, C. BLOUIN and A. J. ROGER, 2003 Estimation of rates-across-sites distributions in phylogenetic substitution models. *Syst. Biol.* **52**: 594–603.
- UPHOLT, W. B., 1977 Evolution of DNA sequence divergence from comparison of restriction endonuclease digests. *Nucl. Acids Research* **4**: 1257–1265.
- WAKELEY, J., 1994 Substitution rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* **11**: 436–442.

- YANG, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**: 105–111.
- YANG, Z., N. GOLDMAN and A. FRIDAY, 1994 Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**: 316–324.
- YANG, Z., R. NIELSEN, N. GOLDMAN and A. PEDERSEN, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431–449.

Chapter 5

CONCLUSION

Final thoughts

The progression of models of molecular evolution has been to more realistically capture the mechanisms of evolutionary change. Relaxing the assumption of rate homogeneity was a major step in this process, but commonly used models of rate variation are still lacking. Synonymous rates have long been assumed constant across sites, but as we have seen this assumption does not hold. Knowing that for many genes synonymous rates are not constant, we can examine potential correlation between this rate and the nonsynonymous rate.

While the model described for the rate correlation functions well and detects correlations in many genes, we have no reason to limit ourselves to using this particular bivariate distribution. Any bivariate discrete distribution can test for correlations, although the number of parameters in such a model necessary to detect correlation may be a limiting factor. Applying a different continuous bivariate distribution is straightforward from the methods described in Chapter 3. The algorithm is capable of discretizing any distribution, even if some details were fine tuned towards this particular model.

As we have seen, the inclusion of site-to-site rate variation is necessary and provides difficult problems. We wish to develop models of heterogeneity that best answer the questions of interest. In certain situations some distributions of rate variation can break down (see Chapter 4) and knowledge of weaknesses of our approaches is essential, lest we make inferences that our model does not truly support.