

PHYLOGENETIC INFERENCE OF COMPLEX, EVOLUTIONARY MODELS: A BAYESIAN APPROACH

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

In

The Department of Biological Sciences

by
Neeraja M. Krishnan
B.E., University of Mumbai, INDIA, 2001
August, 2004

ACKNOWLEDGEMENTS

I wish to extend my sincere gratitude to Dr. David D. Pollock, my advisor and committee chair in Biological Sciences Department, without whose help and guidance, my work that I present here would have been impossible. I would like to express my heart-felt thanks to Dr. Donald Kraft, my major professor in the Computer Science Department for his kind support and for agreeing to serve on my committee. I would also like to thank my other committee members from Biological Sciences Department, Dr. Marcia Newcomer and Dr. Michael Hellberg, both of whom I had my first memorable experiences, in the classes that they taught.

I further wish to thank Dr. Thomas Moore, Associate Chair of Grad. Studies, Biol. Sc. Department and Dr. Pamela Monroe, Asst. Dean of Graduate School for helping me overcome difficulties during my dual program.

I deeply thank my parents for their continued faith in me, and my brother for standing by me and helping my parents in their efforts. I thank Yash Rachakonda for being my best friend and, for his good will and understanding. Last but not the least; I would like to thank Hervé Seligmann for being a good personal and professional companion. The nature walks with him and useful discussions of the biological significance of our analyses were indeed huge motivating factors for my successful work.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	ii
LIST OF TABLES.....	v
LIST OF FIGURES.....	vi
ABSTRACT.....	vii
CHAPTER 1.GENERAL INTRODUCTION AND BACKGROUND	1
1.1. Introduction.....	1
1.2. Background.....	9
1.2.1. Mitochondrion and Its Genome Structure in Animals.....	9
1.2.2. Mitochondrial Replication Mechanism.....	10
1.3. Literature Cited.....	12
CHAPTER 2. ANCESTRAL SEQUENCE RECONSTRUCTION IN PRIMATE MITOCHONDRIAL DNA: COMPOSITIONAL BIAS AND EFFECT ON FUNCTIONAL INFERENCE.....	18
2.1. Introduction.....	19
2.2. Materials and Methods.....	24
2.2.1. Genome Sequences and Phylogeny.....	24
2.2.2. Likelihood Calculations.....	25
2.2.3. Running the Markov Chain.....	27
2.2.4. Chain Convergence Diagnostics.....	30
2.2.5. Parsimony, Maximum Likelihood, and Bayesian Estimation.....	31
2.2.6. Functional Test.....	32
2.2.7. Simulations of Constant and Variable Evolution.....	33
2.3. Results.....	34
2.3.1. Chain Convergence.....	34
2.3.2. Differences in Base Frequencies of Reconstructed Ancestral Sequences.....	34
2.3.3. Simulation Results.....	38
2.3.4. Comparison of Base Frequencies and Structure Stabilities of Reconstructed tRNAs.....	39
2.4. Discussion.....	43
2.5. Literature Cited.....	47
CHAPTER 3. DETECTING GRADIENTS OF ASYMMETRY IN SITE-SPECIFIC SUBSTITUTIONS IN MITOCHONDRIAL GENOMES.....	54
3.1. Introduction.....	55
3.2. Materials and Methods.....	58
3.2.1. Genome Sequences, Alignment, Phylogenetics, and Filtering the Data.....	58
3.2.2. Calculation of Time Spent Single-Stranded.....	59
3.2.3. Posterior Predictive of Reconstructed Ancestral States.....	61

3.2.4.	Incorporation of a Different Asymmetric Mutation Component at Each Site.....	62
3.2.5.	Average Site-Specific mRNA Secondary Structure.....	64
3.3.	Results.....	66
3.3.1.	Analysis of Transition Substitution Response Gradients.....	66
3.3.2.	Analysis of Transversion Substitution Response Gradients.....	67
3.3.3.	Correlation of Secondary Structure and Residual Transition Bias.....	68
3.4.	Discussion.....	69
3.5.	Literature Cited.....	72
CHAPTER 4. SITE-SPECIFIC MODELS: A COMPLETE POSTERIOR-PREDICTIVE APPROACH		75
4.1.	Overview.....	75
4.2.	Partitioning of “Species” and “Sequence” Datasets.....	76
4.3.	Linear Models.....	77
4.4.	MCMC Analyses with Linear and HMM Models on Simulated Data.....	78
4.5.	Results.....	79
4.5.1.	Posterior Means and 95% Confidence Intervals of Slope and Intercept Parameters and Log-likelihoods Under the Linear Model for All Transitions.....	79
4.5.2.	Relative Substitution Rate Responses Profiled Versus Time-spent Single-Stranded in the Genome for the Various “Sequence” and “Species” Partitions.....	81
4.5.3.	Methodological Bias Estimation After Simulating Data Under “Linear” and “Asymptote” Models.....	86
4.6.	Posterior Predictive Analyses.....	87
4.7.	Literature Cited.....	90
CONCLUSIONS.....		92
APPENDIX A: SUPPLEMENTARY DATA FOR CHAPTER 2.....		93
APPENDIX B: SUPPLEMENTARY DATA FOR CHAPTER 3.....		96
VITA.....		97

LIST OF TABLES

Table 2.1	Nucleotide Frequencies and Frequency Ratios for Extant Sequences (Tips) and Ancestral States in the COI Gene.....	36
Table 2.2	Nucleotide Frequencies and Frequency Ratios for Extant Sequences (Tips) and Ancestral States in the Cyt-B Gene.....	37
Table 2.3	Maximum Likelihood Values for Different Methods with the COI and Cyt-B Datasets.....	38
Table 2.4	Biases for Each Nucleotide Averaged Over All the Internal Nodes and MSEs for Various Methods for Simulations Performed with Constant and Variable Models of Evolution.....	39
Table 2.5	Proportion of Base Pairs for which B2 had Higher Complementarity than ML, Classified By Node Ambiguity and Nucleotide Variation at Each Site.....	40
Table 2.6.	Average Nucleotide Frequencies at Tips and Internal Nodes for tRNAs Coded on the Heavy Strand (HS) and Light Strand (LS).....	43
Table 4.1.	Posterior Means and 95% Credible Intervals of Slopes and intercepts for All the Substitutions, Sub-Groups, and Datasets	80
Table 4.2.	ML Estimates and 95% Credible Intervals of Log-Likelihoods of the Variable Asymmetric Model for All the Substitutions, Sub-Groups and Datasets	82
Table 4.3.	Bias and MSEs for A→G and C→T Substitutions from Simulated Data.	86

LIST OF FIGURES

Figure 1.1.	A Schematic Map of the Circular Mitochondrial Genome Showing the Relative Positions of the Protein-Coding Genes, Transfer RNA Genes (tRNAs), Ribosomal RNA (rRNAs) Genes and Origins of Light-Strand (O_L) and Heavy Strand (O_H) Replication.....	10
Figure 1.2.	Replication Mechanism Diagrammed in Steps.....	11
Figure 2.1.	Posterior Probability Density Distributions of the Sixteen Substitution Probabilities for Cyt-B.....	35
2.1.A.	$T \Rightarrow A$, $A \Rightarrow T$, $A \Rightarrow C$, $C \Rightarrow A$, $G \Rightarrow A$, and $A \Rightarrow G$	35
2.1.B.	$T \Rightarrow C$, and $C \Rightarrow T$	35
2.1.C.	$G \Rightarrow T$, $C \Rightarrow G$, $G \Rightarrow C$, and $T \Rightarrow G$	35
2.1.D.	$C \Rightarrow C$, $T \Rightarrow T$, $A \Rightarrow A$, and $G \Rightarrow G$	35
Figure 2.2.	Neighbor-Joining Phylogeny with Ancestral C/T Frequency Ratios of B2, Parsimony and ML Mapped to the Internal Nodes and Observed C/T Frequency Ratios Mapped to the Tips.....	41
Figure 2.3.	Differences between B2 and ML (B2-ML) in tRNA Base-Pairing Compatibility of Predicted Ancestral Sequences (Δ Complementarity) as a Function of the Nucleotide Variability Observed at Site.....	42
Figure 3.1.	Time Spent Single-Stranded During Replication of Vertebrate Mitochondria	56
Figure 3.2.	Phylogeny of Sixteen Primate Species and Two Near Outgroups Used in This Study.....	60
Figure 3.3.	Relative Asymmetric Substitution Response Profiles Versus Time Spent Single-Stranded.....	65
Figure 3.4.	Expanded Views of Relative Asymmetric Substitution Response Profiles Versus Time Spent Single-Stranded.....	70
Figure 3.5.	Excess Purine Transition Asymmetry as a Function of Loopiness	71
Figure 4.1	Site-Specific Relative Rate Responses From HMM Analyses with Respect to Time Spent Single-Stranded	83
Figure 4.2	Site-Specific Bias and MSE Profiles Under HMM Analyses	88
Figure 4.3	Schematic Representation of Model Relationships.....	89
Figure 4.4	Distribution of Likelihood Ratio Test Statistic For $A \Rightarrow G$ and $C \Rightarrow T$ Substitutions For Simulated Data Under Two Types of Models	90

ABSTRACT

Molecular evolution recovers the history of living species by comparing genetic information, exploring genome structure and function from an evolutionary perspective. Here we infer substitution rates and ancestral reconstructions, to better understand mutation responses to some known biochemical phenomena. Mutation processes are commonly inferred using parsimony, maximum likelihood and Bayesian. Parsimony is not explicitly model-based, and is statistically biased due to unrealistic assumptions. The model-based maximum likelihood approaches become computationally inefficient while analyzing large or high-dimensional datasets, leaving little opportunities to incorporate complex evolutionary models.

We implemented a posterior probability (Bayesian) approach that evaluates evolutionary models, applying it to primate mitochondrial genomes. The species nucleotide sequence data were augmented with ancestral states at the internal nodes of the phylogeny. We simplified probability calculations for substitution events along the branches by assuming that only up to one or two substitution events occurred per branch per site. These *conditional pathway* calculations introduce very little bias into the inferred reconstructions, while increasing the feasibility of incorporating complex evolutionary models with higher dimensions. Compositional bias tests, including functional predictions of ancestral tRNAs, show that ancestral sequences from the Bayesian approach are more biologically realistic than those reconstructed by maximum likelihood.

To explore other model complexity, we allowed substitution rates to vary among sites by having a different model at each site. With a strand-symmetric model as the *base* model, asymmetric substitution probabilities for specific substitution types were varied

among sites. This model would not be feasible with standard matrix exponentiation methods, particularly maximum likelihood. We observed for A→G and C→T substitutions almost linear, respectively, almost asymptotic responses (with some regional deviations). Note that the HMM models had no *a priori* response built in them. Observed responses fitted predictions from earlier gene by gene likelihood analyses. For A→G substitutions, deviations from the expected linear response correlated positively with the loop-forming propensity of the corresponding site in the mRNA secondary structure. In the COI region, C→T substitutions have a prominent dip, suggesting protection against mutations. The C→T substitution responses differed significantly between primate sub-groups defined based on their single genome A→G responses.

CHAPTER 1: GENERAL INTRODUCTION AND BACKGROUND

1.1. Introduction

Classification of organisms should be understood not only to study the phylogenetic relationships between them but also to know how changes accumulate over time in the light of the evolutionary processes that occur (Holder and Lewis, 2003; O'Donoghue and Luthey-Schulten, 2003). Individual mutations occur in populations result in variation. This is followed by fixation of various states in different individuals causing divergence. The magnitude of accumulated changes is particularly large when the organisms are reproductively isolated from one another over a period of time (e.g. by a geographic barrier) (Espinola, 1971; Graves, 1991; Schwartz, 1999; Shine et al., 2002; Hurt and Hedrick, 2003; Lehmann et al., 2003; Linn et al., 2003; Martin and McKay, 2004; Takehana et al., 2004)

Detecting and analyzing substitution patterns in organisms with the help of their available genetic information, especially genomes, and knowing how they differ between different phylogenetic groups is central to understanding their evolutionary dynamics (Holland, 2003; Karlin et al., 2003; Sawa et al., 2003; Stone and French, 2003; Xie et al., 2003; Gabaldon and Huynen, 2004). Such comparative genome-based analyses also play a key role in the identification of highly conserved sequences that could potentially be responsible for protein function or structure (Aravind et al., 2002; Hedges, 2002; Cavalli-Sforza and Feldman, 2003; Soltis and Soltis, 2003; Nobrega and Pennacchio, 2004).

Homology between genome sequences needs to be defined before performing any valid comparative analyses (De Pinna, 1991). Sequence identity is often confused with

sequence homology. Homology refers to similarities due to inheritance from a common ancestor (Hillis, 1994). An *alignment* is a hypothesis of homology for a set of sequences. One common way of efficiently carrying out multiple sequence alignments is by making pairwise comparisons between sequences to construct a dendrogram (hierarchical clustering) and computing the final multiple alignments using this dendrogram as a guide tree (Thompson et al., 1997). Each position in this alignment is referred to as a *site*. A *phylogeny* depicts the series of events believed to have occurred during the evolution of a group of organisms where the leaf nodes or *tips* represent the contemporary species or sequences and the internal nodes represent the points where the ancestral lineages separated or began to diverge.

Given a phylogeny and a sequence alignment, it is possible to reconstruct ancestral states at the internal nodes. Ancestral reconstructions, both of nucleotide as well as amino acid sequences, are valuable in deciphering substitution patterns and deriving useful inferences such as adaptation, functional divergence and correlation of substitution events to geographic or environmental factors. It is important to consider the accuracy of these reconstructed sequences, however, before making any valid inference. Ancestral states can be reconstructed by various methods, including parsimony (Fitch, 1971; Swofford and Maddison, 1992; Maddison, 1994), maximum likelihood (ML) (Stewart et al., 1987; Malcolm et al., 1990; Messier and Stewart, 1997; Hassanin and Douzery, 1999; Hibbett and Binder, 2002; Richard et al., 2003; Soltis et al., 2003) and Bayesian (Huelsenbeck and Ronquist, 2001; Huelsenbeck et al., 2001; Bollback 2002; Douady et al., 2003).

Parsimony is not explicitly based on any model and reconstructs ancestral states by minimizing changes on the phylogeny. The idea is largely encapsulated by Ockham's

Razor, all things being equal, the simplest explanation invoking the fewest ad hoc hypotheses will tend to be the correct one. When there are two equally parsimonious explanations, parsimony chooses randomly between the two. A common problem with parsimony is that when the sequences have highly skewed compositions, the ancestral reconstructions are not accurate and are deterministically biased. The bias is such that the most frequent nucleotide or amino acid in the tip sequences becomes more frequent in the ancestral reconstructions, while the rarest one becomes rarer (Collins et al., 1994; Yang, 1996; Zhang and Nei, 1997; Eyre-Walker, 1998; Yang 1998; Sanderson et al., 2000; Conant and Lewis, 2001; Alvarez-Valin et al., 2004).

Ancestral states reconstructed by ML methods maximize the probability of the inferred ancestral state under a pre-specified stochastic model of evolution (Schluter et al., 1997; Pagel, 1999). A typical *marginal* reconstruction performed by ML is such that for each node under consideration, the reconstructed state maximizes the likelihood of the observed tip sequences while allowing the ancestral states at the other nodes to vary (Swofford, 2002). A *joint* reconstruction on the other hand, maximizes the likelihood of the observed tip sequences having reconstructed the entire pathway simultaneously, i.e. ancestral states at all internal nodes, thus evaluating likelihoods of the data given different pathway configurations.

ML methods are statistically more well-founded and perform better than the parsimony methods. Unlike parsimony, information from all the sites is used for calculating likelihood and their models take into account the timing of the substitution event. They are usually robust to sampling errors and violations of model assumptions. However, they are computationally intensive and slow while analyzing large datasets.

The amount of complexity that models explored by ML methods can carry is limited due to the method of likelihood calculations, which typically proceeds by performing a series of matrix exponentiations along each branch for the entire phylogeny. There are severe computational costs while exploring a higher-dimensional parameter space due to multiplicative exploration over each dimension. Therefore, amino-acid models and codon-substitution models are difficult to explore using ML methods.

Bayesian or posterior probability methods are also popularly used nowadays for reconstructing ancestral sequences. While Bayesian methods are also centered on likelihood evaluations, they consider the entire posterior frequency distribution for each parameter explored, including the ancestral states. They are particularly used for exploring multi-dimensional parameter spaces and capable of evaluating all the dimensions simultaneously. It is, therefore, more feasible to build in complexity into the models of evolution using these methods, where more complexity reflects more biological reality. Theoretically, there could be a large number of factors responsible for observed and predicted patterns of changes at both the evolutionary level as well as the physiological level. Modeling at least some of these predicted factors increases our chances of obtaining more accurate answers significantly and also leads to predictions of other previously unknown elements. Even with Bayesian methods, however, there is a trade-off between model complexity and speed and efficiency of the method. Therefore, considering this aspect of computational limitations, a balance must be obtained between the amount of possible realism in the evolutionary models and the CPU limits on the exploratory method.

A Bayesian approach of evaluating posterior distributions of ancestral sequences and model parameters, with computational simplifications is described here (Chapter 2; Krishnan et al., *in press*). The tip sequence data was *augmented* by mapping ancestral states at the internal nodes. Further computational simplification was achieved by restricting the number of substitution events to only one or two per site per branch, as opposed to integrating over all possible ancestral states, and calculating substitution probabilities as probabilities of specific events. Ancestral states and model parameters were explored by running a Markov chain Monte Carlo (MCMC) analyses on them. For comparisons, an approach closely resembling parsimony (B0) where the branch lengths were totally ignored was also implemented. The model parameters were the same for all the sites and each site was treated independently. In a different and more complex analysis, this assumption was relaxed by introducing context-dependence and varying the substitution rates among sites.

We applied these conditional pathway calculations to a nucleotide sequence set of primate mitochondrial genes (see Background section for more information on mitochondria) and found that restricting the substitution events to two or fewer per site per branch (B1 and B2) yielded fairly realistic and unbiased ancestral reconstructions. Surprisingly, ML reconstructions were as much compositionally biased as parsimony and sometimes, even more so. This was true particularly for the sites with a higher compositional variability. The restriction of ML's choice to an ancestral state that gives the best likelihood and not accounting for the entire frequency distribution of other possible alternatives appear to cause this dramatic reconstruction bias.

The functionality of predicted ancestral sequences was performed by comparing helix-forming propensities of reconstructed transfer-RNAs. For highly variable sites with ambiguously reconstructed ancestral sequences, B2 outperformed ML significantly. ML performed only slightly better than B2 for the remaining sites. The fact that Bayesian methods perform significantly better for highly variable sites and more ambiguously reconstructed nodes suggests that it is indeed advantageous to consider the entire posterior distribution of ancestral states as opposed to only the most likely ancestral state. The good performance of restricted pathway methods using simpler models motivated us to build more complexity into these models. This was done by varying the substitution probabilities among sites (Chapter 3; Chapter 4; Krishnan et al., 2004; Krishnan et al., *submitted to DNA and Cell Biology*). Rate heterogeneity among sites has been modeled in the past by accounting for the average rate variation between sites using gamma distributions (Yang 1994). Site-specific models, in which the substitution matrix parameters are allowed to vary among sites (and are not pre-specified), have seldom been incorporated (Koshi and Goldstein, 1995; Dimmic et al., 2001; 2002; Fornasari et al., 2002) and are difficult to incorporate using ML methods, due to computational costs. Additionally, building models at each site after accounting the effects of evolutionary rates at neighboring sites and thereby introducing context-dependence among sites using ML would cause powered matrix exponentiations, which becomes computationally infeasible.

The second important motivation for building these complex models was to profile the site-specific substitution responses of particular substitution types to time-spent single-stranded during mitochondrial replication. The asymmetric replication mechanism

leaves the heavy strand exposed for a period of time. Different genes, therefore, spend different amounts of time in the single-stranded state, according to their distance from the heavy and light strand origins of replication. Mutations result from deaminations occurring from $A \rightarrow G$ and $C \rightarrow T$ and accumulate during the time spent single-stranded. Based on a gene by gene likelihood analyses, it was previously predicted that $A \rightarrow G$ substitutions increase linearly over the time spent single-stranded, whereas $C \rightarrow T$ substitutions increase rapidly to high numbers during initial single-strandedness and then saturate for the remaining time (Faith and Pollock, 2003). To test these predictions by obtaining a more detailed response and to capture any regional deviations from the predicted response, we added to a particular substitution type, a site-specific asymmetric component that was proportional to the time-spent single stranded by that site. Ancestral reconstructions obtained using a simple general-time reversible model served as a first stage of a posterior predictive approach for quick evaluation of complex models at each site. For later analyses, ancestral reconstructions were obtained using site-specific models mainly to calculate likelihoods suitable for model comparisons.

In one case, a strand-symmetric model was used as the base model and asymmetry at each site was proposed in a specific substitution by adding to it a linear asymmetric component proportional to the time spent single stranded by that site. An MCMC chain was run over the slope and intercept parameters of the asymmetric component (Chapter 4). In the other approach, we used hidden Markov Models (HMMs) where the substitution probability at one site was correlated with that of the previous site by a 'hidden' component (Chapters 3 and 4). No *a priori* linear or asymptotic response was programmed into the models.

An almost linear response was observed for the A→G substitutions using HMMs, whereas for the C→T substitutions, a quick increase was observed followed by saturation. These results confirmed earlier predictions (Faith and Pollock, 2003). Comparisons between the responses of transversions and transitions potentially indicated mild residual effects of the high transition responses on the transversions (Chapter 3). C→T substitution responses of two primate sub-groups initially classified based only on their single genome A→G substitution responses (Raina et al., *in review*), also showed remarkable differences in various portions of the genome (Chapter 4). A full posterior predictive approach was later implemented for assessing significance of the model's fit to the data by evaluating distributions of various test-statistics. In particular, likelihood ratio test statistic distributions for A→G and C→T substitution responses based on simulated data under two different models are described in Chapter 4.

While the thesis mainly presents explorations of various nucleotide models, we intend to fully extend our analyses to amino acid sequences and to be able to build realistic amino acid and codon substitution models. Preliminary efforts on reconstructing ancestral sequences using B2 methods and a Jones-Taylor-Thornton (JTT) (Jones et al., 1992) amino acid model have been met with success. The advantage of being able to obtain distributions of evolutionary changes occurring at each site on a given branch gives us the potential to infer the complete set of substitutions leading to any particular phylogenetic group and thereby infer positive selection or adaptation. Preliminary codon substitution models built using an underlying nucleotide model, overlaid by an amino-acid model for the non-synonymous sites successfully evaluated posterior distributions of ancestral reconstructions at a speed 2.5 times slower than that for the simple nucleotide models.

For relatively more diverged phylogenetic groups than primates, restricting the number of substitutions to up to only two per branch might be a bit over-simplified. As an alternative to a theoretical “B4” method, adding additional internal nodes on longer branches allows us to infer ancestral reconstructions at those hypothetical nodes and adequate realism can be achieved. Testing the performance of these methods on simple nucleotide models and assessing *a priori* the behavior of nucleotide substitutions in particular phylogenetic groups gives us a strong foothold for stepping further to higher-dimension models and modeling protein evolution.

1.2. Background

1.2.1. Mitochondrion and Its Genome Structure in Animals

The mitochondrion is a sub-cellular organelle, presumably from endosymbiotic origin. It contains its own circular genome. It is the primary site for oxidative phosphorylation. Its gene sequences suggest close relationship with α -proteobacteria (Lang et al., 1999; Litoshenko, 2002; Tielens et al., 2002; Andersson, 2003; Burger and Lang, 2003; Burger et al., 2003; van Hellemond et al., 2003). Plant and animal mitochondria have very different characteristics, and the focus here will be mainly on animal mitochondria. Animal mitochondrial genomes vary in their sizes and forms, depending on the phylogenetic group they belong to (Nosek and Tomaska, 2003). They are typically between 14 – 42 kb long and generally consist of a variable control region and thirty-seven genes (Figure 1): 13 protein-coding (COI, COII, COIII, CytB, NDI, NDII, NDIII, NDIV, NDIVL, NDV, NDVI, ATP6 and ATP8), 22 transfer-RNAs (Ala, Arg, Asp, Asn, Cys, Gln, Gly, Glu, His, Ile, Leu2, Leu4, Lys, Met, Phe, Pro, Ser2, Ser4, Thr, Trp, Tyr, and Val) and 2 ribosomal RNAs (12S and 16S). There has been yet no evidence for recombination occurring in their DNA

(mtDNA). There are very few non-coding DNA or inter genic spacers, if at all and there are usually no introns. They frequently also follow their own genetic code. Oxygen free radicals produced by the respiratory chain in mitochondria often damage mtDNA, resulting in increased mutation rates.

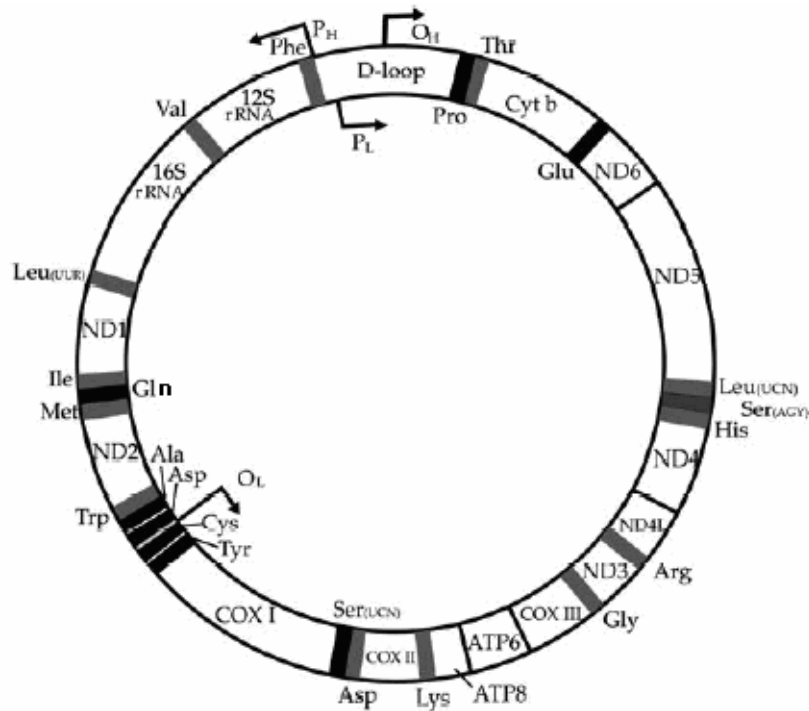


Figure 1.1: A Schematic Map of the Circular Mitochondrial Genome Showing the Relative Positions of the Protein-Coding Genes, Transfer RNA Genes (tRNAs), Ribosomal RNA (rRNAs) Genes and Origins of Light-Strand (O_L) and Heavy Strand (O_H) Replication. The outer circle represents the light strand and the inner circle represents the heavy strand. All the protein-coding genes except NDVI are coded on the light strand and all the tRNAs except tRNA-glu, tRNA-pro, tRNA-ser4, tRNA-asn, tRNA-cys, tRNA-gln, tRNA-gly, tRNA-Ala.

1.2.2 Mitochondrial Replication Mechanism

Currently, there are ~400 vertebrate mitochondrial genomes available in NCBI's GenBank. Genomic biodiversity, as previously mentioned, is a major requirement for studying molecular evolution. In mitochondria, the contiguous stretch of genes with no

intergenic spacers allows easy filtering of usable genetic information. Mitochondria also have a special replication mechanism that results in gradients and trends for particular substitution types across the genome. The asymmetric replication of vertebrate mitochondria leaves different portions of the heavy strand exposed to mutations for different periods of time. During this time spent single-stranded, mutations caused by deaminations, typically those from A→G and C→T, accumulate. Faith and Pollock (2003) predicted that A→G substitutions respond linearly to time spent single stranded whereas C→T substitutions response increased rapidly at first, followed by saturation for the remaining period of time spent single-stranded, suggesting some form of repair. It was also predicted that substitutions other than A→G and C→T respond differently to single-strandedness. Accounting for these genome gradients in the models of evolution significantly increases the accuracy of ancestral state reconstructions and other functional inferences based on these reconstructions.

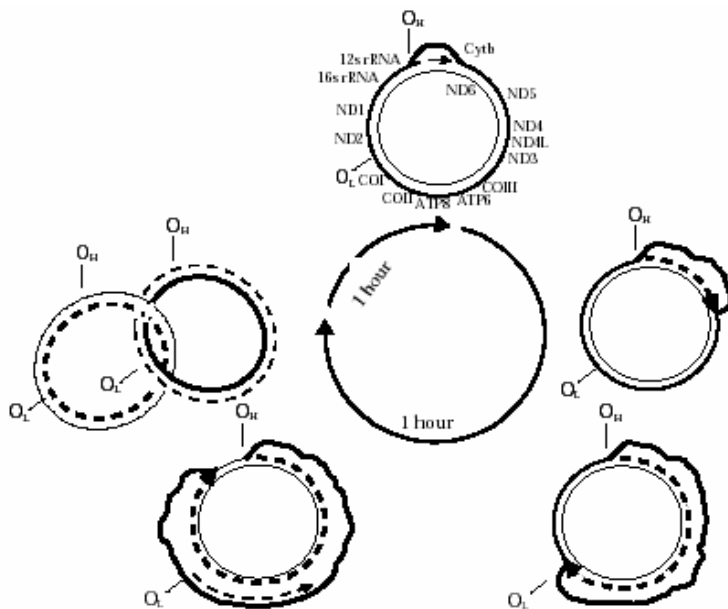


Figure 1.2: Replication Mechanism Diagrammed in Steps (Faith and Pollock, 2003).

1.3. Literature Cited

- ALVAREZ-VALIN, F., CLAY, O., CRUVEILLER, S., BERNARDI, G. (2004). Inaccurate reconstruction of ancestral GC levels creates a "vanishing isochores" effect. *Mol. Phylogenet. Evol.* 31: 788-93.
- ANDERSSON, S.G., KARLBERG, O., CANBACK, B., KURLAND, C.G. (2003). On the origin of mitochondria: a genomics perspective. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358:165-77; discussion 177-9.
- ARAVIND, L., MAZUMDER, R., VASUDEVAN, S., KOONIN, E.V. (2002). Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.* 12:392-9.
- BOLLBACK, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol* 19:1171-1180.
- BURGER, G., GRAY, M.W., LANG, B.F. (2003). Mitochondrial genomes: anything goes. *Trends Genet.* 19:709-716.
- BURGER, G., LANG, B.F. (2003). Parallels in genome evolution in mitochondria and bacterial symbionts. *IUBMB Life.* 55:205-12.
- CAVALLI-SFORZA, L.L., FELDMAN, M.W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nat Genet.* 33 Suppl:266-75.
- COLLINS, T. M., WIMBERGER, P. H., NAYLOR, G. J. P. (1994). Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. *Syst. Biol.* 43:482-496.
- CONANT, G.C., LEWIS, P.O. (2001). Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. *Mol Biol Evol.* 18:1024-33.
- DE PINNA, M.C.C. (1991). Concepts and tests of homology in the cladistic paradigm. *Cladistics.* 7: 367-394.
- DIMMIC, M.W., MINDELL, D.P., GOLDSTEIN, R.A. (2000) Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pac. Symp. Biocomp.*
- DIMMIC, M. W., REST, J. S., MINDELL, D. P., GOLDSTEIN, R. A. (2002). rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J. Mol. Evol.* 55:65-73

- DOUADY, C. J., DELSUC, F., BOUCHER, Y., DOOLITTLE, W. F., DOUZERY, E. J. (2003). Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* 20:248-254.
- ESPINOLA, H.N. (1971). Reproductive isolation between *Triatoma brasiliensis* Neiva, 1911 and *Triatoma petrochii* Pinto & Barreto, 1925 (Hemiptera Reduviidae). *Rev Bras Biol.* 31:277-81.
- EYRE-WALKER, A. (1998). Problems with parsimony in sequences of biased base composition. *J. Mol. Evol.* 47:686-90.
- FAITH, J. J., AND POLLOCK, D. D. (2003). Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* 165:735-745.
- FITCH, W. M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* 20:406-416.
- FORNASARI, M.S., PARISI, G., ECHAVE, J. (2002). Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Mol Biol Evol.* 19:352-356.
- GABALDON, T., HUYNEN, M.A. (2004). Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci.* 61:930-44.
- GELMAN, A. RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7: 457-72.
- GRAVES, G.R. (1991). Bergmann's rule near the equator: latitudinal clines in body size of an Andean passerine bird. *Proc Natl Acad Sci U S A.* 88:2322-5.
- GRAY, M.W. (2003). Diversity and evolution of mitochondrial RNA editing systems. *IUBMB Life.* 55:227-33.
- HASSANIN, A., DOUZERY, E. J. P. (1999). Evolutionary affinities of the enigmatic saola (*Pseudoryx nghetinhensis*) in the context of the molecular phylogeny of Bovidae. *Proc Royal Soc London B-Biol. Sci.* 266:893-900.
- HEDGES SB. (2002). The origin and evolution of model organisms. *Nat. Rev. Genet.* 3:838-49.
- HIBBETT, D. S., BINDER, M. (2002). Evolution of complex fruiting-body morphologies in homobasidiomycetes. *Proc. R. Soc. Lond. B. Biol. Sci.* 269:1963-1969.

- HILLIS, D.M., HUELSENBECK, J.P. (1992). Signal, noise, and reliability in molecular phylogenetic analyses. *J. Hered.* 83:189-95.
- HILLIS, D.M. (1994). Homology in molecular biology. In: *Homology: the hierarchical basis of comparative biology*, ed: Hall, B.K. Academic press, San Diego.
- HOLDER M, LEWIS PO. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet.* 4:275-84.
- HOLLAND, P.W. (2003). More genes in vertebrates? *J. Struct. Funct. Genomics.* 3:75-84.
- HUELSENBECK, J. P., RONQUIST, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.
- HUELSENBECK, J. P., RONQUIST, F., NIELSEN, R., BOLLBACK, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310-2314.
- HURT, C.R., HEDRICK, P.W. (2003). Initial stages of reproductive isolation in two species of the endangered Sonoran topminnow. *Int. J. Org. Evolution.* 57:2835-41.
- JONES, D.T., TAYLOR, W.R., THORNTON, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8:275-82.
- JONSSON, H., SODERBERG, B. (2003). An approximate maximum likelihood approach, applied to phylogenetic trees. *J. Comput. Biol.* 2003;10(5):737-49.
- KARLIN, S., MRAZEK, J., GENTLES, A.J. (2003). Genome comparisons and analysis. *Curr. Opin. Struct. Biol.* 13:344-52.
- KOSHI, J. M., GOLDSTEIN, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *J Mol Evol* 42:313-320.
- KRISHNAN, N. M., SELIGMANN, H., STEWART, C. B., DE KONING, A.P.J., POLLOCK, D. D. Ancestral sequence reconstruction in primates: Compositional Bias and effect on functional inference. *Molecular Biology and Evolution*, in press.
- KRISHNAN, N. M., SELIGMANN, H., RAINA, S. Z., AND POLLOCK, D. D. (2004). Phylogenetic analysis of site-specific perturbations in asymmetric mutation gradients. *Currents in Computational Molecular Biology*. A. Gramada, and P.E. Bourne eds. Pp. 266-267.

- KRISHNAN, N. M., SELIGMANN, H., RAINA, S. Z., POLLOCK, D. D. Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. *DNA and Cell Biology (in review)*
- LANG, B.F., GRAY, M.W., BURGER, G. (1999). Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.*33:351-397.
- LAURIE, D.A., DEVOS, K.M. (2002). Trends in comparative genetics and their potential impacts on wheat and barley research. *Plant Mol. Biol.* 48:729-40.
- LEHMANN, T., LICHT, M., ELISSA, N., MAEGA, B.T., CHIMUMBWA, J.M., WATSENGA, F.T., WONDJI, C.S., SIMARD, F., HAWLEY, W.A. (2003). Population Structure of *Anopheles gambiae* in Africa. *J. Hered.* 94:133-47.
- LINN, C. JR, FEDER, J.L., NOJIMA, S., DAMBROSKI, H.R., BERLOCHER, S.H., ROELOFS, W. (2003). Fruit odor discrimination and sympatric host race formation in *Rhagoletis*. *Proc. Natl. Acad. Sci. U S A.* 100:11490-3.
- LITOSHENKO, A.I.A. (2002). Evolution of mitochondria. *Tsitol. Genet.* 36:49-57.
- MADDISON, D. R. (1994). Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters. *Ann. Rev. Entomol.*39:267–292.
- MALCOLM, B. A., WILSON, K. P. , MATTHEWS, B. W. , KIRSCH, J. F. , WILSON, A. C. (1990). Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* 345:86-89.
- MARTIN, P.R., MCKAY, J.K. (2004). Latitudinal variation in genetic divergence of populations and the potential for future speciation. *Evolution Int J Org Evolution.* 58:938-45.
- MESSIER, W., STEWART, C. B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature* 385:151-154.
- MURPHY, S.K., JIRTLE, R.L. (2003). Imprinting evolution and the price of silence. *Bioessays.* 25:577-88.
- NOBREGA, M.A., PENNACCHIO, L.A. (2004). Comparative genomic analysis as a tool for biological discovery. *J. Physiol.* 554:31-39.
- NOSEK, J., TOMASKA, L. (2003). Mitochondrial genome diversity: evolution of the molecular architecture and replication strategy. *Curr. Genet.* 44:73-84.

- O'DONOGHUE, P., LUTHEY-SCHULTEN, Z. (2003). On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiol. Mol. Biol. Rev.* 67:550-573.
- RAINA, S., FAITH, J.J., DISOTELL, T., STEWART, C.B., SELIGMANN, H., POLLOCK, D. D. (2004) Evolution of base substitution gradients in primate mitochondrial genomes. *in review*
- RICHARD, F., LOMBARD, M., DUTRILLAUX, B. (2003). Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res* 11:605-618.
- SANDERSON, M. J., WOJCIECHOWSKI, M. F., HU, J. M., KHAN, T. S., BRADY, S. G. (2000). Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Mol. Biol. Evol.* 17:782-797.
- SANDERSON, M. J., SHAFFER, H. B. (2002). Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics.* 33: 49-72
- SAWA, G., DICKS, J., ROBERTS, I. N. (2003). Current approaches to whole genome phylogenetic analysis. *Brief Bioinform.* 4: 63-74.
- SCHWARTZ, J.H. (1999). Homeobox genes, fossils, and the origin of species. *Anat Rec.* 257:15-31.
- SHINE, R., REED, R.N., SHETTY, S., LEMASTER, M., MASON, R.T. (2002). Reproductive isolating mechanisms between two sympatric sibling species of sea snakes. *Evolution Int J Org Evolution.* 56:1655-62.
- SOLTIS, D. E., SENTERS, A. E., ZANIS, M. J., KIM, S., THOMPSON, J.D., SOLTIS, P.S., DE CRAENE, L. P. R. , ENDRESS, P. K., FARRIS, J. S. (2003). Gunnerales are sister to other core eudicots: Implications for the evolution of pentamery. *Amer. J. Bot.* 90:461-470.
- SOLTIS, D.E., SOLTIS, P.S. (2003). The role of phylogenetics in comparative genetics. *Plant Physiol.* 132:1790-800.
- STEWART, C. B., SCHILLING, J. W., WILSON, A. C. (1987). Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* 330:401-404.
- STONE, G., FRENCH, V. (2003). Evolution: have wings come, gone and come again? *Curr. Biol.* 13:R436-8.
- SWOFFORD, D.L, MADDISON, W.P. (1992). Parsimony, character-state reconstructions, and evolutionary inferences. Pages 186–223 *in* Systematics, historical ecology, and North American freshwater fishes(R. C. Mayden, ed.). Stanford Univ. Press, Palo Alto, California.

- SWOFFORD, D.L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0. Sinauer Associates, Sunderland, Massachusetts.
- TAKEHANA, Y., UCHIYAMA, S., MATSUDA, M., JEON, S.R., SAKAIZUMI, M. (2004). Geographic variation and diversity of the cytochrome b gene in wild populations of medaka (*Oryzias latipes*) from Korea and China. *Zoolog. Sci.* 21:483-91.
- THOMPSON, J.D., GIBSON, T.J., PLEWNIAK, F., JEANMOUGIN, F., HIGGINS, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25:4876-82.
- TIELENS, A.G., ROTTE, C., VAN HELLEMOND, J.J., MARTIN, W. (2002). Mitochondria as we don't know them. *Trends Biochem Sci.* 27:564-72.
- XIE, G., KEYHANI, N.O., BONNER, C.A., JENSEN, R.A. (2003). Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiol. Mol. Biol. Rev.* 67:303-42
- VAN HELLEMOND, J.J., VAN DER KLEI, A., VAN WEELDEN, S.W., TIELENS, A.G. (2003). Biochemical and evolutionary aspects of anaerobically functioning mitochondria. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358:205-13; discussion 213-5.
- WEISS, G., VON HAESELER, A. (2003). Testing substitution models within a phylogenetic tree. *Mol Biol Evol.* 20:572-8.
- YANG, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39: 306-314.
- YANG, Z. (1998). On the best evolutionary rate for phylogenetic analysis. *Syst Biol.* 1998 Mar;47(1):125-33.
- YANG, Z. (1996). Phylogenetic analysis using parsimony and likelihood methods. *J Mol Evol.* 42:294-307.
- ZHANG, J., NEI, M. (1997). Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol* 44:S139-146.

CHAPTER 2: ANCESTRAL SEQUENCE RECONSTRUCTION IN PRIMATE MITOCHONDRIAL DNA: COMPOSITIONAL BIAS AND EFFECT ON FUNCTIONAL INFERENCE

Reconstruction of ancestral DNA and amino acid sequences is an important means of inferring information about past evolutionary events. Such reconstructions suggest changes in molecular function and evolutionary processes over the course of evolution, and are used to infer adaptation and convergence. Maximum likelihood (ML) is generally thought to provide relatively accurate reconstructed sequences compared to parsimony, but both methods can be used to infer multiple directional changes in ancestral primate mitochondrial DNA (mtDNA) nucleotide frequencies. To better understand this surprising result, as well as to better understand how parsimony and ML differ, we constructed a series of computationally simple restricted likelihood methods that result in calculations intermediate between those of parsimony and likelihood. We also evaluated the entire Bayesian posterior frequency distribution of reconstructed ancestral states. The restricted likelihood methods differed in the number of substitutions allowed along each branch at each site, and in whether branch lengths were considered. We analyzed primate mitochondrial cytochrome b (Cyt-b) and cytochrome oxidase subunit I (COI) genes and found that ML reconstructs ancestral frequencies that are often more different from tip sequences than are parsimony reconstructions. The differences between ancestral and tip frequencies are greater at the more rapidly evolving 3rd codon positions and for genes with less even frequency distributions. In contrast, frequency reconstructions based on the posterior ensemble more closely resemble extant nucleotide frequencies. Simulations indicate that these differences in ancestral sequence inference are probably due to deterministic biases in ML and parsimony reconstructions; Bayesian posterior frequency distributions yield much less biased estimates of ancestral nucleotide frequencies. The methods involving simpler restricted

likelihood calculations have slightly reduced likelihood values compared to full likelihood calculations, but can provide fairly unbiased nucleotide reconstructions and may be useful in more complex phylogenetic analyses than considered here due to their speed and flexibility. To determine whether biased reconstructions might affect inferences of functional properties, ancestral primate mitochondrial tRNA sequences were inferred and helix-forming propensities for conserved pairs were evaluated *in silico*. For ambiguously reconstructed nodes at sites with high variability, ancestral tRNA sequences from Bayesian analyses were more compatible with canonical base pairing than were those inferred by other methods. Thus, nucleotide bias in reconstructed sequences apparently can lead to serious bias and inaccuracies in functional predictions.

2.1.Introduction

Reconstructions of ancestral nucleotide and amino acid sequences are useful in many forms of comparative biology (Karlin, Mocarski, and Schachtel 1994; Maddison and Maddison 2000; Zhang et al. 2003). Accurate reconstruction of ancestral sequences enables us to infer evolutionary pathways, study adaptation, behavioral changes and functional divergences, and correlate site-specific changes with geography or known paleontological events (Bleiweiss 1998; Giannasi, Thorpe, and Malhotra 2000; Beardsley, Yen, and Olmstead 2003). Reconstructions are also at the core of experimental paleo-molecular biochemistry, a pursuit in which sequences of extant taxa are used to predict and resurrect the sequences and functions of ancestral macromolecules (Pauling and Zuckerkandl 1963; Krawczak, Wacey, and Cooper 1996; Benner 2002; Zhang and Rosenberg 2002; Gaucher et al. 2003).

Parsimony and maximum likelihood (ML) methods of reconstruction have been used extensively in various ancestral sequence analyses (Stewart, Schilling, and Wilson 1987; Malcolm et al. 1990; Messier and Stewart 1997; Hassanin and Douzery 1999; Hibbett and

Binder 2002; Richard, Lombard, and Dutrillaux 2003; Soltis et al. 2003), and can sometimes be reliable. For example, ancestral reconstructions using parsimony were 98% accurate in predicting ancestral sequences from experimental phylogenies created by serial propagation of bacteriophage T7 in the presence of a mutagen (Hillis et al. 1992; Bull et al. 1993). Ancestral reconstruction of sequences using parsimony is, however, known to be biased for skewed base compositions (Collins, Wimberger, and Naylor 1994; Zhang and Nei 1997; Eyre-Walker 1998; Sanderson et al. 2000). The bias in parsimony-reconstructed ancestral sequences deterministically decreases the frequency of the rare base and increases that of the most common base.

Although it has been generally assumed that ML sequence reconstruction does not suffer from the same problems (Collins, Wimberger, and Naylor 1994; Zhang and Nei 1997; Eyre-Walker 1998; Sanderson et al. 2000), both ML and parsimony can sometimes fail when reconstructing quantitative traits (Schluter et al. 1997; Hormiga, Scharff, and Coddington 2000; Oakley and Cunningham 2000; Webster and Purvis 2002). ML reconstructions of continuous ancestral traits can be particularly uncertain for traits with frequent changes (Schluter et al. 1997; Cunningham, Omland, and Oakley 1998), but continuous trait reconstruction is arguably hindered much more by modeling inadequacies than problems with inference techniques. Even with discrete traits, however, ML reconstruction has limitations: in a recent “experimental phylogenetics” analysis using PCR-generated mutations, comparisons between known ancestral sequences and those reconstructed using ML showed that while most ancestral sequences were accurately reconstructed, errors increased with the depth of the sequence in the tree (Sanson et al. 2002). Although the models used are still imperfect (Yang, Kumar, and Nei 1995; Koshi and Goldstein 1996), and reconstruction is clearly not error-free, ML is more commonly used in

ancestral reconstruction, mostly due to the large biases of parsimony.

For phylogenetic analyses, ML is generally preferred over parsimony and distance methods due to its greater accuracy and incorporation of more realistic models of evolution (Huelsenbeck 1995; Yang 1996a; Yang 1996b; Huelsenbeck and Rannala 1997; Pollock and Bruno 2000). This is especially true for highly divergent sequences, such as vertebrate mtDNAs. Posterior probability (Bayesian) methods using Markov chain Monte Carlo (MCMC) simulations have, however, recently gained considerable attention in phylogenetic analysis, since they are computationally more efficient and faster than ML methods particularly for analyzing more complex evolutionary models and larger datasets (Huelsenbeck and Ronquist 2001; Huelsenbeck et al. 2001; Bollback 2002; Douady et al. 2003). They also allow nuisance parameter integration, generation of credibility intervals, and analysis of parameter distributions, rather than only the most likely parameters (Antezana 2003). Posterior probability methods are therefore a potentially useful alternative to parsimony and ML methods for reconstructing ancestral sequences (Koshi and Goldstein 1996; Nielsen 2002; Huelsenbeck, Nielsen, and Bollback 2003).

Statistical biases may exist even in Bayesian methods, and the behavior of Bayesian methods in ancestral sequence reconstruction (Koshi and Goldstein 1996) is not presently well known. We have implemented a modification of Nielsen's Bayesian approach (Nielsen 2002; Nielsen and Huelsenbeck 2002) whereby internal states are mapped onto the phylogeny as augmented data during the course of the Markov chain, and we use this method here to address the differences between the Bayesian and ML approaches to ancestral reconstruction. We consider a simplification of this approach in which internal states are mapped only to internal nodes (not within branches), and the number of substitutions between nodes is limited to one or two per branch per site. Although this simplification is unlikely to be formally correct (that is, more than

two substitutions will almost certainly occasionally occur at a single site on a single branch during the course of evolution), it is likely to be a good approximation under many circumstances and may not affect results dramatically in any case, since the probability of substitution between any two states with only two substitutions separating them may not be much different than the probability given many more intervening substitutions. For comparison, we also implement an extremely simple approach that is independent of branch length.

Although Bayesian methodologies are relatively efficient in phylogenetics, they can still become slow when the complexity of the model increases (Huelsenbeck and Ronquist 2001), so considering this aspect of computational limitations is important. The potential benefits of our implementation include increased computational speed and a dramatic increase in the feasibility of incorporating more complex models of evolution than are currently feasible, particularly those in which instantaneous rates matrices vary among gene positions, over time, or with changing sequence context. Computational costs for standard matrix exponentiation methods will increase linearly with the number of matrices, and will increase with the square of the number of states in the matrices, whereas the methods described here will not. In our experience, it is also much easier to program new models with the methods described here, and there is no need to incorporate complicated memorization schemes to save computational time. For example, in work to be described elsewhere we have implemented a model in which the instantaneous rates matrices is different at every site over the entire length of the mitochondrial genome (Krishnan et al., *in review*). Our purpose here, however, is not to demonstrate the implementation of such complex models, but to demonstrate the accuracy of this restricted likelihood implementation by comparing it to full likelihood calculations as implemented by standard programs. The series of computationally simple restricted likelihood methods that we implemented also result in a series

of calculations intermediate between those of parsimony and full likelihood calculations, and this helps to clarify the reasons and conditions under which bias in ancestral reconstruction may occur.

We tested our program by analyzing its ability to infer ancestral sequence distributions from primate mtDNA sequences and from simulated data. Parsimony was recently used to study evolutionary changes of nucleotide composition in primate mtDNA genomes (Schmitz, Ohme, and Zischler 2002), and it was suggested that nucleotide frequencies had changed from their ancestral states. Further preliminary analysis with ML in addition to parsimony supported this result, but also indicated that nucleotide frequencies had changed many times along the primate lineage, always in the same direction. We present evidence that these results may have been strongly influenced by bias in the analytical methods. In an analysis of the cytochrome b (Cyt-b) and cytochrome oxidase I (COI) gene sequences from selected primates, we find that frequencies estimated from the posterior distribution are dramatically more similar to extant sequences than either parsimony or ML, and surprisingly, that ML reconstructions are sometimes less similar than parsimony reconstructions. We simulated primate mtDNA evolution under plausible conditions of stationary and changing mutation processes, and found that considering the entire posterior distribution produced more accurate reconstructions with methods involving very simple calculations; the simplifying assumption of one or two substitutions per site per branch introduces very little bias. While there was little difference between the ML and Bayesian approaches to estimating parameters of the substitution model, the ML approach of estimating a specific ancestral sequence was considerably worse than the Bayesian approach of considering the entire posterior frequency distribution. Using the predicted folding of tRNAs into cloverleaf structures, we also considered the strong possibility that bias in reconstructed sequences can

affect functional inferences, a potentially important consideration for paleo-molecular biochemistry.

2.2 Materials and Methods

2.2.1. Genome Sequences and Phylogeny

Thirteen complete primate mitochondrial genomes were available from GenBank when this study was initiated: *Cebus albifrons* (NC_002763, Arnason et al. 2000), *Gorilla gorilla* (NC_001645, Horai et al. 1995); *Homo sapiens* (NC_001807, Ingman et al. 2000); *Hylobates lar* (NC_002082, Arnason, Gullberg, and Xu 1996); *Lemur catta* (NC_004025, Arnason et al. 2002); *Macaca sylvanus* (NC_002764, Arnason et al. 2000); *Nycticebus coucang* (NC_002765, Arnason et al. 2000); *Pan paniscus* (NC_001644, Horai et al. 1995); *Pan troglodytes* (NC_001643, Horai et al. 1995); *Papio hamadryas* (NC_001992, Arnason, Gullberg, and Janke 1998); *Pongo pygmaeus pygmaeus* (NC_001646, Horai et al. 1995); *Pongo pygmaeus abelii* (NC_002083, Xu and Arnason 1996); and *Tarsius bancanus* (NC_002811, Schmitz, Ohme, and Zischler 2002). Three other primate genomes, *Cercopithecus aethiops*, *Colobus guereza* and *Trachypithecus obscurus* came from colleagues (Raaum et al., submitted) and two non-primate outgroups from GenBank were used, *Tupaia belangeri* (NC_002521, Schmitz, Ohme, and Zischler 2000) and *Cynocephalus variegatus* (NC_004031, Arnason and Janke 2002). Alignments of all tRNAs, rRNAs, and protein-coding genes were created using ClustalW (Thompson, Higgins, and Gibson 1994), concatenated using in-house PERL scripts, and a neighbor-joining tree was determined with the BioNJ algorithm using ML distances based on the general time reversible (GTR) model in PAUP* 4.0 (Swofford 2000). This phylogeny conforms to most expectations for primate phylogeny (Goodman et al. 1998), with the exception of the placements of *Tupaia* and *Tarsius* (Schmitz, Ohme, and Zischler 2000). Since this tree has a greater likelihood than the “true” primate species tree according to both DNA and amino acid complete mitochondrial data, it was

deemed approximately correct and thus used in all further analyses presented in this paper.

Optimization of branch lengths on this topology under the ML criterion in PAUP* (using the lscores command) did not produce substantially different branch lengths or ancestral reconstructions. Questions regarding the reasons for topological inaccuracies of mtDNA-based phylogenies are complex, involving gradients of different mutation types along the genome (Faith and Pollock 2003) and will be dealt with in detail for primates in a subsequent manuscript. Ancestral sequence reconstructions were carried out using the Cyt-b and COI alignments. These genes were chosen for our analysis because they are positioned at the two extremes of a linear G/A gradient on the heavy strand of the mtDNA genome, which increases with the time spent single-stranded during replication (Faith and Pollock 2003). They therefore have the most distinctly different nucleotide frequencies possible in this dataset.

2.2.2. Likelihood Calculations

Classical phylogenetic likelihood methods integrate over all possible ancestral states and all possible branch-specific substitution histories, which requires matrix multiplications and decompositions into Eigenvalues and Eigenvectors. We avoid this here by augmenting the sequence data with mapped ancestral states and calculating probabilities of occurrences of specific events, which simplifies calculations and avoids the matrix multiplication calculations along each branch required for matrix exponentiation methods. The states at internal nodes are treated as hyperparameters and updated over the course of the Markov chain. The probability of a substitution event occurring at time t and not before is given (Rice 1995) as

$$P(E | t) = \lambda e^{-\lambda t} \quad (1)$$

where λ is the rate at which the event (or set of events) occurs, and the probability that no substitution events occur until t is $e^{-\lambda t}$. If we consider two nodes in a tree with states x and z and

separated by a branch of length t_b , and we assume that a single event occurred at time t_1 , with no events occurring over time $t_2 = t_b - t_1$, then the probability of this transition is given as:

$$P(E | t_1, t_2, x, z) = \lambda_{xz} e^{-\Lambda_x t_1} e^{-\Lambda_z t_2} \quad (2)$$

where λ_{xz} is the transition rate from state x to state z , based on the current values of the model parameters, and $\Lambda_j = \sum_{k \neq j} \lambda_{jk}$.

Since there is almost no information concerning the timing of the event, we integrate this probability over all possible times such that

$$P(E | t_b, x, z) = \lambda_{xz} \int_0^{t_b} e^{-\Lambda_x t_1} e^{-\Lambda_z (t_b - t_1)} \partial t_1 = \lambda_{xz} \frac{e^{-\Lambda_x t_b} - e^{-\Lambda_z t_b}}{\Lambda_z - \Lambda_x}. \quad (3)$$

This calculation will be referred to as the B1 method, since there is one substitution per branch.

A similar equation was recently independently derived by D. Robinson and colleagues (J.

Thorne, personal communication) and a method based on equation (2) was used in a different scenario where they relax the assumption of independence among sites and map all substitution events along branches (Robinson et al. 2003)

If we assume that two substitutions occurred between the nodes rather than one, such that state x changes to state y changes to state z , then similar calculations and integrations can be made to obtain

$$P_E(E | t_b, x, y, z) = -\lambda_{xy} \lambda_{yz} e^{-t_b (\Lambda_x + \Lambda_y + \Lambda_z)} * \frac{(\Lambda_y - \Lambda_x) e^{t_b (\Lambda_y + \Lambda_z)} + (\Lambda_y - \Lambda_z) e^{2t_b \Lambda_y} + (\Lambda_x + \Lambda_z - 2\Lambda_y) e^{t_b (\Lambda_y + \Lambda_x)}}{-\left(\Lambda_y - \Lambda_x\right)\left(\Lambda_y - \Lambda_z\right)\left(\Lambda_x + \Lambda_z - 2\Lambda_y\right)}. \quad (4)$$

This will be referred to as the B2 method. The above calculation was summed over all possible states of y to obtain $P(E | t_b, x, z)$ for the B2 method. Further calculations could be made for more than two substitutions between nodes in some cases (J. Thorne, personal communication),

but the calculations become excessive, as suggested by the difference in complexity between Equations 3 and 4, and there is no simple formula for the general case. Alternatively, extra nodes could be inserted between branch points for particularly long branches, where the states at these extra nodes would be treated as a part of the augmented data; this is simpler to program, if not faster to calculate, than a theoretical “B4” method. We do not consider these alternatives here, but rather focus on whether these simplified calculations can be used effectively in some cases to speed computation without great loss in accuracy. In addition, we consider a method (BL-) in which the probability of substitution is independent of branch length, such that

$$P(E | t_b, x, z) = \lambda_{xz} \quad (5)$$

The cumulative probability for all events, D , along a branch, b , is

$$P(b | D) = \prod_x \prod_z C_{xz}^b \sum_y P(E | t_b, x, y, z) \quad (6)$$

where C_{xz}^b is the counted number of changes from state x to state z along branch b over the entire augmented dataset. For B2, during the summation over internal states, y , if $x = y$ or $y = z$ then Equation 3 is used, and if $x = y = z$ the calculation made is the probability that no substitutions occurred. For B1 and BL-, the summation over y is irrelevant, and for BL- the t_b are ignored. Calculations are generally made as sums of log likelihoods of each internal event for computational accuracy, and the log likelihood over the entire tree is the sum of log likelihoods for each branch in the tree.

2.2.3. Running the Markov Chain

Markov chains were run using Monte Carlo techniques that included a mixture of the Metropolis Hastings algorithm (Metropolis et al. 1953; Hastings 1970) and Gibbs sampling (Geman and Geman 1984; Gelfand and Smith 1990). Parameters and augmented data states were initialized in a sequence of approximations similar to the steps in an expectation maximization

(EM) algorithm (Little and Rubin 1983; Meng and Rubin 1991). A first set of states at internal nodes was obtained by moving from the tips of the tree upwards, randomly choosing a state for each internal node from the states of the two immediate descendant nodes. The model parameters were then initialized by summing substitutions over the entire tree based on the initial augmented internal states and calculating the frequencies of these substitutions as proportions of the counts for each nucleotide,

$$\lambda_{xz}^0 = C_{xz}^0 / C_X^0 \quad (7)$$

where $C_x = \sum_y C_{xy}$. For the simplest method (Equation 5), these initial estimates are close to the final ML value under a non-reversible model. For most of our calculations we utilized a general time reversible (GTR) model in which the rates are constrained such that $\lambda_{xz} = \alpha_{xz} \pi_z$, where the rate parameters $\alpha_{xz} = \alpha_{zx}$, and π_x is the equilibrium frequency of state x . In this case, the total forward and backward substitutions are averaged to obtain initial estimates of the rate parameters,

$$\alpha_{xz}^0 = (\lambda_{xz}^0 / \pi_z + \lambda_{zx}^0 / \pi_x) / 2 \quad (8)$$

where the π_i values are estimated independently as $\pi_i^0 = C_i^0 / \sum_y C_y^0$.

After initialization of the model parameters and augmented states, a Markov chain was run in which the internal states or rate parameters were updated with equal probability at each step. The full rate matrix was updated using the Metropolis-Hastings algorithm, in which each set of parameters in the chain, θ_t at step t , depended only on the parameters in the previous step, θ_{t-1} . The parameter values for a new step were proposed based on a proposal density, $q(\theta' | \theta_{t-1})$, and this proposal was accepted or rejected based on the Metropolis-Hastings acceptance function,

$$\theta_t = \begin{cases} \theta' & \text{if } (a \geq 1 \text{ or } a > \text{rand}(0,1)) \\ \theta_{t-1} & \text{otherwise} \end{cases}, \quad (9)$$

where

$$a = \frac{L(D | \theta')P(\theta')q(\theta_{t-1} | \theta')}{L(D | \theta_{t-1})P(\theta_{t-1})q(\theta' | \theta_{t-1})}. \quad (10)$$

In the Markov chains run for this study, the parameter priors, $P(\theta)$, were uniform such that $P(\theta') = P(\theta_{t-1})$ for all θ , and the proposals were symmetric such that $q(\theta' | \theta_{t-1}) = q(\theta_{t-1} | \theta')$ for all θ ; the acceptance probability therefore reduced to the likelihood ratio. The chains work best if the proposal densities match the shape of the target distribution, $P(x)$, but this density is unknown. Here, the proposed changes for the rate parameters followed a normal distribution with variance determined by the acceptance probabilities, and were thus symmetric and not biased towards any parameter values. Proposals of values out of range (e.g., rates less than zero) were reflected about the range boundary. If proposal steps are too small, the chain will mix slowly, i.e., it will move around the space slowly and converge slowly to $P(x)$. If the proposal steps are too large the acceptance rate will be low because the proposals are likely to land in regions of much lower probability density. Since the appropriate size of the proposal step depends on the dataset being used, short simulations with 50 different window range values were run for 200 iterations prior to starting each chain to determine appropriate parameter proposal window sizes. The window sizes for proposals were fixed at values for which 60-80% of proposals from the initial point were accepted. A full matrix update was proposed for the rate matrix in an MCMC generation, and accepted according to the Metropolis-Hastings criterion. States at internal nodes were updated using a Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990; Wang, Rutledge, and Gianola 1994; Liu, Neuwald, and Lawrence 1995; Firat,

Theobald, and Thompson 1997). An initial internal node was picked randomly and a new state was calculated from the probability density of substituting to or from the states at the three surrounding nodes. The remaining internal nodes were then updated in a similar fashion, moving outward from the initially chosen node. Since each new state was sampled from the conditional posterior density, the randomly sampled state was always accepted.

2.2.4. Chain Convergence Diagnostics

After initialization, the Markov chain was run for two thousand iterations until equilibrium, at which point the initial values no longer affect the current values of the model parameters. These “burn-in” samples prior to chain convergence were discarded and excluded from analyses. Chain convergence was confirmed for likelihood and all substitution parameters. To determine whether the chains indeed converged to a stationary distribution, we ran three parallel chains with over-dispersed starting values for the transition matrix. Convergence was confirmed (Gelman et al. 1992; Gelman and Rubin 1996) when the within-chain variance (W_T) was equal to the estimated asymptotic variance ($\hat{\sigma}_T^2$). If T is the number of points generated in a chain and N is the total number of chains, then the among-chain variance is

$$B_T = \frac{1}{N} \sum_{k=1}^N (\bar{\delta}_k - \bar{\delta})^2 \quad (12)$$

and the within-chain variance is

$$W_T = \frac{1}{N} \sum_{k=1}^N s_k^2 = \frac{1}{N} \sum_{k=1}^N \frac{1}{T} \sum_{t=1}^T (\delta_k^{(t)} - \bar{\delta}_k)^2 \quad (13)$$

where

$$\bar{\delta}_k = \frac{1}{T} \sum_{t=1}^T \delta_k^{(t)} \quad \text{and} \quad \bar{\delta} = \frac{1}{N} \sum_{k=1}^N \bar{\delta}_k, \quad (14)$$

and the estimated asymptotic variance is

$$\hat{\sigma}_T^2 = \left(\frac{T-1}{T} W_T \right) + \left(\frac{B_r}{T} \right). \quad (15)$$

For BL-, sampling continued for 25,000 generations, while for B1 and B2 it continued for 50,000 generations. Nucleotide frequencies and nucleotide ratios were calculated at each internal node and averaged across all sampled points. The effective sample size (N_{Eff}) was calculated as

$$N_{Eff} = (N - B) \left(\frac{1 - r_1}{1 + r_1} \right), \quad (16)$$

where N is the total sample size, B is the size of the sample removed for burn-in, and r_1 is a lag one autocorrelation function such that

$$r_1 = \frac{\sum_{i=B}^{i < N} (D_i - \mu)(D_{i+1} - \mu)}{\sum_{i=B}^{i \leq N} (D_i - \mu)^2}, \quad (17)$$

where D_i is the i^{th} sampled data point, and μ is the sample mean with burn-in excluded. To determine a sampling frequency that represented a good trade-off between independence of points and the length of the chain, a test chain (using B2 on the COI data) was sampled at different frequencies between 1 and 10. For sampling every four generations, the proportion of independent data points was ~0.92 (versus ~0.95 for sampling every 10th generation) but the time required to collect these points was less than half that for sampling every 10th generation; we therefore chose every fourth generation as a reasonable sampling interval. Most of the results on convergence diagnostics are presented as supplementary data.

2.2.5. Parsimony, Maximum Likelihood, and Bayesian Estimation

To contrast results from the Markov chains and methods described above with more familiar methods, we performed parsimony and ML ancestral reconstructions using PAUP* 4.0 (Swofford 2000). In addition to estimating the frequencies for each site, we recorded the

maximum likelihood value for the chains run under the BL-, B1, and B2 approaches. Although we present primarily the ML and parsimony results from PAUP*, and the posterior distribution estimates from our own program, there was not a qualitative difference between the biases produced from the ML estimate with our method and parsimony or the GTR model assuming either constant rates or Gamma rates under PAUP*. A technical point worth clarifying is that bias from the optimization methods is a result of choosing a particular nucleotide as ‘best’ as opposed to tracking the entire distribution. We used PAUP*’s choice for parsimony reconstruction without considering alternative equally parsimonious solutions; consideration of these alternatives should not change parsimony’s bias because PAUP* chooses randomly among equally parsimonious ‘solutions’, and hence when one looks across many sites one gets a fair estimate of the performance of the method.

2.2.6. Functional Test

Primate mitochondrial tRNAs were aligned using ClustalX (Thompson, Higgins, and Gibson 1994), and tRNAscan-SE (<http://www.genetics.wustl.edu/eddy/tRNAscan-SE/>) was used to obtain predicted secondary structures (Lowe and Eddy 1997). Only perfectly aligned and consistently paired sites were considered in our analyses, meaning that sites in the alignment were discarded if they included gaps, if they included loops in any of the predicted secondary structures, or if they were paired with different sites in predicted secondary structures from different species. These alignment and pairing criteria were necessary to avoid alignment ambiguity and to avoid changes in the base-pairing context, which we were unprepared to accommodate. Thus, out of about 22,000 aligned sites, 13,803 sites did not have gaps, and only 7,740 sites were consistently paired in predicted secondary structures. Of these, only 3,360 were variable across the primates. The alignments for six tRNAs (tRNA-gln, tRNA-glu, tRNA-ile,

tRNA-met, tRNA-leu4 and tRNA-pro) were used in their entirety, whereas tRNA-tyr was poorly aligned and contributed few sites. The base composition variability at a site was measured with the Shannon index at that site across the species in the study, $S = -\sum_i p_i \ln(p_i)$, where p_i is the frequency of nucleotide i (A, C, G, or T) at that site. The Shannon index was also used to estimate the ambiguity of the posterior probability distribution for each internal node at each site.

2.2.7. Simulations of Constant and Variable Evolution

For the constant evolution simulations, evolution was stationary along each branch on the primate phylogeny. Hence, simulations were performed under the most likely model for a gene by starting at the deepest node and keeping the rate matrix and equilibrium frequencies constant. Under the variable model of evolution, the average of all inferred ancestral node frequencies was used for all the internal branches, while the external branches were simulated using the nearest tip frequencies. The ML rate parameters were kept constant throughout. The frequencies observed in the simulations were recorded for each base and for each internal node (θ_{bn}), and reconstructions ($\hat{\theta}_{bn}$) were made using the parsimony, ML, BL-, B1, and B2 methods. The differences between the reconstructed and simulated frequencies for each base (b) and for each internal node (n) were used to estimate the bias ($\hat{\theta}_{bn} - \theta_{bn}$). The total bias in the frequency reconstruction was summarized by the mean squared error (MSE):

$$MSE = \sum_{n=1}^k \sum_{b=1}^4 \left(\frac{(\hat{\theta}_{bn} - \theta_{bn})^2}{4k} \right), \quad (17)$$

where k is the total number of internal nodes.

2.3. Results

2.3.1. Chain Convergence

For primate COI and Cyt-b alignments, burn-in was achieved after 200 and 500 sampled generations, respectively (Supplementary Data). Apparent convergence can be seen by the lack of change in equilibrium values, made clearer in the expanded windows, for which the noise is greater than any directional trend in the data. Samples were graphed for all 16 transition matrix parameters for confirmation of convergence of each rate parameter (data not shown), and posterior probability distributions were calculated for each parameter (e.g., Figure 2.1). Chain diagnostics confirmed that convergence had been reached, since differences among chains and estimated asymptotic variance were generally less than 1%. (Supplementary Data: Appendix A). After excluding burn-in, the effective sample sizes were 24,111 for COI and 47,692 for Cyt-b.

2.3.2. Differences in Base Frequencies of Reconstructed Ancestral Sequences

Ancestral reconstructions by both parsimony and ML had different base frequencies than the extant taxa (tips), particularly for the less frequent bases (Tables 2.1 and 2.2; values shown are for the heavy strand). For all codon positions together, the use of a model with gamma-distributed rates (gML) does not change the inferred ancestral nucleotide frequencies very much, and in some cases for COI it is slightly worse than the GTR without gamma. In contrast, ancestral frequencies estimated by tracking the entire posterior distribution using any of the three intermediate methods were generally more similar to the extant sequence frequencies. The ancestral state frequencies are most similar to the extant frequencies when up to two substitutions per branch were allowed, indicating that there is little or no bias (95% credible intervals for state frequencies are always within 0.2% of the mean values). The low-frequency base biases in parsimony and ML reconstructions were more noticeable at the more variable 3rd codon sites, which had more uneven frequency distributions (Tables 2.1 and 2.2). The most extreme

frequencies were seen for C on the heavy strand at 3rd codon positions, where average COI frequencies at the tips were 0.065, and Cyt-b frequencies were 0.037. Posterior ancestral frequency estimates with two substitutions per branch (B2) were 0.061 and 0.032, respectively, but for parsimony they were 0.028 and 0.011 and for ML they were 0.026 and 0.009, substantially less than the extant species. Reducing the allowable number of substitutions per branch to one (B1) only marginally increased the difference between the Bayesian estimates and tip frequencies, but omitting the influence of branch lengths entirely (BL-) produced estimates that had half the apparent bias of the parsimony and ML estimates.

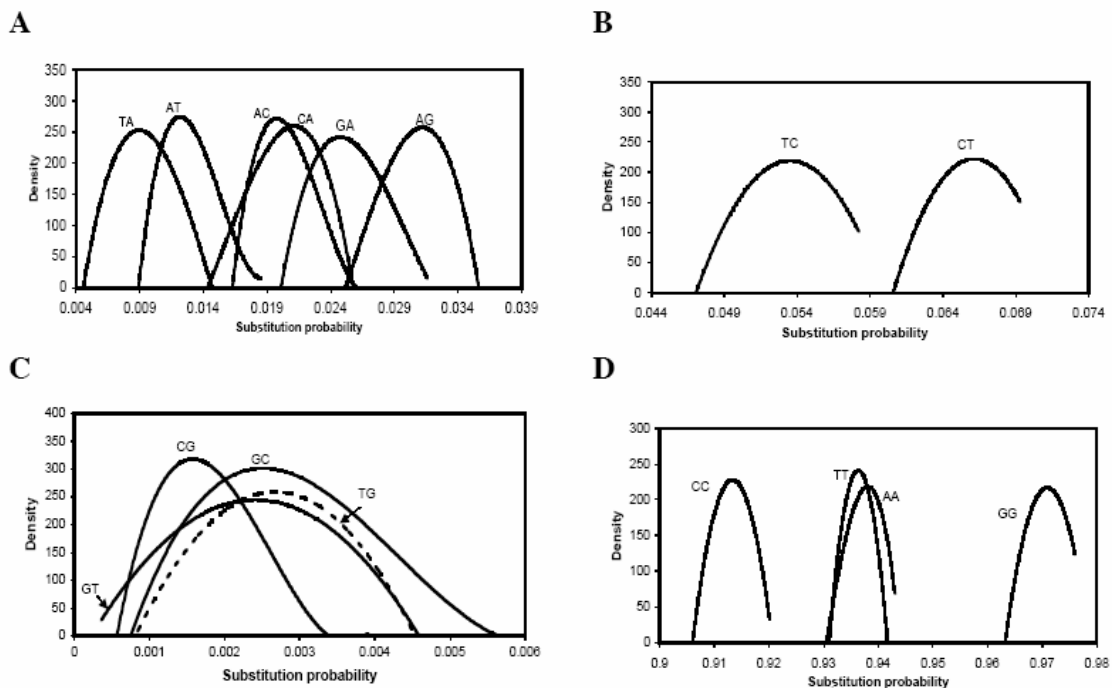


Figure 2.1. Posterior Probability Density Distributions of the Sixteen Substitution Probabilities for Cyt-B. The example shown was calculated using the GTR model and the B2 restricted likelihood method. The substitution probabilities are shown on four different scales: (A) T \Rightarrow A, A \Rightarrow T, A \Rightarrow C, C \Rightarrow A, G \Rightarrow A, and A \Rightarrow G, (B) T \Rightarrow C, and C \Rightarrow T, (C) G \Rightarrow T, C \Rightarrow G, G \Rightarrow C, and T \Rightarrow G, and (D) C \Rightarrow C, T \Rightarrow T, A \Rightarrow A, and G \Rightarrow G. Since the model is reversible, the substitution probabilities at any time point are equal to the rate parameter times the equilibrium frequency of the nucleotide being substituted to.

Table 2.1. Nucleotide Frequencies and Frequency Ratios for Extant Sequences (Tips) and Ancestral States in the COI Gene². Internal node frequencies for all analyses shown were calculated using reversible models ¹assuming constant rates among sites or ²accounting for among-site rate variation using a gamma distribution. For tips, the observed frequency is shown. ³Bold indicates the least and bold italics indicates the most biased method for each nucleotide frequency and frequency ratio. ⁴ML estimates of ancestral node states were used for the GTR model, while for the intermediate restricted likelihood methods BL-, B1, and B2, the posterior distribution at each node was used.

<i>All positions</i>						
Method ^{1,2,4}	T	C	A	G	C/T	G/A
Pars	0.285	0.152	0.274	0.289	0.535	1.06
ML¹	0.282	<i>0.151</i>	<i>0.274</i>	0.293	0.537	1.07
gML²	<i>0.287</i>	0.152	0.272	0.292	<i>0.531</i>	<i>1.072</i>
BL-¹	0.278	0.152	0.275	<i>0.295</i>	0.546	1.07
B1¹	0.277	0.155	0.284	0.283	0.56	0.996
B2¹	<i>0.269</i>	<i>0.164</i>	<i>0.29</i>	<i>0.277</i>	<i>0.608</i>	<i>0.963</i>
Tips	0.268	0.165	0.292	0.275	0.615	0.941

<i>Third codon positions</i>						
Method ^{1,4}	T	C	A	G	C/T	G/A
Parsimony	0.397	0.028	0.195	0.380	0.070	2.10
ML	0.389	<i>0.026</i>	0.198	0.386	<i>0.067</i>	1.95
BL-	0.375	0.044	0.254	0.335	0.119	1.32
B1	0.352	0.061	0.231	0.355	0.174	1.54
B2	<i>0.352</i>	<i>0.061</i>	<i>0.231</i>	<i>0.355</i>	<i>0.177</i>	<i>1.54</i>
Tips	0.349	0.065	0.230	0.356	0.188	1.68

For COI, at 3rd codon positions the average C/T ratio 0.177 for B2, 0.070 for parsimony, 0.067 for ML, and 0.188 at the tips. The differences in C/T ratios were similar for 3rd codon positions in Cyt-b, whereas ML was similar to parsimony (both were around 75% lower than extant sequence frequencies). B2 and B1 were most similar to the tips, off by only about 15%. It is worth noting that the estimates from the posterior with the simple method were much more similar to the tips than ML estimates (using the GTR model) despite the fact that the likelihood

maxima in these runs were considerably lower (Table 2.3). Presumably, the lower maxima were due to the limitation on the number of substitutions per branch per site, but the biases of ML in this situation were overwhelming. Ignoring branch lengths (BL-) produced an even larger drop in likelihood maxima, and differences from the tips were about half as large as those of ML and parsimony. In contrast, the likelihood maxima for the B2 approaches using a non-reversible model were significantly higher than for the GTR model (Table 2.3).

Table 2.2. Nucleotide Frequencies and Frequency Ratios for Extant Sequences (Tips) and Ancestral States in the Cyt-B Gene². Internal node frequencies for all analyses shown were calculated using reversible models ¹assuming constant rates among sites or ²accounting for among-site rate variation using a gamma distribution. For tips, the observed frequency is shown. ³Bold indicates the least and bold italics indicates the most biased method for each nucleotide frequency and frequency ratio. ⁴ML estimates of ancestral node states were used for the GTR model, while for the intermediate restricted likelihood methods BL-, B1, and B2, the posterior distribution at each node was used.

<i>All positions</i>							
Method^{1,2,4}		T	C	A	G	C/T	G/A
Parsimony¹	0.308	0.109	0.238	0.346		0.353	1.46
ML¹	0.305	0.109	0.235	0.352		0.357	1.50
gML²	0.306	0.109	0.235	0.351	0.355		1.496
BL-¹	0.299	0.109	0.265	0.327	0.365		1.24
B1¹	0.291	0.119	0.261	0.328	0.41		1.26
B2¹	0.291	0.119	0.261	0.328	0.41		1.26
Tips	0.292	0.120	0.265	0.323	0.412		1.22

<i>Third codon positions</i>						
Method^{1,4}	T	C	A	G	C/T	G/A
Parsimony	0.410	0.011	0.092	0.486	0.027	5.287
ML	0.409	0.009	0.082	0.500	0.021	6.109
BL-	0.392	0.021	0.158	0.437	0.053	2.759
B1	0.372	0.031	0.158	0.438	0.085	2.777
B2	0.372	0.032	0.158	0.438	0.085	2.777
Tips	0.375	0.037	0.155	0.434	0.098	2.803

For 3rd codon positions, C/T frequency ratio estimates using parsimony, ML, and B2 were

mapped to each node in the primate mitochondrial phylogeny (Figure 2. 2). There was considerable variation in frequency ratios among both extant and ancestral nodes, but the B2 ancestral C/T ratios generally reflected the C/T ratios of nearby nodes, whereas the parsimony and ML frequencies deviated in the direction of their apparent bias.

2.3.3. Simulation Results

For simulations with variable evolutionary rates (Table 2.4), the average bias over all the Ancestral nodes was highest for the most frequent nucleotide (i.e., T): about 0.113 for ML, followed by parsimony at 0.09. For the least frequent nucleotide (i.e., C), the frequency was lower by 0.14 for ML and 0.08 for parsimony. The biases for the Bayesian methods were much lower, in the range of 0.008 to 0.012 for T and -0.02 to -0.03 for C. The mean squared errors (MSEs; Table 2.4) were lowest for B2 (0.0017) and highest for ML (0.0089). For constant rate simulations, parsimony was more biased than ML for C, the rare nucleotide, and less biased for T (Table 2.4). In comparison, B2 deviated by less than half a percent for all four nucleotides. The MSEs for the Bayesian methods were ~4 times less for the constant evolution than for variable evolution, but those for parsimony and ML were about twice as big under constant evolution.

Table 2.3. Maximum Likelihood Values for Different Methods with COI and Cyt-B. ¹For BL-, B1, and B2, maxima were calculated from the optimum encountered during MCMC runs. ²Calculated using a reversible model. ³Calculated using a non-reversible model. All differences are extremely significant based on likelihood ratio tests.

Method/Model¹	COI	Cyt-b
ML²	-13654.9	-11203.3
BL⁻²	-17364.4	-15452.7
B1²	-14845.0	-12474.8
B2²	-13934.8	-11644.8
B2³	-13452.2	-10913.0

Table 2.4. Biases for Each Nucleotide Averaged Over All Internal Nodes and MSEs for Various Methods for Simulations Performed with Constant and Variable Models of Evolution. ¹Equilibrium frequencies varied along the tree during simulation, rate parameters did not. ²Equilibrium frequencies and rate parameters constant during entire simulation. ³ML estimates of ancestral node states used for the GTR model, while for the intermediate methods BL-, B1, and B2, the posterior distribution at each node was used.

Method/Model ³	Variable Evolution ¹				
	C	A	T	G	MSE
Parsimony	-0.080	-0.006	0.090	-0.005	0.005
ML	-0.141	-0.006	0.113	0.034	0.009
BL-	-0.032	0.001	0.012	0.019	0.002
B1	-0.030	0.001	0.008	0.015	0.001
B2	-0.021	0.001	0.008	0.012	0.001
	Constant Evolution ²				
	C	A	T	G	MSE
Parsimony	-0.040	-0.013	0.025	0.048	0.01
ML	-0.024	-0.013	0.043	0.035	0.02
BL-	-0.015	0.004	0.004	0.004	0.0005
B1	-0.011	0.004	0.002	0.003	0.0002
B2	-0.005	0.003	0.003	0.001	0.0002

2.3.4. Comparison of Base Frequencies and Structure Stabilities of Reconstructed tRNAs

To evaluate the effect of base frequency bias on functional inferences, we reconstructed ancestral sequences for all primate mitochondrial tRNAs and examined the compatibility of canonically paired sites in consistently paired ancestral tRNA helices. A similar approach was used to detect sequencing errors by showing that within-species variants that decreased the stabilities of folded sequences were often conserved among other species, and thus probably erroneous (Noor and Larkin 2000). Here, the reconstructed variants of tRNAs that did not retain canonical base pairing were less likely to fold into stable structures, thus indicating errors in reconstruction. We evaluated the canonical base pairing for all methods, but present only the comparison of the ML method (calculated using PAUP*; parsimony results were similar) with

the joint set of posterior probabilities for the B2 method (B1 was slightly worse, but similar to B2) both calculated using the GTR model of evolution. Since the B2 method is only marginally biased (based on the simulations), it can reasonably represent posterior estimates in general, and the importance of the comparison is between optimization methods and posterior estimation of ancestral states, not between full or restricted likelihood, or between Bayesian and ML methods for parameter estimation.

There were 3,360 consistently paired nucleotides at 15 internal nodes for variable sites, and Bayesian integrations were more compatible with base pairing than ML in 1,096 cases (32.6%). To understand which sites were contributing to this effect, we classified reconstructions according to the variability of the site and how ambiguously the node and site combination was reconstructed (Table 2.5). The percentage of cases in which the integrated posterior compatibility was better than the ML reconstruction varied according to the extent of nucleotide variability at a site and the ambiguity of node reconstruction (Table 2.5).

Table 2.5. Proportion of Base Pairs for which B2 had Higher Complementarity than ML, Classified by Node Ambiguity and Nucleotide Variation at Each Site. ¹Low ambiguity sites have ambiguities less than 0.0001, high ambiguity sites are greater than 0.01, and intermediate sites are in between. ²Low variability sites have variabilities less than 0.22, high variability sites are greater than 0.708, and intermediate sites are in between.

Node Ambiguity ¹	Nucleotide Variability at Site ²		
	Low	Intermediate	High
Low	0.39 (593/1540)	0.2 (199/992)	0.42 (44/106)
Intermediate	0.34 (17/50)	0.39 (5/13)	0.6 (3/5)
High	0.27 (96/360)	0.39 (93/240)	0.85 (46/54)

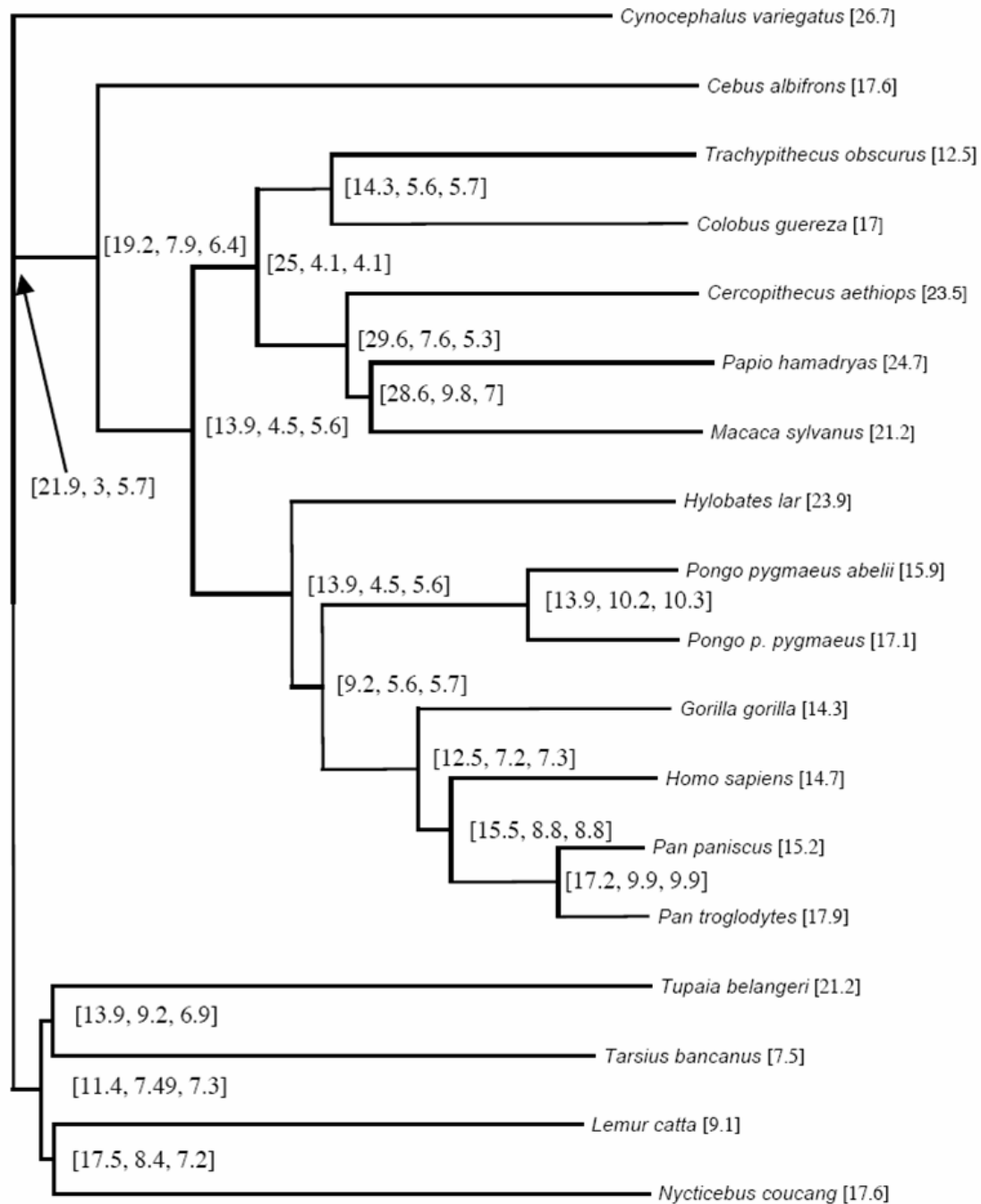


Figure 2.2. The Primate Phylogeny Most Compatible with the Mitochondrial Sequences, Along with the Ancestral State C/T Frequency Ratios of B2, Parsimony and ML Mapped to the Internal Nodes, Along with Observed Ratios for the Sequences at the Tips. Data shown are for the 3rd codon positions of COI. This phylogeny were estimated using the neighbor-joining algorithm with the BioNJ option, with distances calculated using ML and the GTR model. Further optimization of branchlengths with the *lscores* option using ML yielded different branch lengths but did not change reconstruction results. This phylogeny is probably slightly inaccurate in some details with respect to species divergences (see Methods).

At low nucleotide variability, the integrated posterior compatibility was slightly less than ML, but this trend was substantially reversed for sites with high nucleotide variability. The effect of ambiguity also changed, such that for sites with low variability, ML did relatively better with increasing node ambiguity, whereas for sites with high variability ML did considerably worse with increasing node ambiguity. These results make a reasonable amount of sense, in that bias in ancestral base frequencies away from low frequency nucleotides is unlikely to influence results until a moderate level of nucleotide variability is achieved. This was clear from the average amount of improvement in degree of base pairing complementarity with different levels of variability (Figure 2.3). Although ML has a small advantage when variability is low, the disadvantage of ML when variability is high can be quite large.

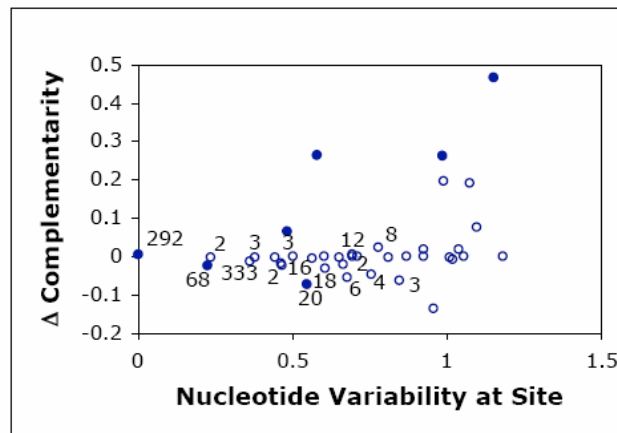


Figure 2.3. Differences between B2 and ML (B2-ML) in tRNA Base-Pairing Compatibility of Predicted Ancestral Sequences (Δ Complementarity) as a Function of the Nucleotide Variability Observed at A Site. Most data points are averages of all 15 internal nodes at a single site, but averages over internal nodes at multiple sites are noted by a label next to the point indicating the number of sites contributing to that point. Filled circles indicate significant differences from zero ($P < 0.05$, 2 tailed t-test).

The observed effects on functional inferences occurred despite the fact that nucleotide frequencies in this dataset of structurally conserved helix pairs were only moderately different

than the tips for parsimony and even more different for ML, whereas B2 integrated posterior frequencies were barely different than the tips (Table 2.6). Some differences depended on which strand encodes the tRNA, but the ordering of the methodologies was similar.

Table 2.6. Average Nucleotide Frequencies at Tips and Internal Nodes for tRNAs Coded on the Heavy Strand (HS) and Light Strand (LS). ¹Internal node frequencies for all analyses were calculated using reversible models. For tips, the observed frequency is shown. ML estimates of ancestral node states were used for the GTR model, while for the intermediate methods BL-, B1, and B2, the posterior distribution at each node was used

Method/Model¹	tRNA	T	C	G	A
Parsimony	HS	0.247	0.272	0.125	0.36
	LS	0.374	0.153	0.221	0.287
ML	HS	0.250	0.262	0.139	0.349
	LS	0.361	0.148	0.211	0.280
BL-	HS	0.254	0.259	0.138	0.349
	LS	0.368	0.154	0.240	0.273
B2	HS	0.255	0.258	0.146	0.341
	LS	0.355	0.154	0.211	0.280
TIPS	HS	0.255	0.246	0.147	0.352
	LS	0.356	0.156	0.211	0.277

2.4. Discussion

Methods that reconstruct an optimal ancestor (parsimony and ML) create large nucleotide frequency differences between reconstructed ancestral sequences and true ancestral sequences, and are therefore biased. Ancestral frequencies estimated by tracking the entire posterior distribution do not show such differences, and are much less biased even when the evolutionary process varies over time. It was surprising that the bias in ML reconstruction was usually similar or more extreme than in parsimony reconstruction. The bulk of the bias seems to arise from the use of optimization methods on these ambiguously determined discrete hyperparameters (ancestral states), rather than from whether the method or model is statistically or theoretically well-founded. Our results do not indicate fundamental differences between the performance of ML and Bayesian analyses for estimating substitution model parameters; but instead show that

biases occur when the most likely ancestor is chosen rather than tracking the entire ancestral distribution. Thus, being “most likely is not enough” (Antezana 2003). When incorporated into Bayesian analyses, features of parsimony, including consideration of only one substitution per branch per site and ignoring branch lengths, produce up to half as much bias as seen in parsimony. It is possible that parsimony’s inability to make anything but a random choice between equally parsimonious reconstructions is solely responsible for making it slightly less biased than ML in our simulations.

We infer that the cumulative effects of ancestral reconstruction biases can be important for functional inference, using the example of the predicted effect on tRNA structure. Comparisons among evolutionary models (GTR, and “parsimony”), restricted likelihood calculations (BL-, B1, B2, ML), methods of inferring ancestors (Bayesian, ML, parsimony), and programs (PAUP*, our programs) help to clarify the nature of the bias, and show that it is not an artifact of any particular set of procedures. Differences between tip sequences and ancestral reconstructions in primates were consistent with expected biases produced by ML and parsimony. Clearly, the idea that ancestral primates have evolved from radically different frequencies than those seen today (Schmitz, Ohme, and Zischler 2002) is no longer tenable, since Bayesian estimates of ancestral frequencies are similar to extant sequences, and there is only a small amount of bias in Bayesian reconstructions whether the evolutionary process is variable or constant. Nucleotide frequencies have clearly changed during primate evolution (Figure 2. 2), but not by nearly as much as are inferred from ML and parsimony reconstructions, and not in consistent and convergent directions along lineages leading to tip sequences.

The effects of reconstruction bias are not limited to errors in reconstructing nucleotide frequencies, but can lead to serious bias and inaccuracies in functional predictions. All else being

equal, one would normally predict that integrating base-pairing potential over posterior probabilities would yield considerably less complementarity than optimization, since with canonical base-pairing three out of four of the possible matches are sub-optimal. The observation that Bayesian estimates of ancestral tRNA base-pairs are better than ML estimates in 20-40% of the cases is disturbing enough, but the fact that at faster-evolving sites they can be better in 85% of the cases has sobering implications for reconstruction enthusiasts. We can make predictions that faster sites and more ambiguous nodes are likely to suffer from the greatest amount of bias, but it does not seem possible to accept reconstruction of ancestral conditions without question, even when the posterior probability of a particular reconstruction is high. If the measured functional features are correlated with particular nucleotides (for example, RNA secondary structure stability is likely correlated with GC content), then functional interpretations will be biased. Any situation where physico-chemical properties must be matched or balanced, as is the case with nucleotide pairing in RNA secondary structure, will also be biased.

Although we analyzed nucleotide content here partly because it is simpler and therefore easier to interpret than amino acid content, and because the simple predictions of canonical base-pairing provided a convenient test (Noor and Larkin 2000) to analyze function in reconstructed sequences, there is no reason to believe that the results cannot be generalized to amino acid sequences, and therefore to reconstruction of functional properties in ancestral proteins. For example, Gaucher *et al.* (2003) recently concluded not only that the common ancestor of all elongation factors of the bacterial Tu family proteins (Ef-Tu) was thermophilic rather than mesophilic, but surprisingly that the common ancestor of all mesophiles was thermophilic, too (Pauling and Zuckerkandl 1963; Benner 2002; Zhang and Rosenberg 2002; Gaucher et al. 2003). Many people may believe that mesophiles are derived from thermophiles, but if the last common

ancestor of mesophiles was thermophilic, mesophily must have arisen in parallel at least twice among the descendents of this ancestor, and all thermophilic descendents must have gone extinct (or at least, not have been sampled in this study).

Although great care was taken in this study to consider alternative reconstructions at ambiguous nodes, our results strongly imply that extremely biased reconstructions can appear certain precisely because of the bias. If thermostability is correlated with whichever amino acids are favored in a biased reconstruction, then the inference that the ancestral mesophile was thermophilic (and thus the inference of multiple parallel derivations of mesophily) would be false. If this is the case, consolation may be found in that reconstruction of ancestors may then be a profitable means to produce thermotolerant proteins from relatively less stable descendants. An obvious means to alleviate some (but not all) of these considerations in future studies would be to take care to maintain amino acid frequencies for all classes or conservation levels within the protein. This will reduce frequency bias, but problems with incorrect functional inference unfortunately may still remain due to interactions among sites (e.g., (Pollock, Taylor, and Goldman 1999)).

Our results also provide an interesting comparison concerning the effects of different assumptions on likelihood maxima and on reconstruction biases. Our simplest approach was similar to parsimony in that branch lengths were ignored, although incorporation of variation of rates among substitution types provided more flexibility than the standard parsimony algorithm. For both protein-coding genes, the likelihood maxima for this method were around four thousand log likelihood units worse than the maxima for the methods with branch lengths, providing strong evidence to reject the hypothesis that branch lengths don't matter. Allowing two substitutions per branch per site rather than only one improved the log likelihood

maxima by about 800-900 units, and allowing an infinite number of substitutions per branch improved the maxima by another 300-400 units. In many ways this is not surprising, since there is no theoretical justification for limiting the number of substitutions per branch, but it is interesting to note that incorporating a non-reversible model of evolution while limiting the substitutions to two per site per branch results in likelihood maxima that are 200 units better than the maxima for reversible models with an infinite number of substitutions allowed. The assumption of a reversible model is usually made for computational convenience, rather than because of any compelling theoretical justification. For the methodology developed here, non-reversible models do not have any greater computational burden than reversible models, and so a non-reversible model limited to two substitutions per branch per site may be both computationally and statistically more justified. We have developed this approach to allow incorporation of more complex and biologically realistic models without undue computational burden, so it is encouraging that the assumptions made result in small likelihood reductions that are easily compensated by other means, and that the reconstructions are only slightly divergent from extant or simulated nucleotide frequencies.

2.5. Literature Cited

- Antezana, M. 2003. When being "most likely" is not enough: Examining the performance of three uses of the parametric bootstrap in phylogenetics. *J Mol Evol* **56**:198-222.
- Arnason, U., J. A. Adegoke, K. Bodin, E. W. Born, Y. B. Esa, A. Gullberg, M. Nilsson, R. V. Short, X. Xu, and A. Janke. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc Natl Acad Sci U S A* **99**:8151-8156.
- Arnason, U., A. Gullberg, A. S. Burguete, and A. Janke. 2000. Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. *Hereditas* **133**:217-228.
- Arnason, U., A. Gullberg, and A. Janke. 1998. Molecular timing of primate divergences as estimated by two nonprimate calibration points. *J Mol Evol* **47**:718-727.

- Arnason, U., A. Gullberg, and X. Xu. 1996. A complete mitochondrial DNA molecule of the white-handed gibbon, *Hylobates lar*, and comparison among individual mitochondrial genes of all hominoid genera. *Hereditas* **124**:185-189.
- Arnason, U., and A. Janke. 2002. Mitogenomic analyses of eutherian relationships. *Cytogenet Genome Res* **96**:20-32.
- Beardsley, P. M., A. Yen, and R. G. Olmstead. 2003. AFLP phylogeny of *Mimulus* section *Erythranthe* and the evolution of hummingbird pollination. *Evolution Int J Org Evolution* **57**:1397-1410.
- Benner, S. A. 2002. The past as the key to the present: resurrection of ancient proteins from eosinophils. *Proc Natl Acad Sci U S A* **99**:4760-4761.
- Bleiweiss, R. 1998. Origin of hummingbird faunas. *Biol J Linnean Soc* **65**:77-97.
- Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol* **19**:1171-1180.
- Bull, J. J., C. W. Cunningham, I. J. Molineux, M. R. Badgett, and D. M. Hillis. 1993. Experimental Molecular Evolution of Bacteriophage-T7. *Evolution* **47**:993-1007.
- Collins, T. M., P. H. Wimberger, and G. J. P. Naylor. 1994. Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. *Syst Biol* **43**:482-496.
- Cunningham, C. W., K. E. Omland, and T. H. Oakley. 1998. Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution* **13**:361-366.
- Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol Biol Evol* **20**:248-254.
- Eyre-Walker, A. 1998. Problems with parsimony in sequences of biased base composition. *J Mol Evol* **47**:686-690.
- Faith, J. J., and D. D. Pollock. 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* **165**:735-745.
- Firat, M. Z., C. M. Theobald, and R. Thompson. 1997. Univariate analysis of test day milk yields of British Holstein-Friesian heifers using Gibbs sampling. *Acta Agric Scand Sect A, Anim Sci* **47**:213-220.
- Gaucher, E. A., J. M. Thomson, M. F. Burgan, and S. A. Benner. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **425**:285-288.

- Gelfand, A. E., and A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* **85**:398-409.
- Gelman, A., and D. B. Rubin. 1996. Markov chain Monte Carlo methods in biostatistics. *Stat Methods Med Res* **5**:339-355.
- Gelman, A., D. B. Rubin, J. B. Carlin, and H. S. Stern. 1992. *Bayesian Data Analysis*. Chapman and Hall, London, New York.
- Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Ieee Transactions on Pattern Analysis and Machine Intelligence* **6**:721-741.
- Giannasi, N., R. S. Thorpe, and A. Malhotra. 2000. A phylogenetic analysis of body size evolution in the *Anolis roquet* group (Sauria: Iguanidae): character displacement or size assortment? *Mol Ecol* **9**:193-202.
- Goodman, M., C. A. Porter, J. Czelusniak, S. L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C. P. Groves. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Molecular Phylogenetics and Evolution* **9**:585-598.
- Hassanin, A., and E. J. P. Douzery. 1999. Evolutionary affinities of the enigmatic saola (*Pseudoryx nghetinhensis*) in the context of the molecular phylogeny of Bovidae. *Proc Royal Soc London B-Biol Sci* **266**:893-900.
- Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**:97-109.
- Hibbett, D. S., and M. Binder. 2002. Evolution of complex fruiting-body morphologies in homobasidiomycetes. *Proc R Soc Lond B Biol Sci* **269**:1963-1969.
- Hillis, D. M., J. J. Bull, M. E. White, M. R. Badgett, and I. J. Molineux. 1992. Experimental phylogenetics: generation of a known phylogeny. *Science* **255**:589-592.
- Horai, S., K. Hayasaka, R. Kondo, K. Tsugane, and N. Takahata. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc Natl Acad Sci U S A* **92**:532-536.
- Hormiga, G., N. Scharff, and J. A. Coddington. 2000. The phylogenetic basis of sexual size dimorphism in orb-weaving spiders (Araneae, Orbiculariae). *Syst Biol* **49**:435-462.
- Huelsenbeck, J. P. 1995. The performance of phylogenetic methods in simulation. *Syst Biol* **44**.
- Huelsenbeck, J. P., R. Nielsen, and J. P. Bollback. 2003. Stochastic mapping of morphological characters. *Syst Biol* **52**:131-158.

- Huelsenbeck, J. P., and B. Rannala. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**:227-232.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754-755.
- Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**:2310-2314.
- Ingman, M., H. Kaessmann, S. Paabo, and U. Gyllensten. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**:708-713.
- Karlin, S., E. S. Mocarski, and G. A. Schachtel. 1994. Molecular evolution of herpesviruses: genomic and protein sequence comparisons. *J Virol* **68**:1886-1902.
- Koshi, J. M., and R. A. Goldstein. 1996. Probabilistic reconstruction of ancestral protein sequences. *J Mol Evol* **42**:313-320.
- Krawczak, M., A. Wacey, and D. N. Cooper. 1996. Molecular reconstruction and homology modelling of the catalytic domain of the common ancestor of the haemostatic vitamin-K-dependent serine proteinases. *Hum Genet* **98**:351-370.
- Little, R. J. A., and D. B. Rubin. 1983. On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *Amer Stat* **37**:218-220.
- Liu, J. S., A. F. Neuwald, and C. E. Lawrence. 1995. Bayesian models for multiple sequence alignment and Gibbs sampling strategies. *J Amer Stat Assoc* **90**:1156-1170.
- Lowe, T. M., and S. R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**:955-964.
- Maddison, D. R., and W. P. Maddison. 2000. *MacClade 4: Analysis of phylogeny and character evolution*. Sinauer Associates, Sunderland, Massachusetts.
- Malcolm, B. A., K. P. Wilson, B. W. Matthews, J. F. Kirsch, and A. C. Wilson. 1990. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* **345**:86-89.
- Meng, X. L., and D. B. Rubin. 1991. Using EM to obtain asymptotic variance - covariance matrices - the SEM algorithm. *J Amer Stat Assoc* **86**:899-909.
- Messier, W., and C. B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* **385**:151-154.
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953.

- Equations of state calculations by fast computing machines. *J Chem Phys* **21**:1087-1092.
- Nielsen, R. 2002. Mapping mutations on phylogenies. *Syst Biol* **51**:729-739.
- Nielsen, R., and J. P. Huelsenbeck. 2002. Detecting positively selected amino acid sites using posterior predictive P-values. *Pac Symp Biocomput*:576-588.
- Noor, M. A., and J. C. Larkin. 2000. A re-evaluation of 12S ribosomal RNA variability in *Drosophila pseudoobscura*. *Mol Biol Evol* **17**:938-941.
- Oakley, T. H., and C. W. Cunningham. 2000. Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. *Evolution* **54**:397-405.
- Pauling, L., and E. Zuckerkandl. 1963. Molecular 'restoration studies' of extinct forms of life. *Acta Chem Scand* **17**:9-16.
- Pollock, D. D., and W. J. Bruno. 2000. Assessing an unknown evolutionary process: Effect of increasing site-specific knowledge through taxon addition. *Mol. Biol. Evol.* **17**:1854-1858.
- Pollock, D. D., W. R. Taylor, and N. Goldman. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol* **287**:187-198.
- Rice, J. A. 1995. *Mathematical statistics and data analysis*. Duxbury Press, Belmont, California.
- Richard, F., M. Lombard, and B. Dutrillaux. 2003. Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res* **11**:605-618.
- 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.
- Sanderson, M. J., M. F. Wojciechowski, J. M. Hu, T. S. Khan, and S. G. Brady. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Mol Biol Evol* **17**:782-797.
- Sanson, G. F., S. Y. Kawashita, A. Brunstein, and M. R. Briones. 2002. Experimental phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions. *Mol Biol Evol* **19**:170-178.
- Schluter, D., T. Price, A. O. Mooers, and D. Ludwig. 1997. Likelihood of ancestor states in adaptive radiation. *Evolution* **51**:1699-1711.
- Schmitz, J., M. Ohme, and H. Zischler. 2000. The complete mitochondrial genome of *Tupaia belangeri* and the phylogenetic affiliation of scandentia to other eutherian orders. *Mol Biol Evol* **17**:1334-1343.

- Schmitz, J., M. Ohme, and H. Zischler. 2002. The complete mitochondrial sequence of *Tarsius bancanus*: evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. *Mol Biol Evol* **19**:544-553.
- Soltis, D. E., A. E. Selters, M. J. Zanis, S. Kim, J. D. Thompson, P. S. Soltis, L. P. R. De Craene, P. K. Endress, and J. S. Farris. 2003. Gunnerales are sister to other core eudicots: Implications for the evolution of pentamery. *Amer J Bot* **90**:461-470.
- Stewart, C. B., J. W. Schilling, and A. C. Wilson. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**:401-404.
- Swofford, D. L. 2000. *Phylogenetic analysis using parsimony (*and other methods)*. Sinauer Associates, Sunderland, Massachusetts.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**:4673-4680.
- Wang, C. S., J. J. Rutledge, and D. Gianola. 1994. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genet Sel Evol* **26**:91-115.
- Webster, A. J., and A. Purvis. 2002. Testing the accuracy of methods for reconstructing ancestral states of continuous characters. *Proc R Soc Lond B Biol Sci* **269**:143-149.
- Xu, X., and U. Arnason. 1996. A complete sequence of the mitochondrial genome of the western lowland gorilla. *Mol Biol Evol* **13**:691-698.
- Yang, Z. 1996a. Phylogenetic analysis using parsimony and likelihood methods. *J Mol Evol* **42**:294-307.
- Yang, Z. 1996b. Among-site rate variation and its impact on phylogenetic analyses. *Tree* **11**:367-371.
- Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**:1641-1650.
- Zhang, C., M. Zhang, J. Ju, J. Nietfeldt, J. Wise, P. M. Terry, M. Olson, S. D. Kachman, M. Wiedmann, M. Samadpour, and A. K. Benson. 2003. Genome diversification in phylogenetic lineages I and II of *Listeria monocytogenes*: identification of segments unique to lineage II populations. *J Bacteriol* **185**:5573-5584.
- Zhang, J., and M. Nei. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol* **44**:S139-146.

Zhang, J., and H. F. Rosenberg. 2002. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proc Natl Acad Sci U S A* **99**:5486-5491.

CHAPTER 3: DETECTING GRADIENTS OF ASYMMETRY IN SITE-SPECIFIC SUBSTITUTIONS IN MITOCHONDRIAL GENOMES

During mitochondrial replication, spontaneous mutations occur and accumulate asymmetrically during the time spent single-stranded by the heavy strand (D_{ssH}). The predominant mutations appear to be deaminations from adenine to hypoxanthine ($A \rightarrow H$, which leads to an $A \rightarrow G$ substitution) and cytosine to thymine ($C \rightarrow T$). Previous findings indicated that $C \rightarrow T$ substitutions accumulate rapidly and then saturate at high D_{ssH} , suggesting protection or repair, whereas $A \rightarrow G$ accumulates linearly with D_{ssH} . We describe here the implementation of a simple hidden Markov model (HMM) of among-site rate correlations to provide an almost continuous profile of the asymmetry in substitution response for any particular substitution type. We implement this model using a phylogeny-based Bayesian Markov chain Monte Carlo (MCMC) approach. We compare and contrast the relative asymmetries in all twelve possible substitution types, and find that the observed transition substitution responses determined using our new method agree quite well with previous predictions of a saturating curve for $C \rightarrow T$ transition substitutions and a linear accumulation of $A \rightarrow G$ transitions. The patterns seen in transversion substitutions show much lower among-site variation and are non-linear and more complex than those seen in transitions. We also find that, after accounting for the principal linear effect, some of the residual variation in $A \rightarrow G/G \rightarrow A$ response ratios is explained by the average predicted nucleic acid secondary structure propensity at a site, possibly due to protection from mutation when secondary structure forms.

3.1. Introduction

Vertebrate mitochondrial DNA (mtDNA) have an asymmetric replication mechanism that leads to asymmetry in nucleotide frequencies (Clayton, 1992a; Tanaka and Ozawa, 1994; Reyes et al., 1998; Yang et al., 2002). Recent studies suggest that these nucleotide asymmetries are caused by asymmetries in the probabilities of various, specific substitutions types (Reyes et al., 1998; Bielawski and Gold, 2002; Faith and Pollock, 2003). The major enzyme responsible for DNA replication in mitochondria is gamma polymerase (Copeland and Longley, 2003; Copeland et al., 2003), which initiates DNA synthesis directionally and asymmetrically from origins of heavy- and light-strand replication (O_H and O_L) that are separated by about two-thirds of the length of the genome (Clayton, 1992b; Graziewicz et al., 2002).

According to the classic model, replication of the heavy strand begins first, and proceeds from the O_H towards Cytochrome b (Cyt-b) along the circular genome (Figure 3.1). When the replication fork reaches the O_L , a short ~30bp stem-loop structure, replication of the light strand starts in the opposite direction, back towards Cytochrome C oxidase I (COI). After the heavy strand replication fork has passed, the original heavy strand remains single-stranded until the light strand replication fork passes, and as a result different portions of the genome spend different amounts of time single-stranded, depending on their location on the circular mitochondrial chromosome and their distance from the O_L . Since COI is close to the O_L and the first gene passed by the light-strand replication fork, it spends the least amount of time single-stranded, whereas Cyt-b is farthest and spends the longest time single-stranded. Assuming constant average movement of the replication forks, estimates of the times spent single-stranded (D_{ssH}) can easily be made (see Materials and Methods for further details). Hydrolytic

is constant, then the time that point X spends single-stranded will be proportional to twice the distance from X to O_L (solid lines.) If a site is past O_L , however (e.g., point Y; case 2), then it will be single-stranded for the time it takes for the light-strand replication fork to travel from wherever it is at that time to that point. If one assumes that the rate of movement of the replication forks is constant, then for example in the case of point Y, which is as distant from O_L as is point X, then the time that point Y spends single-stranded will be proportional to the distance from X to Y, in the direction of light-strand synthesis (dashed line). Since X and Y are equidistant from O_L , this distance is proportional to the length of the genome minus twice the distance from X (or Y) to O_L . The direction of replication from each of the origins is indicated by an arrow, protein-coding and rRNA genes are labeled by their standard abbreviations, and tRNAs are labeled with their standard single-letter abbreviations.

Different types of substitutions respond differently to time spent single-stranded (Faith and Pollock, 2003). The number of A→G substitutions appears to increase almost linearly with D_{ssH} (Faith and Pollock, 2003), and we have used linear models to estimate the slope and intercept for A→G gradients, and to provide credible intervals for these estimators (Krishnan et al., 2004a; Raina et al., *in review*). For C→T substitutions, there is apparently a steep initial rise with increasing D_{ssH} followed by saturation for the rest of the genome (Faith and Pollock, 2003). The gene-level analysis in this study did not provide a clear description of the C→T substitution response curve because the quick increase in substitution rates occurs mainly within COI.

Here, we introduce a method that allows the asymmetric component of the substitution probability matrix to differ among sites without imposing a linear relationship (or any other pre-specified relationship) on the D_{ssH} response curve. We assume that sites with similar D_{ssH} values tend to evolve at similar rates, and this is embodied in our model through a simple hidden Markov model (HMM) component, with the strength of the asymmetric component as the hidden state, and a transition probability such that the difference between substitution probabilities at adjacent sites in the alignment is distributed as a Gaussian with mean zero. Since our purpose is to understand

the substitution process, we employed a Bayesian analysis that assumed a constant phylogeny, and which relied on likely distributions for ancestral reconstructions at each node and each site considered (Krishnan et al., 2004c). The model used to obtain ancestral reconstructions was the simpler general-time-reversible (GTR) model, which introduces an unknown but conservative degree of bias towards reversibility and equal substitution probabilities among sites. The use of these simply generated ancestral state distributions allowed us to build more complicated models that varied at each site with relatively little computational costs. The base model (Bielawski and Gold, 2002) assumed symmetric rates on both strands (Lobry and Sueoka, 2002; Sueoka, 1995), and asymmetry (Bielawski and Gold, 2002) was incorporated into probabilities for specific substitution types as the “hidden” component that was variable among sites. We evaluated results using available complete primate mitochondrial genomes (plus two near outgroups) for all eight transversion types as well as the four transition substitutions.

3.2. Materials and Methods

3.2.1. Genome Sequences, Alignment, Phylogenetics, and Filtering the Data

We used eighteen complete mitochondrial genomes for our study, mostly primates, fifteen of which (thirteen primates and two outgroups) were available from GenBank when this study was initiated. These are: *Cebus albifrons* (NC_002763, Arnason et al., 2000), *Gorilla gorilla* (NC_001645, Horai et al., 1995); *Homo sapiens* (NC_001807, Ingman et al., 2000); *Hylobates lar* (NC_002082, Arnason et al., 1996); *Lemur catta* (NC_004025, Arnason et al., 2002); *Macaca sylvanus* (NC_002764, Arnason et al., 2000); *Nycticebus coucang* (NC_002765, Arnason et al., 2000); *Pan paniscus* (NC_001644, Horai et al., 1995); *Pan troglodytes* (NC_001643, Horai et al., 1995)); *Papio hamadryas* (NC_001992, Arnason et al., 1998); *Pongo pygmaeus pygmaeus*

(NC_001646, Horai et al., 1995); *Pongo pygmaeus abelii* (NC_002083, Xu and Arnason, 1996); and *Tarsius bancanus* (NC_002811, Schmitz et al., 2002) and outgroups *Tupaia belangeri* (NC_002521, Schmitz et al., 2000) and *Cynocephalus variegatus* (NC_004031, Arnason and Janke, 2002). Our colleagues (Raaum et al., *in review*) provided us with three other primate genomes, *Cercopithecus aethiops*, *Colobus guereza* and *Trachypithecus obscurus*.

Gene alignments were obtained using ClustalW (Thompson et al., 1994), and a neighbor-joining (NJ) tree was obtained in PAUP* 4.0 (Swofford, 2000) using a GTR model applied to concatenated alignments of all tRNAs, rRNAs, and protein-coding genes. This tree (Figure 3.2) was used in all further analyses. To analyze substitution rates that were as unaffected by selective processes as possible, we used 3rd codon positions in all thirteen aligned regions of protein-coding genes and only from codons that were in the four-fold redundancy class in all species in the alignment. Since asymmetric mutations appear to occur on the heavy strand, we considered the substitution process for the heavy strand, rather than the coding strand (in contrast to Faith and Pollock, 2003, and many other publications).

3.2.2. Calculation of Time Spent Single-Stranded

Each aligned site, s , in each genome, g , is associated with a position number, p_g , which is based on the arbitrary designation of the beginning of the tRNA adjacent to the control region as position 1. The calculation of time spent single-stranded at a site in a genome (D_{ssHg}^p) depends upon whether that site is located before (case 1) or after (case 2) the light-strand origin of replication, with respect to the direction of movement of the heavy-strand replication fork (Figure 3.1).

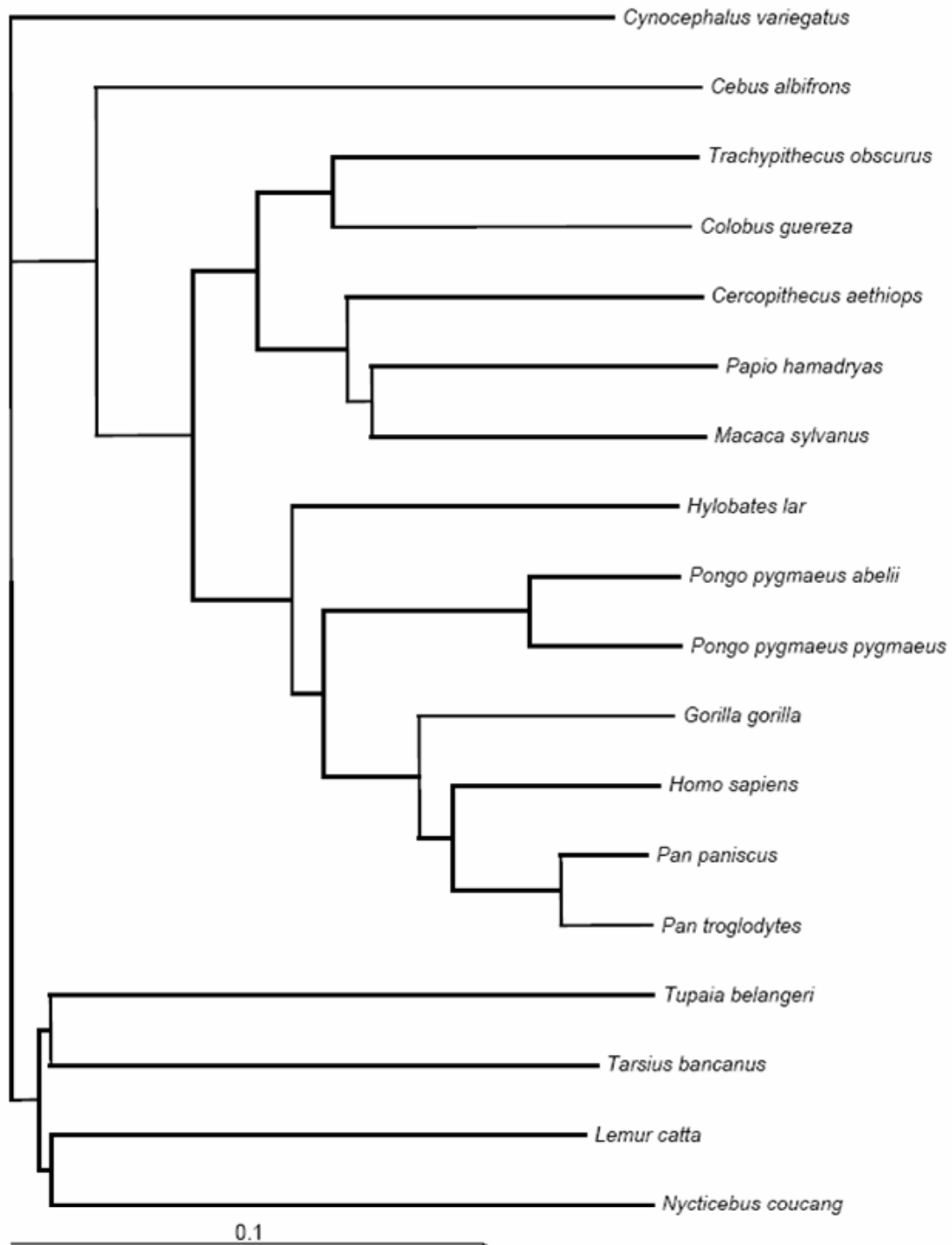


Figure 3.2. Phylogeny of Sixteen Primate Species and Two Near Outgroups Used in this Study. The neighbor-joining algorithm and distances based on the general-time reversible (GTR) model were used on a concatenation of all tRNA, rRNA, and protein coding genes. This tree, including topology and branch lengths, was used in all analyses.

Calculations used the following rules:

$$\begin{aligned} \text{if (1), } D_{ssHg}^p &= \frac{2|p_g - O_{Lg}|}{L_g} \\ \text{if (2), } D_{ssHg}^p &= 1 - \frac{2|p_g - O_{Lg}|}{L_g}, \end{aligned} \quad (1)$$

where O_{Lg} is the position of the origin of light-strand replication in genome g , L_g is the length of genome g , and units are proportional to genome length (that is, in units of genomes divided by the unknown replication rate per genome). If the first position in the arbitrary numbering system of the circular genome lies between O_{Lg} and a position under consideration (Figure 3.1), then a correction must be made to get the distance in nucleotides between O_{Lg} and the position. For each site, the D_{ssHg}^p were then averaged over all 18 species. For simplicity's sake, in future reference we will define D_{ssH} to mean D_{ssH}^p .

3.2.3. Posterior Predictive of Reconstructed Ancestral States

Bayesian analyses that simplify computational complexity by assuming two substitutions per branch have been developed and used to obtain a posterior distribution of ancestral sequences, and have been tested and described elsewhere (Krishnan et al., 2004c). Here, we used this method with a GTR model and the NJ phylogeny to reconstruct a posterior distribution of ancestral states for all nodes for the 3rd codon positions of conserved four-fold redundant codons (see Materials and Methods section). The ancestral states mapped onto internal nodes were treated as augmented data along with the original sequence data and updated during MCMC runs using a Gibbs Sampling scheme (Krishnan et al., 2004c). The distribution of ancestral sequences obtained in this

fashion has been compared with other methods [i.e., ML and parsimony in PAUP* (Swofford, 2000)], and relatively low frequency biases were found (Krishnan et al., 2004c).

Samples from the posterior probability distribution of these augmented data then served as the first stage of a posterior predictive distribution to relatively easily calculate our more complex model, in which the substitution matrix varied among all the sites (Krishnan et al., 2004a; 2004b; Nielsen, 2002). A full posterior predictive approach involves evaluating the accuracy of the model-based data by obtaining distributions of test-statistics such as likelihood ratio (Gelman et al., 1996). Previous studies on among-site or among-gene variation of average rates assume context-dependence (Pedersen and Jensen, 2001; Robinson et al., 2003), correlation of rates among genes (Thorne and Kishino, 2003), or adjust for rate-variation among genes and lineages (Hasegawa et al., 2003). None of these actually vary the substitution probability matrix (as opposed to the average rate) among sites, since with most standard methods there would be an exorbitant computational expense involved in doing so. Here, the calculation of likelihoods is feasible since the ancestral states at internal nodes are known (that is, the posterior predictive ancestral states from the simpler model are known) and used for all calculations involving the complex model.

3.2.4. Incorporation of a Different Asymmetric Mutation Component at Each Site

Our complex model starts with a symmetric “base” model that is the same at all sites and assumes strand symmetric rates of evolution (Bielawski and Gold, 2002; Lobry and Sueoka, 2002; Sueoka, 1995). This symmetric model is not necessarily reversible (in contrast to most commonly used models of evolution), and has fewer free parameters than the reversible model (Yang, 1994) since it assumes equal rates of complementary

substitutions. An asymmetric component was included by adding a site-specific parameter, c_p , to a particular pre-specified substitution rate (e.g., $A \rightarrow G$). This “hidden” component was also subtracted from the rate of self-change at each site for the appropriate nucleotide (e.g., $A \rightarrow A$). The values of c_p at each site were not dependent upon the magnitudes of D_{ssH} , but values of c_p at consecutive sites considered were related by a Markovian component dependent on Δ , the difference in D_{ssH} at consecutive sites (when sites are ordered according to D_{ssH}), such that

$$c_{p'} \sim N(c_p, \alpha \Delta) \quad (2)$$

where N is the normal distribution and α is a variable parameter that determined the magnitude of the variance; it is adjusted over the course of the Markov chain and is constant among sites. Thus, the probability distribution of the asymmetric component at a site was a normally distributed random variable with mean equal to the asymmetric component at the previous site, and variance estimated as a function of Δ and the free parameter α . There was no specific *a priori* linear or non-linear response built into this HMM. An MCMC analysis was run on the parameters of the symmetric model, the site-specific hyperparameters c_p , and the HMM component α . Posterior distributions of each of these parameters were obtained using flat, uninformative priors for all parameters and symmetric, uniform, and bounded proposal distributions centered around the previous parameter value for each step in the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). To determine the optimal proposal ranges, simulations were run from the starting values based on the average of a short preliminary MCMC run, and using different proposal ranges. The range for each parameter proposal distribution that

corresponded to 60-80% acceptance was then fixed for the subsequent runs (Krishnan et al., 2004c).

MCMC analyses with this HMM model were performed on all twelve substitution types in twelve separate runs, and site-specific response curves of substitution rates (relative to the symmetric rate for the same substitution type) versus D_{ssH} were obtained (Figure 3.3).

3.2.5. Average Site-Specific mRNA Secondary Structure

The online interface *mfold* (Zuker, 2003) was used to predict secondary structures for each predicted mRNA for each of the eighteen species, and the “loopiness” of each site in the alignment considered (see Materials and Methods section) was estimated as the proportion of alternative structures for which that site forms a “loop” rather than a “stem”. Although biological effects may occur based on DNA structure, we used RNA structure predictions because they are more developed and probably more accurate, and take into account higher order interactions among sites (Zuker, 2003). We considered all predicted alternative secondary structures that were at least half as stable as the optimal structure, and the ‘loopiness’ of a site was averaged across the 18 species. For analysis of the correlation between loopiness and the asymmetric substitution component for A→G substitutions, the expectation for the component was calculated based on a linear regression of the posterior average A→G/G→A ratio at each site, treating D_{ssH} as an independent variable. Residuals were then calculated as deviations from expectation. Sites were grouped into 17 categories based on degree of average loopiness, and residuals were averaged for all sites in a category. We tested for association between loopiness and average residual A→G/G→A ratio using a weighted regression analysis with loopiness as the independent variable.

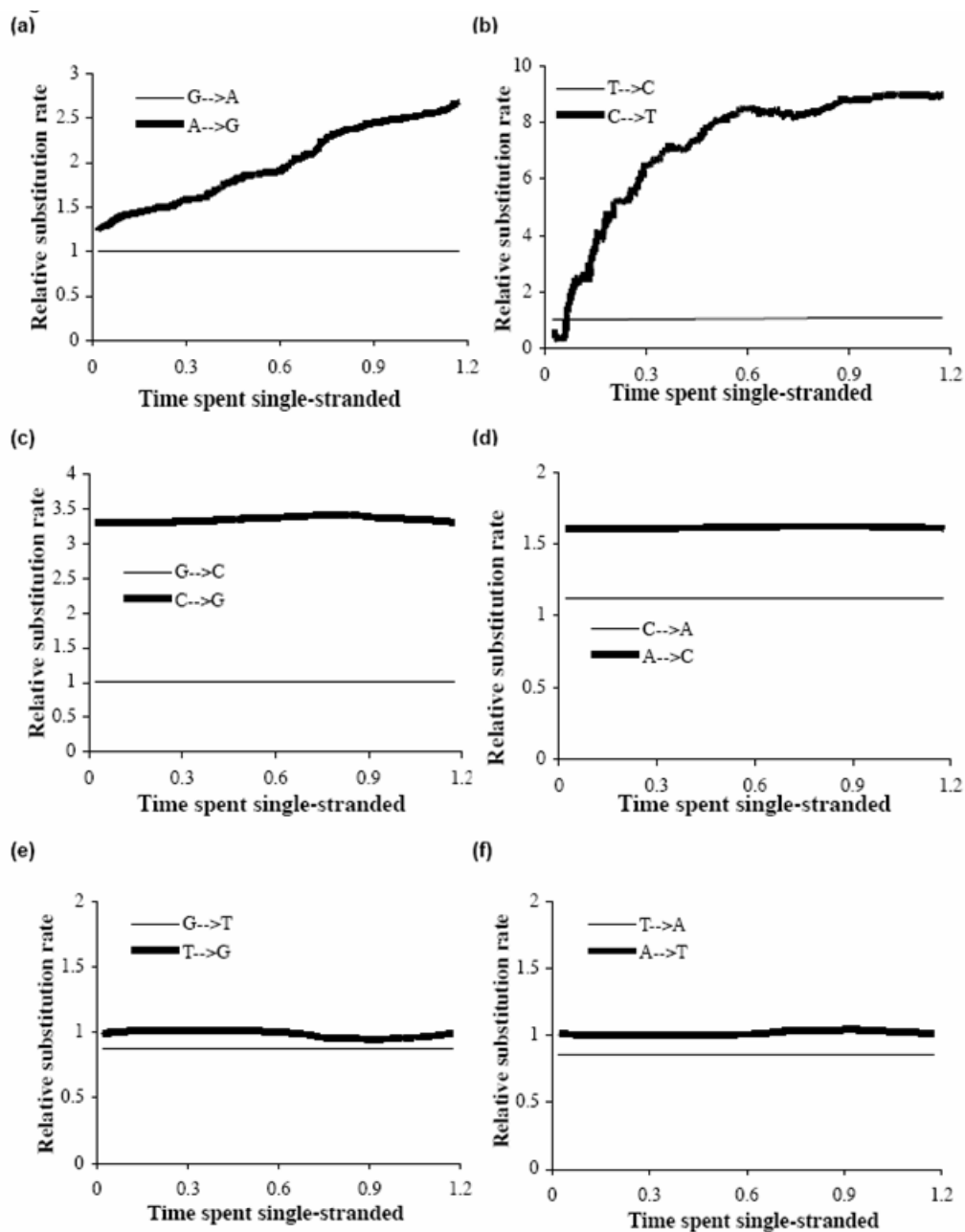


Figure 3.3. Relative Asymmetric Substitution Response Profiles Versus Time Spent Single-Stranded. Asymmetric substitution rates, relative to the magnitude of the same substitution rate in the symmetric “base” model (see Materials and Methods), are plotted versus predicted time spent single-stranded. The analysis of each substitution type was done separately. Substitution profiles are plotted with their reverse substitution profiles as follows (a) $A \rightarrow G$ and $G \rightarrow A$; (b) $C \rightarrow T$ and $T \rightarrow C$; (c) $C \rightarrow G$ and $G \rightarrow C$; (d) $A \rightarrow C$ and $C \rightarrow A$; (e) $T \rightarrow G$ and $G \rightarrow T$ and (f) $A \rightarrow T$ and $T \rightarrow A$. The substitution in each pair with the highest average relative asymmetric rate was arbitrarily designated the “forward” substitution.

3.3. Results

3.3.1. Analysis of Transition Substitution Response Gradients

The approach outlined here is innovative in that it does not specify the precise nature of the relationship between substitution rates and time spent single-stranded, and in that it allows for variation in the rate of individual substitution types at every point in the genome. We were able to evaluate a continuous response that did not require arbitrary choice of window-sizes for averaging and which enabled us to detect both site-specific deviations and regional trends. The nature of each substitution response is not built into the HMM model *a priori*, and there is no equation, linear or otherwise, which determines the type of response visualized. Therefore, any linear or non-linear response observed in our results is a direct reflection of trends in the data. Our use of a posterior distribution of ancestral states from a simpler model makes it feasible to create this site-specific model complexity with relatively little computational effort, and thus allow exploratory statistical analysis of complex site-specific substitution behavior.

We first demonstrated the utility of this method by analyzing transition substitution probabilities. As previously predicted (Faith and Pollock, 2003), substitution probabilities from A→G increase linearly with time spent single-stranded, and C→T substitutions increase rapidly at low single-strandedness values and then remain approximately constant over the rest of the genome (Figure 3.3a and 3.3b). In contrast, the reverse substitutions (G→A and T→C) remain relatively constant and the asymmetric components remain at relatively low levels (Figure 3.3a and 3.3b). Expanded views of the number of reverse transition substitutions relative to the symmetric base model (Figure 3.4a and 3.4b) show that there are trends in these substitutions: G→A substitutions decrease in approximately linear fashion with increasing time spent single-stranded ,

whereas T→C substitutions increase approximately linearly, then appear to level off at about $D_{ssH} = 0.9$. We must caution here that relative rates below 1.0 do not make clear biological sense, since this implies that fewer mutations occur due to time spent in the single-stranded state; the observed trend in G→A substitutions may instead be due to a tendency to confound backward (G→A) substitutions with the much more prevalent forward (A→G) substitutions.

In addition to the main trends, there is also variation in the ratios of substitution probabilities in the form of local dips and rises in the average posterior probability. It is likely that various factors, including codon bias, dinucleotide bias, and nucleic acid secondary structural features, affect substitution rates in addition to the effect of time spent single-stranded. We are currently addressing these factors by combining them into even more complex phylogeny-based evolutionary models; we show preliminary results on the effect of one of these factors (nucleic acid secondary structure) below.

3.3.2. Analysis of Transversion Substitution Response Gradients

The relative levels of transversion substitutions tend to be much closer to 1.0 than for transitions (Figure 3.3c-f), with the notable exceptions of C→G and A→C substitutions, which average around 3.3 and 1.6 times as much (respectively) as their rates in the symmetric base model and vary somewhat along the genome (Figure 3.3c and 3.3d). Arbitrarily referring to the transversion substitutions with a greater asymmetric component as “forward” substitutions (C→G, A→C, T→G, and A→T), all four reverse substitutions are nearly constant along the genome. Furthermore, the forward T→G and A→T relative substitution rates, although on average close to 1.0, can be seen to vary considerably with time spent single-stranded when the scale is expanded (Figure 3.4e and 3.4f). Posterior means for α (the parameter that controls correlation between adjacent

sites) are much higher for forward transitions ($C \rightarrow T$ and $A \rightarrow G$) as compared to backward transitions ($T \rightarrow C$ and $G \rightarrow A$) and all transversions.

The variation in forward transversion response curves with time spent single-stranded is intriguing in that they tend to increase after a short lag, peak around $D_{ssH} = 0.9$, then decrease (Figure 3.4). Again, we are cautious about over-interpreting this, since the observed trend may be due to a tendency to confound the transversion substitutions with the more strongly biased and variable $G \rightarrow A$ and $C \rightarrow T$ substitutions. It is also hard to see a plausible biological reason why mutation rates should decrease with longer times spent single-stranded, and a complicated interactive bias caused by both transition types seems more likely. Transcription could be invoked, but there is no clear difference in the transition responses for ND6, which is transcribed on the opposite side as the other protein-coding genes (Figure 3.3 and Faith and Pollock, 2003). This interpretation is supported by the shape of the $T \rightarrow G$ curve, which is almost exactly the opposite of the $A \rightarrow T$ curve, and reaches a minimum at the same point the others reach a maximum. This point corresponds to the ND4 gene, which has been previously noted to have unusually strong asymmetric features for undetermined reasons (Bielawski and Gold, 2002). If taken at face value, the situation with transversions appears to be complex, and in future studies we will incorporate simulations to determine the strength of biases that strongly asymmetric substitution process may have on inferring other substitution processes.

3.3.3. Correlation of Secondary Structure and Residual Transition Bias

To determine whether nucleic acid secondary structure has a detectable effect on asymmetric transition rates, we looked for a correlation between secondary structure and residual transition asymmetry. Since the average $A \rightarrow G/G \rightarrow A$ ratio has an approximately linear relationship with time spent single-stranded, we performed a linear regression on

this average, treating D_{ssH} as an independent variable. Residuals were then taken as differences from the expectation, $y = 1.2839 * D_{ssH} + 1.2242$. The residuals are larger when there is more loopiness (less secondary structure) at a site (Figure 3.5). A possible explanation for this preliminary result is that formation of secondary structure in the “single-stranded” DNA decreases the effective time spent single-stranded (Seligmann et al., *in review*), thus decreasing the A→G mutation rate.

3.4. Discussion

We have shown here that individual asymmetric mutation processes can be detected and evaluated at a site-specific level along the mitochondrial genome. Such evaluation will be important in developing a more complete and unified model of evolution in mitochondrial genomes. We have obtained an indirect indication that secondary structure may modify mutation and substitution processes, and in other work we are also incorporating the effects of adjacent nucleotides (which can strongly modify substitution rates) and codon bias. Many transversions show no indication of asymmetric substitution bias due to single-strandedness, and only two of them show strong asymmetric bias. Interestingly, the two most biased transversions (C→G and A→C) do not correspond to the best-characterized lesion from oxidative damage, the conversion of the guanine to 8-oxoguanosine, which can mispair with adenine leading to a G→T transversion substitution. There are indications, however, that methylglyoxyl (a major product of DNA oxidation), readily produces C→G transversions *in vivo* (Murata-Kamiya et al., 2000). The variation in the transversion substitution response curves is not simple, and may well be the result of fairly simple mutation response curves (linear or saturating) combined with inference bias introduced by the much stronger transition response curves.

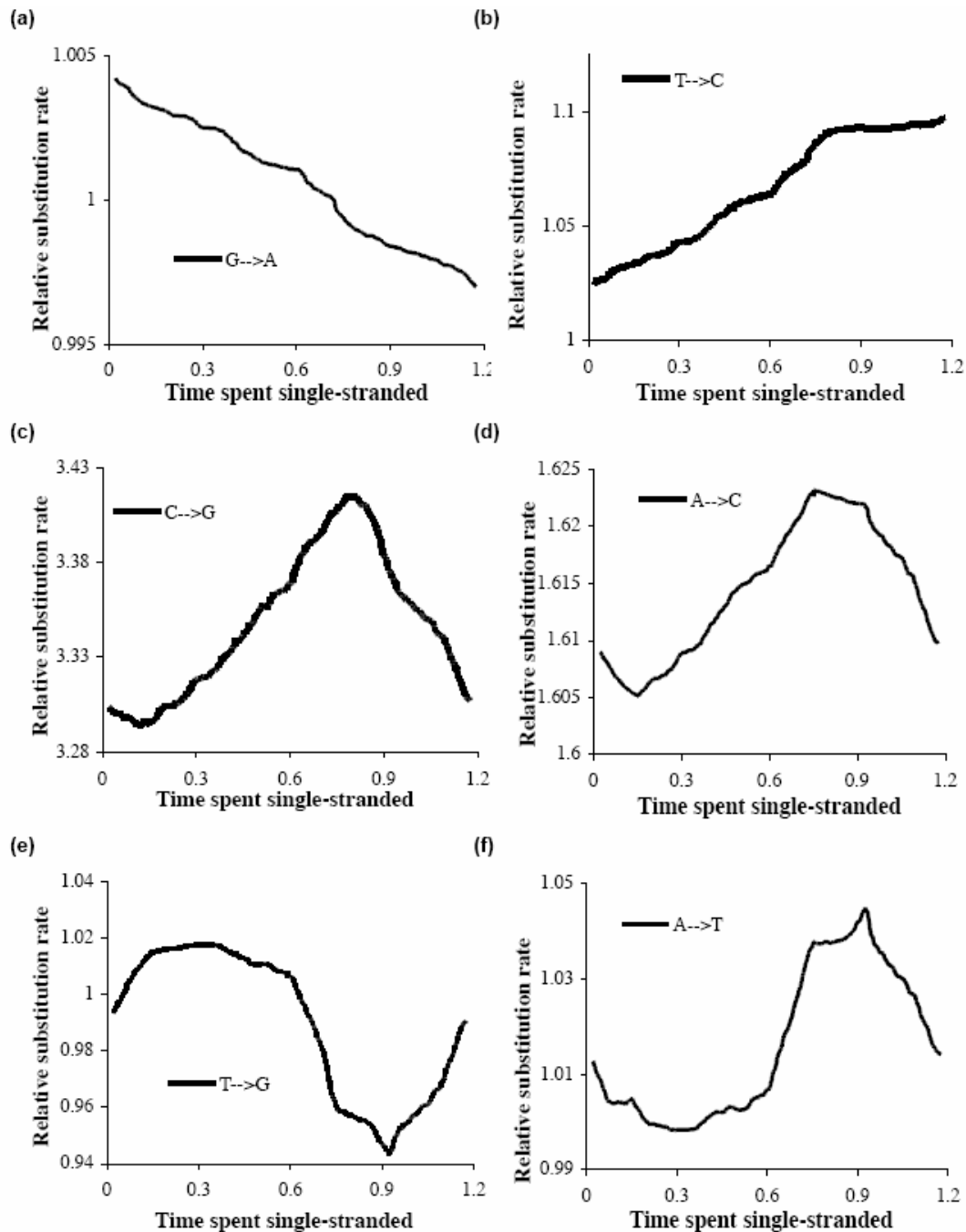


Figure 3.4. Expanded Views of Relative Asymmetric Substitution Response Profiles Versus Time Spent Single-Stranded. We present, on appropriately expanded scales, the profiles that appear approximately flat in Figure 3.3 due to the scale. For the two transitions (a) $G \rightarrow A$ and (b) $T \rightarrow C$, these are the reverse relative rates, whereas for the four transversions (c) $C \rightarrow G$ (d) $A \rightarrow C$ (e) $T \rightarrow G$ (f) $A \rightarrow T$, these are the forward relative rates. The four reverse relative transversion rates are essentially flat on any scale.

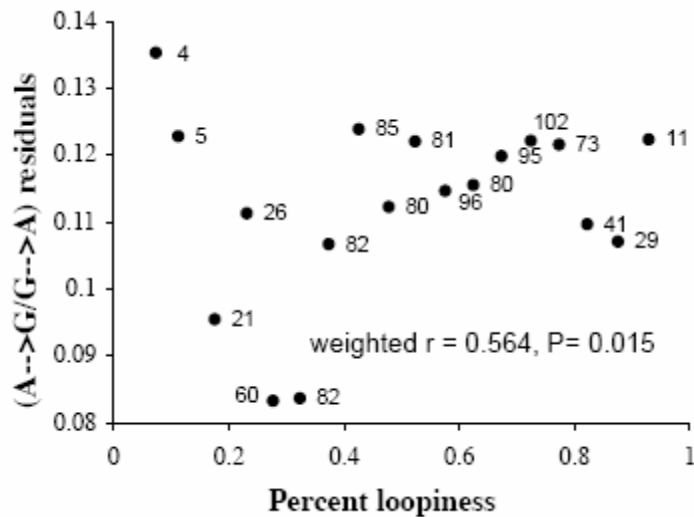


Figure 3.5. Excess Purine Transition Asymmetry as a Function of Loopiness.

Residual $A \rightarrow G/G \rightarrow A$ ratios were calculated assuming D_{ssH} as an independent variable in a standard linear regression analysis. The loopiness at each site was calculated as the fraction of structures in which that site was part of a loop (rather than a stem); all predicted mRNA secondary structures that were at least half as stable as the optimal structure in each species were considered. Sites were grouped into 17 categories based on their loopiness values, and the correlation coefficient ($r = 0.564$, $p = 0.015$) was calculated with points weighted by the number of sites in each category. Points are labeled with the number of sites in each category.

Future work should determine how much inference bias can be expected with these methods, and whether such bias can be corrected for in a combined analysis. It will also be important to expand the analysis to do comparative analysis of other taxon groups and larger taxon groups. The primates were chosen partly because they were the vertebrate family with the largest representation of complete genomes, but presuming that the rapid rate of increase in complete genomes continues (from 67 to over 300 in the last four years), expanded comparative analyses will soon be feasible. The results of the work presented here may be useful for improving phylogenetic analysis, for carrying out refined comparative analysis of substitution processes and replicative mechanisms, and in improving estimates of synonymous DNA substitution processes for incorporation into and comparison with amino acid substitution models, which may allow more accurate

detection of selection and functional divergence.

3.5. Literature Cited

- ARNASON, U., ADEGOKE, J.A., BODIN, K., BORN, E.W., ESA, Y.B., GULLBERG, A., NILSSON, M., SHORT, R.V., XU, X., AND JANKE, A. (2002). Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc. Natl. Acad. Sci. USA* 99, 8151-8156.
- ARNASON, U., GULLBERG, A., BURGUETE, A.S., AND JANKE, A. (2000). Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. *Hereditas* 133, 217-228.
- ARNASON, U., GULLBERG, A., AND JANKE, A. (1998). Molecular timing of primate divergences as estimated by two nonprimate calibration points. *J. Mol. Evol.* 47, 718-727.
- ARNASON, U., GULLBERG, A., AND XU, X. (1996). A complete mitochondrial DNA molecule of the white-handed gibbon, *Hylobates lar*, and comparison among individual mitochondrial genes of all hominoid genera. *Hereditas* 124, 185-189.
- ARNASON, U. AND JANKE, A. (2002). Mitogenomic analyses of eutherian relationships. *Cytogenet. Genome Res.* 96, 20-32.
- BIELAWSKI, J.P., AND GOLD, J.R. (2002). Mutation patterns of mitochondrial H- and L-strand DNA in closely related Cyprinid fishes. *Genetics* 161, 1589-1597.
- CLAYTON, D.A. (1992a). Transcription and replication of animal mitochondrial DNAs. *Int. Rev. Cytol.* 141, 217-232.
- CLAYTON, D.A. (1992b). Structure and function of the mitochondrial genome. *J. Inherit. Metab. Dis.* 15, 439-447.
- COPELAND, W.C., AND LONGLEY, M.J. (2003). DNA polymerase gamma in mitochondrial DNA replication and repair. *Scientific World Journal.* 3, 34-44.
- COPELAND, W.C., PONAMAREV, M.V., NGUYEN, D., KUNKEL, T.A., AND LONGLEY, M.J. (2003). Mutations in DNA polymerase gamma cause error prone DNA synthesis in human mitochondrial disorders. *Acta Biochim Pol.* 50, 155-167.
- FAITH, J.J., AND POLLOCK, D.D. (2003). Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* 165, 735-745.
- FRANCINO, M.P., AND OCHMAN, H. (1997). Strand asymmetries in DNA evolution. *TRIGS* 13, 240-245.
- FREDERICO, L.A., KUNKEL, T.A. AND SHAW, B.R. (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* 29, 160-174.

- FREDERICO, L.A., KUNKEL, T.A. AND SHAW, B.R. (1993). Cytosine deamination in mismatched base-pairs. *Biochemistry* 32, 6523-6530.
- GELMAN, A., MENG, X. AND STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 6, 733-807.
- GRAZIEWICZ, M.A., DAY, B.J., AND COPELAND, W.C. (2002). The mitochondrial DNA polymerase as a target of oxidative damage. *Nuc. Acids Res.* 30, 2817-24.
- HASEGAWA, M., THORNE, J.L., AND KISHINO, H. (2003). Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes Genet. Syst.* 78, 267-283.
- HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97-109.
- HORAI, S., HAYASAKA, K., KONDO, R., TSUGANE, K., AND TAKAHATA, N. (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. U S A* 92, 532-6.
- INGMAN, M., KAESSMANN, H., PAABO, S., AND GYLLENSTEN, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature* 408, 708-13.
- KRISHNAN, N. M., RAINA, S. Z., AND POLLOCK, D. D. (2004a). Analysis of among-site variation in asymmetric substitution patterns. *Biol. Proced. Online*, provisionally accepted.
- KRISHNAN, N. M., SELIGMANN, H., RAINA, S. Z., AND POLLOCK, D. D. (2004b). Phylogenetic analysis of site-specific perturbations in asymmetric mutation gradients. *Currents in Computational Molecular Biology*. A. Gramada, and P.E. Bourne eds. Pp. 266-267.
- KRISHNAN, N. M., SELIGMANN, H., STEWART, C. B., DE KONING, A.P.J., AND POLLOCK, D. D. (2004c). Ancestral sequence reconstruction in primates: Compositional Bias and effect on functional inference. *Mol. Biol. Evol.*, provisionally accepted.
- LOBRY, J.R., AND SUEOKA, N. (2002). Asymmetric directional mutation pressures in bacteria. *Genome Biol.* 3, research0058.1-research0058.14
- METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, AND E. TELLER. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087-1092.
- MURATA-KAMIYA, N., KAMIYA, H., KAJI, H., AND KASAI, H. (2000). Methylglyoxal induces G:C to C:G and G:C to T:A transversions in the supF gene on a shuttle vector plasmid replicated in mammalian cells. *Mutat. Res.* 468, 173-182.
- NIELSEN, R. (2002). Mapping mutations on phylogenies. *Syst. Biol.* 51, 729-739.

- PEDERSEN, J. L., AND JENSEN, J. L. (2001). A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.* 18, 763-776.
- REYES, A., GISSI, C., PESOLE, G., AND SACCONI, C. (1998). Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol. Evol.* 15, 957-66.
- ROBINSON, D.M., JONES, D.T., KISHINO, H., GOLDMAN, N., AND THORNE, J.L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.* 20, 1692-704.
- SCHMITZ, J., OHME, M., AND ZISCHLER, H. (2000). The complete mitochondrial genome of *Tupaia belangeri* and the phylogenetic affiliation of scandentia to other eutherian orders. *Mol. Biol. Evol.* 17, 1334-43.
- SCHMITZ, J., OHME, M. AND ZISCHLER, H. (2002). The complete mitochondrial sequence of *Tarsius bancanus*: evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. *Mol. Biol. Evol.* 19, 544-553.
- SUEOKA, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.* 40, 318-325.
- SWOFFORD, D. L. (2001). *PAUP**. Phylogenetic Analysis Using Parsimony (*And Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- TANAKA, M., AND OZAWA, T. (1994). Strand asymmetry in human mitochondrial DNA mutations. *Genomics* 22, 327-335.
- THOMPSON, J. D., HIGGINS, D.G., AND GIBSON, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nuc. Acids Res.* 22, 4673-4680.
- THORNE, J.L., AND KISHINO, H. (2002). Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* 51, 689-702.
- XU, X., AND ARNASON, U. (1996). A complete sequence of the mitochondrial genome of the western lowland gorilla. *Mol. Biol. Evol.* 13, 691-698.
- YANG, Z. (1994). Estimating the patterns of nucleotide substitution. *J. Mol. Evol.* 10, 1396-1401.
- YANG, M.Y., BOWMAKER, M., REYES, A., VERGANI, L., ANGELI, P., GRINGERI, E., JACOBS, H.T., AND HOLT, I.J. (2002). Biased incorporation of ribonucleotides on the mitochondrial L-strand. *Cell* 111, 495-505.
- ZUKER, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nuc. Acids Res.* 31, 3406-3415.

CHAPTER 4: SITE-SPECIFIC MODELS: A COMPLETE POSTERIOR-PREDICTIVE APPROACH

4.1. Overview

The previous chapter described use of hidden Markov Models (HMMs) to model detailed site-specific substitution responses in mitochondria and the substitution responses of all the twelve possible nucleotide substitutions. In this chapter, we consider different sub-groupings of primate species for additional sequence datasets such as two-fold redundant purines and pyrimidines and demonstrate the differences in their corresponding substitution responses. We also describe another approach, where the response model uses a linear equation with *slope* and *intercept* parameters. An MCMC chain run on these parameters in addition to the model parameters determines their posterior distributions. Similar to the HMMs approach, this approach uses a *base* model with symmetric substitution probabilities and asymmetry of a specific substitution type is added proportionally to single-strandedness. There are considerable differences in the details of the C→T response between two primate sub-groups with different A→G response slopes using both the linear models as well as the HMMs. Simulations were performed under two different models to evaluate the credibility in responses observed for the two methods.

While Chapter 3 was an exploratory analysis which used the ancestral sequence distribution obtained from a simpler model to evaluate more complex site-specific models easily, in this chapter the ancestral states were updated by a similar Gibbs Sampling scheme by using the complex models (Chapter 2; Chapter 3; Krishnan et al., *in press*). This was mainly done in order to obtain comparable likelihoods. A full

posterior predictive approach is considered here where we test for significance of the observed results (Gelman et al., 1996). Sequences were simulated under various null models and distributions of various test statistics such as Likelihood Ratio Test to evaluate the significance of the fit of the model to the data. Other criteria such as Schwarz criterion and Bayes Information criterion (BIC) could also be examined and significance levels compared.

4.2. Partitioning of “Species” and “Sequence” Datasets

Individual sequence datasets were obtained after partitioning the complete mitochondrial sequence alignments into those that contain only the variable third codon positions of (1) conserved two-fold redundant purines (2 X R3), (2) conserved two-fold redundant pyrimidines (2 X Y3), (3) conserved four-fold redundant codons, (4 X 3) and (4) a modified version of the third dataset where the A's and G's were converted to R's and the C's and T's were converted to Y's (4 X RY3). Since the transversions are more likely to be observed in the more variable third codon positions of the four-fold redundant codons, by analyzing this dataset any asymmetry in transversions type substitutions in general, along the genome due to mitochondrial replication could be profiled. Chapter 2 discussed the relative rate response profiles of all eight specific transversions.

For all the sequence datasets mentioned above, the corresponding species dataset of 18 species was separated into three groups: (1) consisting of the *complete* group of all eighteen species, (2) consisting of eleven species (*Cercopithecus aethiops*, *Cynocephalus variegatus*, *Gorilla gorilla*, *Homo sapiens*, *Hylobates lar*, *Macaca sylvanus*, *Pan paniscus*, *Pan troglodytes*, *Papio hamadryas*, *Pongo pygmaeus*, and

Pongo pygmaeus abelii) with higher G/A slopes and intercepts, referred to as the *high* group and (3) consisting of seven species (*Cebus albifrons*, *Colobus guereza*, *Lemur catta*, *Nycticebus coucang*, *Tarsius bancanus*, *Trachypithecus obscurus*, and *Tupaia belangeri*) with lower G/A slopes and intercepts, referred to as the *low* group.

The basis for classifying these species into high and low sub-groups was obtained from a method that developed linear models for estimating slope and intercept parameters for G/A frequency gradients across genomes of individual species (Krishnan et al., 2004; Krishnan et al., *in review*; Raina et al., *in review*) and provided 95% confidence intervals for these estimators. Phylogenies were built separately for the high and low sub-groups and analyses on all the four sequence datasets were performed for these two sub-groups as well.

4.3. Linear Models

The strand-symmetric model is chosen to serve as the independent base model for the dependent site-specific complex models. The variable site-specific asymmetric component to be added to the symmetric rates of any specific substitution was a linear asymmetric component proportional to the D_{ssH} . For each site of the dataset, a linear component was added to the symmetric rate of a specific substitution type. This linear component (c'_p) consisted of *Slope* and *Intercept* as free parameters and was proportional to the D_{ssH}^p of that site “*p*” as shown in equation 1.

$$c'_p = (D_{ssH}^p \times Slope) + Intercept \quad (1)$$

A Markov chain (MCMC) is run on the slope and intercept parameters and 95% confidence intervals are evaluated from the posterior distributions of these parameters.

Parameter priors and proposal distributions were calculated as described in Chapter 2.

4.4. MCMC Analyses with Linear and HMM Models on Simulated Data

Sequence datasets were simulated by adding asymmetry under two different models as follows:

(1) *Linear*: The added asymmetry to the symmetric model rate parameter for A→G rate parameters was a linearly increasing component that was proportional to the D_{ssH} at that site. It followed the equation: $asymmetry = 0.12124 D_{ssH} + 0.1157$ and

(2) *Asymptote*: The added asymmetry to the symmetric model rate parameter for C→T rate parameters had two components: a linearly increasing component until $D_{ssH} = 0.3023$ that was proportional to the D_{ssH} at that site. It followed the equation: $asymmetry = 3.256 D_{ssH} - 0.0122$, and a constant with value equal to -0.9582 added to the symmetric model rate parameter for C→T rate parameters for the remaining D_{ssH} values.

For determining a suitable linear or asymptote model, the posterior mean relative responses from HMM analyses on 4 X 3 dataset were used. The remaining parameters were determined from the posterior means of the four-by-four symmetric model. The phylogeny corresponding to the complete mitochondrial genomes for all the eighteen species was used. Under each model, simulations were performed by starting at the deepest node, with the average of all inferred ancestral node frequencies at a site used for that site for all the internal branches and the nearest tip frequencies for the external branches. The rate parameters were maintained constant along the branches for respective sites.

MCMC analyses were performed on these eighteen simulated sequences using the linear as well as the HMM models. In case of the linear models, the biases were estimated as differences between observed posterior means of slope (\hat{s}) and intercept (\hat{i}) parameters and the slope (s) and intercept (i) of the equation of the line used as a model to simulate the sequences. For the HMM, the bias was estimated for each site as the difference between the observed posterior mean of the substitution rate ($\hat{\lambda}$) and the substitution rate at that site from the model (λ) that was used to simulate sequences. The total bias in the site-specific model evaluation was evaluated by calculating the mean squared error (MSE):

$$MSE = (\hat{\lambda} - \lambda)^2 \quad (4)$$

4.5. Results

4.5.1. Posterior Means and 95% Confidence Intervals of Slope and Intercept Parameters and Log-likelihoods Under the Linear Model for All Transitions

The posterior means and 95% confidence intervals on the slope and intercept parameters of the linear model are shown in Table 4.1. These were obtained from 24 independent runs using three sequence datasets: 2 X R3 for G→A and A→G substitutions, 2 X Y3 for C→T and T→C substitutions and 4 X 3 for G→A, A→G, C→T, and T→C substitutions for each of three species groups (complete, high, and low).

A→G and C→T transitions always have a positive slope, whereas G→A and T→C transitions always have a weaker negative slope. We do not know the biological implications of a decreasing substitution response vs. D_{ssH} across the genome as indicated by the negative slope. It could be a residual effect of the increasing

substitution rates along the D_{ssH} gradient of $A \rightarrow G$ and $C \rightarrow T$ substitutions.

Table 4.1. Posterior Means and 95% Lower and Upper Confidence Intervals (within the square brackets) of slope and intercept parameters for the four substitution types ($G \rightarrow A$, $A \rightarrow G$, $T \rightarrow C$, and $C \rightarrow T$) obtained by performing MCMC analyses using the linear models. Results are shown for three sequence datasets: conserved two-fold purines and pyrimidines (collectively abbreviated as 2 X 3), and conserved four-fold redundant third codon positions) and species groups (complete, high, and low). For $C \rightarrow T$ and $T \rightarrow C$ substitutions, 2 X R3 datasets were used, while for the remaining substitutions, $A \rightarrow G$ and $G \rightarrow A$, 2 X Y3 datasets were used.

	2 X 3		4 X 3	
	Slope	Intercept	Slope	Intercept
$G \rightarrow A^C$	-0.11801 [-0.11802, -0.11801]	-0.00036 [-0.00036, 0.00036]	-0.05988 [-0.05990, 0.05986]	-0.000335 [-0.000337, 0.000333]
$G \rightarrow A^H$	-0.11346 [-0.11348, -0.113432]	0.000506 [0.000503, 0.00051]	-0.05282 [-0.05285, 0.05279]	-0.00136 [-0.00136, 0.00136]
$G \rightarrow A^L$	-0.11863 [-0.11864, -0.11863]	-0.00012 [-0.00012, 0.00012]	-0.01606 [-0.01609, 0.01603]	-0.00014 [-0.00014, 0.00013]
$A \rightarrow G^C$	0.17218 [0.17214, 0.17221]	-0.00027 [-0.00027, 0.00027]	0.06253 [0.062485, 0.062575]	-0.00161 [-0.00161, 0.00161]
$A \rightarrow G^H$	0.20214 [0.20211, 0.20216]	-0.00077 [-0.00077, 0.00077]	0.089628 [0.089581, 0.089675]	-0.00165 [-0.00165, 0.00165]
$A \rightarrow G^L$	0.13457 [0.13432, 0.13482]	0.000506 [0.000503, 0.00051]	0.039318 [0.039256, 0.03938]	-0.00027 [-0.00027, 0.00027]
$T \rightarrow C^C$	-0.11716 [-0.11722, -0.1171]	0.00145 [0.001448, 0.001452]	-0.02794 [-0.02794, 0.02794]	-0.00136 [-0.00136, 0.00136]
$T \rightarrow C^H$	-0.11988 [-0.11992, -0.11983]	0.00218 [0.002176, 0.002184]	-0.03984 [-0.03984, 0.03983]	-0.00157 [-0.00157, 0.00156]
$T \rightarrow C^L$	-0.10575 [-0.10579, -0.10571]	0.000347 [0.000345, 0.00035]	-0.01603 [-0.01603, 0.01603]	-0.00099 [-0.00099, 0.00098]
$C \rightarrow T^C$	0.2996 [0.29945, 0.29976]	0.000621 [0.000619, 0.000624]	0.06111 [0.06107, 0.06115]	-0.00076 [-0.00076, 0.00076]
$C \rightarrow T^H$	0.3195 [0.3196138, 0.31961]	0.000866 [0.000864, 0.000868]	0.08631 [0.08625, 0.08637]	0.001916 [0.001915, 0.001917]

^H **High group:** *Cercopithecus*, *Cynocephalus*, *Gorilla*, *Homo*, *Hylobates*, *Macaca*, *Pan*, *Papio*, *Pongo*; ^L **Low group:** *Cebus*, *Colobus*, *Lemur*, *Nycticebus*, *Tarsius*, *Trachypithecus*, *Tupaia*; ^C **Complete group:** all the eighteen species combining high and low groups.

The slopes and absolute values of intercepts are higher for the high group, intermediate for the complete group and least for the low group for all the substitutions, suggesting a kind of averaging effect between the responses of the high and low groups for the complete group. Also, this fits with observations from previous analyses for

A→G substitutions, based on which the high and low groups were originally identified.

While there are substantial differences between the posterior means of slopes and intercepts of corresponding substitutions within the 2 X 3 and 4 X 3 datasets, the magnitudes of the slopes are in general, much higher for the 2 X 3 dataset as compared to the 4 X 3 dataset. This may as well be resulting from the differences in the sizes of each of these datasets (2 X Y3: 318, 2 X R3: 542 and 4 X 3: 920 nucleotides respectively). The corresponding differences in the magnitudes of the intercepts is comparatively less extreme, and for some substitutions ($G \rightarrow A^C$, $A \rightarrow G^L$, and $C \rightarrow T^C$), the intercepts are even higher for the 4 X 3 dataset.

The likelihoods for the linear models proposing asymmetry in A→G and C→T transitions are much higher than those for asymmetries in G→A and T→C transitions respectively (Table 4.2). Also, within the 4 X 3 dataset, where the same dataset was used to study asymmetry in different substitutions, the likelihood is higher for C→T than A→G substitutions. This difference is about 80 log-likelihood units for the complete group and ~25 log-likelihood units for the high and low groups.

4.5.2. Relative Substitution Rate Responses Profiled Versus Time-spent Single-Stranded in the Genome for the Various “Sequence” and “Species” Partitions

The almost linear relative rate responses for A→G substitutions are steeper for high group, shallower for the low group, and intermediate for the complete group (Figures 4.1.A and 4.1.C), with slight differences between the magnitudes of the responses in the 2 X R3 and 4 X 3 datasets. The details of the C→T response, which in general show a quick increase followed by a saturation for the rest of the genome, are intricate and there are considerable differences between the responses of the various groups and in

the 2 X Y3 and 4 X 3 datasets (Figures 4.1.B and 4.1.D). The slope of the increasing portion of the response varies a lot among the complete, high and low groups, but is higher for the high group and least for the low group.

Table 4.2. Maximum Likelihood Estimates and 95% Confidence Intervals of Log-Likelihoods for MCMC Analyses Run Using Linear Models for the Four Transitions. Results are shown for three sequence datasets: conserved two-fold purines and pyrimidines (collectively abbreviated as 2 X 3), and conserved four-fold redundant third codon positions) and species groups (complete, high, and low). For C→T and T→C substitutions, 2 X R3 datasets were used, while for the remaining substitutions, A→G and G→A, 2 X Y3 datasets were used

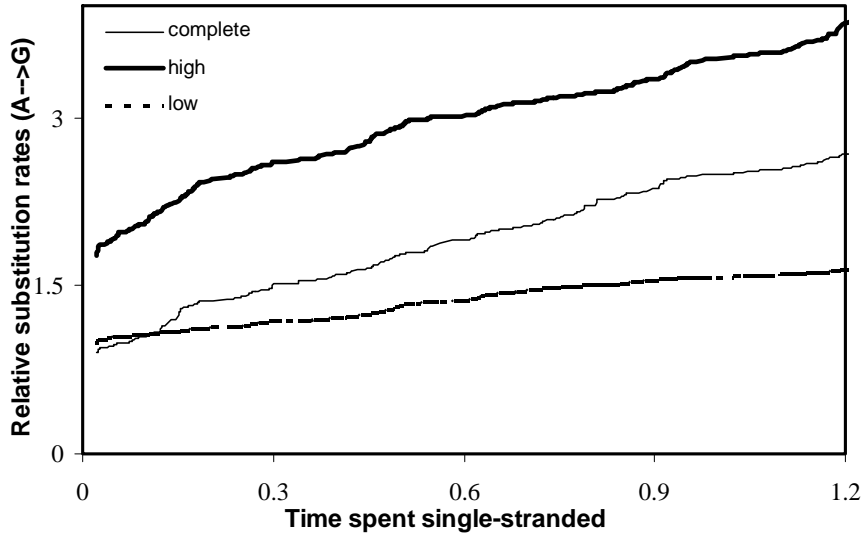
LogLike	2 X 3	4 X 3
G-->A ^C	-3750.85 [-3751.2, -3751.89]	-17183.1 [-17183.1, -17183]
G-->A ^H	-720.3038 [-725.234, -719.134]	-10060.8 [-10060.8, -10060.8]
G-->A ^L	-1887.08 [-1886.23, -1887.93]	-7712.86 [-7712.91, -7712.82]
A-->G ^C	-3507.71 [-3507.9, -3507.53]	-17077.35 [-17077.39, -17077.31]
A-->G ^H	-750.304 [-750.304, -750.304]	-10019.3 [-10019.3, -10019.3]
A-->G ^L	-1828.09 [-1828.17, -1828.01]	-7759.58 [-7759.59, -7759.58]
T-->C ^C	-1042.14 [-1042.39, -1041.88]	-17296 [-17296, -17295.9]
T-->C ^H	-714.442 [-714.615, -714.27]	-10176.4 [-10176.4, -10176.4]
T-->C ^L	-382.999 [-383.083, -382.915]	-7814.97 [-7814.99, -7814.95]
C-->T ^C	-771.128 [-771.256, -771]	-16998.7 [-16998.8, -16998.7]
C-->T ^H	-493.972 [-494.042, -494.902]	-9994.17 [-9994.18, -9994.16]
C-->T ^L	-276.635 [-276.663, -276.607]	-7733.31 [-7733.31, -7733.3]

^H **High group:** *Cercopithecus*, *Cynocephalus*, *Gorilla*, *Homo*, *Hylobates*, *Macaca*, *Pan*, *Papio*, *Pongo*; ^L **Low group:** *Cebus*, *Colobus*, *Lemur*, *Nycticebus*, *Tarsius*, *Trachypithecus*, *Tupaia*; ^C **Complete group:** all the eighteen species combining high and low groups.

The ratio between the slopes of high and low groups for this increasing region of the C→T response curve is almost equal at ~3.56 for the 2 X Y3 and 4 X 3 datasets. The saturation levels of C→T responses among the three “species” groups also vary more for the 2 X Y3 than for the 4 X 3 dataset. The relative rate responses for transversions (R→Y and Y→R) do not show any particular trend of response to D_{ssH} and seem to

average around ~ 1.14 and ~ 1.126 respectively (Figures 4.1.E and 4.1.F).

4.1. A



4.1.B

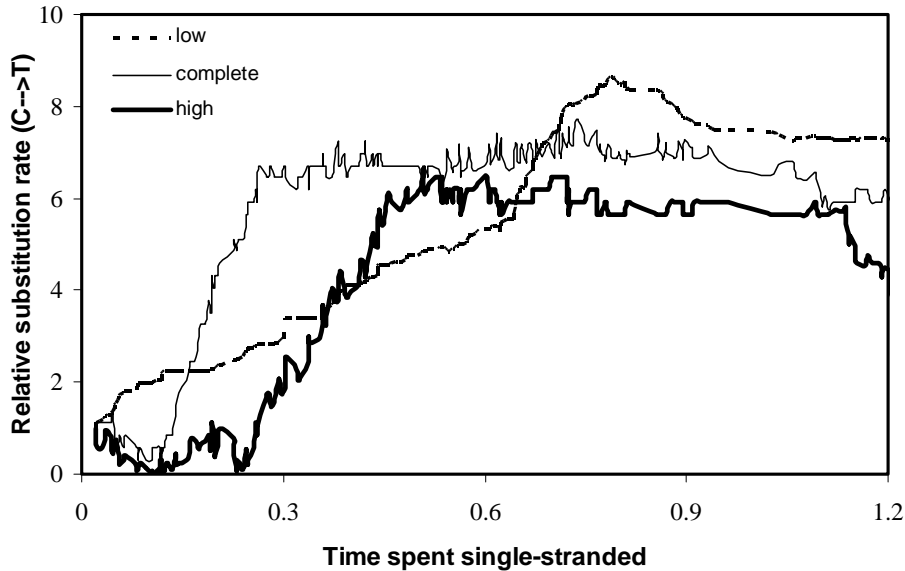
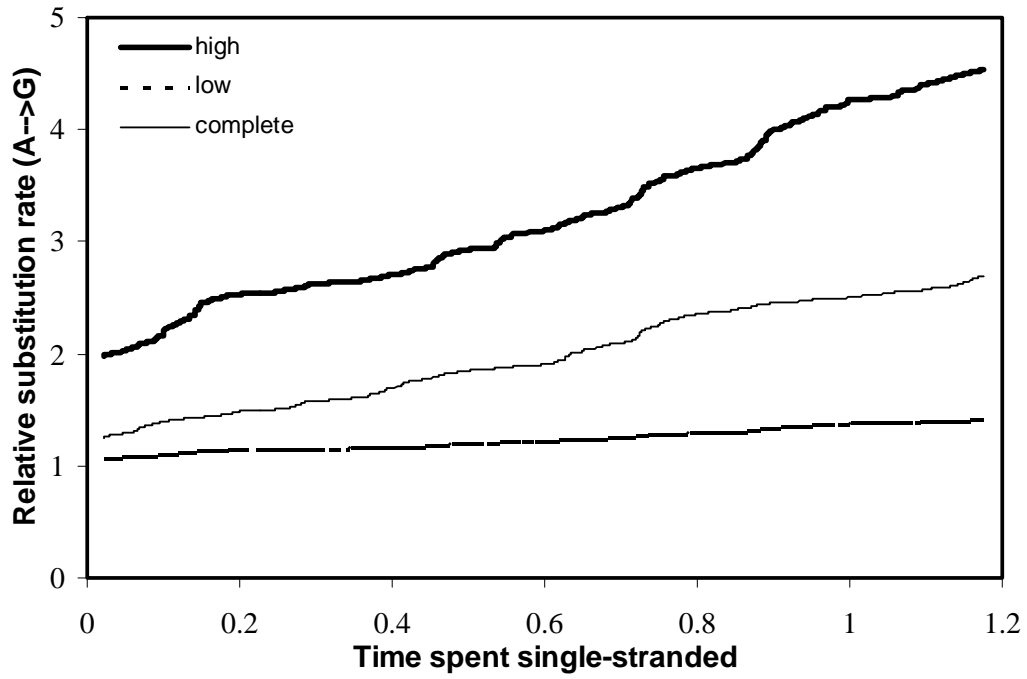


Figure 4.1 Detailed Site-Specific Relative Substitution Rate Responses from the HMM Profiled Vs. Time Spent Single-Stranded for the Various Species Sub-Groups (Complete, High, and Low), and Sequence Datasets. Responses are shown for substitutions of the type: A. $A \rightarrow G$ (2 X R3), B. $C \rightarrow T$ (2 X Y3), C. $A \rightarrow G$ (4 X 3), D. $C \rightarrow T$ (4 X 3), E. $R \rightarrow Y$ (4 X 3RY), and F. $Y \rightarrow R$ (4 X 3RY). For each of the six cases mentioned above, 18 independent MCMC runs using the hidden Markov model were performed on all three species groups.

4.1.C



4.1.D

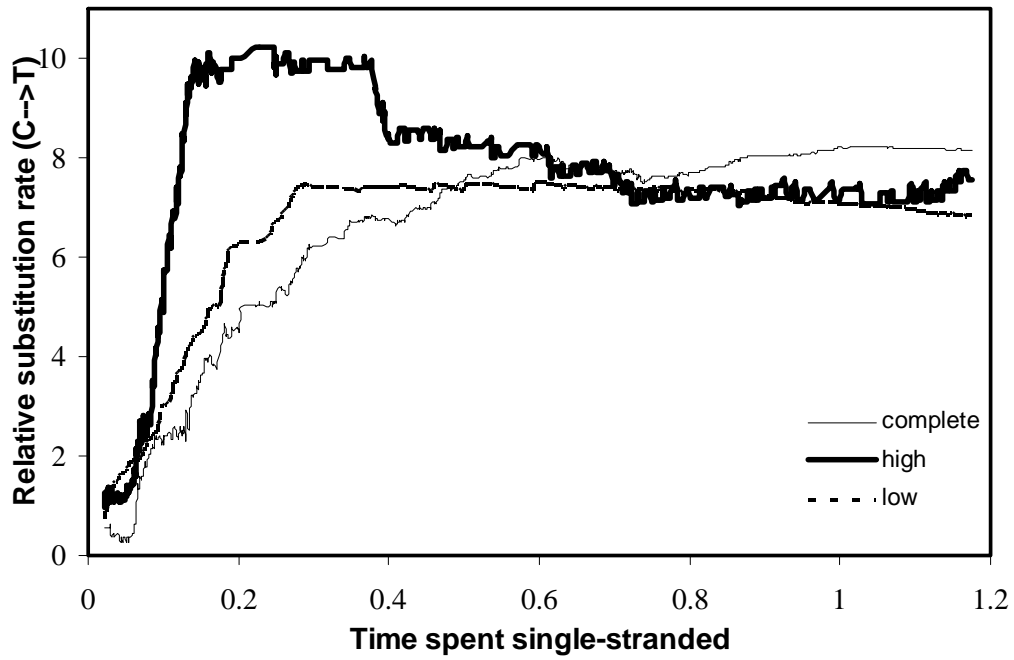
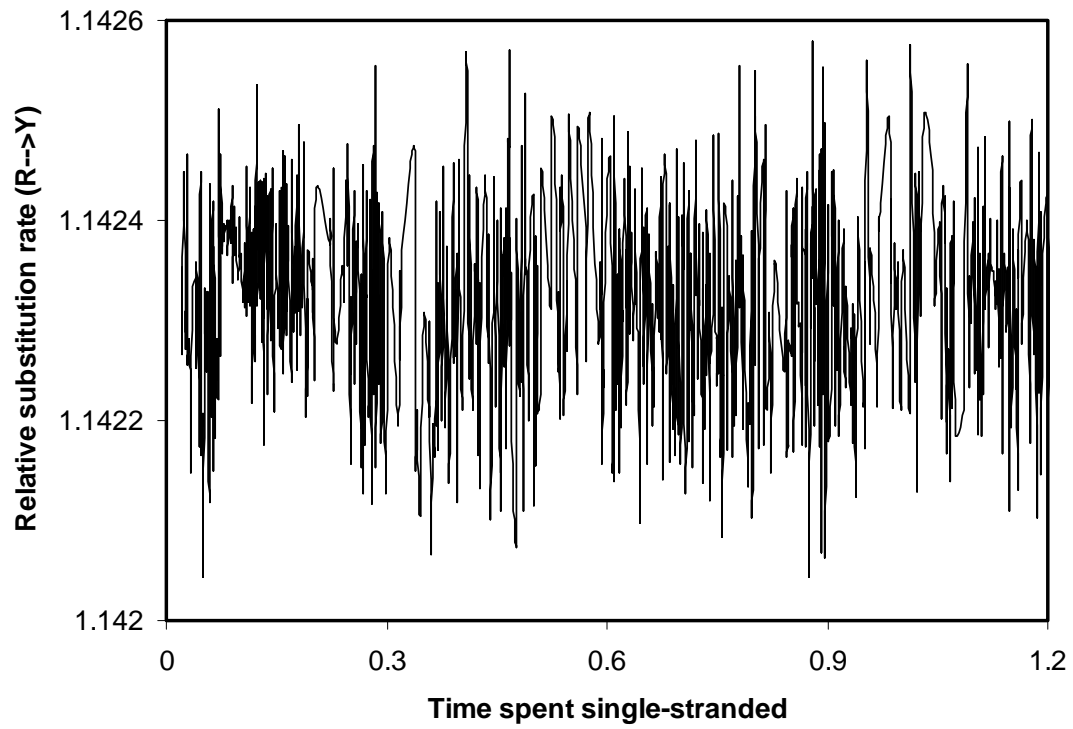


Figure contd...

4.1.E



4.1.F

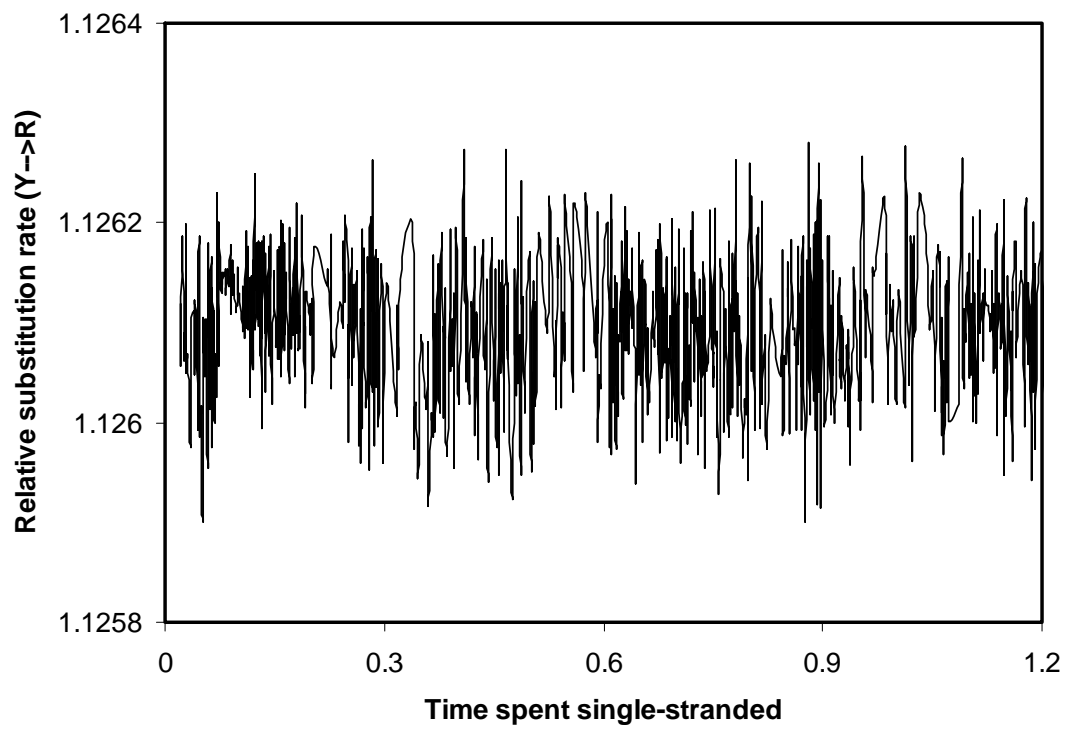


Figure contd..

Table 4.3 Bias and Mean Square Errors (MSEs) in Estimating the Slope and Intercept Parameters for C→T and A→G Substitutions Using the Linear Models.

The bias was calculated as deviation from the expectation of the model under which the sequences were simulated. Simulations were performed using the phylogeny of the complete group of eighteen species under two models: (1) A→G rate parameters increasing linearly among sites, according to the equation of the average linear response from the HMM ($y = 0.12124x + 0.1157$), and (2) C→T rate parameters increasing linearly until $D_{ssH} = 0.3023$ ($y = 3.256x - 0.0122$) and then remaining constant for remaining D_{ssH} values ($y = -0.9582$). This was an approximation to model the asymptotic response of C→T substitutions, with two lines based on the average asymptotic response from the HMM.

	Bias_{Slope}	Bias_{Intercept}	MSE_{Slope}	MSE_{Intercept}
A→G^C	0.00227	-0.00155	0.0000052	0.0000024
C→T^C	0.00461	-0.00041	0.0000213	0.0000002

^C **Complete group:** consists of all the eighteen species: *Cercopithecus*, *Cynocephalus*, *Gorilla*, *Homo*, *Hylobates*, *Macaca*, *Pan*, *Papio*, *Pongo*, *Cebus*, *Colobus*, *Lemur*, *Nycticebus*, *Tarsius*, *Trachypithecus*, *Tupaia*

4.5.3. Methodological Bias Estimation After Simulating Data Under “Linear” and “Asymptote” Models

Table 4.3 shows the bias in estimating slope and intercept parameters for A→G substitutions using sequences simulated under the “Linear” model and C→T substitutions using sequences simulated under the “Asymptote” model. The relatively small mean squared errors (MSEs) indicate that there is a very small bias. The bias in estimating slopes is positive for the two substitutions but is almost twice the amount for A→G as for C→T substitutions. The site-specific bias and MSE profiles after performing HMM analyses are shown in Figure 4.2. These site-specific biases are calculated as differences between the observed substitution rates relative to the symmetric model and the expected relative substitution rates. For A→G substitutions, there is a uniform, linearly increasing bias ranging between -0.0025 and 0.0075, probably introduced by the HMM method. However, for the C→T substitutions, the

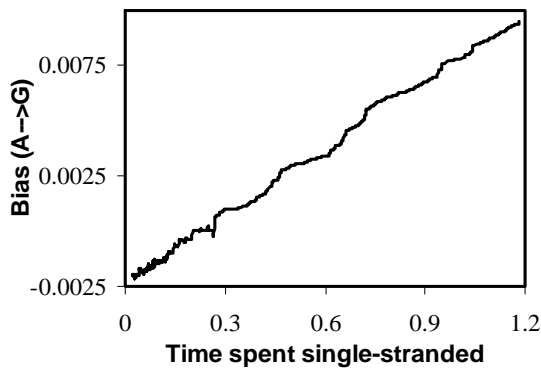
bias increases more steeply from -0.26 to 0.0 until $D_{ssH} \sim 3$, after which it increases steadily with a lesser slope up to 0.06 for the remaining D_{ssH} values. The magnitude of the bias in HMM analyses of asymmetry in C→T substitutions is at least 10 times more than that for the A→G substitutions.

4.6. Posterior Predictive Analyses

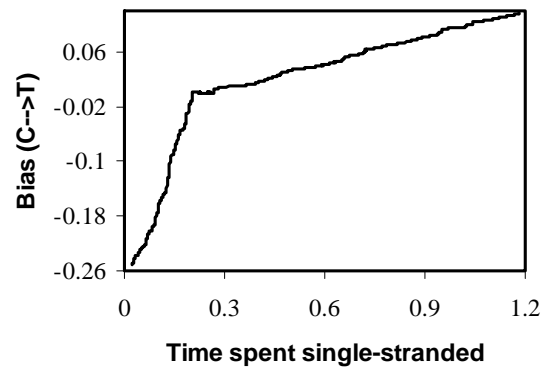
To assess the confidence level in the predicted results, 100 simulations were performed assuming the symmetric model (M1) as well a site-specific model (M2) with constant asymmetry. A schematic representation explaining the differences between a symmetric model (M1), site-specific models with constant (M2) and variable (M3) asymmetry is shown in Figure 4.3. For sequences simulated under the symmetric model, MCMC analyses were performed using the constant asymmetry model (model M2 in Figure 4.1) and for sequences simulated under the constant asymmetry model, MCMC analyses were performed under the variable asymmetry model (Linear [M3] and HMM models). The likelihood ratio test (LRT) statistic was evaluated as twice the log-likelihood difference between the null model and the complex model with asymmetry of a specific substitution type and corresponding distributions of LRT were examined (Figure 4.4). These likelihood ratios were calculated as the difference between posterior averages of the log-likelihoods under the two models considered. The distribution of the LRT statistic for the constant asymmetry model with the symmetric model as the null model shows that there is indeed a significant non-zero intercept for the substitution A→G for the actual data consisting of all 18 species (LRT = 5.654, which is greater than the LRT at 5% level of significance = 3.689 and at 1% level of significance = 4.133). A similar distribution for the variable asymmetry model

as the complex model and the constant asymmetry model as the null model shows that there is a very significant non-zero slope for the A→G substitutions (LRT = 5.781, which is much greater than the LRT at 5% level of significance = 3.511 and 1% level of significance = 3.911). It is important to note that each of these model pairs: M1 and M2, M2 and M3 have a difference of exactly one degree of freedom. According to the chi-square distribution with one degree of freedom, the 5% and 1% levels are 3.84145 and 6.6349, respectively.

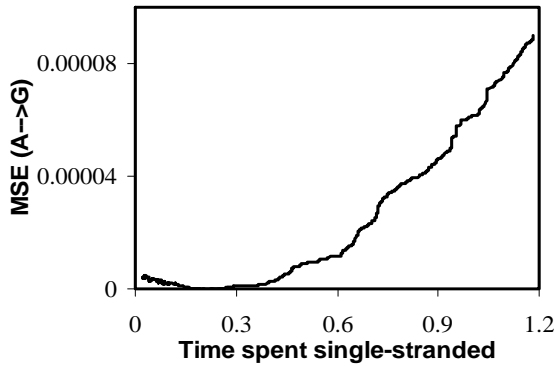
4.2.A



4.2.C



4.2.B



4.2.D

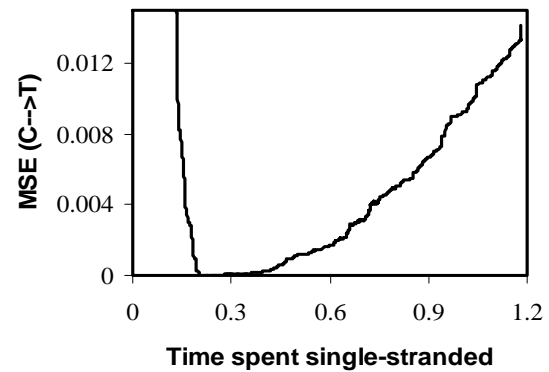
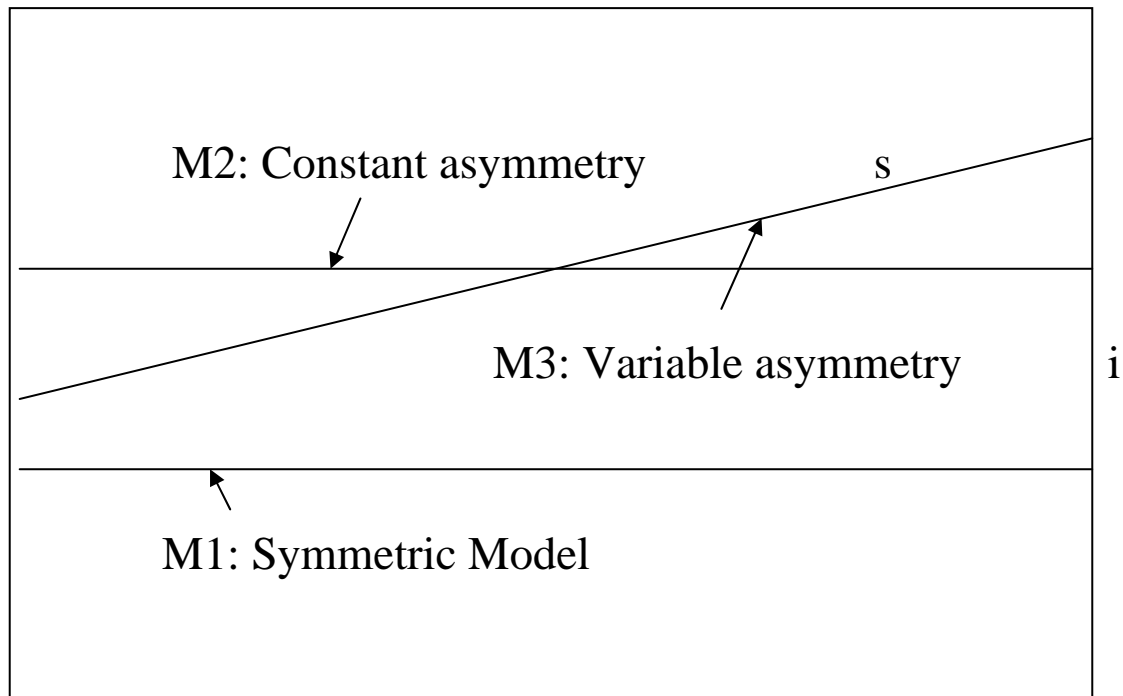


Figure 4.2 Detailed Site-Specific Profiles of the Bias and MSE in Inferring Relative Substitution Rates at that Site Using the HMM Model. For a particular substitution type, the bias was calculated as the difference between the observed relative substitution rate and the expected rate at that site according to the model under which data were simulated. Profiles are shown vs. time spent single-stranded for A. Bias and B. MSE involved in estimating A→G relative substitution rates, C. Bias and D. MSE involved in estimating C→T relative substitution rates

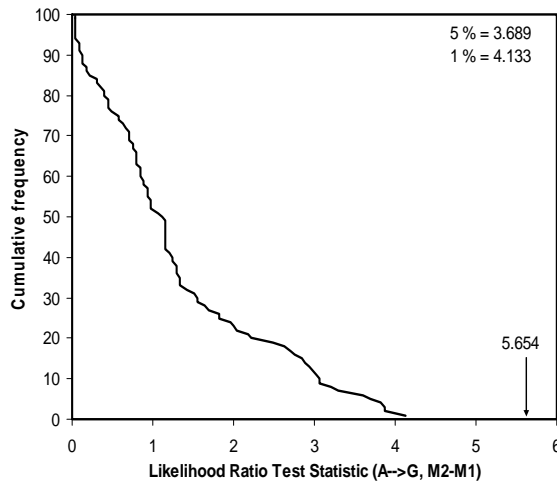


Time spent single stranded

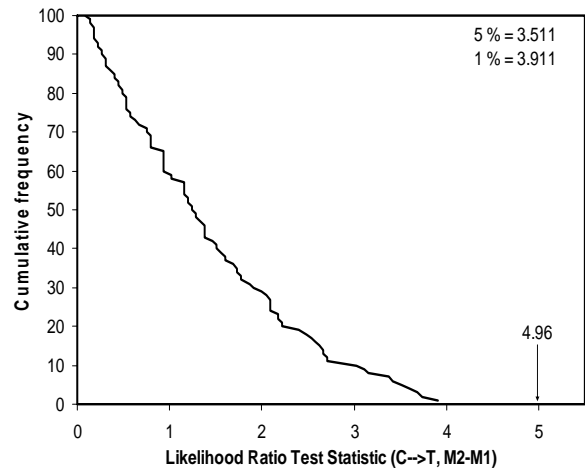
Figure 4.3 Schematic Representations of the Relationships and Differences between Symmetric Model (M1), Site-Specific Models with Constant (M2) and Variable (M3) Asymmetry Plotted Versus the Time Spent Single Stranded. M1 assumes strand-specific symmetry, i.e. complementary substitutions happen with equal probability, M2 is calculated by adding a constant asymmetric component to a particular substitution probability, and M3 is calculated by adding a D_{ssH} -proportional asymmetric component to the same substitution probability.

For the C→T substitutions, while there might be a significant intercept (LRT = 4.96, LRT at 5% = 3.511 and 1% = 3.911), the slope is significant at the 5% level of significance but not at the 1% level of significance. The 5% levels of significance are almost similar for the LRT and chi-square distributions with one degree of freedom. This is expected for nested models. Since these distributions are only from 100 replicates, assessing significance at as far as 1% would need more replicates, perhaps 1000. This might also bring the 1% levels of the LRT distribution closer to that of the chi-square distribution.

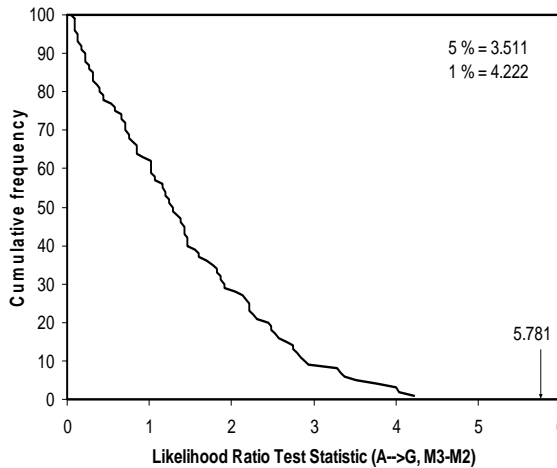
4.4.A



4.4.C



4.4.B



4.4.D

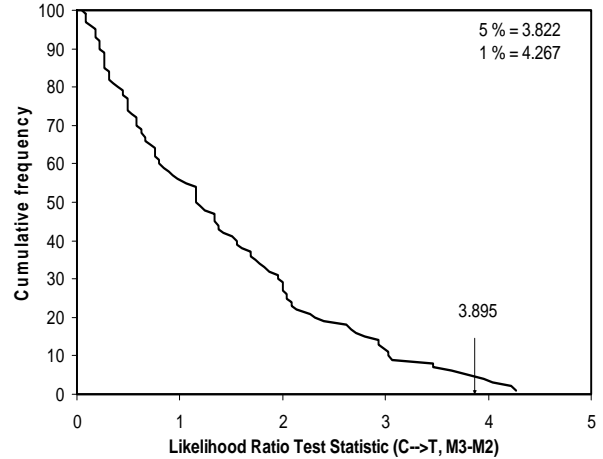


Figure 4.4 Distributions of Likelihood Ratio Test Statistic for Different Substitutions Under Different Null Models (M1 or M2) and Complex Models (M2 or M3). (A) $A \rightarrow G$, null: M1; complex: M2 (B) $A \rightarrow G$, null: M2; complex: M3 (C) $C \rightarrow T$, null: M1; complex: M2 and (D) $C \rightarrow T$, null: M2; complex: M3. The arrows indicate the likelihood ratio test statistic for the real data and each box lists the 5% and 1% levels of significance. Hundred replicates were performed for each set of simulations. For the purpose of comparisons, the 5% and 1% levels of significance according to a chi-square distribution with one degree of freedom are 3.84145 and 6.6349, respectively.

4.7. Literature Cited

GELMAN, A., MENG, X. AND STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 6, 733-807.

KRISHNAN, N. M., SELIGMANN, H., STEWART, C. B., DE KONING, A.P.J., POLLOCK, D. D.

Ancestral sequence reconstruction in primates: Compositional Bias and effect on functional inference. *Molecular Biology and Evolution*, in press.

KRISHNAN, N. M., SELIGMANN, H., RAINA, S. Z., AND POLLOCK, D. D. (2004).

Phylogenetic analysis of site-specific perturbations in asymmetric mutation gradients. *Currents in Computational Molecular Biology*. A. Gramada, and P.E. Bourne eds. Pp. 266-267.

KRISHNAN, N. M., SELIGMANN, H., RAINA, S. Z., POLLOCK, D. D. Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. *DNA and Cell Biology* (*in review*)

RAINA, S., SELIGMANN, H., FAITH, J.J., DISOTELL, T., STEWART, C.B., POLLOCK, D. D. Evolution of base substitution gradients in primate mitochondrial genomes. *in review*

CONCLUSIONS

The conditional pathway methods that make computational simplifications do not dramatically bias results and reflect more biological reality than the conventionally used methods. They also have the potential to carry a lot of complexity into the underlying models of evolution, enabling us to address important biological issues such as: how accurately can we infer information about past evolutionary events, their correlation with geography or paleontological events; how can we infer adaptation or convergence, changes in evolutionary processes or molecular function over time; under what rates do different functional or structural domains of a protein evolve; how better can we understand the mitochondrial replication responses by building site-specific models; what can we tell about protein structure, function, and evolution by evaluating the models at different sites.

APPENDIX A: SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Supplementary Table 1. Chain diagnostics for Cyt-b and COI

Cyt-b	$\bar{\delta}_k$			$\bar{\delta}$	B_N	W_N	$\hat{\sigma}_T^2$
	Chain 1	Chain2	Chain3				
TA	0.011	0.012	0.012	0.01	3.05E-07	6.5E-07	6.52E-07
TT	0.954	0.949	0.949	0.95	5.7E-06	6.9E-05	6.9E-05
TC	0.02	0.024	0.024	0.02	3.22E-06	1.2E-06	1.19E-06
TG	0.015	0.015	0.015	0.01	6.26E-09	8.4E-07	8.42E-07
AT	0.009	0.008	0.009	0.01	2.63E-08	5.1E-07	5.13E-07
AA	0.954	0.947	0.946	0.95	1.19E-05	7.0E-05	6.96E-05
AC	0.006	0.004	0.004	0	9.32E-07	3.4E-07	3.44E-07
AG	0.031	0.040	0.041	0.04	2.07E-05	2.5E-06	2.48E-06
CT	0.016	0.016	0.016	0.02	3.97E-08	1.3E-06	1.31E-06
CA	0.007	0.006	0.006	0.01	5.79E-07	5.9E-07	5.92E-07
CC	0.969	0.972	0.973	0.97	1.98E-06	7.1E-05	7.14E-05
CG	0.007	0.006	0.006	0.01	1.86E-07	5.5E-07	5.51E-07
GT	0.016	0.015	0.014	0.02	7.15E-07	8.3E-07	8.35E-07
GA	0.040	0.044	0.044	0.04	3.46E-06	2.9E-06	2.86E-06
GC	0.006	0.006	0.006	0.01	1.08E-08	3.9E-07	3.91E-07
GG	0.938	0.935	0.936	0.94	1.12E-06	6.8E-05	6.82E-05

COI	$\bar{\delta}_k$			$\bar{\delta}$	B_N	W_N	$\hat{\sigma}_T^2$
	Chain 1	Chain2	Chain3				
TA	0.010	0.009	0.009	0.01	1.81E-07	6.8E-07	6.81E-07
TT	0.957	0.958	0.958	0.96	2.95E-07	0.00038	0.000383
TC	0.018	0.018	0.018	0.02	8.84E-11	5.1E-06	5.05E-06
TG	0.009	0.008	0.008	0.01	3.3E-08	9.7E-07	9.69E-07
AT	0.009	0.008	0.008	0.01	3.30E-08	9.7E-07	9.69E-07
AA	0.951	0.954	0.954	0.95	2.01E-06	0.00041	0.000407
AC	0.005	0.006	0.006	0.01	2.58E-08	6.7E-07	6.68E-07
AG	0.035	0.032	0.032	0.03	1.96E-06	3.4E-05	3.40E-05
CT	0.018	0.017	0.017	0.02	2.75E-07	4.4E-06	4.37E-06
CA	0.006	0.007	0.007	0.01	4.13E-08	7.1E-07	7.10E-07
CC	0.968	0.968	0.968	0.97	4.27E-08	0.00038	0.000377
CG	0.008	0.008	0.008	0.01	1.3E-08	8.0E-07	7.97E-07
GT	0.018	0.018	0.018	0.02	1.49E-09	9.1E-06	9.07E-06
GA	0.036	0.036	0.036	0.04	2.16E-09	5.0E-05	5.00E-05
GC	0.005	0.005	0.005	0	3.81E-09	5.3E-07	5.25E-07
GG	0.941	0.941	0.941	0.94	2.16E-08	0.00045	0.000454

Supplementary Table 2. Bayesian 95% credible intervals for ancestral state frequencies for COI and Cyt-b sequences.

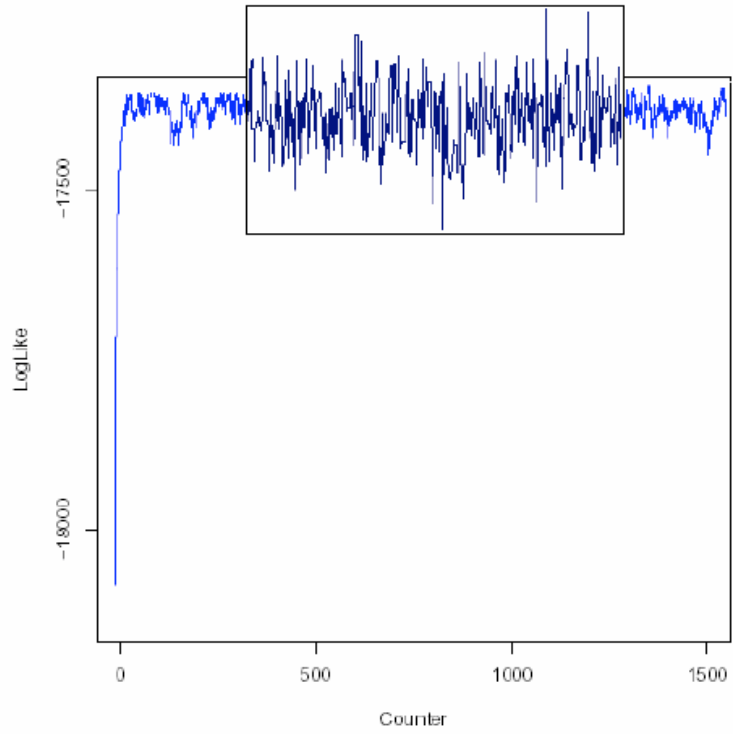
COI, all positions				
Method	T	C	A	G
BL-	[0.277,0.278]	[0.152,0.152]	[0.273,0.275]	[0.294,0.295]
B1	[0.276,0.278]	[0.154,0.155]	[0.282,0.284]	[0.282,0.283]
B2	[0.269,0.270]	[0.164,0.165]	[0.290,0.290]	[0.277,0.278]

COI, 3rd codon positions				
Method	T	C	A	G
BL-	[0.375,0.375]	[0.044,0.044]	[0.254,0.255]	[0.335,0.335]
B1	[0.351,0.352]	[0.061,0.062]	[0.230,0.231]	[0.355,0.355]
B2	[0.351,0.352]	[0.061,0.062]	[0.230,0.231]	[0.355,0.355]

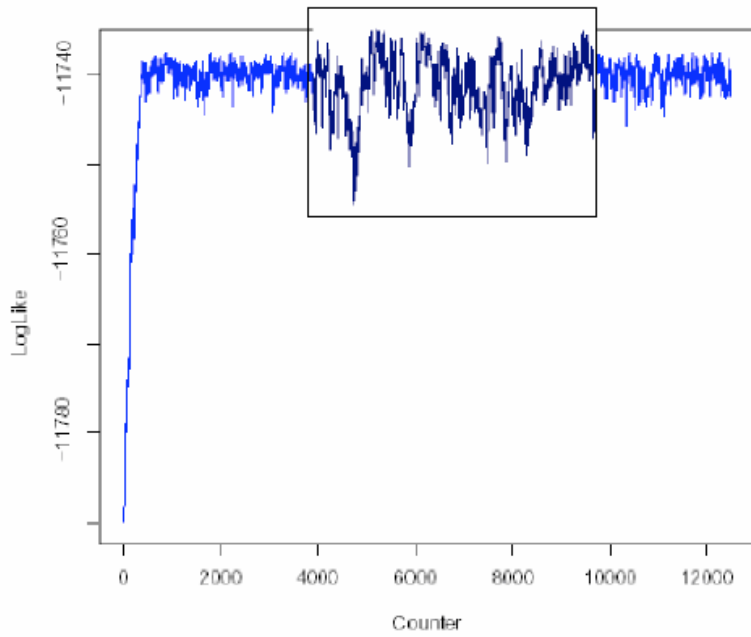
Cyt-b, all positions				
Method	T	C	A	G
BL-	[0.299,0.299]	[0.108,0.109]	[0.264,0.265]	[0.327,0.328]
B1	[0.290,0.291]	[0.119,0.119]	[0.260,0.261]	[0.328,0.328]
B2	[0.290,0.291]	[0.119,0.119]	[0.260,0.261]	[0.328,0.328]

Cyt-b, 3rd codon positions				
Method	T	C	A	G
BL-	[0.391,0.392]	[0.021,0.021]	[0.158,0.158]	[0.437,0.438]
B1	[0.372,0.372]	[0.031,0.032]	[0.157,0.158]	[0.438,0.438]
B2	[0.372,0.372]	[0.031,0.032]	[0.157,0.158]	[0.438,0.438]

Supplementary Figure 1a



Supplementary Figure 1b



APPENDIX B: SUPPLEMENTARY INFORMATION FOR CHAPTER 3

Supplementary Table 1. Posterior Means and Mean Upper and Lower Bounds for 95% Credible Intervals for Variance Parameter α . Posterior averages and 95% credible intervals are calculated for α for the considered sites in the dataset for each of the twelve substitutions, analyzed separately. The table presents the mean of the posterior average and lower and upper 95% credible intervals. The credible range represents the difference between the mean upper and lower bounds and response range is the scale or magnitude of response, difference between the highest and the lowest probability values for all twelve substitutions. The posterior samples for α were obtained off a Markov Chain that was run for a total of 100,000 generations, and sampled every ten generations after discarding the burn-in sample.

Alpha	Mean	Mean 95% lower	Mean 95% upper	Credible Range	Response Range
T-->C ^C	0.0000908	0.0000002	0.0009700	0.0009698	0.072885
C-->T ^C	0.0336243	0.0000003	0.3896753	0.3896750	8.703984
A-->G ^C	0.0016196	0.0000005	0.0186307	0.0186303	1.438235
G-->A ^C	0.0000082	0.0000000	0.0000932	0.0000932	0.007245
C-->G ^C	0.0002765	0.0000006	0.0009310	0.0009304	0.120585
G-->C ^C	0.0000000	0.0000000	0.0000006	0.0000006	4.35E-05
G-->T ^C	0.0000002	0.0000000	0.0000023	0.0000023	0.000174
T-->G ^C	0.0001644	0.0000007	0.0004242	0.0004235	0.073788
A-->C ^C	0.0000413	0.0000000	0.0005860	0.0005859	0.018055
C-->A ^C	0.0000001	0.0000000	0.0000006	0.0000006	4.13E-05
A-->T ^C	0.0001135	0.0000000	0.0003745	0.0003745	0.046726
T-->A ^C	0.0000002	0.0000000	0.0000020	0.0000020	0.000172

^C **Complete group:** consists of all the eighteen species: *Cercopithecus*, *Cynocephalus*, *Gorilla*, *Homo*, *Hylobates*, *Macaca*, *Pan*, *Papio*, *Pongo*, *Cebus*, *Colobus*, *Lemur*, *Nycticebus*, *Tarsius*, *Trachypithecus*, *Tupaia*

VITA

Neeraja Mohan Krishnan was born to Mythili Krishnan and Mohan Krishnan in Warangal, Andhra Pradesh, India, on 27th of July 1980. Most of her upbringing and education was in Bombay (Mumbai), India. She graduated in July 2001 with bachelor degree in computer science from University of Mumbai. She came to Baton Rouge, Louisiana in August 2001 when she started her master's in systems science in Louisiana State University. In August 2002, she enrolled in a dual program with biological sciences and will be graduating with two master's degrees from computer sciences and biological sciences departments in summer 2004. She further intends to pursue a doctoral degree with concentration in computational biology.