

# **EVOLUTION OF BASE SUBSTITUTION GRADIENTS IN PRIMATE MITOCHONDRIAL GENOMES**

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 Interdepartmental Program  
in  
Engineering Science

by

Sameer Raina  
B.S., Louisiana State University, 2001  
August, 2004

## **ACKNOWLEDGEMENTS**

I want to thank Dr. David Pollock for the opportunity to do research in his laboratory. I also want to thank all the members of Pollock Laboratory, especially Dr. Herve Seligmann, Neeraja Krishnan and Judith Beekman. I would also like to extend my sincere gratitude to the other members of my committee, Dr. Subhash Kak and Dr. Ramachandran Vaidyanathan for the interest and support they showed.

## TABLE OF CONTENTS

|                                       |    |
|---------------------------------------|----|
| ACKNOWLEDGEMENTS.....                 | ii |
| LIST OF TABLES.....                   | iv |
| LIST OF FIGURES.....                  | v  |
| ABSTRACT.....                         | vi |
| CHAPTER 1. GENERAL INTRODUCTION.....  | 1  |
| CHAPTER 2. SURVEY OF LITERATURE.....  | 17 |
| CHAPTER 3. MATERIALS AND METHODS..... | 22 |
| CHAPTER 4. RESULTS.....               | 28 |
| CHAPTER 5. DISCUSSION.....            | 42 |
| REFERENCES.....                       | 47 |
| APPENDIX : SUPPLEMENTARY TABLES.....  | 51 |
| VITA.....                             | 58 |

## LIST OF TABLES

|   |    |
|---|----|
| 1. The Vertebrate Mitochondrial Genetic Code.....   | 4  |
| 2. Species used in Study.....   | 22 |
| 3. ML values & CI for parameters of G/A gradients.....  | 29 |
| 4. $\delta\text{LnL}$ between independent and paired analyses of G/A gradient for all pairs.....              | 51 |
| 5. $\delta\text{LnL}$ for hierarchical clustering analyses with G/A gradient.....                             | 52 |
| 6. $\delta\text{LnL}$ , ML values and CIs for parameters for mixture models with G/A gradient.....            | 53 |
| 7. ML values & 95% CI for parameters of C/T gradients.....  | 54 |
| 8. $\delta\text{LnL}$ , MLEs and CIs for hierarchical clustering analyses with C/T gradient.....              | 55 |
| 9. ML values and CIs for parameters of Y/R gradient at 4X sites.....  | 56 |
| 10. $\delta\text{LnL}$ , MLEs and CIs for hierarchical clustering analyses with Y/R gradient at 4X sites..... | 57 |

## LIST OF FIGURES

|  |    |
|--|----|
| 1. Schematic of a Mitochondrion.....   | 6  |
| 2. Schema of a Primate Mitochondrial Genome.....   | 7  |
| 3. Stages in Mitochondrial DNA Replication.....  | 10 |
| 4. Phylogenetic tree of 16 primates and 2 related species.....   | 12 |
| 5. G/A Ratio vs DssH for all species.....  | 30 |
| 6. Important groups from the likelihood-difference based clustering schemes<br>shown on the MLE slopes versus MLE intercepts graph.....          | 32 |
| 7. Posterior probabilities for each species to belong to each model for the<br>five-model mixture.....   | 33 |
| 8. G/A mixture model groups mapped onto the NJ phylogenetic tree.....  | 35 |
| 9. Graph of MLE slopes versus MLE intercepts along with major groups<br>showing a summary interpretation of G/A evolution.....                   | 36 |
| 10. Regression of slope plus intercept for 1 <sup>st</sup> and 2 <sup>nd</sup> codon positions versus<br>the 3 <sup>rd</sup> codon position..... | 39 |
| 11. Likelihood comparison of the most likely trees.....  | 40 |
| 12. Linear regression of G/A intercept and R/Y slope versus gestation time.....  | 45 |

## **ABSTRACT**

The availability of large amounts of genetic data from the mitochondrial DNA of species has created an unprecedented opportunity for the study of evolutionary processes. Being our closest relatives on the evolutionary tree the primates are a prime candidate for the study of evolutionary processes. The availability of large amounts of genetic data from the primates allows us to study and compare results from different phylogenetic reconstruction methods and to study and trace rudimentary evolutionary processes within the primate lineage. The evolutionary process studied here is the response of the nucleotide frequency ratios to single-strandedness of sites during mitochondrial DNA replication. This response curve is shown to be linear where the slope and intercept of the curve are related to the efficacy of the replication mechanisms and the binding capacity of the gamma-polymerase responsible for mitochondrial DNA replication. A Bayesian analysis of the response curves of the species is conducted and clustering schemes are developed to partition the species based on their response curves. These partitions are then mapped on the phylogenetic tree of the species to trace the evolution of the response curve within the primates.

## **CHAPTER 1. GENERAL INTRODUCTION**

### **1.1 Evolution**

Living creatures contain genetic information or a genotype that determines how their bodies or phenotype develop, function and perish. The molecule called DNA (deoxyribonucleic acid) is the carrier of genetic information in living creatures. In eukaryotic cells, or cells that have a nucleus and organelles, the DNA exists within the nucleus. In prokaryotic cells, or cells that don't have a well-defined nucleus, the DNA could exist anywhere within the cell. Eukaryotic cells also have organelles called mitochondria that have their own DNA. Similarly plant cells have chloroplasts which have their own DNA.

The genotype is hereditary and is transferred to the offspring. Sometimes mutations cause changes in the genotype. Most mutations are deleterious and jeopardize the survival of the mutant. Those mutations that allow the mutant to live the course of its life could give rise to novel traits in the phenotype. These traits could be favorable or unfavorable depending on whether they make the mutant better or worse suited to its environment. Natural selection is the edge in survivability of individuals that have favorable traits and the rejection, often extreme, of individuals with unfavorable traits. The new trait could be favorable because it makes the individual stronger, faster, more clever, or simply better looking. We cannot yet predict the phenotypic outcome of a mutation and cannot thus predict the selective advantage or disadvantage a mutation can offer. We can, and do, however, study the rate at which mutations take place and the circumstances in which mutations become more or less probable.

In order to study mutation rates we can either conduct *in vitro* multi-generational experiments to actually witness mutations or take advantage of the existing biodiversity and its broad evolutionary canvass. The trouble with sequencing existing individuals from different species is that they have probably lost most of the unfavorable traits and retain only the favorable ones. When a mutation gets accepted or fixed in a population it is called a substitution. Thus the existing biodiversity can tell us only about substitution rates.

Now if a mutation does not change the survivability of the individual then it is selectively neutral. If the probability of a mutation occurring in an individual is  $\mu$  and the size of the population is  $n$  then the expected number of mutants in the population is  $\mu * n$ .

According to the drift theory of evolution (which applies in scenarios with no selection) if a mutation exists in  $x$  number of individuals in a population of size  $y$  then the

probability of the mutation getting fixed is  $\frac{x}{y}$ . Thus with an expected  $\mu * n$  number of

mutants in a population of size  $n$  the probability of the mutation getting fixed, or the

probability of substitution, is  $\frac{\mu * n}{n}$  or  $\mu$ . Therefore, in the absence of selection, the

mutation rate and substitution rate are the same.

As it happens, there are sites in the DNA molecule that are free or nearly free of selection. The explanation for this lies in the genetic code.

The DNA molecule is a long arrangement of units called nucleotides. There are four kinds of nucleotides that are distinguished by the bases they carry. The bases are adenine, cytosine, guanine and thymine. The respective nucleotides, that also contain a phosphate and a sugar group, are called adenosine, cytidine, guanosine and thymidine. Both the

bases and the nucleotides can be represented by their first letters. Structurally speaking the bases fall into the categories of purines (R) and pyrimidines (Y). A and G are purines and C and T are pyrimidines.

The full DNA molecule is double-stranded or it consists of two DNA strands, each a chain of nucleotides with each nucleotide binding with a corresponding nucleotide on the other strand. An A on one strand binds with a T on the other and a C binds with a G.

In the making of proteins the two strands come apart and an mRNA molecule is created with the same base sequence as the gene on the DNA. This process is called transcription.

The mRNA molecule is like the DNA molecule except it is single-stranded and has the base uracil (U) instead of thymine (T). Then the mRNA travels to a ribosome where

tRNAs bind to the bases on the mRNA. A tRNA binds to three bases on the mRNA (called a codon) and has the amino acid corresponding to the codon on the other end of

its structure. As the tRNAs line up with one of their ends binding to the mRNA their other ends build a chain of amino acids that form the protein. This chain ends upon

encountering a stop codon or a codon that signals the end of the protein. The rules that associate the amino acids with the codons are tabulated in the genetic code (Table 1).

In the mitochondrial genetic code shown here there are some amino acids that are coded by four codons and some that are coded by two codons. The number of codons that

encode an amino acid is called the redundancy level of the amino acid. Thus Proline (P) is four-fold (or 4X) redundant. It so happens that when multiple codons code for an

amino acid it is mostly the third codon position that is different between the different codons. Thus if the base in the third codon position of a codon for a 4X amino acid

mutated, the new codon would code for the same amino acid. If the third codon position

of a codon for a 2X amino acid was a purine and it mutated to another purine then the codon would still code for the same amino acid. The same goes for a pyrimidine in a third codon position of a 2X codon.

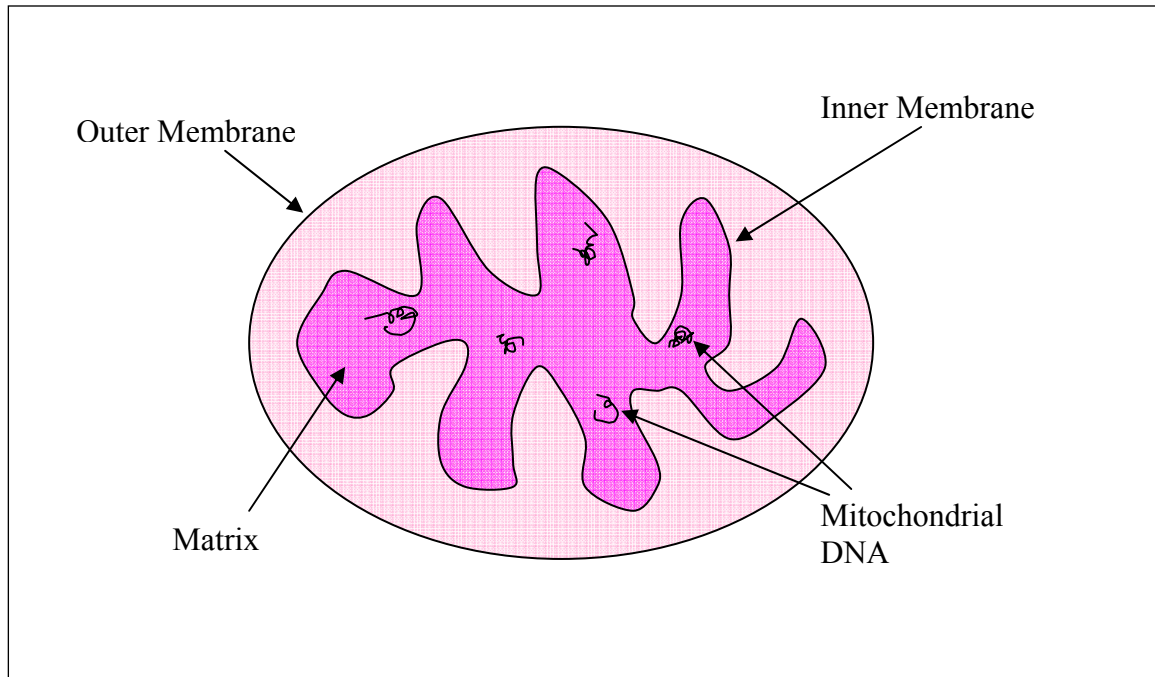
**Table 1.** The Vertebrate Mitochondrial Genetic Code. For every codon the corresponding Amino Acid is shown with the letter representation of the AA in between. The ‘\*’ represents the terminating codon.

|           |           |           |           |
|-----------|-----------|-----------|-----------|
| TTT F Phe | TCT S Ser | TAT Y Tyr | TGT C Cys |
| TTC F Phe | TCC S Ser | TAC Y Tyr | TGC C Cys |
| TTA L Leu | TCA S Ser | TAA * Ter | TGA W Trp |
| TTG L Leu | TCG S Ser | TAG * Ter | TGG W Trp |
| CTT L Leu | CCT P Pro | CAT H His | CGT R Arg |
| CTC L Leu | CCC P Pro | CAC H His | CGC R Arg |
| CTA L Leu | CCA P Pro | CAA Q Gln | CGA R Arg |
| CTG L Leu | CCG P Pro | CAG Q Gln | CGG R Arg |
| ATT I Ile | ACT T Thr | AAT N Asn | AGT S Ser |
| ATC I Ile | ACC T Thr | AAC N Asn | AGC S Ser |
| ATA M Met | ACA T Thr | AAA K Lys | AGA * Ter |
| ATG M Met | ACG T Thr | AAG K Lys | AGG * Ter |
| GTT V Val | GCT A Ala | GAT D Asp | GGT G Gly |
| GTC V Val | GCC A Ala | GAC D Asp | GGC G Gly |
| GTA V Val | GCA A Ala | GAA E Glu | GGA G Gly |
| GTG V Val | GCG A Ala | GAG E Glu | GGG G Gly |

The kinds of mutations that don't change the amino acid coded for are called synonymous or silent-site mutations. The kind of mutations, at any codon position, that would entail a change in the amino acid coded for are called nonsynonymous or replacement mutations. Sites that are free from selection are called neutral sites and those that are somewhat free from selection are called nearly-neutral sites. Since a change in the third codon position of a codon may not be reflected in the amino acid, and thus the protein, it may not have an effect on the phenotype. Thus third codon positions are relatively free of selection. According to the somewhat simplistic picture presented here the third codon positions of codons that code for 4X amino acids are neutral. In reality they are nearly neutral but it is convenient to think of them as neutral because then we can apply the drift theory to these sites.

## **1.2 Mitochondrion**

The mitochondrion is an organelle found in nearly all eukaryotic (ones that have a nucleus surrounded by a membrane and have organelles) cells. A cell could have either a single large mitochondrion or, more often, hundreds or thousands of mitochondria. The mitochondria are the sites of cellular respiration. They take in sugars, fats and other fuels and using oxygen, break them down to generate energy in the form of ATP. They are semi-autonomous organelles with their own DNA that grow and reproduce within a cell. A mitochondrion is about 1 to 10  $\mu\text{m}$  long (Figure 1). Its structure consists of a double-layered envelope that contains the mitochondrial matrix. Most of the enzymes responsible for its function are located in the matrix or are embedded in the inner layer of the

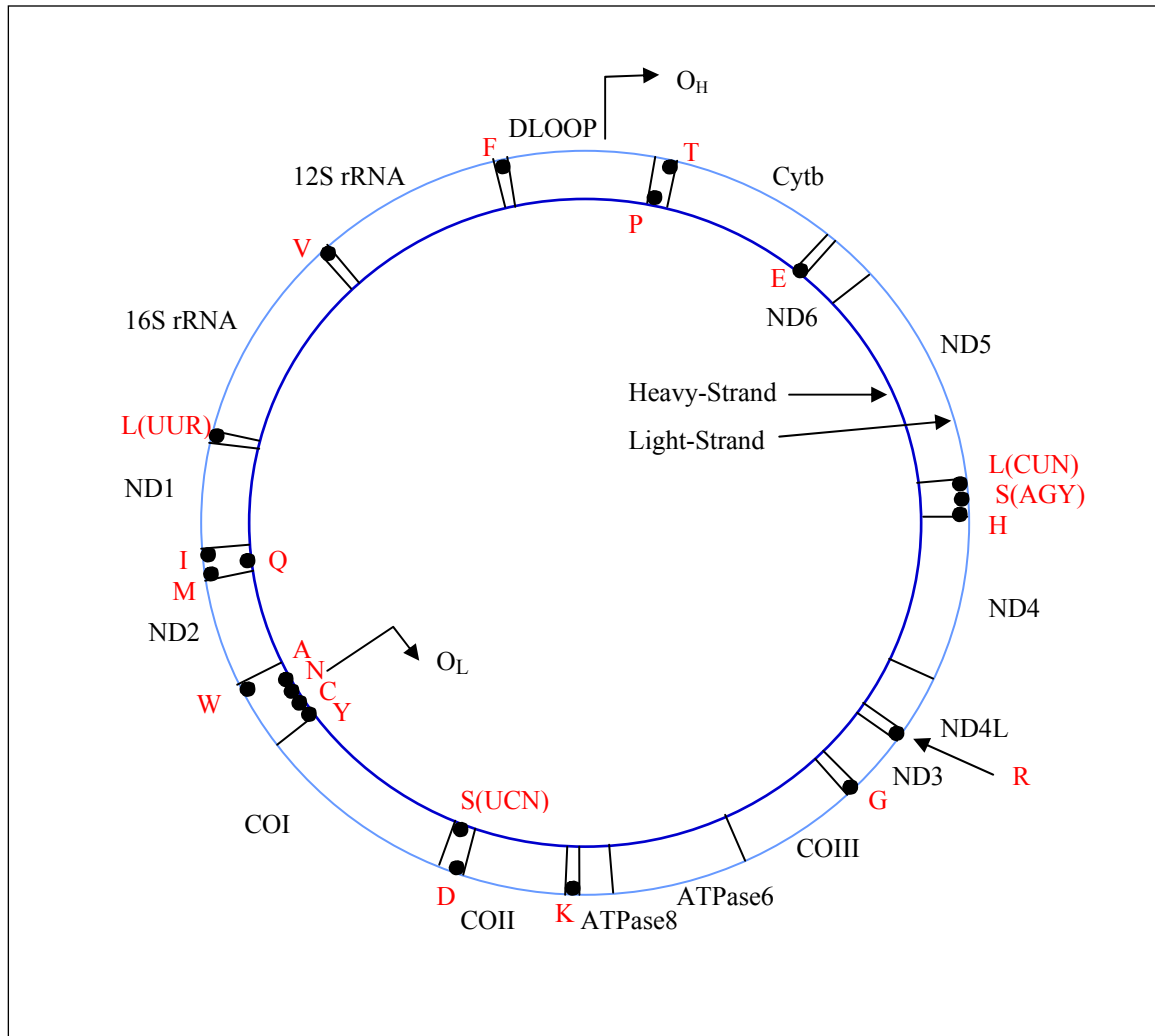


**Figure 1.** Schematic of a Mitochondrion.

envelope. Because of a superficial similarity in the structures of mitochondria and bacteria it was believed that the mitochondria originated as bacteria that lived within eukaryotic cells as symbiotic partners. Later when genes coded by the mitochondria were sequenced and their phylogeny studied they were found to be closer to bacterial genes than to anything else, thus substantiating the belief.

Like bacterial DNA mitochondrial DNA (mtDNA) exists in a circular, double-stranded genome (Figure 2). The two strands are called heavy and light because of imbalanced nucleotide composition in the two strands - the heavy-strand is rich in guanines(G) and the light-strand is rich in cytosines(C). The replication of each strand starts at their respective origins of replication.

There is a lot of variation in the gene composition and arrangement in mitochondrial genomes from different species. This is due to higher mutation rates in mitochondrial



**Figure 2.** Schema of a Primate Mitochondrial Genome. All genes are shown on the strand that they are expressed on. The tRNA genes are represented by the letter representing the corresponding amino acid and are shown in red.

genomes and due to horizontal gene transfers that carry mitochondrial genes into nuclear DNA and subsequently leave the mitochondrial genomes smaller than before. In fact the size of mitochondrial genomes range from an average of around 16,000 base-pairs in animals to 200,000-2,500,000 base-pairs in plants. The gene composition and arrangement within vertebrates are highly conserved though with few gene rearrangements known. This, and the absence of recombination in mitochondrial

genomes causes the vertebrate mitochondrial genomes to be a good candidate for the study of mutation rates. The small size of a vertebrate mitochondrial genome makes them fast and easy to sequence, hence we have many complete vertebrate mitochondrial genomes available to us with a dense sampling of the primates (16 primate genomes) and other closely related species.

Most vertebrate mitochondrial genomes and all primate genomes have the same gene arrangement (Figure 2) with 13 protein coding genes, 12 of which are coded on the light-strand and 1 on the heavy-strand. They also have two genes coding for ribosomal RNAs and 22 genes coding for transfer RNAs or tRNAs. They also have a DLOOP region that contains the origin of heavy-strand replication  $O_H$ . The origin of light-strand replication  $O_L$  lies within the WANCY region that contains five tRNA genes.

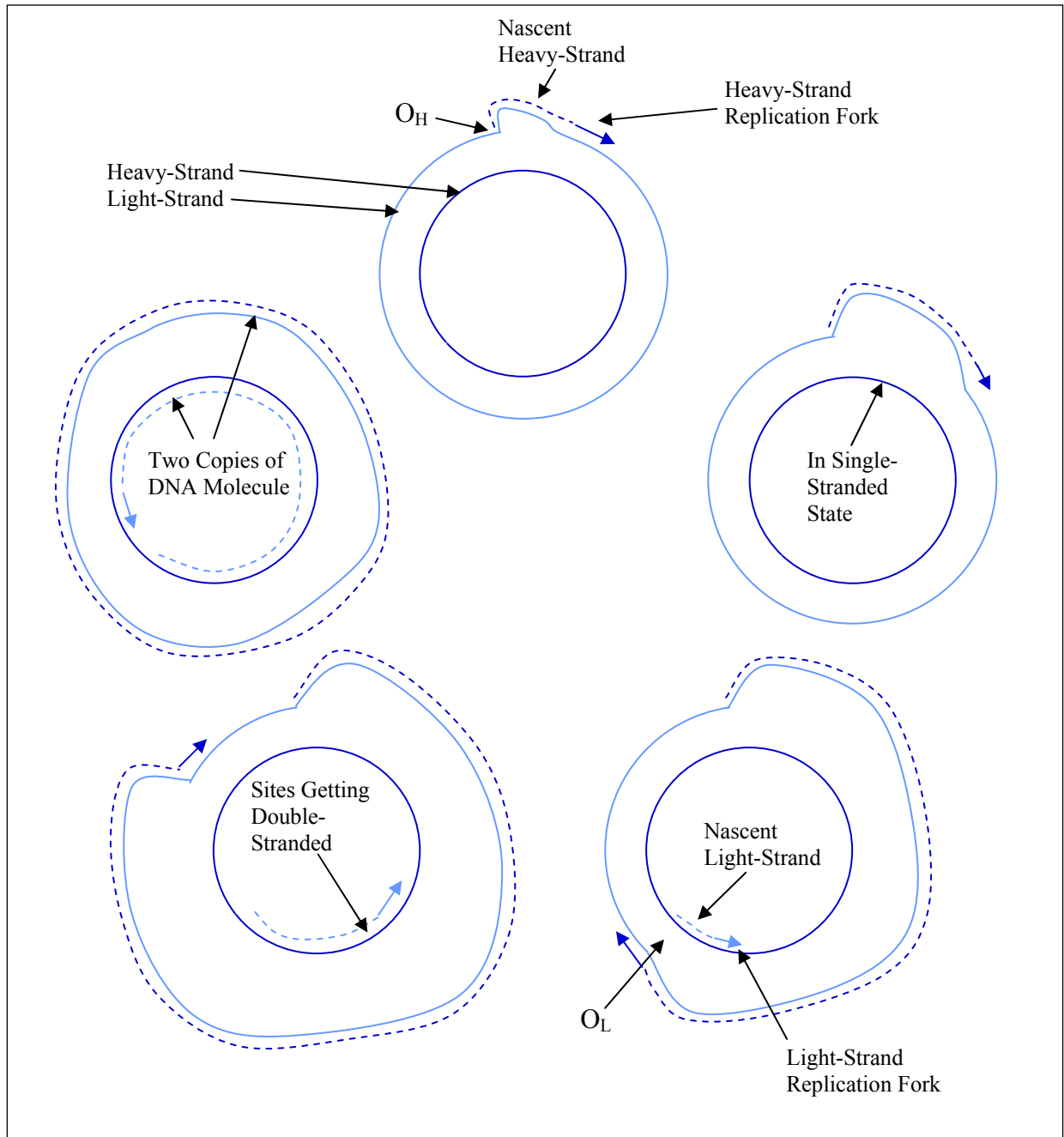
The replication of mtDNA starts when a  $\gamma$ -polymerase binds to the origin of heavy-strand replication ( $O_H$ ). The polymerase then starts constructing a nascent heavy-strand to complement the parental light-strand as it moves in the clockwise direction as shown (Figure 3). As the heavy-strand replication fork moves along it makes the parental heavy-strand single-stranded. When this heavy-strand replication fork passes the origin of light-strand replication ( $O_L$ ) another polymerase binds to the  $O_L$  to start construction of the nascent light-strand to complement the parental heavy-strand. This light-strand replication fork proceeds in the opposite direction than that of the heavy-strand replication fork. The two replication forks are assumed to travel at the same rate. In the course of the replication process different sites along the heavy-strand remain single-stranded for different amounts of time. Given the equal and steady rates of progression of the two replication forks we get a steady gradient in the duration of single-strandedness

for sites on the heavy-strand. This duration is called the Duration of Single-Strandedness of the Heavy-strand or *DssH*.

During the single-stranded state of a site on the heavy-strand of mitochondrial DNA the base at the site is more susceptible to mutations than when the site is in the double-stranded state. The kinds of mutations that change one purine to another or one pyrimidine to another are called transitions. The mutations that change a purine to a pyrimidine or vice versa are called transversions. The mutations that are most likely to occur in the single-stranded state in mitochondrial DNA are deaminations that lead to transitions in the bases. In this state transitions are much more likely to occur than transversions. If we, then, focus our attention on 3<sup>rd</sup> codon positions and assume that only transitions take place then the mutations don't cause the amino acid to change and the sites are free of selection. Thus the drift theory could be applied to these sites and the observed substitution rates can be equated to mutation rates.

### **1.3 Molecular Phylogenetic Reconstruction**

We are witnessing an explosion in sequencing of genetic data and the consequent knowledge of proteins that brings. The biological macromolecules – DNA, RNA and proteins have replaced morphological and paleontological information as the indices to study and measure evolution with. Thus the similarities and differences in these macromolecules either between species or between individuals within a species are now used to determine the evolutionary or phylogenetic relationships and the extent of divergence between them. A taxon (plural : taxa) is defined as a taxonomical unit. In phylogenetic analysis a taxon could be a species or an individual that represents a subset within the species. A taxon could even represent an individual itself. Each of the many



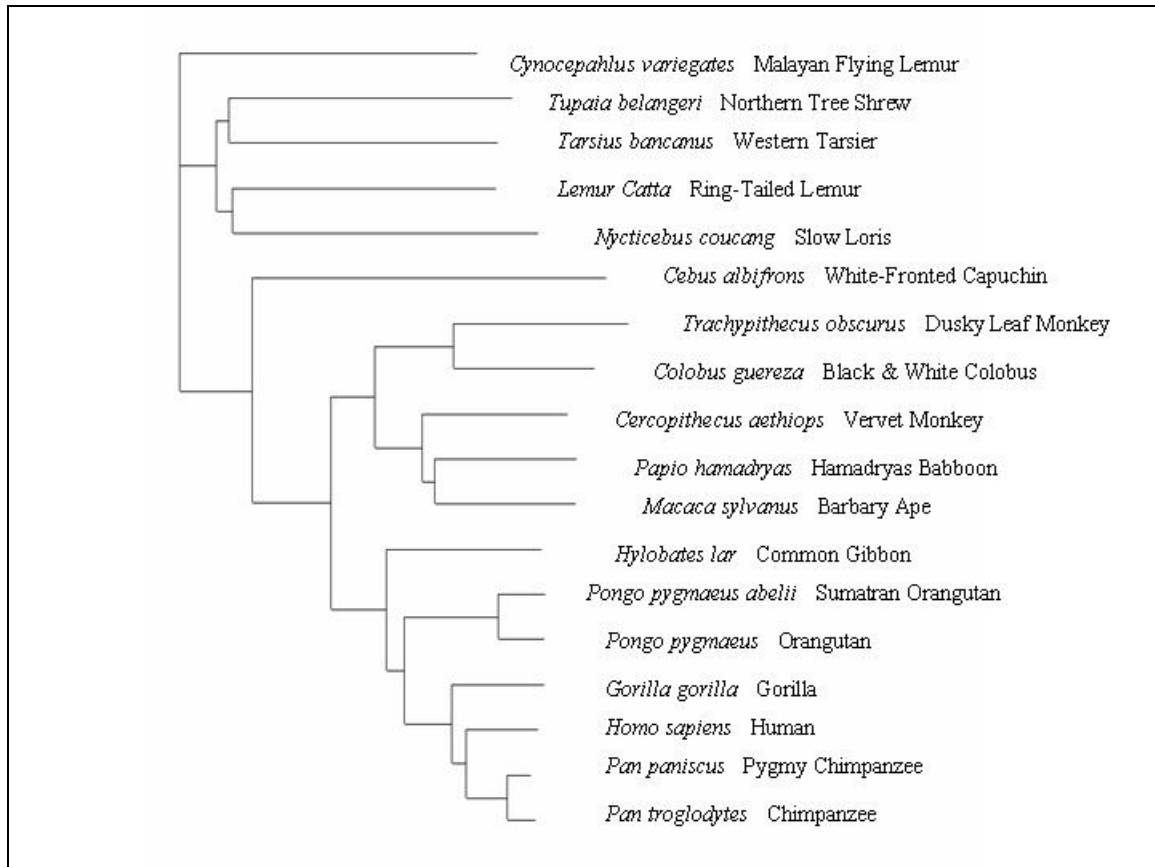
**Figure 3.** Stages in Mitochondrial DNA Replication.

species/individuals that the phylogenetic analysis is being conducted on is referred to as a taxon.

The most commonly used visual representation of phylogenetic relationships are the phylogenetic trees (Figure 4) with the taxa as leaves or tips and the topology representing the phylogenetic relationships between the taxa. The length of a branch represents the extent of change or the number of substitutions between the parent node and child node. The DNA and RNA macromolecules consist of nucleotides and can be represented by a string of bases that make up the molecule. Similarly the proteins can be represented by a string of amino acids that make up the protein. Thus the macromolecules are reduced to simple string of information that can be subjected to mathematical algorithms for comparative study for phylogenetic reconstruction.

There are many classes of algorithms that predict the phylogenetic relationships between taxa using some genetic or amino acid data from the taxa. Parsimony algorithms try to minimize the total number of substitutions along the various branches of the phylogenetic tree. Distance based algorithms find the genetic distances between all pairs of taxa and then find a suitable topology to satisfy the requirements of genetic distance. The genetic distance between two taxa is just the Euclidean distance between the strings of genetic data of the two taxa. Neighbor joining is a distance based algorithm. Likelihood based algorithms employ a matrix of substitution rates between the different nucleotides/amino acids and explore the state space of the different topologies and branch lengths based on the likelihood of each proposed solution. Likelihood based methods employ Markov chains or other heuristic methods to explore the state space of solutions.

Most phylogenetic analyses ignore the possibility of coevolution. Thus it is assumed that each site evolves independently and that substitutions along one branch of the tree are



**Figure 4.** The phylogenetic tree of 16 primate and 2 related species

independent of substitutions along another branch of the tree. The possibility of the rates of substitution varying over time is also ignored.

A problem that sometimes occurs in phylogenetic analyses is that of convergence.

Similar genetic features could evolve along different branches of the tree but this could be mistakenly seen to be an indication of evolutionary relatedness. Another problem faced in phylogenetic reconstruction is that of long branch attraction where two or more unrelated taxa could be so different from the rest that this common difference is taken as a sign of evolutionary relatedness. Results of some phylogenetic reconstruction algorithms can also be confounded by a convergence in base frequencies, thus if taxa

along different lineages evolve similar base frequencies the algorithm might place them together on the tree. All of the above problems could be a result of coevolution.

To provide a point of reference when studying the phylogenetic relationships between taxa one or more outgroups may be used. Outgroups are taxa that lie outside the group being studied that are included in the phylogenetic analysis to provide a point of reference and a sense of perspective to the tree.

#### **1.4 Primates**

Primates include humans, apes, monkeys, prosimians and some related animals. There are about 190 primate species known today. The habitat of non-human primates is limited to tropical and sub-tropical areas in South and Central America, Africa, South and Southeast Asia. Despite their evolutionary relatedness their size ranges from just a few ounces for a mouse lemur to as much as 400 pounds for a full grown gorilla. Although they are not physically specialized for any particular activity or sensory mode they are remarkably clever and adaptive. The grasping or prehensile ability of their hands is very advanced and with the exception of humans they have prehensile feet too. With the exception of the spider monkey, which has four fingers, all primates are pentadactyl or have five fingers and toes. Primates have a tendency for erectness in their upper bodies. They can also be characterized by their large brains (compared to their body size), long gestation periods and life spans. Most primates are arboreal or tree dwelling. Even the terrestrial ones usually sleep on trees with the exception of humans and gorillas. Most of them are diurnal animals, and all of them are highly social with a complex repertoire of vocalizations and displays. They are very flexible in their diet and almost all of them are omnivorous.

There are three suborders of primates – the *prosimii* (lemurs and lorises), the *anthropoidea* (old world Monkeys, new world monkeys, apes and humans) and the *tarsiodea* (tarsiers). The tarsiers are supposed to lie midway between the *prosimii* and the *anthropoidea*.

Because of the natural curiosity one might have towards primates, or the monkeys and apes at least, and because of the insights that we might gain about humans, the evolutionary study of primates is both desirable and necessary. It is for this reason that the primates are the most densely sequenced order. There are 16 complete primate mitochondrial genomes available to us. Many species closely related to the primates have also been sequenced.

The schema of a primate mitochondrial genome is shown in Figure 2. Because of the conserved gene arrangement within primates, the high mutation rates in mitochondrial genomes and the dense sampling of the primates the mitochondrial genomes of primates are a good candidate for evolutionary study.

### **1.5 The Metropolis-Hasting Algorithm**

A Markov chain is a stochastic process where, given the current state of the variable, the past state and the future state of the variable are independent. In the Metropolis-Hasting algorithm the first state or generation of a variable is chosen at random. The proposal for the next generation of the variable is picked from a uniform distribution centered at the current state. If the proposal has a higher likelihood than that of the current generation then the proposal becomes the next generation . If the proposal has a lower likelihood than the current generation then the probability that the proposal will be the next generation is equal to the ratio of the likelihoods of the proposal and the current

generation. Otherwise the next generation is the same as the current generation.

Equilibrium is reached when the likelihood stops getting significantly better with more iterations. The likelihood calculation depends on the system being studied.

The method of proposing the next generation is called the transition kernel of the chain.

The width of the uniform distribution in the transition kernel of the parameter should be large enough to explore different regions in its state space and not get stuck in a local maxima. At the same time the width should not be so large that most of the proposals are in 'bad' areas and are not accepted. It is recommended that the acceptance rate of proposals be between 30 and 80 %.

The process of reaching equilibrium is called burn-in. Once equilibrium is reached further iterations of the chain explore the posterior probability space of the variable.

Enough iterations of the chain after burn-in can thus yield a lot of insight into the maxima and the distribution of a variable in a system. A sampling of the chain in equilibrium can be used to get the 95 % credibility or confidence intervals (CI). This is done by eliminating the 2.5 % largest and 2.5 % smallest values from the samples. The range of the samples left behind is then the 95 % credible interval. The 99 % CIs can be obtained similarly by eliminating the 0.5 % extreme values from either end.

The best estimate for the parameter is that value of the parameter that gives the highest likelihood. This is found by simply monitoring the chain for the maximum likelihood (ML) and the value of the parameter associated with it. The value of the parameter that yields the ML is called the maximum likelihood estimator (MLE).

The method described above can be used to estimate and explore the posterior probability space of more than one parameter. In that case proposals for all the parameters are

accepted or rejected together. However the transition kernel, or the method of generation of the proposals, could be different for different parameters.

## CHAPTER 2. SURVEY OF LITERATURE

### 2.1 Single-Strandedness and $D_{ssH}$

The heavy-strand at a site becomes single-stranded when the heavy-strand replication fork passes over it, and stays so until the light-strand replication fork passes in the other direction. Assuming that the replication forks both travel at the same constant speed a site remains single-stranded for a time that is proportional to the distance the replication forks need to travel to make the site double-stranded again. Sites that lie before the  $O_L$  on the path of the heavy-strand replication remain single-stranded for as long as it takes for the heavy-strand replication fork to travel from the site in question to the  $O_L$  and then for the light-strand replication fork to come back again, a time proportional to twice the distance to the  $O_L$ . When the heavy-strand replication fork passes the  $O_L$  it initiates the light-strand replication fork, and when it traverses further to a site to make it single-stranded the light-strand replication fork has traversed the same distance from the  $O_L$  but in the other direction. Thus the site has to wait for a time proportional to the length of the genome minus twice the distance from the site to the  $O_L$  for the light-strand replication fork to make it double-stranded again.

In the approach described below, the possibility of a delay in initiation of light-strand replication is accounted for. This formula is often divided by the length of the genome to obtain a normalized measure for the time spent single-stranded,  $D_{ssH}$  (Tanaka and Ozawa 1994). Although there has been some recent controversy regarding this mechanism of replication (Holt, Lorimer, and Jacobs 2000; Yang et al. 2002; Bowmaker et al. 2003; Holt and Jacobs 2003), a preponderance of biochemical evidence supports this “classic” model (Bogenhagen and Clayton 2003a; Bogenhagen and Clayton 2003b),

and the evolutionary outcome of substitution rates is itself compelling supporting evidence (Faith and Pollock 2003).

## **2.2 Phylogenetic Reconstruction**

Nucleotide frequencies in mitochondrial DNA vary considerably across mammalian lineages (Honeycutt et al. 1995; Gissi et al. 2000), and such variation may create considerable difficulties for phylogenetic inference, including biased attraction of branches leading to species with similar base frequencies (Van Den Bussche et al. 1998; Reyes, Pesole, and Saccone 2000; Wiens and Hollingsworth 2000). Rates of evolution also appear to vary (Honeycutt et al. 1995; Gissi et al. 2000), but it is often unclear how rates and nucleotide frequencies are related; few studies have gone into these processes in detail. In reconstruction of deep primate phylogeny, variation in frequencies and rates is believed to cause consistent biases (Felsenstein 1978; Lockhart et al. 1992; Graybeal 1993; Meyer 1994; Yoder, Vilgalys, and Ruvolo 1996; Felsenstein 2001), but the reasons for this variation are unclear (Philippe and Laurent 1998), and it is uncertain how it should be taken into account during phylogenetic reconstruction. The underlying evolutionary mechanism has presumably changed, but how? One important factor, only recently clarified, is that different types of mutation rates respond differently to a gradient of single-strandedness that is generated during mitochondrial replication (Faith and Pollock 2003). Thus, it is clearly insufficient to assume that relationships among substitution types are constant across sites or across evolutionary time, and targeted methods are needed to evaluate the response to single-strandedness for different mutation rates in individual genomes.

### 2.3 Nucleotide Frequency Gradients

The single-stranded state is particularly prone to deaminations, especially deaminations of cytosine (C) and adenine (A), which cause transitions to thymine (T) and guanine (G) on the heavy-strand (Asakawa et al. 1991; Tanaka and Ozawa 1994; Reyes et al. 1998). Since transition rates are much greater than transversion rates and therefore dominate equilibrium processes, these excess transitions lead to higher G/A and T/C ratios than in the absence of single-strand mutations. Frederico found that C is very unstable (Frederico, Kunkel, and Shaw 1990; Frederico, Kunkel, and Shaw 1993), and the T/C ratio (or conversely, the A/G ratio on the light-strand) increases quickly with increasing *DssH*, apparently saturating at low values of *DssH* (Faith and Pollock 2003). The deamination of A⇒Hypoxanthine (which is replaced by G) is a slower process (Tarr and Comer 1964; Parham, Fissekis, and Brown 1966; Krasuski et al. 1997), and the gradient in *DssH* causes differences among genes in the rate of A⇒Hypoxanthine deaminations on the heavy-strand, which results in differences in the C/T ratio along the light-strand (Limaïem and Henaut 1984a; Delorme and Henaut 1991), and in differences in GC and AT skew (compositional bias), particularly at third codon positions and non-coding sites (Jermini et al. 1994; Tanaka and Ozawa 1994; Jermini, Graur, and Crozier 1995; Reyes et al. 1998).

Although skew is a sensitive means of detecting differences among genes, both skew measures confound the results of the two major single-stranded transitions, C⇒T and A⇒G (Perna and Kocher 1995). Faith and Pollock (2003), using maximum likelihood analyses of 45 vertebrates with the same gene arrangement and relatively consistent evolutionary rates, found strong evidence that the (heavy-strand) A⇒G substitution rate

per gene increases linearly with *DssH*, while other substitutions do not. C $\Rightarrow$ T substitutions are more prevalent, but are uniformly high along the genome and thus contribute little to differences in nucleotide content along the genome. Although in previous work (Limaiem and Henaut 1984b; Tanaka and Ozawa 1994; Reyes et al. 1998; Faith and Pollock 2003) it has been traditional to refer to substitutions and base frequencies with respect to the light-strand (i.e., the direction in which twelve out of thirteen genes are coded), in this work they are referred to as they are on the complementary heavy-strand. Since the excess mutations occur on the heavy-strand, this simplified complementary notation reduces the potential for confusion in the results and discussion, but readers must be aware of this distinction when comparing the discussion to other works.

## **2.4 Gamma Polymerase**

Our current understanding of the evolutionary processes leading to mutational asymmetry in mitochondria suggests a means to better understand it. The slope of the G/A gradient is presumably an inverse function of the rate of replication and therefore inversely proportional to the efficiency of gamma polymerase (the replicating enzyme in vertebrate mitochondria); the intercept of the gradient is presumably a function of the G/A ratio sans the effect of single-strandedness and the rate at which light-strand synthesis is initiated (which in turn might be affected by both the shape of the origin of replication and the binding abilities of the gamma polymerase accessory subunit). For other substitution types, particularly C $\Rightarrow$ T, repair mechanisms (Meyer 1994) may alter the slope and intercept, and probably the linearity of response; when functioning efficiently they may completely eliminate any detectable response to single-strandedness.

## 2.5 Outline of Research

In the research presented here the variation in nucleotide ratio gradients among primates and two outgroups was studied. The primates, with 16 complete mitochondrial genomes, are the most densely sampled vertebrate order, and generally have an increased rate of evolution relative to other mammals (Gissi et al. 2000). The focus was on the heavy-strand G/A gradient at 3<sup>rd</sup> codon positions, since there is a strong expectation that it will increase linearly with *DssH*, but the heavy-strand C/T and pyrimidine/purine ( $Y/R=(C+T)/(A+G)$ ) ratios, and the G/A gradients at the 1<sup>st</sup> and 2<sup>nd</sup> codon positions are also reported on (C/T ratios are reported rather than T/C ratios because they have lower variances). Likelihood-based methods were developed to evaluate the response to single-strandedness. A joint Bayesian and maximum likelihood approach was used to evaluate the among-species differences in response to *DssH*, and both mixture model and hierarchical clustering methodologies were utilized to evaluate whether different species evolved in similar fashions. These tools created the ability to detect and explain divergence and convergence of base frequencies among primates, and in addition were able to provide a causal explanation for phylogenetic reconstruction bias in parts of the tree: the tree shrew falsely clusters with the tarsier, well within the primates, and the tarsier falsely clusters with the prosimians, rather than as a sister taxon to the anthropoid primates (Schmitz, Ohme, and Zischler 2001). To maintain the clarity of the results narrative, a great deal of the raw results from the likelihood analysis is placed in the Appendix, and the figures and tables presented in the main text are reserved for critical interpretive information.

## CHAPTER 3. MATERIALS AND METHODS

### 3.1 Materials

**Table 2.** Common names, scientific names, abbreviations used in figures, and accession numbers for sequences used.

| <b>Common name</b>     | <b>Species</b>                 | <b>Abb.</b> | <b>Accession</b>        |
|------------------------|--------------------------------|-------------|-------------------------|
| Human                  | <i>Homo sapiens</i>            | Hsa         | NC_001807 <sup>1</sup>  |
| Chimpanzee             | <i>Pan troglodytes</i>         | Ptr         | NC_001643 <sup>2</sup>  |
| Pygmy Chimpanzee       | <i>Pan paniscus</i>            | Ppa         | NC_001644 <sup>2</sup>  |
| Gorilla                | <i>Gorilla gorilla</i>         | Ggo         | NC_001645 <sup>2</sup>  |
| Sumatran Orangutan     | <i>Pongo pygmaeus abelii</i>   | Pab         | NC_002083 <sup>3</sup>  |
| Orangutan              | <i>Pongo p. pygmaeus</i>       | Ppy         | NC_001646 <sup>2</sup>  |
| Common Gibbon          | <i>Hylobates lar</i>           | Hla         | NC_002082 <sup>4</sup>  |
| Barbary Ape            | <i>Macaca sylvanus</i>         | Msy         | NC_002764 <sup>5</sup>  |
| Hamadryas Baboon       | <i>Papio hamadryas</i>         | Pha         | NC_001992 <sup>6</sup>  |
| Vervet Monkey          | <i>Cercopithecus aethiops</i>  | Cae         | NC_006669 <sup>7</sup>  |
| Black & White Colobus  | <i>Colobus guereza</i>         | Cgu         | NC_006670 <sup>7</sup>  |
| Brown-Ridged Langur    | <i>Trachypithecus obscurus</i> | Tob         | NC_006671 <sup>7</sup>  |
| White-Fronted Capuchin | <i>Cebus albifrons</i>         | Cal         | NC_002763 <sup>5</sup>  |
| Slow Loris             | <i>Nycticebus coucang</i>      | Nco         | NC_002765 <sup>5</sup>  |
| Ring-Tailed Lemur      | <i>Lemur catta</i>             | Lca         | NC_004025 <sup>8</sup>  |
| Western Tarsier        | <i>Tarsius bancanus</i>        | Tba         | NC_002811 <sup>9</sup>  |
| Northern Tree Shrew    | <i>Tupaia belangeri</i>        | Tbe         | NC_002521 <sup>10</sup> |
| Malayan Flying Lemur   | <i>Cynocephalus variegatus</i> | Cva         | NC_004031 <sup>8</sup>  |

(Horai et al. 1995; Arnason, Gullberg, and Xu 1996; Xu and Arnason 1996; Arnason, Gullberg, and Janke 1998; Arnason et al. 2000; Ingman et al. 2000; Schmitz, Ohme, and Zischler 2000; Arnason et al. 2002; Schmitz, Ohme, and Zischler 2002).

All complete primate mitochondrial genomes available at the time this study was initiated were used (Table 2). As outgroups, the complete genomes of the flying lemur and the tree

shrew were used. For all genomes, individual protein-coding genes were extracted, concatenated, and codon positions determined automatically using C programs.

### 3.2 Analysis of Single Genomes

Likelihoods of slopes and intercepts for individual species were calculated as follows: based on a model ( $M$ ) and set of parameters ( $\theta$ ), the likelihood of a particular genome was calculated by multiplying across sites in a sequence ( $S^m$ ) from species  $m$  of length  $N$ ,

$$L(S^m | M, \theta) = \prod_{i=1}^N P(S_i^m | M, \theta) \Delta(C_i) \quad (3.1)$$

where  $\Delta(C_i)$  is a delta function equal to zero or one depending on whether the site was in the class of interest (3<sup>rd</sup> codon positions). For simplicity and clarity, the  $M$  will henceforth be dropped from equations and considered implicit, as will the  $\Delta(C_i)$ .

Synonymous third codon positions were generally used to obtain sites that were least likely to have been affected by selection, although first and second codon positions were also analyzed for comparison. Frequency ratios arising from each pair of reciprocal transitions ( $G \leftrightarrow A$  and  $T \leftrightarrow C$ ) were analyzed separately, as was the ratio arising from transversions between nucleotide classes ( $Y \leftrightarrow R$ ) for 4x redundant 3<sup>rd</sup> codon positions. Due to the nature of the mitochondrial genetic code (i.e., there are no 3x redundant codon classes), the 2x and 4x redundant codon classes could be studied jointly rather than separately for the G/A and C/T ratios, whereas Y/R analyses were restricted to 4x redundant positions (the probability of being a purine or pyrimidine in a 2x redundant class is dependent on selection at the amino acid level, so combined analysis of all 3<sup>rd</sup>

codon positions would improperly include amino acid selective effects in the analysis of transversion rates).

Since G/A ratios are thought to increase linearly with  $DssH$ , it is reasonable, particularly in the case of G/A ratio, to build a simple linear model of increase in these ratios, and determine what plausible values are for the slope ( $\zeta$ ) and intercept ( $\iota$ ). Thus, if  $DssH_i^m$  is the calculated  $DssH$  value at site  $i$  for sequence  $m$ , and  $\theta$  is the vector of unknown parameters in the model, then

$$P(S_i^m | \theta) = P(S_i^m | DssH_i^m, \zeta, \iota) \quad (3.2)$$

For an example using the G/A ratio,  $f(G/A)_i^m = \zeta * DssH_i^m + \iota$ ,

$P(G)_i^m = f(G/A)_i^m / [1 + f(G/A)_i^m]$ , and  $P(A)_i^m = 1 - P(G)_i^m$ . For each individual genome, a Markov chain was run using the Metropolis-Hastings Monte Carlo algorithm to sample the posterior probability space (Metropolis et al. 1953; Hastings 1970),

$$P(\theta | S^m) = \frac{P(S^m | \theta)P(\theta)}{\int_{\theta} P(S^m | \theta)P(\theta)} \quad (3.3)$$

The prior probabilities,  $P(\theta)$ , were assumed to be flat, uninformative priors, with  $\zeta$  ranging from  $-\infty$  to  $\infty$ , and  $\iota$  ranging from 0 to  $\infty$ . Proposals for  $\zeta$  and  $\iota$  where  $f(G/A) < 0$  for some  $DssH_i^m$  were excluded. Parameter proposals in the Markov chain were distributed  $\sim U[-\delta, +\delta]$  about the current state, with the magnitude of  $\delta$  equal to 0.3 for both  $\zeta$  and  $\iota$ ; values of  $\delta$  were chosen so that between 30% and 80% of the proposals were accepted. The 95% credibility interval was obtained by excluding the 2.5% most extreme values on either side, and the maximum for the run was taken as an estimate of

the maximum likelihood value. The chain was run for 100,000 generations where the first 1000 generations were removed as burn-in. The rest of the generations were sampled at every 100<sup>th</sup> spot in the chain. All chains were run ten times with different seed values to detect any differences in maximum likelihood values or distributions across runs. All likelihood values were stored and reported as natural logarithms.

### **3.3 Hierarchical Clustering**

To determine the similarity of genomes in their evolutionary patterns, Markov chains were also run over multiple genomes simultaneously in hierarchical and mixture model clustering schemes. In the hierarchical clustering scheme, single sets of maximum likelihood estimators (MLEs) of slope and intercept for a group of genomes were determined jointly. The MLs from single-genome analyses were taken first. Then chains over all possible pairs of genomes were run and their MLs found. The loss in likelihood or  $\delta \ln L$  was the difference between the sum of the MLs (in log-values) of the individual genomes and the ML (also in log-values) of the pair.  $\delta \ln L$  represented the extent to which the gradients in nucleotide ratios in the two genomes were similar. The smaller the  $\delta \ln L$ , the more similar the two genomes were. That pair or union of genomes that had the smallest  $\delta \ln L$  was combined into one cluster for subsequent stages. In the subsequent stages the  $\delta \ln L$ s for all unions between all uncombined individual genomes and all combined genomes were found and the union with the smallest  $\delta \ln L$  combined into a cluster again. This was continued until all the genomes were combined into one cluster. Since twice the  $\delta \ln L$  for combining sets can be approximated as a chi-square distribution with two degrees of freedom,  $\chi_2^2$  (Rice 1995), the log likelihood differences were used as a measure of confidence in the formation of clusters.

### 3.4 Mixture-Model Clustering

In another clustering scheme, a Markov chain was run on 3<sup>rd</sup> codon positions in the complete primate dataset using a series of mixture models (the outgroups were not included in this scheme). In any one implementation of this method, a predetermined number of models ( $K$ ) were allowed to exist, with the constraint that the models were ordered by strength of intercept to avoid problems of identifiability. The mixture density for a genome can be written as,

$$P(S^m | \Psi) = \sum_{k=1}^K \pi_k P(S^m | \theta_k) \quad (3.4)$$

where  $\Psi$  is the vector containing all the unknown parameters in the mixture model, i.e., all  $\pi_k$  and  $\theta_k$ , and the different models were given even and constant mixing proportions,  $\pi_k = 1/K$ . The  $\delta$  value for updating both the  $\zeta$  and  $\iota$  parameters was  $0.3/\sqrt{K}$ , and overall likelihoods were calculated by multiplying the likelihoods for each genome. At any time point (i.e., for any set of parameters,  $\theta$ ) it is possible to calculate the posterior probability that a particular model applies to a particular species

$$P(M_k | S^m) = \frac{P(S^m | \theta_k)P(\theta_k | M_k)P(M_k)}{\sum_{k=1}^K P(S^m | \theta_k)P(\theta_k | M_k)P(M_k)} \quad (3.5)$$

Mixture models were run with two to eight mixed models. The log likelihoods for these models are presented, but the  $\delta \ln L$ s for mixture models are not necessarily distributed as  $\chi^2$  (McLachlan and Peel 2000), and determining the appropriate number of mixture models is one of the more difficult problems in statistics. In this study, however, twice the improvement in log likelihood going from five to six models was slightly below 5.74

(just below significance under a  $\chi^2_2$  assumption) while the improvement going to seven models was only 4.12, and there was a reduction in likelihood moving to eight models (the model with the largest or smallest intercept tended to wander off into irrelevance). For two to five models, and oftentimes with six models, the posterior probabilities for each sequence belonging to one of the models were often close to one, whereas with seven or eight models many sequences had mixed affiliation among models, which was taken as another sign that seven or more models were not useful. Accordingly the results for up to six mixed models are presented.

### **3.5 Phylogenetic Analysis**

Phylogenetic trees were obtained using the combined sequences of all 12 proteins coded on the light-strand. Neighbor joining tree was obtained from DNA sequences using the general time reversible (GTR) model in Paup\* (Swofford 2000). Both DNA and amino acid sequences were used with Poisson models in MrBayes, and the tree with the highest likelihood was selected (Huelsenbeck and Ronquist 2001). The topological structures of these trees are similar and largely uncontroversial except for the deeper nodes (Schmitz, Ohme, and Zischler 2002). To obtain comparative likelihood values, a maximum likelihood analysis was also run on these topologies (based on DNA sequences and the GTR model) using the lscore function in Paup\*. In addition topologies intermediate between these and what appears from the literature were also evaluated (Schmitz, Ohme, and Zischler 2002) to be a good estimate of the “true” phylogeny, where the two non-primates are constrained to be outgroups, and tarsier is constrained to be a sister group of the anthropoid apes.

## CHAPTER 4. RESULTS

### 4.1 Evolution of G/A Gradients

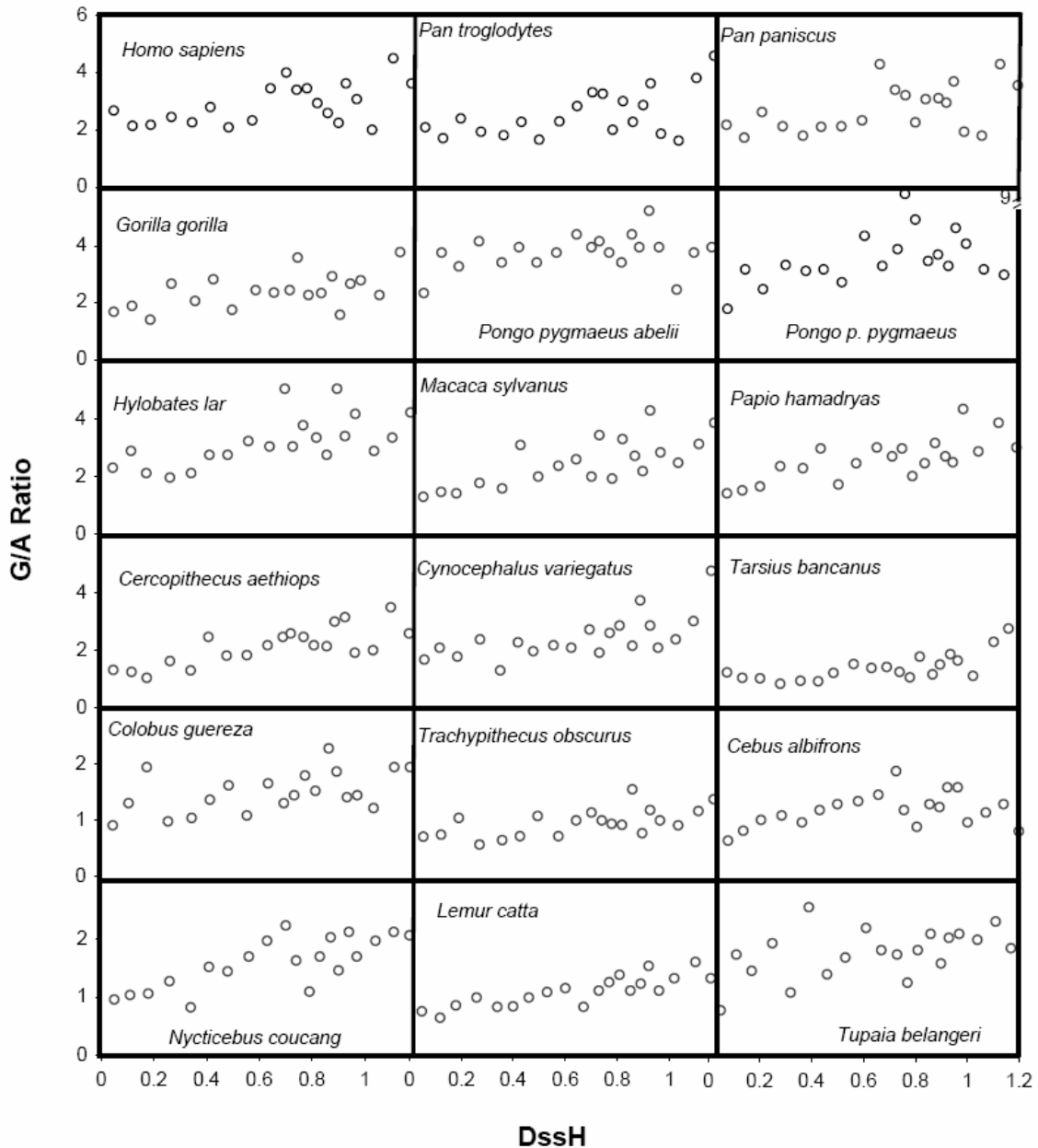
The expectation, based on a simultaneous analysis of complete vertebrate genomes (Faith and Pollock 2003) was that synonymous sites in individual primate genomes would have a linear relationship between the heavy-strand G/A ratio and the time spent single-stranded. MCMC runs on individual genomes showed significantly positive slopes in all cases (Figure 5, Table 3). There was considerable variation among genomes in both slope and intercept, and values for many pairs of species were apparently different in that they lay outside their respective 95% credible intervals (Table 3). Comparisons of null models with one response curve for a pair of genomes to models with independent response curves for each genome in a pair showed that, based on the  $\chi^2_2$  distribution, most pairs of genomes have significantly different responses to time spent single-stranded (Appendix Table A). To obtain a better idea of the meaning of this variation, species were clustered based on their G/A ratio responses according to a hierarchical clustering approach and according to mixture model analyses with between two and eight mixture models. It is useful to compare and combine the two approaches, since hierarchical clustering may be order dependent, while significance levels for the mixture models have uncertain validity (McLachlan and Peel 2000).

In the hierarchical clustering (Figure 6a), clusters that were not rejected at the 0.05% significance level included one large set of species (Group 10: an outgroup, Cva, plus chimpanzees and gorillas, Ptr, Ppa, and Ggo, and two old world monkeys, Msy, and Pha), and a few pairs (the two Pongo species, Ppy and Pab; colubines and lorises, Cgu and

**Table 3.** Maximum likelihood values & 95% CI for slopes and intercepts of G/A gradients in primates and two outgroups.

| Species                        | Max Like | Slope                | Intercept            |
|--------------------------------|----------|----------------------|----------------------|
| <i>Homo sapiens</i>            | -1275.61 | 0.860 [0.228, 1.561] | 2.204 [1.768, 2.710] |
| <i>Pan troglodytes</i>         | -1339.08 | 0.925 [0.363, 1.490] | 1.761 [1.403, 2.176] |
| <i>Pan paniscus</i>            | -1335.41 | 1.061 [0.491, 1.645] | 1.686 [1.326, 2.126] |
| <i>Gorilla gorilla</i>         | -1332.45 | 1.187 [0.578, 1.794] | 1.622 [1.266, 2.056] |
| <i>Pongo pygmaeus abelii</i>   | -1169.74 | 0.661 [0.110, 1.740] | 3.096 [2.443, 3.636] |
| <i>Pongo p. pygmaeus</i>       | -1189.91 | 1.541 [0.502, 2.543] | 2.417 [1.853, 3.155] |
| <i>Hylobates lar</i>           | -1214.29 | 1.544 [0.735, 2.331] | 2.077 [1.643, 2.623] |
| <i>Macaca sylvanus</i>         | -1297.84 | 1.729 [1.216, 2.319] | 1.197 [0.906, 1.531] |
| <i>Papio hamadryas</i>         | -1284.19 | 1.586 [0.962, 2.179] | 1.451 [1.134, 1.832] |
| <i>Cercopithecus aethiops</i>  | -1353.94 | 1.494 [1.039, 2.018] | 1.087 [0.830, 1.384] |
| <i>Colobus guereza</i>         | -1425.30 | 0.525 [0.195, 0.904] | 1.104 [0.893, 1.351] |
| <i>Trachypithecus obscurus</i> | -1469.87 | 0.415 [0.190, 0.630] | 0.695 [0.567, 0.847] |
| <i>Cebus albifrons</i>         | -1405.69 | 0.344 [0.091, 0.642] | 0.947 [0.743, 1.144] |
| <i>Nycticebus coucang</i>      | -1335.30 | 0.965 [0.609, 1.329] | 0.906 [0.709, 1.147] |
| <i>Lemur catta</i>             | -1408.20 | 0.607 [0.359, 0.883] | 0.688 [0.536, 0.870] |
| <i>Tarsius bancanus</i>        | -1422.08 | 0.708 [0.420, 0.994] | 0.844 [0.673, 1.048] |
| <i>Tupaia belangeri</i>        | -1263.74 | 0.694 [0.303, 1.122] | 1.258 [1.006, 1.557] |
| <i>Cynocephalus variegatus</i> | -1269.62 | 1.132 [0.582, 1.658] | 1.553 [1.224, 1.955] |

Nco; humans and gibbons, Hsa and Hla; langurs and lemurs, Tob and Lca; and capuchins and tarsiers, Cal and Tba). At moderately large costs ( $\delta\text{LnL} < 10$ ), the outgroup *Tupaia* joined Cgu and Nco to form Group 11, the third new world monkey, Cae, joined the large chimp/gorilla/old world monkey group to form Group 12, the orangutans joined the human and gibbon to form Group 13, and langurs, lemurs, capuchins and tarsiers all



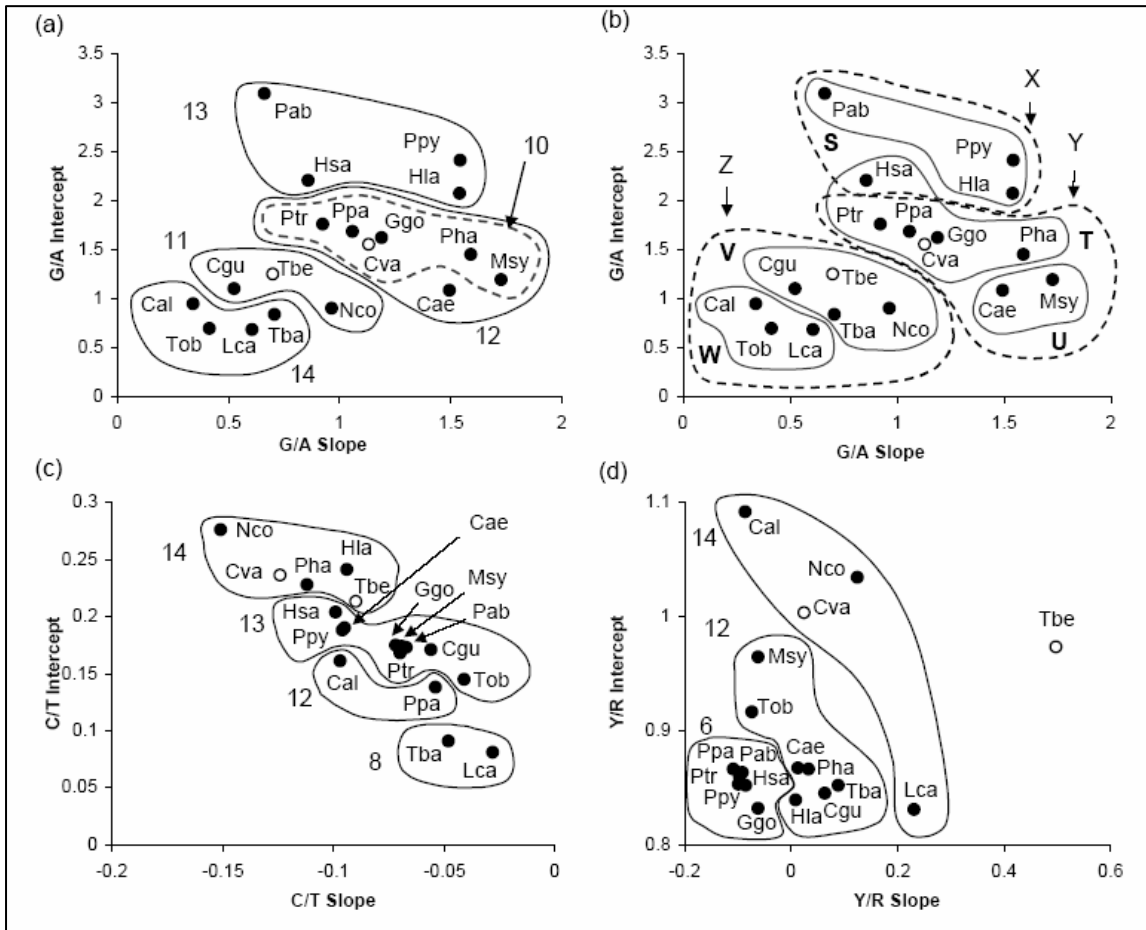
**Figure 5.** G/A ratios for complete primate mitochondrial genomes and two near outgroups. Third codon positions containing G/A were grouped into twenty equal-size bins for each genome, and the ratio of G/A in each bin is graphed versus the average *DssH* for that bin.

joined together to form Group 14. The next two mergers had larger likelihood costs ( $10 > \delta L_n L > 60$ ), with the deep-branching primates and outgroups (Groups 11 and 14) joining together first, followed by the great apes and old world monkeys (Groups 12 and

13). The primates and outgroups could be merged together as one group, but only at an extreme cost of  $\delta\text{LnL} = 497$ .

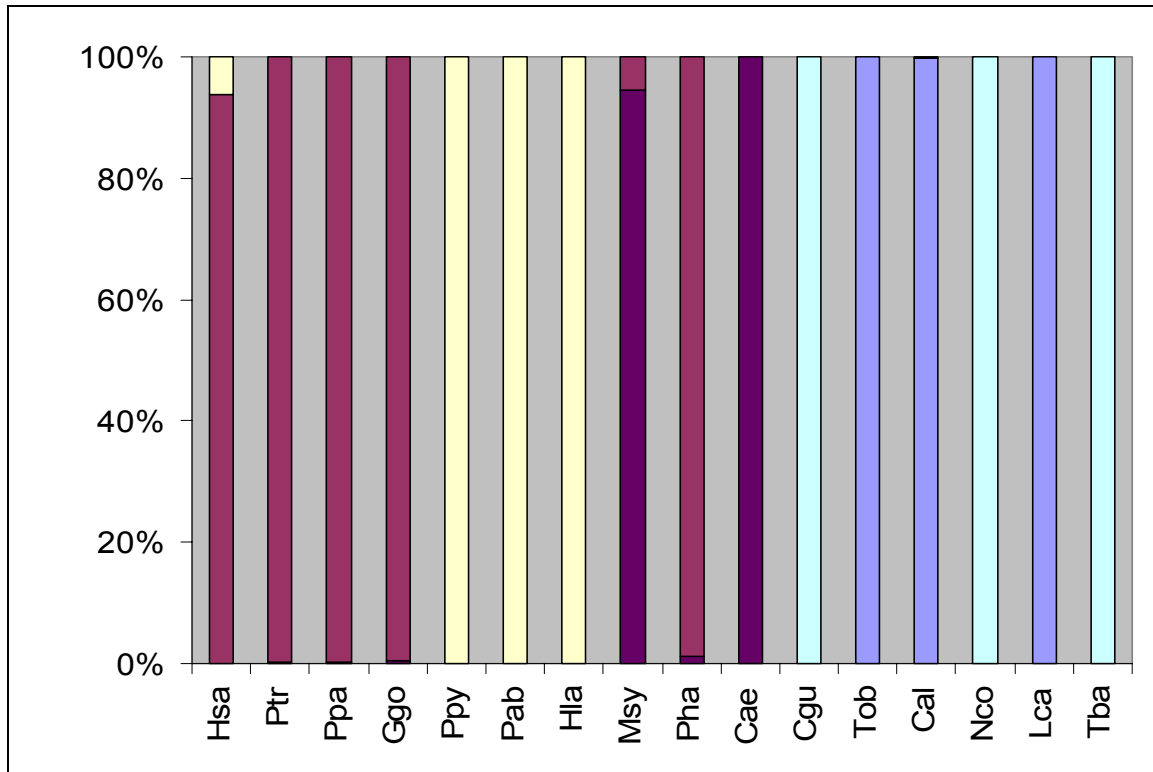
One difficulty in interpreting these results is that the order of clustering can strongly affect whether particular species join together early or late in the hierarchical clustering process. For example, in pairwise comparisons, humans could easily have joined together with the gorilla, chimpanzee, and pygmy chimpanzee at only a small decrease in likelihood ( $\delta\text{LnL} \sim 3$ ). They joined most easily with the gibbon, however, thus being led away from the other great apes, and this order of clustering meant that the baboon/macaque cluster was subsequently slightly more likely to join the gorilla/chimp/pygmy chimp cluster than was the human/gibbon cluster. The human/gibbon cluster was then more likely to join the orangutans than this combined cluster, and all the great apes and old world monkeys joined together at the rather high cost of 60 log likelihood units (Appendix, Table B). Other interesting points are that the intercept tended to matter more in clustering than the slope, and as expected, clusters were more easily joined when a slightly smaller intercept was balanced with a slightly bigger slope.

Mixture model analysis offers an alternative means of assessing similarity among responses to the gradient that is not order dependent. In such analyses, all species were evaluated simultaneously (the outgroups were excluded), and the best set of models was determined (Appendix, Table C). Although individual species were not deterministically linked to a specific model, the posterior probability that data from a particular species was generated by each model can be calculated (Equation 3.5), and for six models or



**Figure 6.** Graph of MLE slopes versus MLE intercepts along with major clusters in ratio cluster analyses. Results are shown for hierarchical (a) and mixture analyses (b) of G/A ratios, and hierarchical analyses of (c) C/T, and (d) Y/R ratios. Groups are labeled by their order of clustering.

fewer, the posterior probability for each species was approximately one for one of the models and approximately zero for the others, although in ten replicates there was some variance in the posterior for the five and six model cases (data not shown). Clustering is obviously related to the results from the hierarchical analysis, but due to the non-hierarchical nature, switches in alliances among groups can occur for different numbers of clusters in the mixture analysis. For example, with three models (Figure 6b), humans clustered with the orangutans and gibbons (Group X), as before, while the other great



**Figure 7.** Posterior probabilities for each species to belong to each model for the five-model mixture. The posterior probabilities are averaged across ten independent chains. The models in descending order of magnitude of intercept are yellow (Group S), brown (Group T), dark brown (Group U), sky blue (Group V), and blue (Group W). Group identifications are the same as in Figure 6b.

apes clustered with the old world monkeys (Group Y), and the remaining primates all clustered together (Group Z). With five models, the deeper primates split into two groups (Groups W and V), as did the great ape/old world monkey mixed group (Groups T and U). In the latter case, two of the old world monkeys split off, but the baboons remained in a cluster with the hominids, which included humans, as expected based on phylogenetic relatedness. In Figure 7 are shown the posterior probabilities that each species belongs to each of these models; it is clear that although the ML results discussed above definitively place the species with particular models, the posterior allegiances are often shared between models when they are adjacent to one another. If these clusters are mapped onto

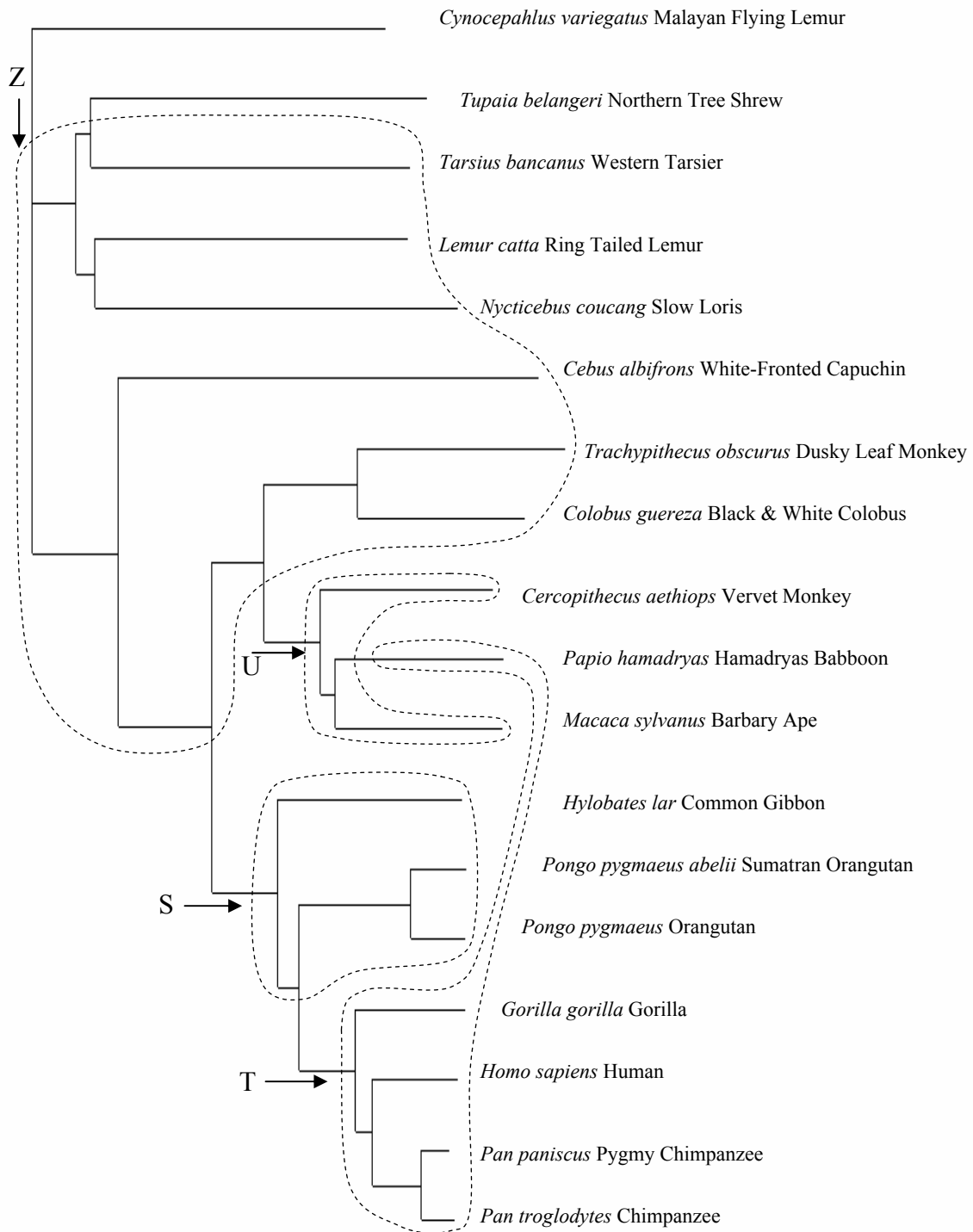
a phylogenetic tree (Figure 8), it is clear that the baboons, and to some extent all of the old world monkeys, have converged to a similar response curve as the hominids.

An interpretation of the evolution of the G/A response curves can now be made (Figure 9). The three deepest diverging primates, Lemur, Nycticebus, and Tarsius, (prosimians and tarsier), have similar slopes and intercepts, with some variation, in the range of 0.61-0.96 for the slope and 0.69-0.91 for the intercept. In the transition to the anthropoid primates (including cebids and colubines), intercepts remained similar (0.69 - 0.95) but the slopes notably decreased to a range of 0.34–0.53. In an apparent convergence, the old world monkeys (baboon, mangabey, and macaque) increased their slopes (1.5-1.7; the largest among the primates) and intercepts (1.1-1.5), and the lesser and great apes increased their slopes and intercepts to the ranges of 0.66-1.5 and 1.6-3.1. The hominids are tightly clustered in intercepts (with the exception of Homo), and fairly clustered in slopes, but the orangutans and gibbon have the highest intercepts among the primates, and their slopes cover the extremes of the range among greater and lesser apes.

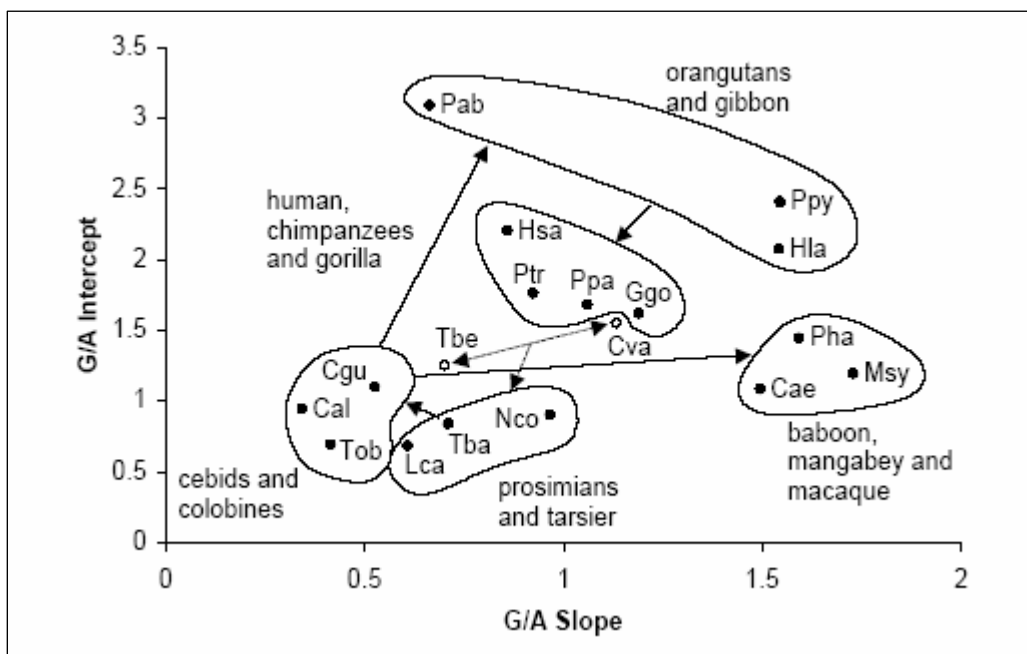
Interestingly, the outgroup Cynocephalus is very similar to the gorilla, while the other outgroup, Tupaia, is closest to Tarsier.

#### **4.2 Evolution of C/T and R/Y Gradients**

Although the C/T ratio did not show a clear slope in the earlier study (Faith and Pollock 2003), individual and hierarchical analyses on the C/T ratio response to single-strandedness were performed to determine if there was any variation in the level of asymmetry or the existence of a slope among the primates (Appendix, Tables D and E). These analyses were also performed on the Y/R ratio at 4x redundant 3<sup>rd</sup> codon positions to see if there was detectable variation in slopes and intercepts for transversions



**Figure 8.** G/A mixture model groups mapped onto the NJ phylogenetic tree. Arrows indicate possible locations of large changes in the response curve. Clusters shown are for the model with five clusters, except that clusters V and W have similar slopes and intercepts, and are grouped into cluster Z as in the three-cluster analysis.



**Figure 9.** Graph of MLE slopes versus MLE intercepts along with major groups showing a summary interpretation of G/A evolution. Arrows indicate possible changes in response curves, and are discussed in the text.

(Appendix, Tables F and G). In the C/T analysis, there are three discrete groups that required only small  $\delta \ln L$  penalties to form (less than 2.5), but required substantial penalties (8.5 – 67.5) to merge (Figure 6c). The largest group (Group 13) includes most of the apes and old world and new world monkeys, and has a strong bias against C (C/T intercept = 0.17 [0.16-0.19]) and a slightly negative but not significant slope (slope = -0.076 [-0.09 - -0.06]) indicating increasing bias against C with increasing single-strandedness. Two non-anthropoid apes, Lemur and Tarsius, form the smallest group (Group 8), with a slope only slightly less than zero, and already a very strong bias against C at the intercept (0.09 [0.06-0.11]). The third group (Group 14) is an odd assortment of the two outgroups plus Papio, Hylobates, and Nycticebus, with a similar but slightly more

negative slope compared to the large group (-0.12 [-0.15 - -0.09]), but with substantially less bias against C (0.24 [0.21 – 0.27]). The phylogenetic separation of these species indicates that there may be a recurrent mechanism by which bias against C may be reduced, presumably by increasing protection against or repair of the causative mutation. Results with the C/T ratio are very tentative because of the non-linear response, and indeed, studies currently underway indicate that there is considerable complexity in the evolution of this response curve.

The Y/R ratio analysis of individual genomes also proved interesting, in that Tupaia was the only organism with a significant slope (Figure 6d, Appendix, Table F). Tupaia had an even ratio of pyrimidines to purines at zero *DssH*, but had a positively increasing bias toward pyrimidines with increasing *DssH* (slope = 0.50 [0.11, 0.82], intercept = 0.97 [0.78, 1.25]). In addition to Tupaia, there were three groups (6, 12, and 14) that required moderate or large likelihood penalties (11.4 – 61.5) to merge with one another (all penalties in the process of creating these groups were less than 1.9; see Appendix, Table G), and which all had slightly positive or non-significant slopes.

The final merging of all species incurred a very large likelihood penalty (61.5), because there was a large difference in the intercepts between Group 16 (Tupaia plus Group 14), which had a positive slope and an equal ratio of pyrimidines to purines at the intercept, and the clustered remainder of the primates (Group 15 = Group 12 plus Group 6), which had a slightly negative but non-significant slope, and a Y/R ratio of 0.867 [0.82-0.92] at the intercept, and were thus significantly biased towards purines. The generally flat slopes in the primates provided little evidence for excess transversion mutations in response to single-strandedness, although the significant slope in Tupaia (and the

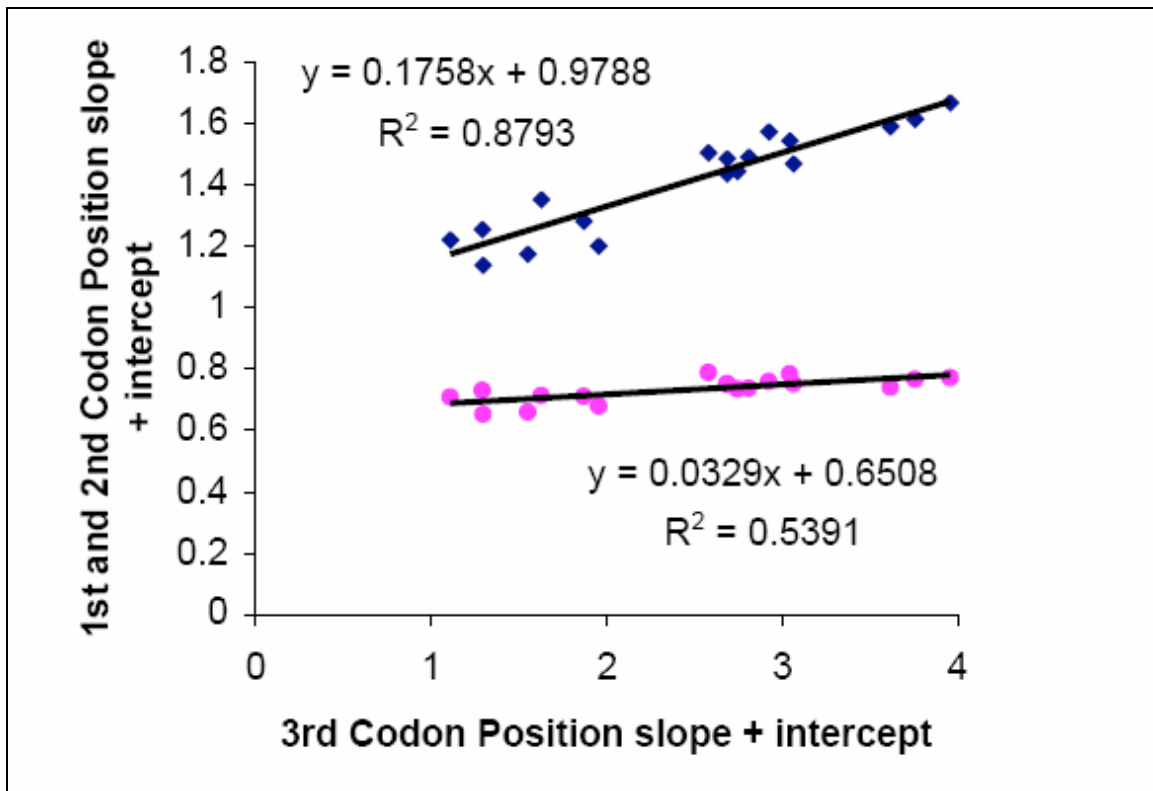
significant slope for the combined members of Group 16) is preliminary evidence that such a response can exist in some organisms (it is perhaps usually controlled by efficient repair mechanisms). Interestingly, Tarsius did not group with the prosimians and outgroups based on the Y/R ratio, while the deepest-branching monkey, Cebus, did, although the differences between the tarsier and Lemur were not large (Supplementary Data, Tables F and G).

The bias towards purines in the apes and most monkeys indicates a derived trend. Although such a bias cannot occur in a perfectly symmetric mutation model (where the mutation processes are equivalent on both strands), the strong and consistent transition bias against C (described above) could conceivably create a transversion bias through secondary effects without any alteration in transversion rates. The pattern of species with this bias did not match the pattern of species differences in the C/T bias, however, so it seems probable that there may have been a derived change in the rates of at least one type of transversion. It is also possible that these differences could be due to derived changes in the degree of codon bias or some other form of selection on synonymous sites, although it seems implausible that such selective alternatives could explain the positive slope in Tupaia.

### **4.3 Correlation of 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> Codon Positions, and Comparison of Phylogenetic Trees**

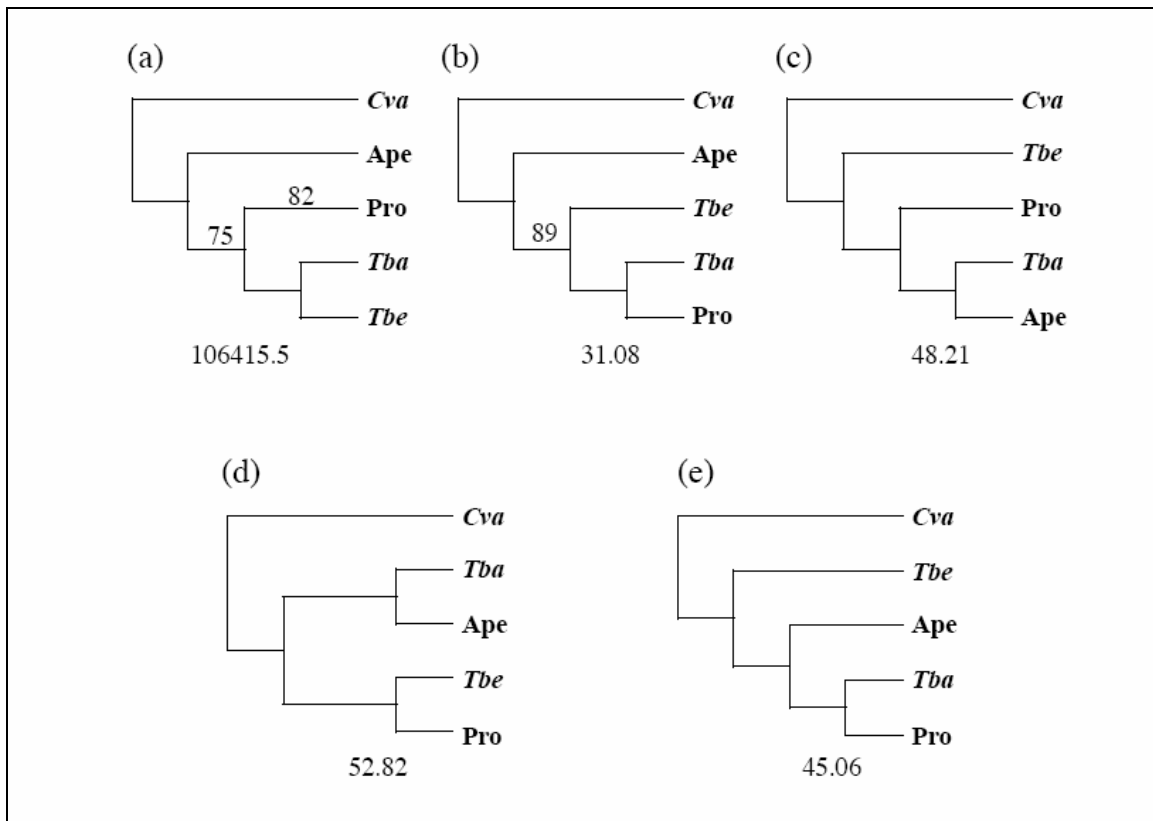
Evolutionary changes in the number of deaminations in the single-stranded state may also affect 1<sup>st</sup> and 2<sup>nd</sup> codon positions, but because many more changes at 1<sup>st</sup> codon positions and all changes at 2<sup>nd</sup> codon positions are non-synonymous, they are constrained by selection at the amino acid level. At 1<sup>st</sup> codon positions, nine out of eighteen slopes are

significantly greater than zero, while for 2<sup>nd</sup> codon positions no individual slopes are significant. Nevertheless, linear regressions of the G/A ratio slope plus intercept of both 1<sup>st</sup> and 2<sup>nd</sup> codon positions on 3<sup>rd</sup> codon positions (Figure 10) are extremely significant (both probabilities are less than 0.001). Although the regression slopes are much less than one, particularly for the slow evolving 2<sup>nd</sup> codon positions, this result indicates, not surprisingly, that nucleotide biases in mutation rates also affect amino acid substitution rates, presumably mostly for neutral or nearly neutral substitutions.



**Figure 10.** Regression of slope plus intercept for different codon positions. The MLE estimators of slope plus intercept response curves for each species in the analysis for 1<sup>st</sup> codon positions (diamonds) and 2<sup>nd</sup> codon positions (circles) versus 3<sup>rd</sup> codon positions. The regression line is shown, and the slope, intercept, and  $R^2$  values are shown adjacent to each line.

Evolutionary changes in biases in nucleotide and amino acid composition may affect phylogenetic reconstruction with mitochondrial data (Felsenstein 1978; Lockhart et al. 1992; Graybeal 1993; Meyer 1994; Yoder, Vilgalys, and Ruvolo 1996; Felsenstein 2001). The nucleotide data strongly supports a tree (Figure 11a) that is not consistent with current views of primate phylogeny (Figure 11c); the joining of *Tarsius* together with *Tupaia*, a non-primate included as an outgroup, and placing this pair as a sister group to the prosimians, hardly seems credible. The amino acid data supports a tree (Figure 11b)



**Figure 11.** Comparison of the most likely trees relating the deeply diverging primate groups and outgroups. Bootstrap values for the DNA-based NJ analysis are shown on (a) when less than 100%. Posterior probabilities for the nucleotide Bayesian analysis were 100%, and the one branch less than 100% in the amino acid analysis is shown on (b). The likelihood is shown for (a), the most likely topology under the DNA-based analysis, and differences from the most likely tree are shown underneath topologies (b) – (e).

that is only slightly improved relative to morphological expectations (Figure 11c), and which is also the second-best tree in terms of likelihood scores. Support for the favored tree is good, both in terms of relative likelihood scores compared to the expected tree and alternative intermediates (Figure 11), and in terms of neighbor joining bootstrap and Bayesian posterior probability support for branches.

## CHAPTER 5. DISCUSSION

The results of this study provide details on the evolution of the response of various substitutions to the gradient of single-strandedness encountered during mitochondrial replication. For simplicity, the evolution of this response will be referred to as “gradient evolution”, and the combined slope and intercept as the “response curve”. Gradient evolution was mostly phylogenetically consistent, but there are clear instances of convergent changes in the response curve. Since changes in equilibrium base frequencies are the necessary outcome of evolution of the mutation spectrum, and because evolution of base frequencies can dramatically mislead phylogenetic analyses (Felsenstein 1978; Lockhart et al. 1992; Graybeal 1993; Meyer 1994; Yoder, Vilgalys, and Ruvolo 1996; Felsenstein 2001), this result may explain some difficulties in primate phylogenies determined by mitochondrial analysis. In particular, the placement of the tree shrew within the primates even though it is believed to be more distantly related than the flying lemur (Schmitz et al. 2002), is likely to be an artifact of mutational convergence in mitochondria. Furthermore, the controversial placement of the tarsier as sister group to the prosimians rather than to the anthropoid primates may well also be an artifact of mutational convergence. By placing these convergences in the context of response to structural aspects of the replication system, considerable explanatory power was provided to what is otherwise a confusing mixture of outcomes of these processes (that is, the average nucleotide frequencies reached at dynamic equilibrium).

The tools presented here are useful for comparative analysis and documenting the extent and range of evolution of mutational responses. The earlier observation of an average linear response of  $A \Rightarrow G$  substitutions in the vertebrates was based on a gene-by-gene

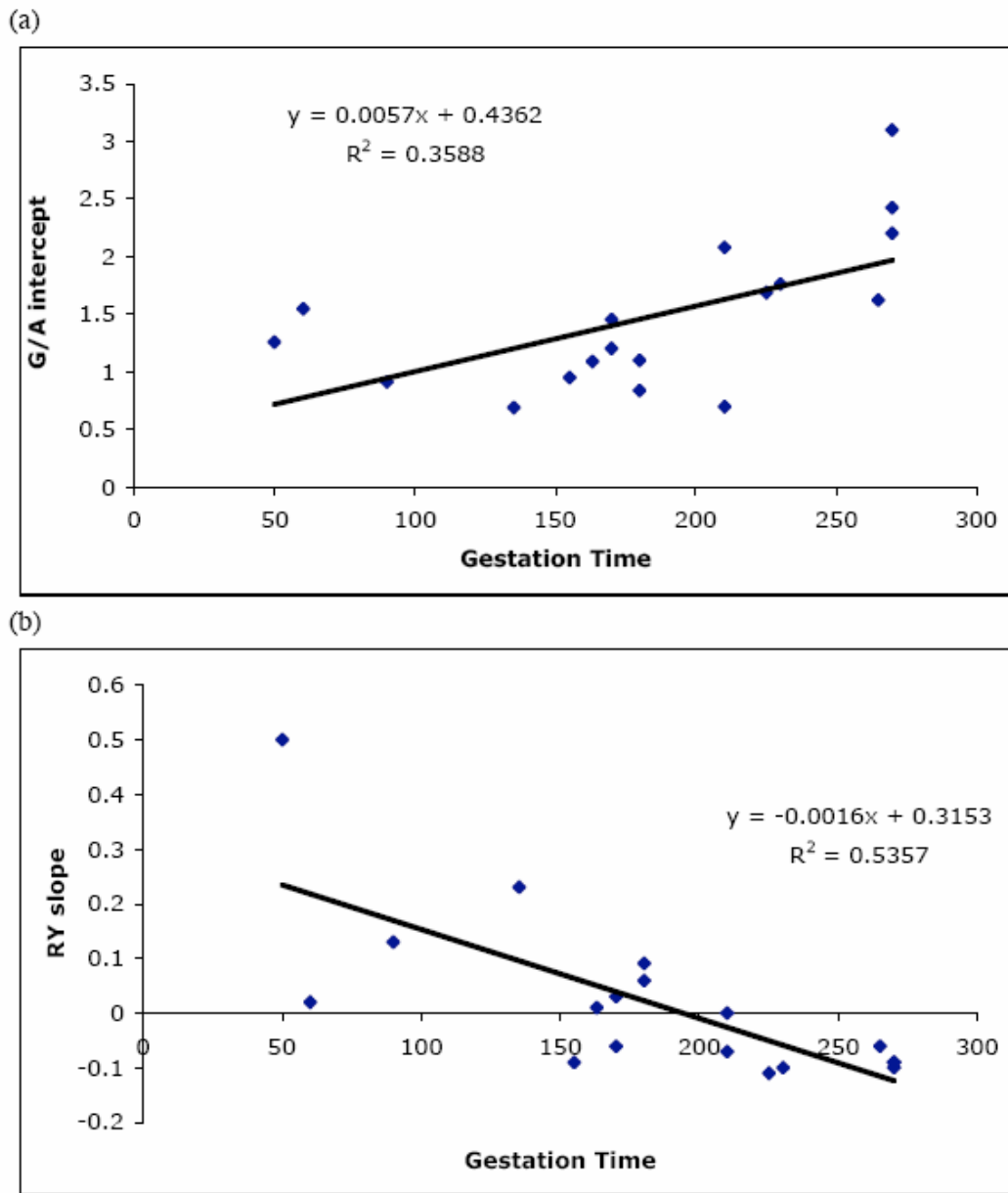
analysis using phylogeny-based maximum likelihood techniques (Faith and Pollock 2003), but given the expectation of a linear response to single-strandedness, the ability to assess the strength of the response in individual genomes with the likelihood approaches is surprisingly good. Based on the current analysis, incorporation of a gradient evolution model directly into phylogeny-based likelihood analysis, which could include allowing for changes in the strength of response along the phylogeny, will be necessary to obtain accurate estimates and variances for topology and divergence times. Although there are considerable challenges in developing such a model, since the mutation process is different at every site in the genome, the expected power and accuracy of such a method will be much greater than existing methods that allow for variable rates of evolution along the tree and among sites, but maintain constant relative evolutionary processes among sites. The consistency of the change in response to the gradient of single-strandedness will potentially allow the development of what would be a unique mixing of non-stationary models with differences in the substitution process at every site in a genome.

The existence of these substitution gradients along the genome that vary with substitution type and over time helps make a strong argument for dense taxonomic sampling, i.e., “genomic biodiversity” (Pollock et al. 2000) even stronger. Higher density sampling allows for more accurate prediction of site-specific rates in complex models, and more accurate prediction of site-specific differences can be extremely beneficial to phylogenetic reconstruction using likelihood-based techniques (Pollock and Bruno 2000). If the taxa sampled are closely related, a more accurate description of the mutation process should be obtained (Bielawski and Gold 1996). Furthermore, since the gradients

appear to change over time, increased taxonomic sampling would allow more precise delineation of whether these changes are gradual or occur in large bursts. A phylogeny-based Bayesian analysis was developed to more precisely model the evolution of these gradients, and greater amounts of taxon sampling will allow better direct inference of ancestral gradients, as well as better descriptions of the response curves for other substitutions besides A $\Rightarrow$ G, which are clearly non-linear (Faith and Pollock 2003). May not be needed for thesis.

Other potentially important effects of these gradients, and the evolution of these gradients, that should be considered are what kind of effect they have had on amino acid substitutions, whether they can be incorporated into codon-based models, and whether they substantially affect the ability to detect selection and adaptation in mitochondria using synonymous versus non-synonymous substitution ratios. They may also affect how synonymous and non-synonymous ratios are used in population genetics to understand how selection affects polymorphism levels.

Since mitochondria are so closely tied to metabolism and energy consumption, it is relevant to consider whether the observed evolutionary changes might be tied to concurrent changes in physiology. The G/A response intercept has a significant positive slope when regressed against gestation time (Figure 12a;  $P < 0.01$ ), and the R/Y response slope versus gestation time is significantly negative (Figure 12b;  $P < 0.01$ ). In both of these cases, there are weaker correlations with other physiological factors that are themselves highly correlated with gestation time, including brain weight, longevity, and body mass at birth. The reasons for these correlations, although interesting, remain highly speculative. To accurately dissect causal factors and determine statistical significance



**Figure 12.** Linear regression of G/A intercept and R/Y slope versus gestation time. The slope, intercept, and  $R^2$  values are shown next to the regression lines.

will require a phylogenetic method for reconstructing ancestral gradients, as well as higher density sampling within primates and among other vertebrates to obtain more accurate reconstructions of ancestral gradient, more examples of large-scale changes in

gradient response curves, and more examples of large changes in brain weight, longevity, body mass at birth, and/or gestation time.

## REFERENCES

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. F. 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.

Asakawa, S., Y. Kumazawa, T. Araki, H. Himeno, K. Miura, and K. Watanabe. 1991. Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. *J Mol Evol* 32:511-520.

Bielawski, J. P., and J. R. Gold. 1996. Unequal synonymous substitution rates within and between two protein-coding mitochondrial genes. *Mol Biol Evol* 13:889-892.

Bogenhagen, D. F., and D. A. Clayton. 2003a. Concluding remarks: The mitochondrial DNA replication bubble has not burst. *Trends Biochem Sci* 28:404-405.

Bogenhagen, D. F., and D. A. Clayton. 2003b. The mitochondrial DNA replication bubble has not burst. *Trends Biochem Sci* 28:357-360.

Bowmaker, M., M. Y. Yang, T. Yasukawa, A. Reyes, H. T. Jacobs, J. A. Huberman, and I. J. Holt. 2003. Mammalian mitochondrial DNA replicates bidirectionally from an initiation zone. *J Biol Chem* 278:50961-50969.

Delorme, M. O., and A. Henaut. 1991. Codon usage is imposed by the gene location in the transcription unit. *Curr Genet* 20:353-358.

Faith, J. J., and D. D. Pollock. 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* 165:735-745.

Felsenstein, J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. *Systematic Zoology* 27:401-410.

Felsenstein, J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J Mol Evol* 53:447-455.

- Frederico, L. A., T. A. Kunkel, and B. R. Shaw. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* 29:2532-2537.
- Frederico, L. A., T. A. Kunkel, and B. R. Shaw. 1993. Cytosine deamination in mismatched base pairs. *Biochemistry* 32:6523-6530.
- Gissi, C., A. Reyes, G. Pesole, and C. Saccone. 2000. Lineage-specific evolutionary rate in mammalian mtDNA. *Mol Biol Evol* 17:1022-1031.
- Graybeal, A. 1993. The phylogenetic utility of cytochrome b: lessons from bufonid frogs. *Mol Phylogenet Evol* 2:256-269.
- Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97-109.
- Holt, I. J., and H. T. Jacobs. 2003. Response: The mitochondrial DNA replication bubble has not burst. *Trends Biochem Sci* 28:355-356.
- Holt, I. J., H. E. Lorimer, and H. T. Jacobs. 2000. Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA. *Cell* 100:515-524.
- Honeycutt, R. L., M. A. Nedbal, R. M. Adkins, and L. L. Janecek. 1995. Mammalian mitochondrial DNA evolution: a comparison of the cytochrome b and cytochrome c oxidase II genes. *J Mol Evol* 40:260-272.
- 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.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.
- Ingman, M., H. Kaessmann, S. Paabo, and U. Gyllensten. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708-713.
- Jermiin, L. S., D. Graur, and R. H. Crozier. 1995. Evidence from Analyses of Intergenic Regions for Strand-Specific Directional Mutation Pressure in Metazoan Mitochondrial-DNA. *Molecular Biology and Evolution* 12:558-563.
- Jermiin, L. S., D. Graur, R. M. Lowe, and R. H. Crozier. 1994. Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome b genes. *J Mol Evol* 39:160-173.
- Krasuski, A., J. Galinski, R. T. Smolenski, and M. Marlewski. 1997. [Deamination of adenine and adenosine in staphylococci]. *Med Dosw Mikrobiol* 49:113-122.

- Limaiem, J., and A. Henaut. 1984a. [Fluctuation of the incidence of the 4 bases along the mitochondrial genome of mammals using correspondence factorial analysis]. *C R Acad Sci III* 298:279-286.
- Limaiem, J., and A. Henaut. 1984b. [Demonstration of a sudden change in the use of codons in the vicinity of transcription termination]. *C R Acad Sci III* 299:275-280.
- Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, and A. W. Larkum. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. *J Mol Evol* 34:153-162.
- McLachlan, G., and D. Peel. 2000. *Finite mixture models*. Wiley-Interscience.
- 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.
- Meyer, A. 1994. Shortcomings of the Cytochrome-B Gene as a Molecular Marker. *Trends in Ecology & Evolution* 9:278-280.
- Parham, J. C., J. Fissekis, and G. B. Brown. 1966. Purine-N-oxides. 18. Deamination of adenine-N-oxide derivatives. *J Org Chem* 31:966-968.
- Perna, N. T., and T. D. Kocher. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J Mol Evol* 41:353-358.
- Philippe, H., and J. Laurent. 1998. How good are deep phylogenetic trees? *Curr Opin Genet Dev* 8:616-623.
- 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., J. A. Eisen, N. A. Doggett, and M. P. Cummings. 2000. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol Biol Evol* 17:1776-1788.
- Reyes, A., C. Gissi, G. Pesole, and C. Saccone. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* 15:957-966.
- Reyes, A., G. Pesole, and C. Saccone. 2000. Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. *Gene* 259:177-187.
- Rice, J. A. 1995. *Mathematical statistics and data analysis*. Duxbury Press, Belmont, California.
- Schmitz, J., M. Ohme, B. Suryobroto, and H. Zischler. 2002. The colugo (*Cynocephalus variegatus*, Dermoptera): the primates' gliding sister? *Mol Biol Evol* 19:2308-2312.

- 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.
- Schmitz, J., M. Ohme, and H. Zischler. 2001. SINE insertions in cladistic analyses and the phylogenetic affiliations of *Tarsius bancanus* to other primates. *Genetics* 157:777-784.
- Swofford, D. L. 2000. *Phylogenetic analysis using parsimony (\*and other methods)*. Sinauer Associates, Sunderland, Massachusetts.
- Tanaka, M., and T. Ozawa. 1994. Strand asymmetry in human mitochondrial DNA mutations. *Genomics* 22:327-335.
- Tarr, H. L., and A. G. Comer. 1964. Deamination of Adenine and Related Compounds and Formation of Deoxyadenosine and Deoxyinosine by Lingcod Muscle Enzymes. *Can J Biochem Physiol* 42:1527-1533.
- Van Den Bussche, R. A., R. J. Baker, J. P. Huelsenbeck, and D. M. Hillis. 1998. Base compositional bias and phylogenetic analyses: a test of the "flying DNA" hypothesis. *Mol Phylogenet Evol* 10:408-416.
- Wiens, J. J., and B. D. Hollingsworth. 2000. War of the Iguanas: conflicting molecular and morphological phylogenies and long-branch attraction in iguanid lizards. *Syst Biol* 49:143-159.
- Xu, X., and U. Arnason. 1996. The mitochondrial DNA molecule of Sumatran orangutan and a molecular proposal for two (Bornean and Sumatran) species of orangutan. *J Mol Evol* 43:431-437.
- Yang, M. Y., M. Bowmaker, A. Reyes, L. Vergani, P. Angeli, E. Gringeri, H. T. Jacobs, and I. J. Holt. 2002. Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for apparent strand-asymmetric DNA replication. *Cell* 111:495-505.
- Yoder, A. D., R. Vilgalys, and M. Ruvolo. 1996. Molecular evolutionary dynamics of cytochrome b in strepsirrhine primates: the phylogenetic significance of third-position transversions. *Mol Biol Evol* 13:1339-1350.

## APPENDIX : SUPPLEMENTARY TABLES

**Table A.** Difference in log likelihood values ( $\delta\text{LnL}$ ) between independent and paired analyses of the G/A gradient for all species pairs.

|     | Ptr         | Ppa         | Ggo         | Ppy   | Pab         | Hla         | Msy         | Pha         | Cae         | Cgu   | Tob    | Cal    | Nco         | Lca         | Tba         | Tbe         | Cva         |
|-----|-------------|-------------|-------------|-------|-------------|-------------|-------------|-------------|-------------|-------|--------|--------|-------------|-------------|-------------|-------------|-------------|
| Hsa | <b>2.88</b> | <b>2.96</b> | <b>3.01</b> | 4.68  | 6.27        | 1.59        | 7.60        | 3.38        | 14.43       | 49.85 | 137.57 | 87.69  | 43.08       | 108.00      | 69.46       | 25.81       | 4.62        |
| Ptr |             | <b>0.05</b> | <b>0.19</b> | 14.30 | 17.62       | 7.46        | <b>2.48</b> | <b>1.18</b> | 5.03        | 29.33 | 101.91 | 60.10  | 24.13       | 76.58       | 44.54       | 11.91       | <b>0.32</b> |
| Ppa |             |             | <b>0.05</b> | 14.01 | 17.69       | 7.15        | <b>1.86</b> | <b>0.75</b> | 4.52        | 29.77 | 102.55 | 60.78  | 24.03       | 76.84       | 44.70       | 12.11       | <b>0.21</b> |
| Ggo |             |             |             | 13.58 | 17.60       | 6.79        | <b>1.45</b> | <b>0.43</b> | 4.31        | 30.65 | 103.99 | 62.04  | 24.39       | 77.85       | 45.45       | 12.62       | <b>0.22</b> |
| Ppy |             |             |             |       | <b>0.94</b> | <b>1.17</b> | 18.96       | 11.99       | 31.30       | 83.58 | 190.27 | 130.73 | 72.67       | 154.01      | 107.34      | 50.34       | 16.82       |
| Pab |             |             |             |       |             | 3.31        | 25.17       | 16.84       | 38.35       | 90.28 | 200.05 | 138.32 | 80.76       | 164.05      | 116.15      | 56.12       | 21.09       |
| Hla |             |             |             |       |             |             | 10.88       | 5.70        | 20.41       | 65.11 | 161.38 | 107.46 | 55.55       | 128.31      | 86.04       | 36.57       | 9.22        |
| Msy |             |             |             |       |             |             |             | <b>0.88</b> | <b>1.61</b> | 27.09 | 94.17  | 56.11  | 19.19       | 68.49       | 38.83       | 10.81       | <b>1.19</b> |
| Pha |             |             |             |       |             |             |             |             | 4.52        | 34.28 | 109.01 | 66.68  | 26.57       | 81.82       | 48.78       | 15.07       | <b>0.92</b> |
| Cae |             |             |             |       |             |             |             |             |             | 16.03 | 71.76  | 39.43  | 9.89        | 49.66       | 24.91       | 4.81        | <b>2.78</b> |
| Cgu |             |             |             |       |             |             |             |             |             |       | 21.68  | 5.66   | <b>1.55</b> | 11.79       | <b>2.12</b> | 3.34        | 24.87       |
| Tob |             |             |             |       |             |             |             |             |             |       |        | 5.32   | 27.10       | <b>1.99</b> | 11.92       | 40.42       | 91.71       |
| Cal |             |             |             |       |             |             |             |             |             |       |        |        | 10.03       | <b>2.09</b> | <b>2.36</b> | 17.02       | 53.28       |
| Nco |             |             |             |       |             |             |             |             |             |       |        |        |             | 14.67       | 3.20        | <b>2.51</b> | 19.47       |
| Lca |             |             |             |       |             |             |             |             |             |       |        |        |             |             | 4.43        | 26.10       | 67.75       |
| Tba |             |             |             |       |             |             |             |             |             |       |        |        |             |             |             | 9.51        | 38.14       |
| Tbe |             |             |             |       |             |             |             |             |             |       |        |        |             |             |             |             | 9.28        |

**Table B.** Difference in log likelihood values ( $\delta\text{LnL}$ ) for hierarchical clustering analyses with the G/A gradient.

| Group | ML         | $\delta\text{LnL}$ | Slope | 95% CI        | Intercept | 95% CI        | Members            |
|-------|------------|--------------------|-------|---------------|-----------|---------------|--------------------|
| 1     | -2667.904  | 0.045              | 1.125 | [0.677,1.522] | 1.653     | [1.402,1.948] | Ppa, Ggo           |
| 2     | -4007.134  | 0.148              | 1.056 | [0.702,1.373] | 1.690     | [1.482,1.930] | Ptr, Group 1       |
| 3     | -5277.082  | 0.324              | 1.075 | [0.794,1.376] | 1.657     | [1.468,1.860] | Cva, Group 2       |
| 4     | -2582.905  | 0.885              | 1.663 | [1.238,2.110] | 1.319     | [1.090,1.568] | Msy, Pha           |
| 5     | -2360.598  | 0.941              | 1.116 | [0.357,1.779] | 2.739     | [2.322,3.286] | Ppy, Pab           |
| 6     | -2762.157  | 1.551              | 0.737 | [0.484,0.970] | 1.006     | [0.859,1.185] | Cgu, Nco           |
| 7     | -2491.481  | 1.586              | 1.175 | [0.557,1.736] | 2.146     | [1.806,2.543] | Hsa, Hla           |
| 8     | -2880.060  | 1.986              | 0.502 | [0.325,0.661] | 0.693     | [0.588,0.817] | Tob, Lca           |
| 9     | -2830.129  | 2.363              | 0.537 | [0.332,0.746] | 0.886     | [0.752,1.042] | Cal, Tba           |
| 10    | -7862.619  | 2.632              | 1.265 | [1.025,1.506] | 1.544     | [1.397,1.698] | Group 3, Group 4   |
| 11    | -4029.290  | <u>3.391</u>       | 0.722 | [0.495,0.949] | 1.084     | [0.948,1.235] | Tbe, Group 6       |
| 12    | -9222.544  | 5.990              | 1.307 | [1.079,1.517] | 1.469     | [1.341,1.613] | Cae, Group 10      |
| 13    | -4858.542  | 6.463              | 1.166 | [0.742,1.612] | 2.413     | [2.127,2.715] | Group 5, Group 7   |
| 14    | -5719.941  | 9.753              | 0.518 | [0.385,0.647] | 0.784     | [0.698,0.872] | Group 8, Group 9   |
| 15    | -9794.258  | 45.026             | 0.588 | [0.467,0.702] | 0.901     | [0.827,0.984] | Group 11, Group 14 |
| 16    | -14140.676 | 59.590             | 1.292 | [1.085,1.494] | 1.747     | [1.624,1.879] | Group 12, Group 13 |
| 17    | -24431.846 | 496.912            | 0.928 | [0.821,1.043] | 1.351     | [1.281,1.428] | Group 15, Group 16 |

**Table C.** Difference in log likelihood values ( $\delta\text{LnL}$ ) and maximum likelihood values and CIs for slope and intercept for mixture model analyses with the G/A gradient.

| <b>Models</b> | <b>ML</b> | <b><math>\delta\text{LnL}</math></b> | <b>#</b> | <b>slope</b> | <b>95%CI</b>     | <b>intercept</b> | <b>95%CI</b>    |
|---------------|-----------|--------------------------------------|----------|--------------|------------------|------------------|-----------------|
| 1             | -21889.1  | --                                   | 1        | 0.93         | [ 0.81 , 1.06 ]  | 1.35             | [ 1.26 , 1.43 ] |
| 2             | -21387.1  | 501.961                              | 1        | 0.58         | [ 0.47 , 0.69 ]  | 0.85             | [ 0.78 , 0.93 ] |
|               |           |                                      | 2        | 1.32         | [ 1.09 , 1.52 ]  | 1.76             | [ 1.64 , 1.91 ] |
| 3             | -21337.3  | 49.81                                | 1        | 0.58         | [ 0.47 , 0.70 ]  | 0.85             | [ 0.78 , 0.93 ] |
|               |           |                                      | 2        | 1.35         | [ 1.11 , 1.56 ]  | 1.44             | [ 1.32 , 1.62 ] |
|               |           |                                      | 3        | 1.20         | [ 0.66 , 1.62 ]  | 2.39             | [ 2.13 , 2.78 ] |
| 4             | -21310.9  | 26.41                                | 1        | 0.47         | [ 0.32 , 0.62 ]  | 0.77             | [ 0.67 , 0.86 ] |
|               |           |                                      | 2        | 0.74         | [ 0.54 , 0.90 ]  | 0.95             | [ 0.84 , 1.08 ] |
|               |           |                                      | 3        | 1.38         | [ 1.07 , 1.57 ]  | 1.44             | [ 1.33 , 1.64 ] |
|               |           |                                      | 4        | 1.16         | [ 0.70 , 1.60 ]  | 2.42             | [ 2.16 , 2.77 ] |
| 5             | -21307.0  | 3.88                                 | 1        | 0.49         | [ 0.32 , 0.62 ]  | 0.76             | [ 0.66 , 0.86 ] |
|               |           |                                      | 2        | 0.74         | [ 0.55 , 0.91 ]  | 0.92             | [ 0.84 , 1.07 ] |
|               |           |                                      | 3        | 1.60         | [ 1.14 , 1.88 ]  | 1.12             | [ 0.98 , 1.56 ] |
|               |           |                                      | 4        | 1.20         | [ 0.63 , 1.58 ]  | 1.69             | [ 1.49 , 2.51 ] |
|               |           |                                      | 5        | 1.23         | [ -0.18 , 1.76 ] | 2.51             | [ 2.16 , 3.74 ] |
| 6             | -20304.1  | 2.87                                 | 1        | 0.45         | [ 0.38 , 0.59 ]  | 0.76             | [ 0.68 , 0.81 ] |
|               |           |                                      | 2        | 0.69         | [ 0.54 , 0.82 ]  | 0.96             | [ 0.86 , 1.00 ] |
|               |           |                                      | 3        | 1.58         | [ 1.43 , 1.77 ]  | 1.17             | [ 1.02 , 1.26 ] |
|               |           |                                      | 4        | 1.14         | [ -1.20 , 1.32 ] | 1.64             | [ 1.15 , 1.74 ] |
|               |           |                                      | 5        | 1.14         | [ 0.89 , 1.44 ]  | 2.18             | [ 1.68 , 2.25 ] |
|               |           |                                      | 6        | 1.16         | [ 0.81 , 2.03 ]  | 2.69             | [ 1.97 , 2.99 ] |

**Table D.** Maximum likelihood values & 95% CI for slopes and intercepts of C/T gradients.

| <b>Species</b>                 | <b>Max Like</b> | <b>Slope</b>            | <b>Intercept</b>     |
|--------------------------------|-----------------|-------------------------|----------------------|
| <i>Homo sapiens</i>            | -574.508        | -0.099 [-0.160, -0.035] | 0.204 [0.156, 0.258] |
| <i>Pan troglodytes</i>         | -537.508        | -0.070 [-0.128, -0.014] | 0.168 [0.128, 0.218] |
| <i>Pan paniscus</i>            | -485.382        | -0.054 [-0.103, -0.007] | 0.138 [0.103, 0.181] |
| <i>Gorilla gorilla</i>         | -552.404        | -0.072 [-0.130, -0.013] | 0.175 [0.129, 0.226] |
| <i>Pongo pygmaeus</i>          | -532.055        | -0.096 [-0.148, -0.045] | 0.188 [0.147, 0.235] |
| <i>Pongo pygmaeus abelii</i>   | -553.196        | -0.067 [-0.127, -0.008] | 0.173 [0.128, 0.230] |
| <i>Hylobates lar</i>           | -676.377        | -0.094 [-0.169, -0.025] | 0.241 [0.187, 0.306] |
| <i>Macaca sylvanus</i>         | -579.117        | -0.069 [-0.119, -0.018] | 0.174 [0.134, 0.223] |
| <i>Papio hamadryas</i>         | -629.595        | -0.112 [-0.169, -0.046] | 0.228 [0.179, 0.280] |
| <i>Cercopithecus aethiops</i>  | -560.245        | -0.095 [-0.150, -0.049] | 0.190 [0.152, 0.241] |
| <i>Colobus guereza</i>         | -603.767        | -0.056 [-0.121, 0.006]  | 0.171 [0.124, 0.228] |
| <i>Trachypithecus obscurus</i> | -556.391        | -0.041 [-0.093, 0.015]  | 0.145 [0.106, 0.191] |
| <i>Cebus albifrons</i>         | -503.499        | -0.097 [-0.140, -0.052] | 0.161 [0.123, 0.202] |
| <i>Nycticebus coucang</i>      | -731.419        | -0.151 [-0.217, -0.090] | 0.276 [0.224, 0.340] |
| <i>Lemur catta</i>             | -381.697        | -0.028 [-0.062, 0.004]  | 0.081 [0.057, 0.113] |
| <i>Tarsius bancanus</i>        | -354.41         | -0.048 [-0.084, -0.015] | 0.091 [0.064, 0.124] |
| <i>Tupaia belangeri</i>        | -720.159        | -0.090 [-0.149, -0.034] | 0.213 [0.170, 0.267] |
| <i>Cynocephalus variegatus</i> | -657.057        | -0.124 [-0.185, -0.066] | 0.236 [0.189, 0.292] |

**Table E.** Difference in log likelihood values ( $\delta\text{LnL}$ ), MLEs and CIs for hierarchical clustering analyses with the C/T gradient.

| Group | ML         | $\delta\text{LnL}$ | Slope  | 95% CI          | Intercept | 95% CI        | Members            |
|-------|------------|--------------------|--------|-----------------|-----------|---------------|--------------------|
| 1     | -1132.319  | 0.005              | -0.068 | [-0.108,-0.028] | 0.173     | [0.143,0.210] | Pab, Msy           |
| 2     | -1684.743  | 0.019              | -0.069 | [-0.103,-0.033] | 0.174     | [0.147,0.204] | Ggo, Group 1       |
| 3     | -1092.322  | 0.020              | -0.097 | [-0.132,-0.056] | 0.190     | [0.156,0.223] | Ppy, Cae           |
| 4     | -1286.693  | 0.040              | -0.118 | [-0.167,-0.074] | 0.232     | [0.196,0.275] | Pha, Cva           |
| 5     | -2222.444  | 0.193              | -0.069 | [-0.097,-0.042] | 0.172     | [0.150,0.196] | Ptr, Group 2       |
| 6     | -2007.164  | 0.312              | -0.109 | [-0.145,-0.075] | 0.226     | [0.199,0.260] | Tbe, Group 4       |
| 7     | -2826.595  | 0.383              | -0.066 | [-0.094,-0.042] | 0.172     | [0.152,0.195] | Cgu, Group 5       |
| 8     | -736.590   | 0.476              | -0.037 | [-0.062,-0.009] | 0.085     | [0.061,0.108] | Lca, Tba           |
| 9     | -1667.328  | 0.497              | -0.097 | [-0.132,-0.064] | 0.194     | [0.167,0.225] | Hsa, Group 3       |
| 10    | -3383.627  | 0.642              | -0.062 | [-0.085,-0.037] | 0.167     | [0.147,0.189] | Tob, Group 7       |
| 11    | -1408.715  | 0.918              | -0.126 | [-0.174,-0.076] | 0.261     | [0.219,0.306] | Hla, Nco           |
| 12    | -990.204   | 1.322              | -0.076 | [-0.107,-0.040] | 0.149     | [0.119,0.178] | Ppa, Cal           |
| 13    | -5052.389  | 1.434              | -0.073 | [-0.091,-0.057] | 0.176     | [0.163,0.192] | Group 9, Group 10  |
| 14    | -3418.328  | <u>2.449</u>       | -0.117 | [-0.145,-0.086] | 0.240     | [0.214,0.267] | Group 6, Group 11  |
| 15    | -6051.045  | 8.452              | -0.076 | [-0.090,-0.061] | 0.172     | [0.158,0.185] | Group 12, Group 13 |
| 16    | -9494.782  | 25.408             | -0.087 | [-0.099,-0.074] | 0.192     | [0.180,0.204] | Group 14, Group 15 |
| 17    | -10298.870 | 67.498             | -0.081 | [-0.093,-0.068] | 0.179     | [0.169,0.190] | Group 8, Group 16  |

**Table F.** Maximum likelihood values and CIs for slope and intercept of the Y/R gradient at four-fold redundant sites.

| <b>Species</b>                 | <b>Max Like</b> | <b>Slope</b>           | <b>Intercept</b>     |
|--------------------------------|-----------------|------------------------|----------------------|
| <i>Homo sapiens</i>            | -1400.82        | -0.085 [-0.320, 0.104] | 0.852 [0.709, 1.038] |
| <i>Pan troglodytes</i>         | -1402.11        | -0.097 [-0.284, 0.111] | 0.860 [0.697, 1.022] |
| <i>Pan paniscus</i>            | -1397.57        | -0.108 [-0.309, 0.093] | 0.866 [0.719, 1.047] |
| <i>Gorilla gorilla</i>         | -1390.7         | -0.062 [-0.261, 0.127] | 0.832 [0.699, 0.995] |
| <i>Pongo pygmaeus</i>          | -1424.58        | -0.099 [-0.305, 0.092] | 0.853 [0.715, 1.027] |
| <i>Pongo pygmaeus abelii</i>   | -1425.01        | -0.091 [-0.294, 0.120] | 0.863 [0.696, 1.029] |
| <i>Hylobates lar</i>           | -1422.7         | 0.009 [-0.219, 0.207]  | 0.839 [0.688, 1.014] |
| <i>Macaca sylvanus</i>         | -1386.52        | -0.062 [-0.324, 0.178] | 0.964 [0.789, 1.181] |
| <i>Papio hamadryas</i>         | -1391.75        | 0.033 [-0.211, 0.260]  | 0.866 [0.712, 1.057] |
| <i>Cercopithecus aethiops</i>  | -1380.54        | 0.013 [-0.208, 0.228]  | 0.867 [0.707, 1.052] |
| <i>Colobus guereza</i>         | -1329.45        | 0.064 [-0.176, 0.310]  | 0.845 [0.688, 1.029] |
| <i>Trachypithecus obscurus</i> | -1308.35        | -0.074 [-0.300, 0.143] | 0.916 [0.748, 1.112] |
| <i>Cebus albifrons</i>         | -1329.75        | -0.086 [-0.394, 0.194] | 1.091 [0.888, 1.347] |
| <i>Nycticebus coucang</i>      | -1349.58        | 0.125 [-0.193, 0.410]  | 1.034 [0.830, 1.290] |
| <i>Lemur catta</i>             | -1301.46        | 0.231 [-0.043, 0.487]  | 0.831 [0.661, 1.037] |
| <i>Tarsius bancanus</i>        | -1313.28        | 0.089 [-0.148, 0.315]  | 0.852 [0.689, 1.054] |
| <i>Tupaia belangeri</i>        | -1330.48        | 0.498 [0.110, 0.822]   | 0.973 [0.777, 1.248] |
| <i>Cynocephalus variegatus</i> | -1402.81        | 0.024 [-0.218, 0.291]  | 1.003 [0.820, 1.193] |

**Table G.** Difference in log likelihood values ( $\delta\text{LnL}$ ), MLEs and CIs for hierarchical clustering analyses with the Y/R gradient at four-fold redundant sites.

| <b>Group</b> | <b>ML</b>  | <b><math>\delta\text{LnL}</math></b> | <b>Slope</b> | <b>95% CI</b>   | <b>Intercept</b> | <b>95% CI</b> | <b>Members</b> |          |
|--------------|------------|--------------------------------------|--------------|-----------------|------------------|---------------|----------------|----------|
| 1            | -2802.929  | 0.004                                | -0.089       | [-0.239,0.044]  | 0.855            | [0.753,0.971] | Hsa,           | Ptr      |
| 2            | -4227.95   | 0.01                                 | -0.09        | [-0.216,0.026]  | 0.858            | [0.779,0.958] | Pab,           | Group 1  |
| 3            | -2822.163  | 0.013                                | -0.104       | [-0.239,0.030]  | 0.86             | [0.755,0.972] | Ppa,           | Ppy      |
| 4            | -2721.216  | 0.019                                | 0.049        | [-0.112,0.211]  | 0.856            | [0.739,0.983] | Pha,           | Cgu      |
| 5            | -5618.687  | 0.04                                 | -0.085       | [-0.197,0.010]  | 0.852            | [0.779,0.941] | Ggo,           | Group 2  |
| 6            | -8440.897  | 0.047                                | -0.091       | [-0.185,-0.011] | 0.855            | [0.794,0.926] | Group 3        | Group 5  |
| 7            | -4101.818  | 0.064                                | 0.038        | [-0.104,0.162]  | 0.859            | [0.762,0.973] | Cae,           | Group 4  |
| 8            | -2732.737  | 0.173                                | -0.027       | [-0.228,0.158]  | 1.044            | [0.910,1.196] | Cal,           | Cva      |
| 9            | -2731.247  | 0.203                                | -0.032       | [-0.182,0.120]  | 0.877            | [0.764,1.002] | Hla            | Tob,     |
| 10           | -5415.33   | 0.233                                | 0.05         | [-0.066,0.165]  | 0.857            | [0.772,0.946] | Tba,           | Group 7  |
| 11           | -6802.412  | 0.564                                | 0.027        | [-0.077,0.137]  | 0.878            | [0.799,0.960] | Msy,           | Group 10 |
| 12           | -9534.654  | 0.994                                | 0.01         | [-0.081,0.096]  | 0.878            | [0.814,0.944] | Group 9        | Group 11 |
| 13           | -4035.749  | 1.555                                | 0.064        | [-0.086,0.221]  | 0.968            | [0.857,1.090] | Lca,           | Group 8  |
| 14           | -5387.217  | <u>1.89</u>                          | 0.08         | [-0.058,0.219]  | 0.984            | [0.885,1.093] | Nco,           | Group 13 |
| 15           | -17986.928 | 11.378                               | -0.039       | [-0.100,0.018]  | 0.867            | [0.822,0.916] | Group 6        | Group 12 |
| 16           | -6729.539  | 11.843                               | 0.145        | [0.015,0.268]   | 0.988            | [0.892,1.087] | Tbe,           | Group 14 |
| 17           | -24777.996 | 61.528                               | -0.001       | [-0.052,0.055]  | 0.901            | [0.857,0.942] | Group 15       | Group 16 |

## **VITA**

Sameer Raina is a native of Kashmir in India. He was born in 1979 and he came to the United States in 1997 to pursue a Bachelor of Science in Electrical Engineering. After completing his bachelor's he sought to combine his mathematical and computational background with his interests in evolution and cognitive science. He found the right kind of opportunity and environment in Dr. David Pollock's Evolutionary Biology Laboratory where he worked as a research fellow while simultaneously pursuing a Masters of Science in engineering science. Since then he has become engaged in the field of genomics and phylogenetic reconstruction, especially in the context of mutations and mutation rates. He plans to work in this field for a few years after which he will be considering pursuing a doctorate.