INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Université d'Ottawa · University of Ottawa

Phylogenetic implications of the effect of nucleotide bias on amino acid composition

Peter G. Foster

Thesis submitted to the
School of Graduate Studies and Research
University of Ottawa
in partial fulfillment of the requirements for the
Ph.D. degree in the
Ottawa-Carleton Institute of Biology

Ottawa, Ontario, Canada

October 6, 1997

Acknowledgements

I thank my supervisor, Donal Hickey, and the members of my advisory committee, Guy Drouin and George Carmody. I also thank the members of Dr. Hickey's group, especially Erin Yoshida, Lars Jermiin, and Kaarina Benkel. I also thank my family and friends.

Abstract

This study addresses the problem of amino acid compositional bias in molecular phylogenetic reconstruction based on protein sequences. Compositional bias is a factor that has largely been left out of conventional phylogenetic analyses, but there is now reason to believe that it is a significant factor.

It is shown that in animal mitochondria homologous genes that differ in AT/GC content code for proteins differing in amino acid content in a manner which reflects the AT and GC content of the codons. This relationship is also seen in nuclear and bacterial genes, and is significant even in the highly-conserved *hsp*70 and *ef*1α/*tu* genes.

This bias in the amino acid content of proteins can result in incorrect protein-based phylogenetic trees. A striking example is presented where common phylogenetic tools fail to recover the correct tree from animal mitochondrial protein sequences. The two taxa with the greatest compositional bias continually group together in these analyses, despite a lack of close biological relatedness. It is concluded that, while protein-based phylogenetic analyses have been favoured over analyses using biased DNA, even protein-based trees can be misleading.

Current models of protein sequence evolution used for phylogenetic reconstruction assume that the amino acid composition is stationary. In order to accommodate evolution involving compositional change, directional mutation probability models are proposed and their properties described. In simulation studies with biased and non-biased branches on the same tree, the biased branches "attract" in subsequent analyses, paralleling results above using real mitochondrial sequences. Improved methodology is proposed which can use both stationary and directional models in phylogenetic reconstruction, and which can use different models on different branches of the same phylogenetic tree.

Contents

List of Figures

List of Tables

Chapter 1

General introduction

This study addresses the problem of amino acid compositional bias in molecular phylogenetic reconstruction based on protein sequences. I describe below the relationship between nucleotide AT/GC bias in protein-coding genes and the amino acid content of the proteins for which they code. I show that genomes and homologous genes which differ in AT/GC content code for proteins that differ in amino acid content in a manner which reflects the AT and GC content of the codons of the amino acids. Having established this relationship I then examine its phylogenetic implications. I show that current methods of phylogenetic reconstruction, which generally ignore compositional bias, tend to fail due to this bias. Indeed, it was concern over the robustness of current methods of phylogenetic reconstruction using protein sequences derived from genes of differing AT/GC content that was the motivation for beginning this study. As I describe below, it has recently become generally accepted that biased DNA can affect phylogenetic reconstruction. For this reason some investigators have now opted to use protein sequences for phylogenetic work. That I show that phylogenetic inferences are also affected by protein bias will therefore be of interest.

The following four chapters develop these themes. Each of these chapters is written to stand alone: I ask the reader to forgive the repetition that this strategy entails. Below I will provide some background to introduce the chapters to follow.

1.1 DNA varies in AT/GC content, and this is reflected in the amino acid composition of encoded proteins

The first observation that AT/GC content in DNA varies among different organisms was made over 40 years ago (Lee et al., 1956). Both entire genomes and individual genes can vary in AT/GC content (Muto and Osawa, 1987; Jermiin et al., 1994; Hashimoto et al., 1994). In protein-coding genes, differences in nucleotide AT/GC content can occur at both synonymous (silent) and non-synonymous (replacement) codon sites. If only synonymous codon sites differ between genes then the corresponding protein sequences will be identical. However, if non-synonymous codon sites are affected the corresponding protein sequences will be different. In protein-coding genes, much of the difference in AT/GC content occurs at synonymous sites, so that while the DNA sequences may be quite different the protein sequences are less so. It is assumed that mutational biases affect all codon positions equally within a gene, but because of the countervailing force of natural selection, the effect of these biases accumulates much more rapidly at the synonymous sites.

The constraint of natural selection is not complete, and protein sequences are more or less free to vary—otherwise there would be no protein evolution. Here I would like to focus on the observation that directional mutation pressure on the AT/GC content of protein-coding DNA results in the protein sequences tending to become biased in a predictable direction, reflecting the AT/GC content of the codons. This observation has been made in various ways with various degrees of resolution for many years, as I describe here.

That nucleotide composition is related to amino acid content at the whole organism level in bacteria was shown 36 years ago by Sueoka (1961). Subsequent research has noted relationships between nucleotide composition and changes at non-synonymous codon sites with greater resolution. A study of animal mitochondrial genes showed that AT pressure affected both silent and replacement sites, and that codon changes reflected the AT bias, resulting in more AT-rich codons (Jukes and Bhushan, 1986). The high adenine content in the HIV RNA genome can be related

to high incidence of A-containing codons, resulting in some changes in amino acid use compared to related RNA genomes (Berkhout and van Hemert, 1994). A survey of a large number of human genes has shown a relationship between the AT/GC content at silent sites and amino acid composition (D'Onofrio et al., 1991; Collins and Jukes, 1993). An analysis of mitochondrial cytochrome b genes in several eukaryote orders has shown that there are large differences in the AT/GC content at both synonymous and non-synonymous codon sites (Jermiin et al., 1994). A survey of mammalian cytochrome P-450 genes has shown that the content of amino acids coded for by AT or GC-rich codons correlates with base compositional differences (Porter, 1995). However, work by Hashimoto et al. (1994; 1995) has shown that this is not the case in elongation factors 1α and 2.

In Chapter 2, below, I further delineate this relationship. I compare entire mitochondrial genomes of the chicken and honeybee and show that this effect is significant in all AT- and GC-rich codons, and in all genes in the two genomes. In Chapter 3 I survey the relationship between nucleotide AT/GC content and the content of AT- and GC-rich codons in several mitochondrial genomes, nuclear digestive enzymes and highly-conserved eukaryotic and bacterial genes. The effect is seen in all sequences examined, even in highly conserved *hsp*70 and *ef*1α / *tu* genes.

1.2 Phylogenetic analysis

Biological molecules of DNA and protein contain their history in their sequences. It is the challenge of molecular phylogenetics to infer that history from present-day sequences, to reconstruct the evolutionary tree which gave rise to extant molecules. In that respect it is similar to classical inference of evolution based on morphological or biochemical traits of living organisms, but it differs in that molecular phylogenetics generally cannot make use of fossils. Molecular phylogenetics can use either DNA or protein sequences. Historically, protein sequences were first available and first used. Protein sequences were important in the development of the molecular clock hypothesis and the neutral theory of evolution (Zuckerkandl and Pauling, 1965; Kimura,

1968; Kimura, 1983). With the advent of rapid DNA sequencing, DNA became the molecule of choice for molecular phylogenetics. Ribosomal RNA genes in particular have been very useful for phylogenetics because they are easy to obtain, of high information content, and ubiquitous. However, for reasons such as the desire for long lookback times or because of compositional bias as discussed below, protein sequences have again been favoured.

There are many methods available for phylogenetics (Hillis et al., 1993; Felsenstein, 1988; Felsenstein, 1996; Swofford et al., 1996; Nei, 1996). Three of the main classifications of methods of molecular phylogenetic inference are parsimony methods, distance-based methods, and maximum likelihood methods. Parsimony methods seek the shortest evolutionary path to explain the extant sequences. Distance methods use a two step process: first a matrix of distances is calculated between each sequence pair, and then a tree is inferred from this distance matrix. Maximum likelihood methods use an explicit probability model and seek a tree which will maximize the probability of obtaining the sequences under analysis.

These methods have their strengths and weaknesses. For example, parsimony methods suffer from unequal rate effects, such that long branches, representing rapid evolutionary change along those branches, tend to collapse into one another ("attract") in inferred trees. Weaknesses of model-based methods (this includes maximum likelihood and also the distance matrix computation step of distance-based methods) can stem from the assumptions of the evolutionary model used. The model of evolution used should of course reflect as much as possible the reality of evolution in order to accurately reconstruct it. Below, I discuss two of these assumptions: that of stationarity of composition, and that of homogeneity of the model throughout the tree. I describe phylogenetic implications when the sequences under analysis violate those assumptions, and argue that methodology can be improved by removing those assumptions.

1.3 Phylogenetic implications of composition bias

1.3.1 Phylogenetic implications of composition bias in DNA sequences

In recent years it has become apparent that nucleotide bias, including AT/GC bias and codon bias, can affect phylogenetic reconstruction based on DNA sequences (Hasegawa and Hashimoto, 1993). For example, in trying to trace the origins of cyanelles, the phylogeny inferred from DNA sequences is not in agreement with that obtained from biochemical data, due to fortuitous similarities of distant sequences (Lockhart et al., 1992). It has also been shown that codon usage bias affects phylogenetic tree reconstruction in dipteran actin genes (He and Haymer, 1995).

In a phylogenetic context, protein sequences are often considered more useful than DNA sequences because they are thought to be immune to base compositional differences occurring in the latter, and so phylogenetic information would be unaffected in the protein sequence. DNA sequences may be subject to directional mutation pressure, but by virtue of the redundancy of the genetic code, the amino acids coded for by this DNA can be buffered from this pressure. One can imagine sequences of widely differing AT/GC content yet with the same amino acid sequence, with all of the DNA differences occurring at synonymous codon sites. As mentioned above, it is assumed that directional mutation pressure affects all sites equally, but that many mutations at non-synonymous sites are not accepted because of natural selection. Indeed it is often assumed that the non-synonymous sites are totally free of any bias, an assumption which underlies our confidence in the use of amino acid sequence data as a means of constructing reliable alignments and building unbiased molecular phylogenies.

Loomis and Smith (1990) compared phylogenetic trees based on eight protein sequences with a previous tree based on small subunit rDNA and argued that the rDNA tree was misleading because of nucleotide content bias. Hasesgawa et al. (Hasegawa and Hashimoto, 1993; Hasegawa et al., 1993; Hashimoto et al., 1994; Hashimoto et al., 1995) examined the early branchings of the eukaryotes, and opted for phylogenetic trees based on the protein sequences of elongation factors 1α and 2 because

the genomes were of widely different AT/GC content, which they felt would bias the trees. In these studies, evidence is presented that amino acid composition in these proteins is not affected by widely differing AT/GC contents among the taxa examined.

1.3.2 Phylogenetic implications of composition bias in protein sequences

The studies referred to above provide evidence that nucleotide composition bias can affect phylogenetic reconstruction when using DNA sequences. We should also ask whether the same applies to amino acid composition bias. There is evidence cited above that protein sequences are not immune to compositional differences in the DNA. It can be noted that none of the studies that examined this relationship pursued its implications on phylogenetic reconstruction. The effect of this bias on phylogenetic reconstruction has not been well-studied. Steel et al. (1993), Lockhart et al. (1994), and Steel et al. (1995), focused on the effect of and correction for biased DNA on phylogenetic reconstruction, but make little mention that the same observations and correction techniques might be applied to protein sequences. In justifying the use of elongation factor protein sequences for phylogenetic studies, Hashimoto et al. (1994, 1995) point out that the amino acid composition of these proteins are not affected by extreme AT/GC-bias in the DNA, but do not give an opinion about whether phylogenetic reconstruction would have been affected had the proteins had a biased composition.

I explore phylogenetic implications in Chapter 4, where it is shown that DNA-driven bias in the amino acid content of proteins can result in incorrect protein-based phylogenetic trees. A striking example is presented where common phylogenetic tools fail to recover the correct tree from animal mitochondrial protein sequences. The data set is very extensive, containing several thousand sites per sequence, and the incorrect phylogenetic trees are statistically very well supported. The two taxa with the greatest compositional bias continually group together in these analyses, despite a lack of close biological relatedness. It can be concluded that even protein-based phylogenetic trees can be misleading, and so continued caution is advised in phylogenetic reconstruction

using protein sequences, especially those that are compositionally biased.

1.3.3 Biased patterns of amino acid changes

The probability of one amino acid substituting for another over evolutionary time can be described using mutation probability matrices. These can be used to model evolution as a Markov chain. The archetype is the PAM (accepted point mutation) matrix (Dayhoff et al., 1978). Similar matrices made from larger data sets have been calculated more recently (Jones et al., 1992; Gonnet et al., 1992). These matrices are constructed empirically by aligning homologous protein sequences, tabulating amino acid substitutions, and normalizing to 1 substitution in a protein 100 residues long, which is defined as 1 PAM. Matrices for greater than one PAM can be obtained by multiplying this matrix by itself: for example to obtain a mutation probability matrix for 100 PAM the 1 PAM matrix is raised to the power of 100. The relatedness odds matrix is derived from the mutation probability matrix and can be used as an estimate of the relatedness of two amino acids at the same site in an alignment of protein sequences. The relatedness odds matrix, or the log odds matrix as it is usually stated, is not a Markov process probability matrix and is not used in maximum likelihood tree estimation nor distance estimation by maximum likelihood (Felsenstein, 1996).

Substitution matrices, or relatedness odds matrices derived from them (Dayhoff et al., 1978), are widely used in sequence alignment, database searching, and phylogenetic reconstruction (Thompson et al., 1994; Pearson, 1990; Felsenstein, 1993). Substitution matrices can also be used in evolutionary simulations, which I describe in Chapter 5. All of the models described above are formulated assuming a stationary amino acid composition. That means that the frequency of the amino acids in an evolutionary simulation using those stationary models remain at or approach the average amino acid frequencies of the sequences from which the model was computed. In order to simulate evolution involving changing amino acid composition I develop directional mutation probability matrices. In the simulations in Chapter 5 I test whether the maximum likelihood method of phylogenetic reconstruction using its built-in stationary model is able to recover the correct tree topology in simple

four-taxon simulations where two of the distantly related branches evolve under a directional model of evolution. That the biased branches in these simulations tend to group together in the resulting analyses provides at least a partial explanation for the observation made in Chapter 4 that biased branches in real mitochondrial sequences tend to group together in phylogenetic analyses.

These results should not undermine our belief that biological sequences reflect their history, nor should it undermine our hope that we can infer that history, but it does mean that we should not be simplistic in our analyses. This study addresses the simplifying assumptions of stationarity of amino acid composition, and the assumption of homogeneity of the tree, that all branches of the inferred tree evolve under the same model. When compositional bias is extreme phylogenetic analysis can indeed be compromised by these assumptions, and improvements in methodology are proposed which remove these assumptions.

1.4 Objectives and Hypotheses

The first objective is to establish the relationship between AT/GC bias at the nucleotide level and amino acid composition at the protein level in the protein coding regions of entire mitochondrial genomes, and then to survey the extent to which that relationship holds in other genes. The hypothesis is that AT-rich protein coding genes will tend to be rich in amino acids coded for by AT-rich codons and poor in amino acids coded for by GC-rich codons, and that GC-rich protein coding genes will tend to be rich in amino acids coded for by GC-rich codons and poor in amino acids coded for by AT-rich codons. The second objective is to see if phylogenetic reconstruction using protein sequences is compromised by this amino acid composition bias. The hypothesis is that DNA-driven amino acid composition bias will compromise phylogenetic reconstruction and tend to make similarly-biased protein sequences group together erroneously in phylogenetic analyses.

Chapter 2

Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria*

2.1 Abstract

We show that in animal mitochondria homologous genes that differ in guanine plus cytosine (G+C) content code for proteins differing in amino acid content in a manner that relates to the G+C content of the codons. DNA sequences were analyzed using square plots, a new method that combines graphical visualization and statistical analysis of compositional differences in both DNA and protein. Square plots divide codons into 4 groups based on first and second position A+T (adenine plus thymine) and G+C content and indicate differences in amino acid content when comparing sequences that differ in G+C content. When sequences are compared using these plots, the amino acid content is shown to correlate with the nucleotide bias of the genes. This amino acid effect is shown in all protein-coding genes in the mitochondrial genome, including *cox1*, *cox2*, and *cob*, mitochondrial genes which are com-

* P.G. Foster, L.S. Jermiin, and D.A.Hickey, *J. Mol. Evol.* (1997) 44:282–288. The author thanks Springer-Verlag, New York for permission to reproduce this article.

monly used for phylogenetic studies. Furthermore, nucleotide content differences are shown to affect the content of all amino acids with A+T- and G+C-rich codons. We speculate that phylogenetic analysis of genes so affected may tend erroneously to indicate relatedness (or lack thereof) based only on amino acid content.

Key Words: G+C content, nucleotide bias, amino acid bias, mitochondrial genes, phylogeny.

2.2 Introduction

Whole genomes may differ widely in their average guanine plus cytosine (G+C) content (Lee et al., 1956; Muto and Osawa, 1987). Individual genes from different genomes may also differ in G+C content (Jermiin et al., 1994; Hashimoto et al., 1994). These differences can occur at both synonymous (silent) and non-synonymous (replacement) codon sites (Jukes and Bhushan, 1986). If only synonymous codon sites differ between genes, the corresponding protein sequences will be identical. However, if non-synonymous codon sites are affected the corresponding protein sequences will be different.

A relationship between the nucleotide composition and amino acid content is known from several genomic sources, including bacterial DNA (Sueoka, 1961; Andersson and Sharp, 1996), viral RNA (Berkhout and van Hemert, 1994), animal mitochondrial DNA (Jukes and Bhushan 1986; Jermiin et al. 1994, 1997) and eukaryotic nuclear DNA (D'Onofrio et al., 1991; Collins and Jukes, 1993; Porter, 1995). However, this relationship is not universal (Hashimoto et al., 1994, 1995).

In a phylogenetic context, protein sequences are often considered more useful than DNA sequences because they can be immune to base compositional differences occurring in the latter, and so phylogenetic information would be unaffected in the protein sequence (Loomis and Smith, 1990; Hasegawa and Hashimoto, 1993; Hashimoto et al., 1995). However, the evidence presented above does suggest that protein sequences are not always immune to compositional differences in the DNA and this could have an effect on the success of phylogenetic analysis (Steel et al.,

1993, 1995).

In order to delineate the relationship between compositional bias at the DNA and protein levels, and in view of its potential phylogenetic implications, we develop an analytical method which constitutes a unifying approach to the analysis of compositional bias. We introduce the square plot analysis, a joint graphical and statistical approach which can relate the A+T and G+C content at non-synonymous codon sites with the amino acid content. The statistical significances of the compositional differences are tested. This approach is illustrated by an analysis of mitochondrial protein-coding genes from the chicken (*Gallus gallus*) and the honeybee (*Apis mellifera*).

2.3 Materials and Methods

Sequences for the chicken and honeybee mitochondrial genomes were obtained from Genbank (accession numbers X52392, L06178, respectively). Additionally we used mitochondrial genomes of *Drosophila yakuba*, *Bos taurus*, *Homo sapiens*, *Mus musculus*, *Xenopus laevis*, *Cyprinus carpio*, and *Caenorhabditis elegans* (accession numbers X03240, J01394, J01415, J01420, M10217, X61010, and X54252, respectively). Sequences in each genome were pooled by taking the coding regions of protein-coding genes as specified in the Genbank entries, removing start and stop codons where these existed, and then concatenating the sequences.

We predicted that nucleotide bias would affect non-synonymous sites of codons such that A+T bias would increase the proportion of A+T-rich codons, and G+C bias would increase the proportion of G+C-rich codons. Here we define A+T-rich codons as those codons which have either A or T in both the first and second codon positions, and G+C-rich codons as those codons which have either G or C in both the first and second codon positions. We use the definition of synonymous and non-synonymous sites as given in Jukes and Bhushan (1986). As a tool to test this prediction we wanted to make a graphic representation which would show both the distribution of A+T and G+C content at non-synonymous sites in a gene, and show

the relative amounts of amino acids which are coded for by A+T- and G+C-rich codons. To do this, we simplified a standard ($4 \times 4 \times 4$) genetic code table by removing the third codon position (thereby making a 4×4 table), and rearranged it so that G and C were grouped together and A and T were grouped together, thereby abstracting a 2×2 table. This made four groups of codons/amino acids, arrayed as a 2×2 graphic, a "square plot", as shown in Figure 2.1. The frequency of each codon, and therefore of each amino acid, in these groups is indicated by the area of a square in the appropriate quadrant. In animal mitochondrial genetic codes (Osawa et al., 1992) leucine is found in two quadrants of the square plot, and so this amino acid and its two synonymous codon families need to be analysed separately.

Square plots allow comparison of codon content between sequences which differ in A+T and G+C content. We test the statistical significance of these comparisons in two ways. To compare the A+T- and G+C-rich quadrants of two square plots, we use a standard 2×2 contingency table and χ^2 test. To compare contents of individual amino acids between taxa, or to compare grouped amino acids between taxa, we used difference of proportions tests, either assuming a binomial distribution (Fleiss, 1981), or using the bootstrap (Efron, 1979; Diaconis and Efron, 1983; Efron and Tibshirani, 1991). In the former, it is assumed that the statistical distribution of an amino acid, or an amino acid group, follows the binomial distribution, which for large numbers can be approximated by the normal distribution. In the bootstrap, pseudo-sequences are generated from both taxa by random sampling with replacement, square plot analyses made, and the square plot proportions of one subtracted from another. This is repeated 10000 times, to make a distribution of differences. The position of zero in this distribution, whether in the main body of the distribution, or in the tail of the distribution, or off the distribution altogether, indicates whether the square plot values differ, and the significance of the difference. One-tailed tests are used.

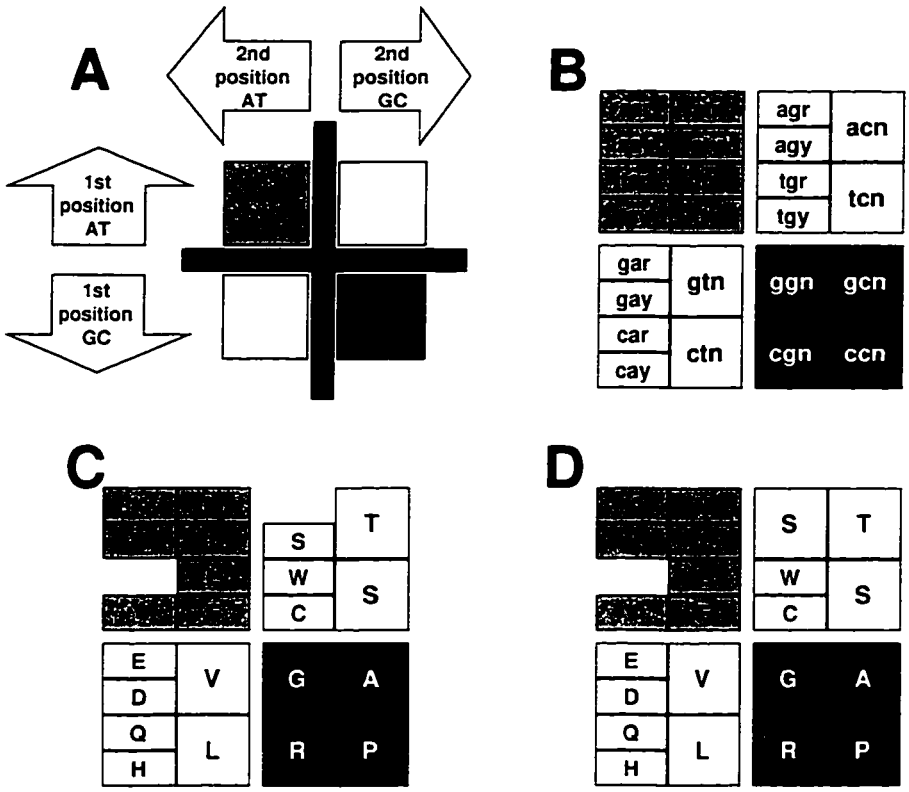


Figure 2.1: Square plot usage and correspondence with codons and amino acids.
A. Those codons with A or T in the first position are placed in the upper quadrants, while codons with G or C in the first position are placed in the lower quadrants. Codons with A or T in the second position are placed on the left side, while codons with G or C in the second position are placed on the right side. This means that the area of the upper left quadrant represents the number of codons with A or T in both the first and second positions, and so on.
B. Codon families of vertebrate and arthropod mitochondria included in square plot quadrants. In addition to nucleotides *a*, *c*, *g*, and *t*, we have *r* (purines, *a* or *g*), *y* (pyrimidines, *c* or *t*), and *n* (any).
C. Amino acids in the square plot quadrants as given by the vertebrate mitochondrial genetic code.
D. Amino acids in the square plot quadrants as given by the arthropod mitochondrial genetic code. Blanks in the quadrants represent stop codons.
Note that in **C** and **D**, leucine codons are found in both the upper and lower left quadrants, and so need to be analysed separately.

Table 2.1: Nucleotide composition differences in mitochondrial genomes of chicken and honeybee.

	Chicken	Honeybee
G+C coding region	0.471	0.168
G+C at synonymous codon sites	0.514	0.050
G+C at non-synonymous codon sites	0.445	0.241

The entire protein-coding regions of the mitochondrial genomes were analysed excluding start and stop codons.

2.4 Results and Discussion

Square plots used in this study divide codons into four groups based on the A+T and G+C content at their first and second positions. The number of codons in each group is plotted as area in the four quadrants of two-axes plots (Figure 2.1). Square plots allow visual and statistical comparison of the distribution of A+T and G+C-containing codons and their corresponding amino acids among genes. In animal mitochondria, codons for all but one amino acid are placed in separate quadrants of the square plot—codons for leucine are found in two quadrants (Osawa et al., 1992). For this reason leucine was analysed separately in this study.

2.4.1 Amino acid bias in pooled mitochondrial genes.

The mitochondrial genomes of the chicken and honeybee have been completely sequenced (Desjardins and Morais, 1990; Crozier and Crozier, 1993). The chicken genome has a G+C content of 46%, while the honeybee genome is only 15% G+C, and this is reflected in the nucleotide composition of the coding regions of these genomes (Table 2.1). While the relatively unbiased chicken mitochondrial genome shows little nucleotide bias at synonymous sites, the A+T bias of the honeybee mitochondrial genome manifests in an extreme A+T bias at synonymous codon sites, in agreement with Jukes and Bhushan (1986), and Jermiin et al. (1994, 1997).

Non-synonymous codon sites also differ in nucleotide composition between the

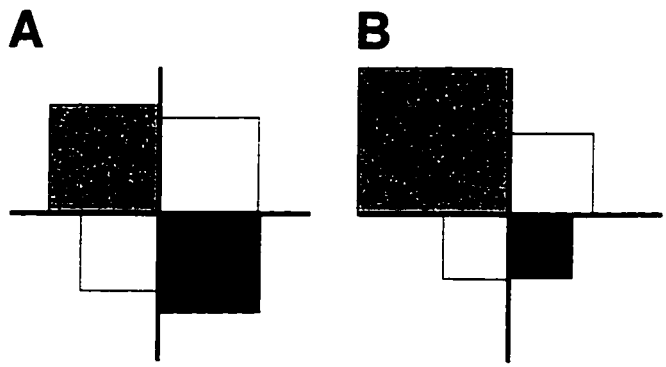


Figure 2.2: Square plots of complete mitochondrial genomes. **A** Chicken **B** Honeybee. In the chicken mitochondrial genome, excluding start and stop codons, there are 3772 codons, of which 664 code for leucine, leaving 3108 codons for the square plot analysis. Excluding start and stop codons, there are 3663 codons in the Honeybee mitochondrial genome, of which 568 code for leucine, leaving 3095 codons for the square plot analysis. For interpretation, see Figure 2.1

chicken and honeybee mitochondrial genomes (Table 2.1). To show this effect with more resolution, we made square plots of these genomes. Pooled protein coding sequences of each genome were analysed for A+T and G+C content at the first and second positions of the codons. Results are shown in square plots in Figure 2.2.

We can see large differences between corresponding quadrants in the two square plots shown in Figure 2.2. Specifically, the honeybee mitochondrion has a large excess of amino acids that lie in the upper left quadrant, and a relative deficiency of amino acids that lie in the lower right quadrant of the square plot. In order to quantitatively compare these differences we tabulate the number of codons in each quadrant in Figure 2.2 in Table 2.2. We can see that codons in the upper left quadrant, with A or T at both the first and second codon positions, comprise 59% of analysed codons in the honeybee, but only 32% in the chicken: this represents an almost two-fold difference in the proportion of this amino acid group between these two species, and is highly significant (see the footnote in Table 2.2 for statistical analyses). The large difference between the G+C contents at synonymous sites compared to the difference

Table 2.2: Codon counts in square plot quadrants in mitochondrial genomes of chicken and honeybee[a]

	Chicken	Honeybee
AT-AT[b]	1001	1830
AT-GC	780	573
GC-AT	516	358
GC-GC	811	334
Total	3108	3095

[a]The number of codons in the indicated square plot quadrant is shown. Stop, start, and leucine codons are excluded.

[b]The quadrants are indicated by AT or GC content at the first and second codon position— thus AT-GC represents the quadrant with A or T at the first position and G or C at the second position, corresponding to the upper right square plot quadrant. Statistical analysis. A 2×2 contingency table of the A+T- and G+C-rich quadrants gives $\chi^2 = 413.5$, for which P is essentially zero. The difference of proportions test of the A+T- and G+C-rich quadrants assuming the binomial distribution gives $z = 21.3$ and 15.5, for both of which P is again essentially zero. The difference of proportions test using the bootstrap gives $I_0/bootstraps < 0.0001$ in both cases. (I_0 is the index, or position, of zero in the distribution of bootstrap differences. Sequences were bootstrapped 10000 times. If zero is not within the distribution of differences, then $I_0/bootstraps < 0.0001$)

between the G+C contents at nonsynonymous sites (Table 2.1) is consistent with the notion that nucleotide bias (for example caused by directional mutation pressure (Jermiin et al., 1996)) determines the amino acid bias, and not with the reverse.

2.4.2 *Amino acid composition differences in individual mitochondrial genes*

From the results presented above, it appears that the overall amino acid composition of the chicken and honeybee mitochondrial genomes is related to the nucleotide composition. We wished to determine whether this was also the case in individual genes, or whether the overall differences that we saw were due to a subset of the mitochondrial genome. To examine this question, we compared each of the 13 protein-coding genes between honeybee and chicken mitochondria. G+C content in the genes was similar to that shown in Table 2.1, specifically (means and standard errors of the mean are given): a) the G+C content of the genes individually was similar to the content of the genomes from which they came (chicken 0.470 ± 0.006, honeybee 0.159 ± 0.010); b) the A+T bias of the honeybee genes manifests in an extreme A+T bias at synonymous sites, while the relatively unbiased chicken does not show this extreme effect (chicken 0.504 ± 0.123, honeybee 0.052 ± 0.006); and c) the A+T bias of the honeybee genes results in a smaller A+T bias at non-synonymous sites, while the relatively unbiased chicken again does not appear to show this effect (chicken 0.450 ± 0.008, honeybee 0.225 ± 0.016).

These 13 genes were also compared using square plots. The differences between the square plots from the two species were similar to that found when comparing the entire genomes (Figure 2.2). The A+T- and G+C-rich quadrants in square plots from each gene were shown to be significantly different between the two species based on the χ^2 test ($P < 0.002$), and the difference of proportions test ($P < 0.02$).

Square plots of 3 of these pairs, those of *cox1*, *cox2*, and *cob* genes, are shown in Figure 2.3. We chose to focus on these genes because they are commonly used in phylogenetic studies. We can examine differences found in the upper left quadrants, which shows codons with A or T in both the first and second codon positions (Table 2.3). Each gene shows a large, and highly significant, difference in proportion of these codons, and therefore of the amino acids for which they code, between the two species.

Table 2.3: Codon count in the A+T-rich quadrant, and statistical analysis of square plots of individual mitochondrial genes.

	A+T-rich quadrant [a]		statistics [b]		
	Chicken	Honeybee	χ^2	z	I_0/*bootstraps*
cox1 [c]	151/452	204/447	17.54	3.75	
	33%	46%	$P < 0.001$	$P < 0.0001$	0.0001
cox2 [c]	51/194	104/202	18.05	5.14	
	26%	51%	$P < 0.001$	$P < 0.000001$	< 0.0001
cob [c]	115/316	194/333	22.30	5.57	
	36%	58%	$P < 0.000001$	$P < 0.0000001$	< 0.0001
atp8	16/45	32/43	8.62	3.66	
	36%	74%	$P = 0.0017$	$P < 0.001$	0.0001
atp6	50/164	111/184	41.09	5.57	
	30%	60%	$P < 0.0000000001$	$P < 0.0000001$	< 0.0001
cox3	64/227	127/226	37.29	6.03	
	28%	56%	$P < 0.000000001$	$P < 0.000000001$	< 0.0001
nad1	88/257	144/253	32.48	5.14	
	34%	57%	$P < 0.00000001$	$P < 0.000001$	< 0.0001
nad2	96/277	201/295	54.56	8.01	
	35%	68%	$P < 0.0000000000001$	$P < 0.000000000000001$	< 0.0001
nad3	27/90	63/96	23.44	4.86	
	30%	66%	$P < 0.000001$	$P < 0.000001$	< 0.0001
nad4	116/361	226/350	52.93	8.65	
	32%	65%	$P < 0.000000000001$	$P \approx 0$	< 0.0001
nad4L	21/79	36/65	17.36	3.52	
	27%	55%	$P < 0.001$	$P < 0.001$	< 0.0001
nad5	176/500	291/464	63.85	8.54	
	35%	63%	$P < 0.000000000000001$	$P \approx 0$	< 0.0001
nad6	30/146	97/137	69.40	8.49	
	21%	71%	$P < 0.0000000000000001$	$P \approx 0$	< 0.0001

[a] Start, stop, and leucine codons are excluded. Data are shown as count / (total codons, excluding start, stop, and leucine), and percent.

[b] χ^2 is from a 2 × 2 contingency table for the A+T- and G+C-rich quadrants. Difference of proportions tests are given for the A+T-rich quadrant. Results for the difference of proportions test assuming the binomial distribution are given by the z statistic and its associated P-value. Results for the difference of proportions test using the bootstrap are given by I_0/*bootstraps* (see Table 2.2 for an explanation of I_0/*bootstraps*)

[c] See Figure 2.3.

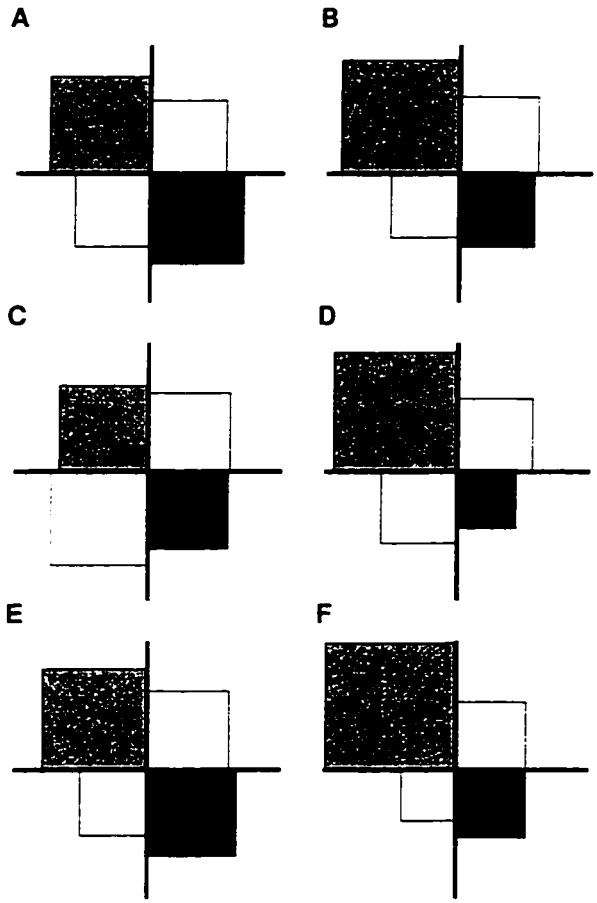


Figure 2.3: Square plots of individual mitochondrial genes. A, C, E: Chicken, B, D, F: Honeybee, A, B: *cox1*, C, D: *cox2*, E, F: *cob*

2.4.3 The frequency of many amino acids is affected by nucleotide content bias

Having shown a relationship between nucleotide and amino acid composition, we wished to determine which amino acids are affected. In order to show which particular amino acids in the square plot groups are affected, we plotted a breakdown of the individual amino acids in bar chart form (Figure 2.4). This allows us to compare the contents of individual amino acids of the mitochondrial genes from the two species.

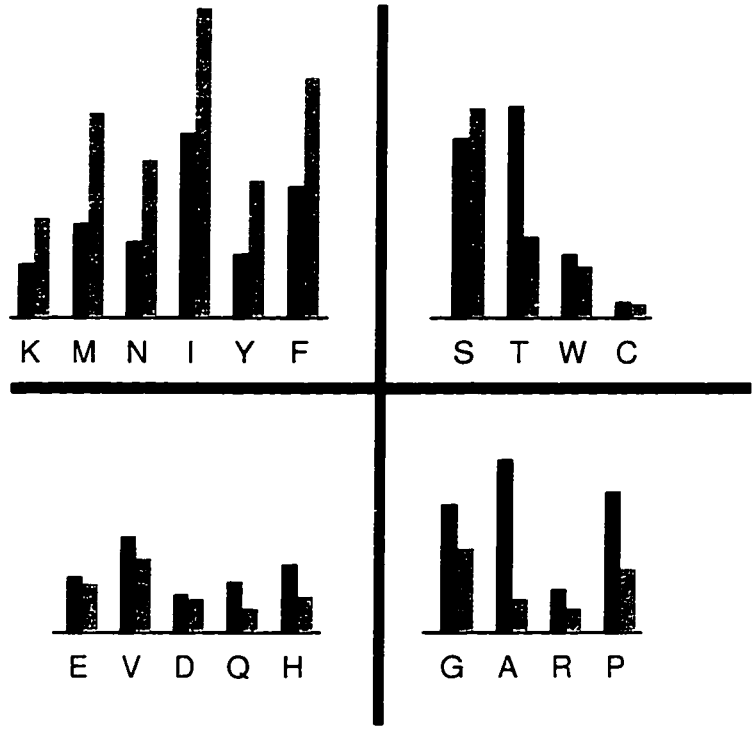


Figure 2.4: Amino acid differences between chicken (black bars) and honeybee (gray bars) mitochondrial genomes. Amino acids are grouped as in square plots, and so this figure corresponds directly to Figure 2.2

This figure shows that the content of all the amino acids in the upper left and lower right quadrants (A+T-rich and G+C-rich codons, respectively) are strikingly different between the two species. In particular, the A+T-rich honeybee genome codes for more of each A+T-rich codon family, and less of each G+C-rich codon family compared to the relatively unbiased chicken mitochondrial genome. The differences are all highly significant, as shown in Table 2.4.

These results are more striking than those that were obtained by comparing the total mitochondrial codons from (A+T-rich) *Drosophila yakuba* and sequence pooled from several (A+T-neutral) vertebrate species (Jukes and Bhushan, 1986). That study

Table 2.4: Amino acid composition in mitochondrial genomes of chicken and honeybee

Amino acid		chicken	honeybee	z	P	I_0/ *bootstraps* [a]
Lys	K	90	160	4.55	< 0.00001	< 0.0001
Met	M	156	328	8.19	< 0.000000000000001	< 0.0001
Asn	N	126	249	6.59	< 0.0000000001	< 0.0001
Ile	I	304	493	7.23	< 0.000000000001	< 0.0001
Tyr	Y	106	219	6.48	< 0.0000000001	< 0.0001
Phe	F	219	381	7.01	< 0.00000000001	< 0.0001
Gly	G	215	135	4.36	< 0.00001	< 0.0001
Ala	A	289	57	12.80	≈ 0	< 0.0001
Arg	R	72	39	3.14	< 0.001	0.0012
Pro	P	235	103	7.34	< 0.000000000001	< 0.0001

Amino acids are grouped as in square plots. Amino acid counts are shown. The z statistic is calculated based on the amino acid proportions relative to the total number of codons in the genome, excluding start, stop, and leucine codons, as given in Table 2.2.

[a] See the footnote to Table 2.2 for an explanation of I_0/ *bootstraps*.

found that in *Drosophila*, there was an increase in phenylalanine, asparagine, and tyrosine but not isoleucine, lysine, or methionine, and less of alanine and proline but not glycine or arginine. We have repeated that analysis, and shown in addition that the differences in the amino acid groups coded for by A+T- and G+C-rich codons are highly significant (χ^2 of the A+T and G+C-rich quadrants is 47.4, $P < 0.00000000001$; the difference of proportions test of the KMNIYF and GARP quadrants assuming the binomial distribution gives $z = 6.9$ and 5.3 respectively, for which $P < 0.00000000001$ and $P < 0.0000001$, respectively; the difference of proportions test of the KMNIYF and GARP quadrants using the bootstrap gives I_0/ *bootstraps* < 0.0001 in both cases, using 10000 bootstraps). Those results are consistent with, but less dramatic than, the results obtained in the present study. For a third data set we repeated the analysis using the A+T-rich mitochondrial genome of *Caenorhabditis elegans* compared to the relatively A+T-neutral carp mitochondrial

Table 2.5: Leucine codon family usage

Codon family	Chicken	Honeybee
TTR	0.021	0.135
CTN	0.155	0.020
total	0.176	0.155

The proportion of indicated codon family in the pool of all codons in the mitochondria of the indicated species is shown.

genome, with similar results. Recently a similar analysis comparing amino acid compositions between several homologous proteins of *Escherichia coli* and the A+T-rich *Rickettsia prowazekii* obtained similar results, showing that this effect exists in bacteria as well as in animal mitochondria (Andersson and Sharp, 1996).

While we are examining individual amino acids of the square plots, it would be appropriate to look at the way that leucine codon family usage changes with A+T-pressure (Sueoka, 1962). In animal mitochondria, leucine is coded for by TTR, found in the upper left square plot quadrant, and by CTN, found in the lower left quadrant. Codon family usage of leucine in pooled mitochondrial genes from the honeybee and the chicken is tabulated in Table 2.5. This table allows us to see the reciprocal way that the two codon families are used in the two species. In the A+T-rich honeybee genome, most of the leucine codons are (T-rich) TTR codons. In the relatively unbiased chicken genome, most of the leucine codons are CTN ($\chi^2 = 703$, $P \approx 0$).

In summary, we have compared the coding sequences of entire mitochondrial genomes to show that nucleotide bias is a significant driving force in amino acid composition of the encoded proteins. We chose the genomes of the chicken and the honeybee because of their large difference in G+C content (Table 2.1). We have shown that the amino acid profile differs between the two sequences in a manner which relates to the nucleotide content of their codons, and we have shown that this difference is statistically significant. The A+T-rich genome of the honeybee mito-

chondrion has significantly more of the amino acids F, Y, M, I, N, and K, as shown in the upper left quadrant of the square plot, while the genome of the chicken mitochondrion has significantly more of the amino acids G, A, R, and P, as shown in the lower right quadrant of the square plot (Figure 2.2, Figure 2.4).

2.4.4 Implications for phylogeny reconstruction

Phylogenetic trees of protein-coding genes may be based on either the DNA sequence or the protein sequence. One reason for using the protein sequences is that composition or codon-usage bias in DNA is a confounding factor in construction of phylogenetic trees (Loomis and Smith, 1990; Hasegawa and Hashimoto, 1993; Lockhart et al., 1992; He and Haymer, 1995). Another reason is that natural selection tends to lessen the effects of nucleotide bias. We assume that the force of mutational and DNA repair biases affects all nucleotide positions equally within a gene. Because of the countervailing force of natural selection, however, the effect of these biases accumulates much more rapidly at the synonymous sites. Indeed it is often assumed that the non-synonymous sites are totally free of any bias, an assumption which underlies our confidence in the use of amino acid sequence data as a means of constructing reliable alignments and building unbiased molecular phylogenies.

Evidence of a relationship between compositional biases at the DNA and protein levels, as has been shown in this study, implies that DNA bias can confound phylogenetic reconstruction based on protein sequences. Models of evolution used in phylogeny reconstruction algorithms, such as the Dayhoff et al. (1978) model, are limited in their ability to accommodate amino acid composition differences. We have shown in particular that the compositions of the *cox1*, *cox2*, and *cob* proteins, three proteins that are commonly used in phylogenetic studies, are affected by compositional differences at the DNA level. However at present it is not known how serious a problem, if any, this effect poses for phylogenetic reconstruction. We can speculate that amino acid bias will affect phylogenetic reconstruction using protein sequences similar to the way that nucleotide bias affects phylogenetic reconstruction using DNA sequences (Lockhart et al., 1994). That is, phylogenetic analysis may falsely indicate

homology (or lack thereof) based only on amino acid content. It is not expected that the effect of amino acid bias will be as pronounced as the effect of nucleotide bias, because much of the bias at the DNA level is manifested in synonymous codon sites (Table 2.1).

Studies are underway to assess the effect of bias in nuclear genes and in highly-conserved genes. Preliminary analyses suggest that indeed some nuclear genes do show the relationship described in this paper, but the highly-conserved elongation factor 1α appears to be immune from this effect, in agreement with Hashimoto et al. (1994).

2.5 Acknowledgements

This research was supported by an operating grant from NSERC (Canada) to DAH, and by the John Curtin School of Medical Research at the Australian National University to LSJ. We wish to thank Joe Felsenstein for his comments concerning the bootstrap.

Chapter 3

Susceptibility of homologous proteins to compositional bias due to AT/GC bias in DNA

3.1 Abstract

That there is a relationship between the AT/GC content of DNA and the amino acid composition of encoded proteins has been known for some time. Recently it was shown that the mitochondrial genomes of the chicken and honeybee, which differ greatly in AT/GC at the DNA level, code for proteins which differ in amino acid composition in a manner which reflects the AT and GC content of the codons. Here I have extended this observation to 11 complete animal mitochondrial genomes, and show that there is a pronounced correlation between the AT/GC content of DNA and the composition of AT- and GC-rich codons and their associated amino acids. This relationship is also seen in the nuclear/bacterial α-amylase and trypsin genes, and is significant even in the highly-conserved *hsp*70 and *ef*1α/*tu* genes. Since common phylogenetic methods assume a stationary composition, this widespread DNA-driven amino acid composition bias is expected to be problematic in protein-based phylogenetic reconstruction.

3.2 Introduction

Both entire genomes (Lee et al., 1956; Muto and Osawa, 1987), and homologous genes (Jermiin et al., 1994; Hashimoto et al., 1994) may differ widely in AT or GC content. Since these differences can occur not only at synonymous (silent) sites but at non-synonymous (replacement) codon sites as well(Jukes and Bhushan, 1986), it can therefore be expected that differences in AT/GC content will have an effect on the amino acid composition. Indeed, a relationship between the nucleotide composition and amino acid content is known from several genomic sources, including bacterial DNA (Sueoka, 1961; Andersson and Sharp, 1996), viral RNA (Berkhout and van Hemert, 1994), animal mitochondrial DNA (Jukes and Bhushan 1986; Jermiin et al., 1994) and eukaryotic nuclear DNA (D'Onofrio et al., 1991; Collins and Jukes, 1993; Porter, 1995). However, this relationship is not universal (Hashimoto et al., 1994, 1995). Recently Foster et al. (1997) analysed the protein-coding sequences of the complete mitochondrial genomes of the chicken and honeybee to further delineate this relationship between AT/GC content at the DNA level and composition at the protein level. It was found that, relative to the AT/GC-neutral chicken genome, the AT-rich honeybee genome was rich in AT-rich codons, coding for the amino acids F, Y, M, I, N, and K, and poor in GC-rich codons, coding for the amino acids G, A, R, and P.

In the present study I have tested the generality of this observation by examining this relationship in several complete mitochondrial genomes. These genomes showed a strong correlation between AT/GC content at the DNA level and compositional bias at the protein level. In addition, I looked for this correlation in bacterial and eukaryotic nuclear α-amylase and trypsin genes. I have found that these genes, and even the conserved *hsp*70 and *ef*1α/*tu* genes, are susceptible to amino acid compositional bias correlated with AT/GC bias in the DNA.

3.3 Methods

Amino acid and nucleotide frequencies were measured using software described in Appendix A. Regression analysis was done using the statistics package in *Mathematica.*

3.4 Results

In the following analyses DNA and protein sequences were examined from coding regions only. Start and stop codons were excluded from the analysis. I then tested for a correlation between the GC composition of the DNA and the composition of AT- and GC-rich codon groups and their corresponding amino acids. The amino acid leucine in the animal mitochondria genetic code and the amino acids leucine and arginine in the universal genetic code are coded for by two codon families. For example, leucine is coded for by the codon families CTN, which is AT/GC-neutral, and TTR, which is AT-rich. Because of the AT/GC-neutral codon family, leucine was excluded from the group of AT-rich codons. For a parallel reason arginine was excluded from the group of GC-rich codons in analyses which used the universal genetic code.

The protein-coding regions of 11 animal mitochondrial genomes were analysed for amino acid composition (Figure 3.1). A strong correlation was found between increasing AT at the nucleotide level and the group of amino acids coded for by AT-rich codons (FYMINK), and also between increasing GC content and the amino acids coded for by GC-rich codons. The chicken and honeybee mitochondrial genomes analysed in Foster et al. (1997) are included in this set at 47 and 17% GC, respectively, and it is evident that these two are part of a general trend among animal mitochondria.

Leucine in animal mitochondria is coded for by two codon families- CTN, and TTR. One might predict that since there is an AT-rich codon family but no GC-rich family that there would be a correlation between the amount of leucine and AT at the nucleotide level. This is not borne out, however, and the leucine content does

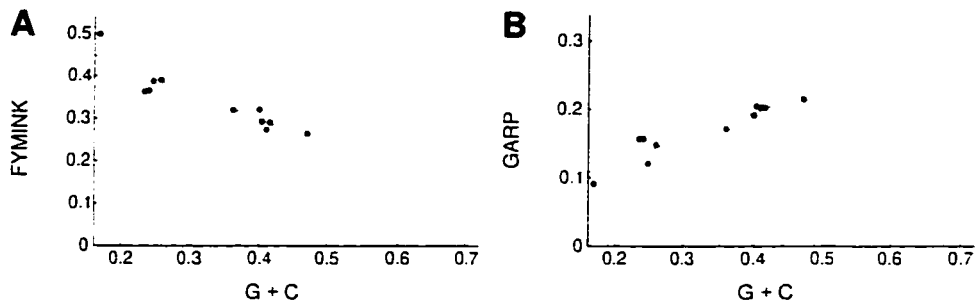


Figure 3.1: Correlation of amino acid composition with nucleotide GC content in proteins coded by animal mitochondria.
A. Negative correlation of amino acids coded for by AT-rich codons with increasing GC content. $r^2 = 0.86$, $P = 0.00004$.
B. Positive correlation of amino acids coded for by GC-rich codons with increasing GC content. $r^2 = 0.88$, $P = 0.00002$.

not change significantly over the range of GC contents of the mitochondrial genomes ($r^2 = 0.17$, $P = 0.21$). However if we look at the codon families individually, we find that as nucleotide GC increases that the CTN codons increase ($r^2 = 0.92$, $P = 0$) at the expense of the TTR codons ($r^2 = 0.87$, $P = 0.00003$).

To see if the same relationship holds in non-mitochondrial genes, the genes of the digestive enzyme α-amylase from a wide range of taxa were examined (Figure 3.2A & B). Neither arginine nor leucine were included, because in the universal genetic code both of these amino acids also have GC-neutral codon families. The same strong relationship was found in α-amylase genes as was found among the mitochondrial genomes.

This relationship was also examined in trypsin genes (Figure 3.2C & D). The taxonomic range of available sequences was not large, composed of mostly animal genes, with one fungal and one bacterial gene. The bacterial gene was 71% GC, and the animal genes were mostly 48 to 62% GC, and there was one insect trypsin-like gene at 43% GC. The correlation was low but significant for the trypsin genes, likely due to the small taxonomic range and small range of GC contents.

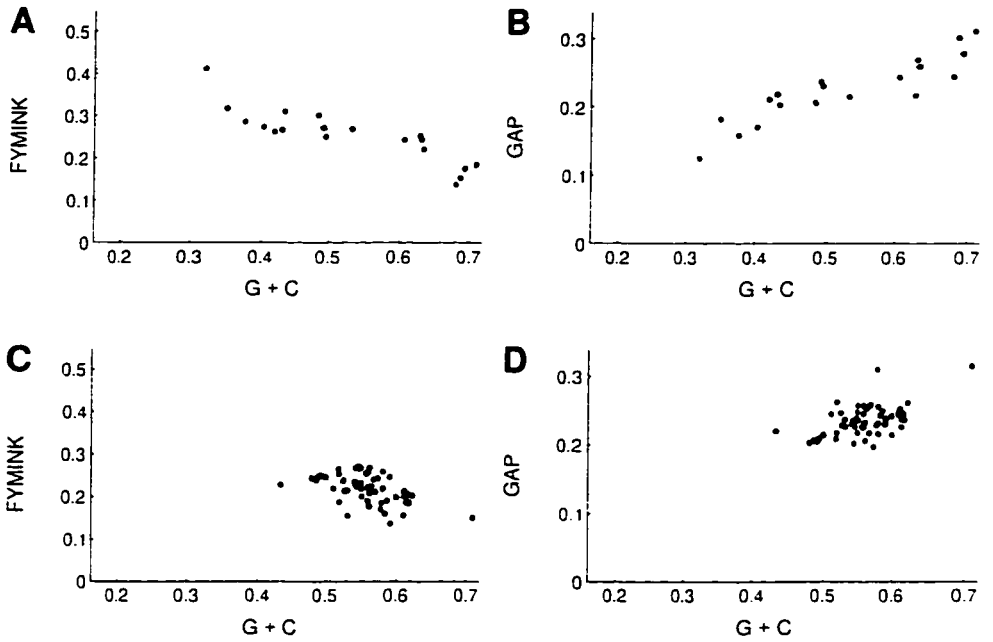


Figure 3.2: Correlation of amino acid composition with nucleotide GC content in α-amylase and trypsin.
A. Negative correlation of amino acids coded for by AT-rich codons with increasing GC content in α-amylase genes. $r^2 = 0.74$, $P = 0$.
B. Positive correlation of amino acids coded for by GC-rich codons with increasing GC content in α-amylase genes. $r^2 = 0.80$, $P = 0$.
C. Negative correlation of amino acids coded for by AT-rich codons with increasing GC content in trypsin genes. $r^2 = 0.27$, $P = 0.00003$.
D. Positive correlation of amino acids coded for by GC-rich codons with increasing GC content in trypsin genes. $r^2 = 0.30$, $P = 0$.

This relationship was further examined in the highly-conserved *hsp*70 and *ef*1α/*tu* genes (Figure 3.3). These sequences were taken from phylogenetically wide sources, including eubacteria, archaea, and eukaryotes. The *hsp*70 genes include *dnaK* genes from bacteria and archaea. The *ef*1α/*tu* genes include *ef*1α from both archaea and eukaryotes, and *ef*1*tu* from bacteria and organelles. The correlation in these genes was not large, but in all cases was significant at the 5% level.

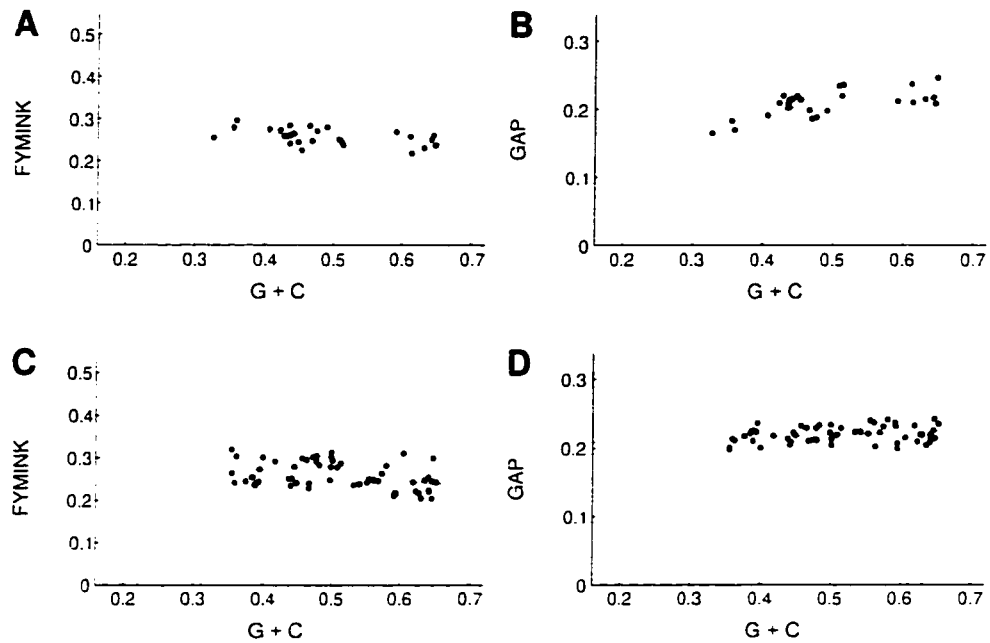


Figure 3.3: Correlation of amino acid composition with nucleotide GC content in *hsp*70 and *ef*1α/*tu*.
A. Negative correlation of amino acids coded for by AT-rich codons with increasing GC content in *hsp*70 genes. $r^2 = 0.23$, $P = 0.011$.
B. Positive correlation of amino acids coded for by GC-rich codons with increasing GC content in *hsp*70 genes. $r^2 = 0.41$, $P = 0.0003$.
C. Negative correlation of amino acids coded for by AT-rich codons with increasing GC content in *ef*1α/*tu* genes. $r^2 = 0.15$, $P = 0.002$.
D. Positive correlation of amino acids coded for by GC-rich codons with increasing GC content in *ef*1α/*tu* genes. $r^2 = 0.068$, $P = 0.04$.

3.5 Discussion

In this study the observation described in Foster et al. (1997), that nucleotide AT/GC bias affects the amino acid composition of encoded proteins in the mitochondria of the chicken and honeybee, was expanded to include 11 complete animal mitochondrial genomes. This effect was also shown in the nuclear and bacterial genes

α-amylase and trypsin, and was significant even in the conserved *hsp*70 and *ef*1α/*tu* genes.

An amino acid composition bias was not found in several *ef*1α sequences examined by Hashimoto et al. (1994), while here I have found a small bias in a superset of those sequences. When only the sequences examined by Hashimoto et al. (1994) were examined, I obtain a borderline/insignificant correlation (FYMINK *vs* GC: r^2 = 0.23, P = 0.05; GAP *vs* GC: r^2 = 0.08, P = 0.27), and so my results are not in disagreement. The results that I present here are from a larger group of sequences and from a more widely diverged group of taxa. Differences in approach can be noted, also, as they examined bias in individual amino acids, while here I look at amino acid groups. Their caution, however, appears to have been well-founded.

This widespread compositional bias may be problematic in protein-based phylogenetic reconstruction. Phylogenetic methods commonly in use assume that the amino acid composition is stationary over evolutionary time, yet clearly this assumption is violated by the sequences that have been analysed in this study. As a case in point, amino acid composition bias is pronounced in animal mitochondrial genomes, the genes of which are commonly used in phylogenetic studies. It has recently been shown that this bias can result in incorrect phylogenies using animal mitochondrial protein sequences (Foster and Hickey, 1997). Furthermore, amino acid composition bias has been shown to occur in eukaryotic nuclear and bacterial genes as well, including a significant bias in the highly conserved *hsp*70 and *ef*1α/*tu* genes. While the bias shown in the example in Foster and Hickey (1997) was so extreme that it resulted in an incorrect tree topology, it is expected that less extreme biases will be problematic as well, resulting in lower statistical confidence and mis-estimated branch lengths in inferred trees. Furthermore, we can expect that compositional bias will be especially troublesome in deep and ambiguous phylogenies, where subtleties such as compositional bias could bear on the result. Continued caution is therefore urged in the use of compositionally biased protein sequences in phylogenetic studies.

3.6 Appendix: sequences used

The mitochondrial genomes (with their accession numbers) were from *Apis mellifera* (L06178), *Cepaea nemoralis* (U23045), *Drosophila yakuba* (X03240 *etc.*), *Strongylocentrotus purpuratus* (X12631), *Anopheles gambiae* (L20934), *Artemia franciscana* (X69067), *Caenorhabditis elegans* (X54252, S93745), *Locusta migratoria* (X80245), *Paracentrotus lividus* (J04815), *Bos taurus* (J01394), and *Gallus gallus* (X52392).

The α-amylase genes were from *Aedes atropalpus* (U01209), *Aeromonas hydrophila* (L19299), *Alteromonas haloplanktis* (X58627), *Anopheles merus* (U01210), *Aspergillus oryzae* (X12725), *Butyrivibrio fibrisolvens* (M62507), *Drosophila melanogaster* (X04569), *Dictyoglomus thermophilum* (X15948 M36505), *Natronococcus sp.* (D26510), *Pseudomonas sp.* (D10769 D01143), *Rattus norvegicus* (M24962), *Debaryomyces occidentalis* (S77586), *Saccharomycopsis fibuligera* (X05791), *Streptomyces griseus* (X57568), *Thermomonospora curvata* (X59159), *Tribolium castaneum* (X06905), *Thermoanaerobacterium thermosulfurigenes* (M57692 M57580 X54982 X54654), *Thermoactinomyces vulgaris* (X69807), *Xanthomonas campestris* (M85252 M32874)

Trypsin genes were from *Aedes aegypti* (X64362, X64363), *Anopheles gambiae* (Z22930 (7 sequences), Z18890), *Bos taurus* (D38507), *Choristoneura fumiferana* (L04749), *Drosophila erecta* (U40653 (8 sequences)), *Drosophila melanogaster* (U04853 (8 sequences), U41476), *Canis sp.* (M11589, M11590), *Gallus gallus* (U15155, U15156, U15157), *Homo sapiens* (X15505, M22612, M27602), *Lucilia cuprina* (L15632), *Mus musculus* (X04574), *Manduca sexta* (L16805, L16806, L16807), *Neobellieria bullata* (X94691), *Pleuronectes platessa* (X56744), *Rattus norvegicus* (M16624, X15679, V01273, V01274), *Rattus rattus* (X59012, X59013), *Fusarium oxysporum* (S63827), *Simulium vittatum* (L08428), *Salmo salar* (X70075, X70073, X70074), *Streptomyces griseus* (M64471), *Xenopus laevis* (X53458, U72330).

The *hsp*70 genes were from the list given in Rensing and Maier (1994), excluding partial and protein-only sequences.

The bacterial elongation factor *tu* sequences were from *Streptomyces cinnamoneus* (X98831), *Planobispora rosea* (X98830), *Mycoplasma pneumoniae* (AE000019 U00089), *Bacillus subtilis* (D64127), *Mycoplasma genitalium* (U39732 L43967), *Chlamydia trachomatis* (M74221), *Wolinella succinogenes* (X76872), *Taxeobacter ocellatus* (X77036), *Thiobacillus cuprinus* (X76871), *Stigmatella aurantiaca* (X76870), *Spirochaeta aurantia* (X76874), *Herpetosiphon aurantiacus* (X76868), *Fibrobacter succinogenes* (X76866), *Flavobacterium ferrugineum* (X76867), *Chlorobium vibrioforme* (X77033), *Cytophaga lytica* (X77035), *Corynebacterium glutamicum* (X77034), *Chloroflexus aurantiacus* (X76864 X76865), *Brevibacterium linens* (X76863), *Ureaplasma urealyticum* (Z34275), *Mycobacterium leprae* (D13869), *Chlamydia trachomatis* (L22216), *Thermus aquaticus thermophilus* (X61957 S53348), *Mycoplasma hominis* (M57675), *Escherichia coli* (J01690), *Thermus aquaticus* (X66322 S42355), *Salmonella typhimurium* (X55117), *Salmonella typhimurium* (X55116), *Mycobacterium tuberculosis* (X63539 S40925), *Mycobacterium leprae* (Z14314), *Mycoplasma hominis* (X57136).

The organellar *eftu* genes were from *Homo sapiens* (mitochondrial origin, now nuclear) (L38995), *Bos taurus* (mitochondrial origin, now nuclear) (L38996), *Arabidopsis thaliana* (chloroplast origin, now nuclear) (X52256), *Euglena gracilis* (chloroplast) (X00044), *Euglena gracilis* (chloroplast) (Z11874 S55425), *Odontella sinensis* (chloroplast) (Z67753), *Porphyra purpurea* (chloroplast) (U38804). It was not clear whether the sequences from *Mus musculus* (M22432) and *Xenopus laevis* (M75873) were *eftu* of organelle origin or nuclear *ef*1α.

The archaeal *ef*1α sequences were from *Halobacterium halobium* (D32120), *Sulfolobus solfataricus*

(X70701), *Thermococcus celer* (X52383), *Thermoplasma acidophilum* (X53866), *Sulfolobus acidocaldarius* (X52382), *Pyrococcus woesei* (X59857), *Methanococcus vannielii* (X05698), *Haloarcula marismortui* (X16677).

The *ef*1α sequences used in Hashimoto (1994) were also used. These were *Absidia glauca* (X54730), *Artemia sp.* (X03349 J01165 X00546), *Arabidopsis thaliana* (X16430), *Drosophila melanogaster* (X06869), *Euglena gracilis* (X16890), *Entamoeba histolytica* (M92073 M34256), *Giardia lamblia* (D14342), *Homo sapiens* (X03558), *Lycopersicon esculentum* (X53043), *Mucor racemosus* (J02605 M16352), *Plasmodium falciparum* (X60488), *Xenopus laevis* (M25504), *Candida albicans* (M29934), *Saccharomyces cerevisiae* (M15666), and the three archaeal sequences X16677, X05698, and X52382 already listed above.

Chapter 4

Compositional bias can affect protein-based phylogenetic reconstruction*

4.1 Abstract

Phylogenetic analyses based on protein sequences are generally considered to be more reliable than those derived from the corresponding DNA sequences because it is believed that the use of encoded protein sequences circumvents the problems caused by nucleotide compositional biases in the DNA sequences. There exists, however, a correlation between AT/GC bias at the nucleotide level and content of AT- and GC-rich codons, and their corresponding amino acids. Consequently, protein sequences can also be affected secondarily by nucleotide compositional bias. Here, we report that this bias in the amino acid content of proteins can result in incorrect protein-based phylogenetic trees. We present a striking example where common phylogenetic tools fail to recover the correct tree from animal mitochondrial protein sequences. The data set is very extensive, containing several thousand sites per sequence, and the incorrect phylogenetic trees are statistically very well supported. Additionally, neither

* P.G. Foster and D.A.Hickey

the use of the LogDet/paralinear transform nor removal of positions in the alignment with AT- or GC-rich codons allowed recovery of the correct tree. The two taxa with the greatest compositional bias continually group together in these analyses, despite a lack of close biological relatedness. We conclude that even protein-based phylogenetic trees can be misleading, and we advise caution in phylogenetic reconstruction using protein sequences, especially those that are compositionally biased.

4.2 Introduction

Hasegawa and Hashimoto (1993) pointed out that phylogenetic analyses based on rRNA genes could be unreliable due to extreme AT or GC nucleotide bias in the rRNA genes of some taxa. They suggested that the inferred amino acid sequences of encoded proteins provide more reliable phylogenies. Many molecular evolutionists now agree that protein sequences are relatively free from the effects of nucleotide bias (Loomis and Smith, 1990; Lockhart et al., 1992). This view is based on the assumption that, while DNA may be driven to extremes of AT or GC bias by directional mutation pressure, the protein composition remains constant, due to the greater functional constraints on the protein sequence. Here we demonstrate that, contrary to this assumption, there can be very significant variation in amino acid composition between homologous proteins and that this variation can result in erroneous phylogenetic reconstruction.

We have recently shown (Foster et al., 1997) that amino acid sequences can be compositionally biased in a manner that parallels the nucleotide composition of the codons. For instance, we showed that those animal mitochondrial genes which are most AT-rich at the DNA level tend to be rich in those amino acids which are encoded by AT-rich codons, ie, codons with either A or T in the first and second codon position; this set includes the codons for phenylalanine (F), tyrosine (Y), methionine (M), isoleucine (I), asparagine (N), and lysine (K). These same proteins are correspondingly poor in amino acids coded for by GC-rich codons: glycine (G), alanine (A), arginine (R), and proline (P). This effect is not limited to animal mitochondrial

genes; it has been reported for a wide range of genes and genomes (Sueoka, 1961; Andersson and Sharp, 1996; Collins and Jukes, 1993; Porter, 1995; Jukes and Bhushan, 1986; Jermiin et al., 1994; D'Onofrio et al., 1991), although there are some genes that appear to be immune to this effect (Hashimoto et al., 1994; Hashimoto et al., 1995). Here, we ask if proteins with similarly-biased amino acid compositions will tend to be grouped together in a protein-based molecular phylogeny, even if they do not share a recent common ancestor.

We performed phylogenetic analyses of the protein coding sequences of several mitochondria, which included taxa with varying amounts of amino acid composition bias. Our available arsenal of phylogenetic tools, including distance and parsimony methods, maximum likelihood, and LogDet distance correction, failed to recover the correct tree. The two taxa with the greatest compositional bias, the honeybee and the nematode, continually grouped together in these analyses, despite lack of close biological relatedness of these taxa.

4.3 Results

We have chosen mitochondrial proteins from animal species for which the entire mitochondrial genome has been sequenced and we have used the concatenated protein sequences for our analyses. The species cover a broad phylogenetic range within the metazoa and the pattern of their true phylogenetic divergences is not in dispute (Figure 4.1A). They include two taxa, the honeybee *Apis mellifera*, and the nematode *Caenorhabditis elegans*, that have very AT-rich mitochondrial genomes and that show the predicted bias in amino acid composition (Foster et al., 1997; Table 4.1). The goal of our study was to test the prediction that the shared amino acid bias could cause these two taxa to group together, despite the fact that they are not true sister taxa.

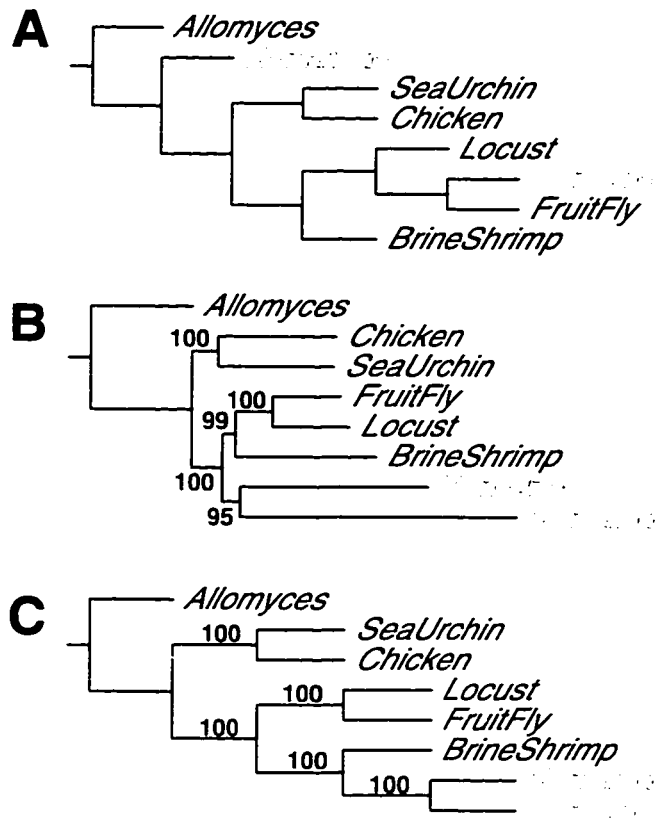


Figure 4.1: Phylogenetic trees of mitochondrial protein sequences. The AT-rich honeybee and nematode are highlighted.
A. Biological tree.
B. Tree computed using the neighbor-joining method.
C. Tree computed using the method of maximum parsimony.
Numbers are bootstrap values, the percent of 100 bootstraps. Using protml v 2.2, with the Dayhoff model with frequencies taken from the sequences using the -f option, to calculate the maximum likelihood, the most likely tree is C, the maximum parsimony tree ($lnL = -54018.4$). The neighbor-joining tree was not significantly different ($\Delta lnL = -2.8 \pm 11.5$). The biological tree was least likely ($\Delta lnL = -177.0 \pm 40.5$).

Table 4.1: Amino acid composition bias in mitochondrial protein-coding genes [a]

	% FYMINK	% GARP	FYMINK / GARP
Honeybee [b]	49.1 ± 0.9	9.5 ± 0.5	5.2
Nematode	38.8 ± 0.8	12.3 ± 0.6	3.2
Locust	38.4 ± 0.8	15.4 ± 0.6	2.5
Fruit fly	35.6 ± 0.8	16.2 ± 0.6	2.2
Brine Shrimp	32.0 ± 0.8	17.2 ± 0.6	1.9
Allomyces	31.4 ± 0.8	20.4 ± 0.7	1.5
Sea Urchin	28.9 ± 0.8	21.1 ± 0.7	1.4
Chicken	26.3 ± 0.7	21.9 ± 0.7	1.2

[a] Protein sequences of the twelve protein-coding genes common to all the taxa were aligned individually using clustalw (14). The ragged ends of the individual alignments were trimmed, and then the alignments were concatenated to make an alignment 3713 amino acids in length. In this alignment, the proportion (± standard error) of AT-rich FYMINK amino acids (Phe, Tyr, Met, Ile, Asn, and Lys), GC-rich GARP amino acids (Gly, Ala, Arg, and Pro), and the ratio between them are shown.

[b] Genbank accession numbers L06178, X54252, X80245, X03240, X69067, U41288, J04815, and X52392, respectively

4.3.1 Distance, parsimony, and maximum likelihood methods fail to find the correct tree

We constructed phylogenetic trees using both the distance-based neighbor-joining method, and the method of maximum parsimony (Felsenstein, 1993). Mitochondrial protein sequences from the fungus *Allomyces macrogynus* were used as an outgroup to the animal mitochondrial sequences. The results (Figure 4.1) show that, although the honeybee and nematode are widely separated in evolutionary time (Figure 4.1A), they are erroneously grouped together in both of the computed phylogenetic trees (Figure 4.1B & C). Despite being incorrect, these computed trees are very well supported as indicated by the high bootstrap values. Using the method of maxi-

mum likelihood (Adachi and Hasegawa, 1992), the true biological tree was shown to be least likely of the three trees shown in Figure 4.1.

Erroneous trees were also obtained when individual mitochondrial genes from these taxa were examined. The protein sequences of the *cob*, *cox1*, and *nad5* genes were analysed by neighbor-joining and maximum parsimony. In the case of both *cob* and *cox1*, for both methods of analysis, a tree with the topology shown in Figure 4.1B was obtained. In the case of the *nad5* gene a tree with the topology (Allomyces, ((Chicken, SeaUrchin), (((Locust, FruitFly), (Honeybee, Nematode), BrineShrimp))) was obtained by neighbor-joining, and a tree with topology (Allomyces, ((Chicken, SeaUrchin), ((Locust, FruitFly), (BrineShrimp, (Honeybee,Nematode))))); was obtained by maximum parsimony. In no case was the correct biological tree obtained, and in all cases the honeybee and nematode grouped together.

The effect of a change of outgroup was examined, again using the full set of sequences common to all the mitochondria examined. When the mitochondrial sequences of the liverwort (*Marchantia polymorpha*, accession number M68929, 32.4% FYMINK, 21.5% GARP, FYMINK/GARP=1.5) were substituted for those of *Allomyces* as the outgroup to the analysis, identical tree topologies were obtained, the same as Figure 4.1B and C for neighbor-joining and maximum parsimony, respectively. The likelihood of these trees, as measured by protml v 2.2, using the Dayhoff model with frequencies taken from the sequences using the -f option, was again parsimony $\approx$ neighbor-joining > biological (parsimony $lnL = -53917.5$; neighbor-joining $lnL = -53925.4$, $\Delta lnL = -7.8$, standard error 13.3; biological $lnL = -54119.9$. $\Delta lnL = -202.4$, standard error 41.4).

A new mutation probability matrix specific to vertebrate mitochondrial proteins has been calculated (Adachi and Hasegawa, 1996) and has been incorporated into the puzzle program (Strimmer and von Haeseler, 1996). When this model was used, the order of the likelihoods of the tree topologies was again parsimony $\approx$ neighbor-joining > biological (parsimony $lnL = -56437.73$; neighbor-joining $lnL = -56455.66$, $\Delta lnL = -17.93$; biological $lnL = -56657.33$. $\Delta lnL = -219.6$).

4.3.2 LogDet/paralinear transform

Phylogenetic analyses commonly assume that trees are homogeneous, that is that the same model applies throughout the tree, and that the sequence composition is stationary. The LogDet/paralinear transform is a method of calculating a distance matrix which is able to recover the correct tree when sequences evolve under non-homogeneous, non-stationary models (Lockhart et al., 1994; Lake, 1994). There are two forms of the LogDet transform, which are given in Equations 1 and 3 in Lockhart et al. (1994). Equation 3 is equivalent, with scaling, to the paralinear distance (Lake, 1994; Swofford et al., 1996).

Care needs to be taken in the choice of which sites of an alignment to include in the calculation of the LogDet/paralinear distance. Inclusion of invariant sites in the distance calculation tends to mis-estimate the amount of change (Lockhart et al., 1994; Lockhart et al., 1996). Additionally, sites which vary a great deal are problematic because of saturation. It has been shown to be useful to exclude both of these extremes by using only parsimony sites (Lockhart et al., 1994). Another area where care is required in these calculations stems from the use of the logarithm of the determinant of the matrix of transitions between the two sequences between which the distance is being calculated. The calculation can result in a negative determinant, for which the logarithm is undefined. The interpretation in this case would be that there is such a large divergence between the two taxa that the sequences are effectively random. The distance between those taxa is then arbitrarily large. In order to tree the distance matrix, using for example the neighbor-joining algorithm, one needs to choose an arbitrarily large number as the distance between these problem taxa. For example, the program PAUP* version 4.0 sets the values of these undefined distances at twice the distance of the largest defined distance in the distance matrix. However, the choice of this distance affects the tree topology, and so caution is needed in interpreting such trees.

We calculated LogDet distances using both Equations 1 and 3 (Lockhart et al., 1994), using all 3713 sites or using only the 1715 parsimony sites. When using parsimony sites there were 9 pairwise comparisons which had negative determinants.

We set these to 1.1 ×, 2 ×, and 10 × the largest defined distance. The resulting distance matrices were then analysed using the neighbor-joining method. When all sites were used, both Equation 1 and Equation 3 resulted in a tree of the same topology as the tree shown in Figure 4.1B. LogDet calculations based only on parsimony sites resulted in the trees shown in Figure 4.2. In no case was the correct biological tree obtained, and so it appears that this data set is intractable to correction by the LogDet/paralinear transform.

4.3.3 Removal of biased amino acids

We can speculate that the honeybee and nematode mitochondrial proteins have independently become "FYMINK-rich" at the amino acid level, due to AT pressure at the nucleotide level, and that many of these FYMINK amino acids happen to be at homologous sites in the two sequences. Phylogenetic algorithms then mistake this correspondence as relatedness due to recent common ancestry, and consequently group the sequences together in the inferred tree. To test this, various sites in the alignment were removed, and the remainder re-analysed. Entire columns were removed, thereby preserving the alignment.

We first removed all sites in the alignment which contained any of the FYMINK amino acids. The remainder was analysed with protdist/neighbor-joining and with protpars (Figure 4.3A). The distance and parsimony tree topologies for this shortened alignment of 993 positions are the same as that for the entire alignment of 3713 positions, although some bootstrap values are lower. The nematode and honeybee still group together. Similar results were obtained when, in addition to the FYMINK set of amino acids, leucine was also removed (Figure 4.3B). Recall that in all these taxa leucine has two codon families, one of which is AT-rich, while the other is AT-neutral. When the 1422 positions which contained any of the GARP amino acids were removed, we obtained the trees shown in Figure 4.3C. Again, the honeybee and nematode are found together or nearby in the tree, and the bootstrap values are somewhat smaller. We then removed all the positions in the alignment which contained any of FLYMINK (including leucine) or GARP amino acids (Figure 4.3D).

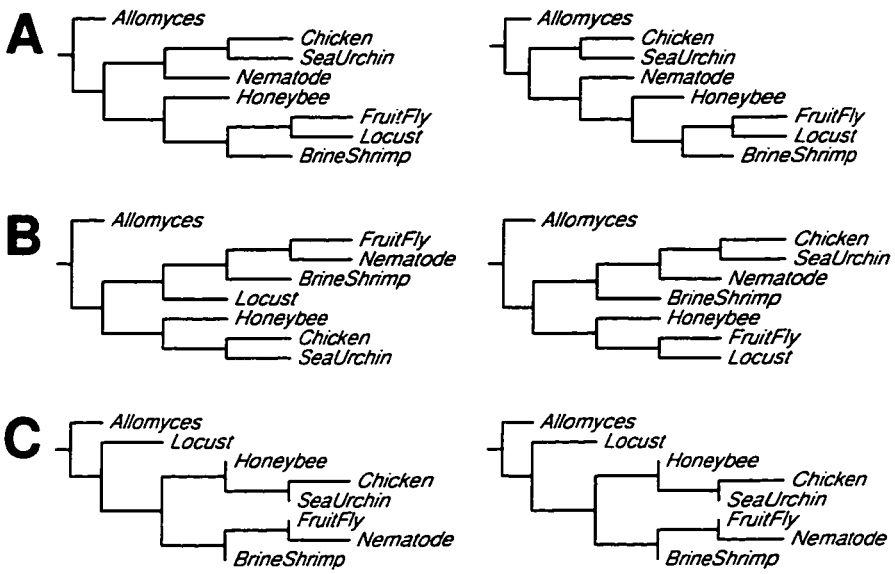


Figure 4.2: Neighbor-joining trees from LogDet/paralinear distances of parsimony sites. Distances of the parsimony sites of the alignment described in Figure 4.1 were calculated using the LogDet transform. Using the equations from Lockhart et al. (1994), calculations using Equation 1 are shown on the left, and calculations using Equation 3 are shown on the right. Distance matrices were then treed using the neighbor-joining method. Resulting trees are shown, with all branch lengths except negative branch lengths equalized. As described in the text, some pairwise sequence comparisons resulted in undefined distances because of negative determinants. These were set to the following factors times the largest defined distance: **A.** 1.1 ×, **B.** 2 ×, **C.** 10 ×.

This perturbed the resulting trees somewhat more, tending to separate the honeybee and nematode. This short alignment of only 360 amino acids did not result in the correct tree, and the bootstap values were low.

4.4 Discussion

The neighbor-joining, maximum parsimony, and maximum likelihood methods all failed to reconstruct the correct phylogeny from entire mitochondrial protein sequences. Not only did they fail, but they indicated incorrect trees with high statistical confidence. No fault can be found with the choice of sequences, as mitochondrial sequences are commonly used in phylogenetics, and the use of the entire genome is considered especially reliable (Russo et al., 1996). In addition, the LogDet/paralinear transform did not allow reconstruction of the correct tree, even when only parsimony sites were used (Lockhart et al., 1994; Lake, 1994). Nei describes a related phylogenetic problem in using total amino acid sequences (Nei, 1996). When 11 vertebrate species were examined, a correct and well-supported phylogeny was obtained. However when lamprey and sea urchin sequences were incorporated, an incorrect phylogeny was obtained with high bootstrap values using several different tree-building methods. The reason for this was not clear.

We tested the possibility that similarly-biased taxa tended to group together ("attract") solely because of an increase in AT-rich codons, or a decrease in GC-rich codons, by removal of positions in the alignment where the corresponding amino acids were found. Again, the correct tree was not obtained. It appears that the amino acid bias does not reside only in those groups of amino acids. We can suppose that there are several signals in the sequences, some of which come from common ancestry and some from amino acid composition bias. It appears that the signal due to composition bias can pervade the sequences and overwhelm or hide the signals due to common ancestry.

We conclude that phylogenetic trees based on amino acid sequences can indeed be misleading because they are subject to the effects of compositional biases. In the case we have described here, the incorrect result is very well supported statistically. This is because we have used a large data set (several thousand amino acids from each taxon) and because we deliberately chose an example where the differences in amino acid composition are pronounced. More subtle biases will cause similar problems, however, in cases where the real phylogenetic distinctions are more difficult, such

as the analysis of very ancient divergences. It will be of special interest to look for the possible effects of compositional bias in those protein-based phylogenies which have been the subject of much recent debate (Golding and Gupta, 1995; Doolittle et al., 1996; D'Erchia et al., 1996). For instance, in one of these studies (D'Erchia et al., 1996), the molecular phylogeny was deemed to be highly reliable based on the consistency of the results obtained by different methodological approaches, the large number of sites included in the analysis, and the very significant bootstrap values obtained. In the example we have given here all of these criteria are also met, but for a molecular phylogeny that is obviously wrong. This indicates that, despite the power of molecular phylogenetic inference, caution is warranted in the interpretation of all molecular phylogenies.

4.5 Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to DAH. We thank Brian Golding and Lars S. Jermiin for their comments on the manuscript, and Peter Lockhart for his comments on the LogDet transform.

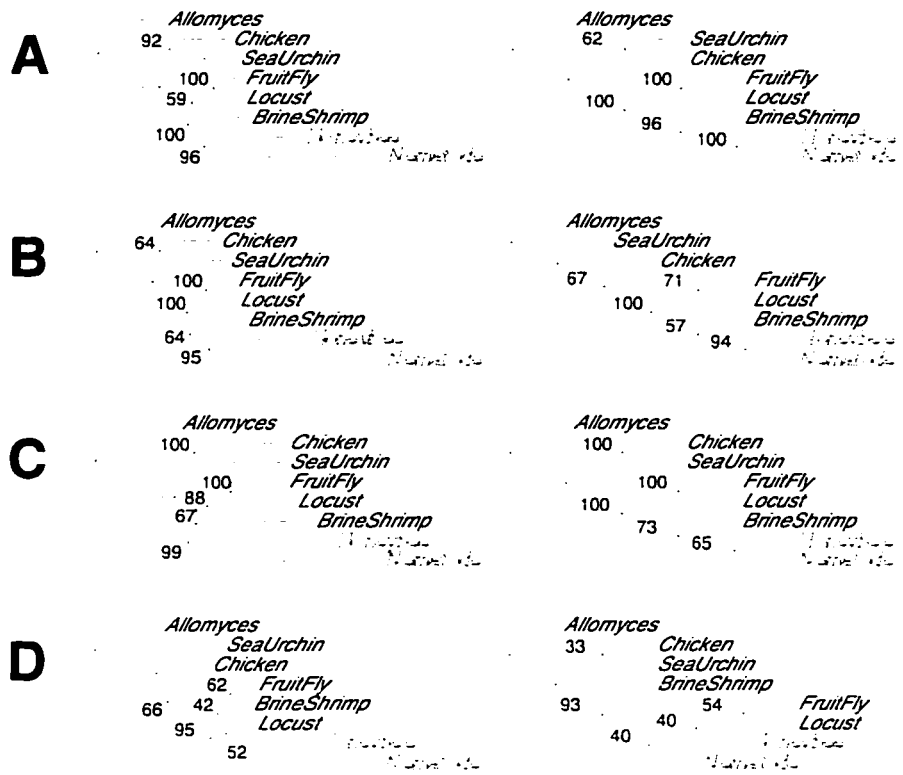


Figure 4.3: Removal of biased amino acids from the alignment before analysis with the distance-based protdist/neighbor-joining on the left, and by maximum parsimony using protpars on the right. Neighbor-joining branch lengths are meaningful, but the branch lengths of the parsimony analysis have been equalized. Numbers are bootstrap values, the percent of 100 bootstraps. The original alignment is 3713 positions.

A. FYMINK removed, leaving 993 positions. B. FLYMINK removed, leaving 806 positions. C. GARP removed, leaving 2291 positions. D. Both FLYMINK and GARP removed, leaving 360 positions.

Chapter 5

Biased branches attract

5.1 Abstract

It has been shown in protein-coding genes that there is a correlation between AT/GC bias at the nucleotide level and content of AT- and GC-rich codons, and their corresponding amino acids. Consequently, protein sequences can also be affected secondarily by nucleotide compositional bias. Recently it has been shown that this bias in the amino acid content of proteins can result in incorrect protein-based phylogenetic trees. In that study, the two taxa with the greatest compositional bias continually grouped together despite a lack of close biological relatedness.

Probabilistic models of protein sequence evolution currently in use assume that the amino acid composition of the protein sequences under analysis is constant. This is not realistic, because we know that amino acid compositions of sequences differ among taxa. To accommodate this, directional mutation models are proposed here, which model the evolution of protein sequences when the amino acid composition changes.

Here I have made evolutionary simulations using both directional models and stationary models on the same tree. The directional branches then become compositionally biased during the simulations. When the sequences resulting from the simulations are then analysed using a maximum likelihood program with its intrinsic

stationary model, the directional branches tend to group together despite being on non-adjacent branches in the simulation. Therefore since the assumption of stationary composition inherent in the analytical method is not in accord with the sequences under analysis, the analysis is compromised and tends to give erroneous results.

5.2 Introduction

Differences in AT/GC content of the DNA of genes and genomes is common (Lee et al., 1956; Muto and Osawa, 1987). In protein-coding genes, these differences in the AT/GC content at the nucleotide level can result in amino acid composition differences. An increase in GC content of the DNA can result in an increase in GC-rich codons and their corresponding amino acids, and an increase in AT content of the DNA can result in an increase in AT-rich codons and their corresponding amino acids (Foster et al., 1997; Foster, 1997; Andersson and Sharp, 1996; Berkhout and van Hemert, 1994; Jukes and Bhushan, 1986; Jermiin et al., 1994; D'Onofrio et al., 1991; Collins and Jukes, 1993; Porter, 1995)(but see Hashimoto et al., 1994, 1995 for exceptions).

When DNA compositional bias is extreme, many now consider phylogenetic reconstruction based on encoded protein sequences from these problem taxa to be more reliable than DNA-based phylogenetic reconstruction (Hasegawa and Hashimoto, 1993; Loomis and Smith, 1990; Lockhart et al., 1992). However, it has recently been shown that the DNA-driven amino acid composition bias described above does indeed have implications for protein-based phylogenetic reconstruction (Foster, 1997). In the example that was used to illustrate this problem well-supported yet incorrect trees were obtained with several phylogenetic methods of analysis, including maximum likelihood. In that example two AT-rich taxa, apparently driven by a similar directional mutation pressure at the DNA level, became compositionally biased in similar ways at the protein level, and consequently grouped together in inferred phylogenetic trees despite lack of close biological relatedness.

Model-based methods for phylogenetic reconstruction from protein sequences

have various assumptions, but the sequences under analysis may or may not have evolved in accord with those assumptions. For example, models currently in use, such as the Dayhoff model (Dayhoff et al., 1978) or the JTT model (Jones et al., 1992), assume a stationary amino acid composition: methods which employ these models necessarily incorporate that assumption. We know however that real sequences do differ in amino acid composition among taxa. The stationary models of evolution that we use for analysis must therefore differ from the processes by which real compositionally biased sequences have evolved. We want to know how robust current methods employing stationary models are to this violation of stationarity. Here the method of maximum likelihood is tested using simulations. Using simulation, since the evolutionary history and the true tree are known, we can ask how well maximum likelihood with a stationary model as it is currently used is able to recover the correct tree in the face of compositional bias.

In the present study I ask if biased branches group together ("attract") in evolutionary simulations. Biased branches are made in these simulations with non-stationary Markov models of evolution, while stationary Markov models such as the Dayhoff model are used on the non-biased branches. Calculation of non-stationary models is described below. The results of the simulations are analysed using the maximum likelihood program `protml` (Adachi and Hasegawa, 1992) using the JTT stationary model to find the best-supported trees. Results show that there is indeed a tendency for biased branches to group together in these analyses despite lack of close relatedness.

5.3 Methods

The matrix powering approach taken by Benner, Cohen, & Gonnet (1994) allows calculation of mutation probability matrices from sequences that are widely separate in evolution, and for this reason it is preferred over the classic Dayhoff approach (Dayhoff et al., 1978). The process begins with tabulation of a transition matrix, $\boldsymbol{T}$, where elements $\boldsymbol{T}_{ij}$ are the number of substitutions from amino acid j to amino

acid i in the alignment(s) from which the matrix is computed. $\boldsymbol{T}$ is like Dayhoff's accepted point mutation table (Dayhoff et al., 1978, Figures 79, 80) except that it includes amino acids that do not change, which are placed on the diagonal of $\boldsymbol{T}$. In the Benner, Cohen, & Gonnet formulation, $\boldsymbol{T}$ is symmetrical, made so by counting substitutions in both directions. From $\boldsymbol{T}$ can be computed $\boldsymbol{N}$, which is the sum of each of the columns of $\boldsymbol{T}$, made into a diagonal matrix. It was recognized that

$$\boldsymbol{M}^{x} \cdot \boldsymbol{N} = \boldsymbol{T}$$

and therefore

$$\boldsymbol{M}^{x} = \boldsymbol{T} \cdot \boldsymbol{N}^{-1} \tag{5.1}$$

where x is the average evolutionary distance of the sequences from which $\boldsymbol{T}$ was made. Having defined a unit of evolution as one accepted mutation in a sequence of 100 amino acids, then

$$\sum \boldsymbol{f}(1 - \boldsymbol{M}_{ii}) = 0.01 \tag{5.2}$$

where $\boldsymbol{f}$ is the vector of frequencies of the amino acids, and $\boldsymbol{M}_{ii}$ is the diagonal of $\boldsymbol{M}$. This equation defines one PAM (accepted point mutation (Dayhoff et al., 1978)), and can be used to find x. Since $\boldsymbol{f}$ can be obtained from either $\boldsymbol{T}$ or $\boldsymbol{N}$, it is only necessary to find, by numerical approximation, an x and $\boldsymbol{M}$ which will make the Equations 5.1 and 5.2 both true.

Some protein sequences evolve toward a biased protein composition, and in order to model this I propose directional (non-stationary) mutation probability matrices. They are made as are other mutation matrices but with the difference that when changes between sequences are tallied to make $\boldsymbol{T}$, they are counted in only one direction. An entry $\boldsymbol{T}_{ij}$ in $\boldsymbol{T}_{A \rightarrow B}$ is the number of times that amino acid j occurs in

sequence A aligned with amino acid i in sequence B. In directional mutation models, $\boldsymbol{T}$ is not symmetrical. Directional mutation matrices have properties required of a proper mutation matrix: the probabilities of every column sum to 1.0, and the relationship $\sum \boldsymbol{f}(1 - \boldsymbol{M}_{ii}) = 0.01$ still holds, which defines one PAM. Additionally the matrix raised to a very high power indicates the characteristic equilibrium frequencies.

In a symmetric $\boldsymbol{T}$, the sum of the columns (and, being equal, the sum of the rows) is proportional to $\boldsymbol{f}$, the frequency vector of the sequences from which the matrix was derived. This is the same frequency vector to which any sequence mutated using the model $\boldsymbol{M}$ implied by that $\boldsymbol{T}$ will evolve. Considering directional mutation matrices, the sum of the columns of $\boldsymbol{T}_{A \to B}$ is proportional to $\boldsymbol{f}_A$, and the sum of the rows of $\boldsymbol{T}_{A \to B}$ is proportional to $\boldsymbol{f}_B$. Note that the B sequence would not be at the equilibrium frequency.

When making models using the matrix powering approach from finite amounts of sequence information, it is a common problem to tabulate a $\boldsymbol{T}$ matrix which contains zeros for transitions between some amino acids pairs. These zeros can be problematic because in calculating $\boldsymbol{M}$ from $\boldsymbol{M}^x$ by $(\boldsymbol{M}^x)^{1/x}$ we can obtain elements of $\boldsymbol{M}$ which are less than zero. This can be corrected by replacing the zeros in $\boldsymbol{T}$ with arbitrary small numbers. The approach that I have adopted is to convert the problem $\boldsymbol{M}$ to $\boldsymbol{T}^1$ by scalar multiplication of the elements of the columns of $\boldsymbol{M}$ by the elements of the frequency vector $\boldsymbol{f}$. Then any positions in $\boldsymbol{T}^1$ which are less than or equal to zero are changed to half the lowest positive entry. Since $\boldsymbol{T}^1$ represents substitutions expected for an evolutionary distance of 1 PAM, a corrected $\boldsymbol{M}$ can then be calculated without compromise from the corrected $\boldsymbol{T}^1$ using the Dayhoff method.

5.3.1 Simulation of evolution

Dayhoff (1978) described how to do evolutionary simulation using mutation probability matrices, as follows. Let us say that we want to mutate a protein sequence a PAM using a model $\boldsymbol{M}$. First, $\boldsymbol{M}$ is raised to the power a. For each amino acid

Table 5.1: Dayhoff mutation probability matrix for 250 PAM, excerpt [a]

Replacement amino acid	Original amino acid			
	A	R	N	D
A	0.132066	0.061058	0.090602	0.093228
R	0.028665	0.166791	0.040891	0.030386
N	0.042045	0.040419	0.064196	0.065109
D	0.050154	0.03482	0.075479	0.114037

[a] Excerpted from Dayhoff (1978) Figure 83

in the sequence, a random number is chosen between 0 and 1. Directing attention to the amino acid's appropriate column in $\boldsymbol{M}^a$, we use the random number to choose the amino acid to which the original mutates. For example, if we use the Dayhoff model at 250 PAM (Table 5.1), and the amino acid that is to be mutated is alanine, we formulate a mechanism which says that if the random number is between 0 and 0.132066 then it will remain alanine, if it is between 0.132066 and 0.132066+0.028665 it will be mutated to arginine, if it is in the following 0.042045 slot it will be mutated to asparagine, and so on. The process is repeated for each amino acid along the length of the sequence.

5.4 Results

5.4.1 Compositional bias from directional mutation models

In order to simulate the evolution of sequences with compositional bias in a plausible manner the simulation models were made from real sequences. Both stationary and directional models were newly calculated from mitochondrial sequences, and these models were used in simulations as described below. The results of the simulations were then analysed using the `protml` maximum likelihood program using the JTT stationary model to test whether this method can recover the simulation tree.

Calculation of a mutation probability matrix by the matrix powering approach described above begins with an alignment of protein sequences. For the first set

of simulations, the alignment described in Foster (1997) was used. This alignment was made from the complete protein sequences of seven animal mitochondria and a fungal mitochondrial outgroup (Genbank accession numbers L06178, X54252, X80245, X03240, X69067, U41288, J04815, and X52392). The protein sequences of the 12 protein-coding genes common to all taxa were aligned individually using `clustalw` (Thompson et al., 1994). The ragged ends of each of these alignments were trimmed and the resulting alignments were concatenated to make a single alignment 3713 amino acids in length. Based on this alignment nematode and honeybee erroneously grouped together in phylogenetic reconstructions (Foster, 1997). The fungal outgroup was used to make the alignment, but was not used to make the mutation probability models below.

This alignment was used to make both a stationary model and a directional model. The seven animal mitochondrial sequences were divided into two groups based on amino acid composition bias, with the nematode and honeybee in the high FYMINK/GARP ratio group, and the sea urchin, chicken, locust, fruit fly, and brine shrimp in the low FYMINK/GARP ratio group (Figure 5.1). The stationary model $\boldsymbol{M}_{SCLFB}$ was calculated from $\boldsymbol{T}_{SCLFB}$, the matrix of all substitutions observed among the sequences in the low FYMINK/GARP group (Table 5.2). In this group there were 17603 substitutions from one amino acid to another, and 19527 positions where the amino acid was the same on both sequences compared. Since this is a stationary model, each of these was counted in both directions, and so the resulting matrix $\boldsymbol{T}_{SCLFB}$ is symmetrical. The average evolutionary distance of 72.4 PAM which makes both Equations 5.1 and 5.2 true is found by numerical approximation, and this allows calculation of the mutation probability matrix $\boldsymbol{M}_{SCLFB}$ (Table 5.3). Note that calculation of $\boldsymbol{M}_{SCLFB}$ using the Dayhoff (1978) method from such widely diverged sequences would not be reliable. The directional model $\boldsymbol{M}_{NH}$ was calculated from all substitutions observed between all the sequence combinations from the low FYMINK/GARP group to the high FYMINK/GARP group (Table 5.4). In this case there were 21760 substitutions and 15370 positions where the amino acid was the same in both the two sequences. Since this is a directional model, each of these was

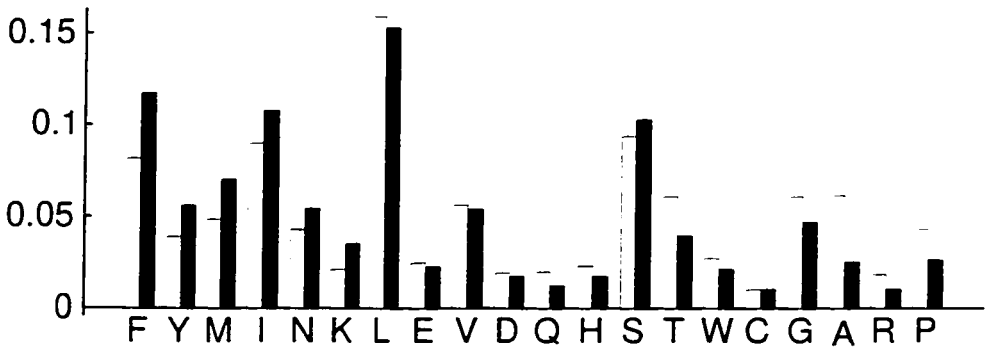


Figure 5.1: Amino acid composition of the two groups of sequences used to make the mitochondrial mutation probability models. Amino acids are ordered such that FYMINK is on the left, coded for by AT-rich codons, and GARP is on the right, coded for by GC-rich codons. The white bars show the composition of the low FYMINK/GARP sequences (sea urchin, chicken, locust, fruit fly, and brine shrimp), and the black bars show the composition of the high FYMINK/GARP sequences (nematode and honeybee). Note that each AT-rich codon (FYMINK) is lower in the low FYMINK/GARP sequences, and each GC-rich codon (GARP) is higher in the low FYMINK/GARP sequences.

counted in the given direction only, and so the resulting matrix $\boldsymbol{T}_{NH}$ is not symmetrical. Finding the average evolutionary distance of 100.7 PAM allows calculation of the directional mutation probability matrix $\boldsymbol{M}_{NH}$ (Table 5.5). The frequencies of the amino acids taken from the models, $\boldsymbol{f}_{SCLFB}$ and $\boldsymbol{f}_{NH}$ (Figure 5.2), are not identical to the frequencies of the amino acids in the sequences (Figure 5.1). The difference is due to positions in the alignment with a gap in one of the sequences—these pairs are not counted in the model.

If we take a sequence with amino acid frequencies $\boldsymbol{f}_{SCLFB}$ (the frequencies of the amino acids used to make the SCLFB model, the white bars in Figure 5.2) and mutate it with $\boldsymbol{M}_{NH}$, the amino acid frequencies will change. The frequencies of the FYMINK group of amino acids will increase at the expense of the GARP group of amino acids. The frequencies of the amino acids at any distance a PAM are proportional to the sum of the rows of $\boldsymbol{C}$ in

Table 5.2: Transition matrix T_{SCHB} [a]

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5634	24	158	32	52	40	48	289	16	211	305	56	135	158	144	724	275	41	65	161
R	24	986	28	1	1	42	7	20	13	14	13	53	20	4	4	24	10	0	14	6
N	158	28	1202	81	19	82	72	91	76	94	148	106	52	85	59	346	148	23	67	55
D	32	1	81	946	0	4	77	15	15	10	20	12	6	16	11	45	17	9	11	8
C	52	1	19	0	262	2	2	21	3	40	75	3	24	25	12	84	53	1	15	18
Q	40	42	82	4	2	688	78	3	67	24	44	48	28	31	20	79	49	9	32	14
E	48	7	72	77	2	78	1074	22	15	31	38	39	19	19	24	79	29	10	31	10
G	289	20	91	15	21	3	22	3006	14	54	101	20	53	43	67	269	85	11	21	63
H	16	13	76	15	3	67	15	14	1114	13	34	13	22	34	3	54	35	10	59	2
I	211	14	94	10	40	24	31	54	13	2358	1262	47	373	343	59	229	293	47	88	694
L	305	13	148	20	75	44	38	101	34	1262	5850	61	761	840	87	384	352	73	186	518
K	56	53	106	12	3	48	39	20	13	47	61	758	29	35	26	87	53	4	21	5
M	135	20	52	6	24	28	19	53	22	373	761	29	1096	199	29	156	144	31	58	149
F	158	4	85	16	25	31	19	43	34	343	840	35	199	2940	46	180	159	65	319	167
P	144	4	59	11	12	20	24	67	3	59	87	26	29	46	2124	140	62	13	39	43
S	724	24	346	45	84	79	79	269	54	229	384	87	156	180	140	2862	526	41	103	156
T	275	10	148	17	53	49	29	85	35	293	352	53	144	159	62	526	1682	24	78	186
W	41	0	23	9	1	9	10	11	10	47	73	4	31	65	13	41	24	1424	48	12
Y	65	14	67	11	15	32	31	21	59	88	186	21	58	319	39	103	78	48	1426	35
V	161	6	55	8	18	14	10	63	2	694	518	5	149	167	43	156	186	12	35	1622

[a] Counts of transitions (substitutions) among amino acids in mitochondrial protein coding genes. Substitutions are counted in both directions among the sea urchin, chicken, locust, fruit fly, and brine shrimp sequences. This transition matrix is used to make the stationary mutation matrix M_{SCHB}.

Table 5.3: Stationary mutation probability matrix M_{SCLEB} [a]

To amino acid	From amino acid																			
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	993922	235	856	282	1375	381	393	1200	23	515	365	561	712	409	800	2580	1185	318	305	710
R	35	996422	151	28	53	599	22	71	93	24	3	714	134	7	13	46	18	20	83	10
N	299	352	986985	1410	529	1524	874	406	1075	264	216	1939	245	264	375	1525	813	199	527	230
D	44	29	629	995517	53	27	874	35	140	6	14	102	11	40	25	103	35	40	28	19
C	114	29	126	28	986090	27	22	88	23	132	149	26	156	66	50	367	336	20	111	58
Q	62	645	705	28	53	990285	1049	9	888	36	47	714	178	79	100	241	248	40	222	38
E	79	29	503	1128	53	1306	993512	71	93	84	34	510	89	26	125	229	88	79	222	10
G	598	235	579	113	529	27	175	995121	93	60	108	153	289	79	350	871	283	40	83	288
H	4	117	579	169	53	1034	87	35	995024	6	27	102	111	79	13	126	141	40	444	10
I	378	117	554	28	1164	163	306	88	23	984348	3147	663	3205	1174	300	550	1750	437	499	5988
L	475	29	806	113	2327	381	218	282	187	5585	989543	561	6677	3430	375	1009	1485	556	1082	2476
K	97	821	957	113	53	762	437	53	93	156	74	990789	156	92	150	264	248	20	111	10
M	281	352	277	28	740	435	175	229	234	1726	2026	357	983451	712	125	516	813	278	361	691
F	272	29	503	169	529	327	87	106	280	1067	1756	357	1202	990309	200	447	690	596	2967	710
P	281	29	378	56	212	218	218	247	23	144	101	306	111	106	995274	413	230	79	250	173
S	1978	235	3348	508	3385	1143	874	1341	514	575	594	1174	1002	515	900	987490	3713	358	693	710
T	589	59	1158	113	2010	762	218	282	374	1186	567	714	1024	515	325	2408	986475	159	555	1113
W	70	29	126	56	53	54	87	18	47	132	95	26	156	198	50	103	71	996226	333	10
Y	97	176	478	56	423	435	350	53	748	216	263	204	289	1412	225	287	354	477	991042	58
V	325	29	302	56	317	109	22	265	23	3739	871	26	801	488	225	424	1025	20	83	986691

[a] Probability ($\times 10^6$) of substitutions among amino acids in mitochondrial protein coding genes for an evolutionary distance of 1 PAM, calculated from the transition matrix T_{SCLEB} (Table 5.2).

Table 5.4: Transition matrix T_{NH} [a]

Amino acid in nematode and honeybee	Amino acid in sea urchin, chicken, locust, fruit fly, and brine shrimp																			
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2396	17	124	30	7	47	43	108	30	123	219	60	83	139	103	223	134	25	53	71
R	6	318	1	0	0	2	0	0	6	5	3	2	2	0	0	2	2	0	0	1
N	108	22	503	65	15	52	68	113	66	84	105	75	45	48	74	222	88	16	44	32
D	14	1	43	352	2	5	37	1	14	14	5	11	11	6	6	32	18	7	6	5
C	26	3	15	0	65	6	2	14	0	22	39	2	18	19	6	52	32	5	11	13
Q	17	18	12	7	9	197	9	7	15	16	15	6	11	7	12	24	14	1	8	10
E	11	0	29	42	1	20	460	10	7	14	25	8	13	12	28	39	16	1	17	12
G	99	6	47	24	6	9	5	1066	4	38	44	9	17	11	19	131	30	7	12	11
H	12	10	13	3	0	14	7	3	450	6	13	6	3	2	2	21	1	0	14	5
I	188	12	78	8	39	15	18	68	10	984	754	41	197	271	78	181	205	44	59	405
L	165	10	78	14	43	49	36	102	31	634	2252	35	326	394	92	236	227	84	97	295
K	66	92	85	11	8	52	25	34	19	57	73	348	35	32	32	102	64	20	29	11
M	122	28	80	11	38	42	15	69	28	257	552	24	476	135	44	143	101	31	51	138
F	185	24	80	17	41	25	28	86	31	297	650	23	183	1356	81	238	168	88	219	160
P	45	5	14	1	1	11	6	14	2	11	8	2	8	5	704	32	12	0	3	11
S	494	19	141	26	36	66	44	285	36	118	197	37	77	103	121	1224	287	27	60	97
T	92	14	54	24	10	29	23	56	2	90	92	14	40	27	45	140	509	4	13	62
W	5	10	14	1	1	3	4	12	1	20	33	1	12	36	6	16	11	518	10	11
Y	80	23	54	23	17	33	22	45	44	120	230	21	76	179	15	95	93	50	630	50
V	153	10	31	9	17	15	10	41	10	232	267	13	59	72	38	131	118	20	22	562

[a] Counts of transitions (substitutions) among amino acids in mitochondrial protein coding genes. Substitutions are counted from any of the sea urchin, chicken, locust, fruit fly, and brine shrimp sequences to both the nematode and honeybee sequences. This transition matrix is used to make the directional mutation matrix M_{NH}.

Table 5.5: Directional mutation probability matrix M_{NH} [a]

To amino acid	From amino acid																			
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	993999	116	1424	503	52	972	607	613	371	440	469	1216	707	707	993	1070	1018	236	440	381
R	17	992842	12	28	52	54	22	9	93	36	3	25	22	7	12	6	9	20	14	10
N	253	233	988571	2014	840	1620	1561	946	1622	476	201	2330	398	170	893	1593	702	158	523	38
D	17	29	600	993455	52	27	737	9	232	59	3	203	110	7	12	137	123	79	28	10
C	70	58	225	28	982632	162	22	70	23	119	114	25	309	118	12	478	456	79	165	76
Q	44	640	150	168	945	987204	173	18	371	95	3	152	133	13	149	125	105	20	110	76
E	4	29	350	1119	52	594	993908	18	46	24	40	101	110	39	298	182	70	20	193	76
G	314	58	575	559	210	54	22	992969	23	202	74	51	110	7	99	819	88	79	83	10
H	26	233	150	28	52	432	87	9	993974	12	27	101	11	7	12	102	9	20	165	38
I	506	29	774	28	2519	27	43	210	23	986072	3197	962	2628	1820	720	546	2035	473	330	6494
L	201	29	300	56	1574	1134	520	490	417	5138	988909	456	4793	2527	720	910	1807	1340	798	2247
K	166	2793	1224	56	210	1674	390	123	232	274	134	992605	331	105	248	512	544	315	303	10
M	349	873	1199	112	3358	1458	43	438	556	1784	2567	355	986029	602	298	649	597	355	523	1238
F	445	407	600	168	2204	27	260	350	324	1392	2071	152	1855	991797	695	1104	1105	1380	3165	1028
P	166	116	175	28	52	324	87	70	23	36	3	25	66	7	992494	182	70	20	14	95
S	2294	175	1924	224	2834	1944	650	2767	603	333	415	557	663	484	1414	989241	3877	355	661	876
T	314	407	874	727	525	1134	477	438	23	737	194	203	508	26	571	1138	985192	20	55	724
W	4	233	175	28	52	27	43	70	23	83	60	25	88	209	25	46	53	993930	83	57
Y	183	466	550	503	735	864	304	210	881	535	590	355	729	1139	12	330	790	828	992211	152
V	628	233	150	168	1049	270	43	175	139	2153	925	101	398	209	323	831	1351	276	138	986364

[a] Probability ($\times 10^6$) of substitutions among amino acids in mitochondrial protein coding genes for an evolutionary distance of 1 PAM, calculated from the transition matrix T_{NH} (Table 5.4).

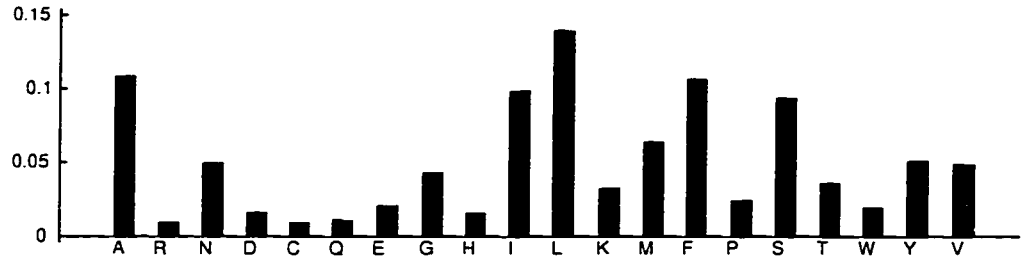


Figure 5.2: Amino acid frequencies from the transition matrices of the models $\boldsymbol{M}_{SCLFB}$ (white bars) and $\boldsymbol{M}_{NH}$ (black bars). The gray bars show the frequency of amino acids originally $\boldsymbol{f}_{SCLFB}$ but which have been mutated 100 PAM under the model $\boldsymbol{M}_{NH}$.

$$\boldsymbol{M}_{NH}^{a} \cdot \boldsymbol{N}_{f.SCLFB} = \boldsymbol{C} \tag{5.3}$$

where $\boldsymbol{N}_{f.SCLFB}$ is a diagonal matrix with elements proportional to the starting frequency $\boldsymbol{f}_{SCLFB}$. If we take a protein sequence originally of composition $\boldsymbol{f}_{SCLFB}$, as shown in the white bars in Figure 5.2, and mutate it 100 PAM using the model $\boldsymbol{M}_{NH}$, the resulting amino acid frequencies are as shown in the gray bars in Figure 5.2. The frequencies of two amino acids, phenylalanine (F, Phe) and glycine (G, Gly), as they change from $\boldsymbol{f}_{SCLFB}$ levels over 500 PAM are plotted in Figure 5.3. Notice that the frequency of phenylalanine increases while the frequency of glycine decreases, as expected. The frequencies approach an asymptote, the equilibrium frequencies of the model.

5.4.2 Biased branches attract

A simple four taxon unrooted tree was used for the simulation (Figure 5.4). The simulation was done with the terminal branches all evolving at the same rate, all 100 PAM long, and with no rate heterogeneity among sites. The composition of the ancestral sequences was taken from the stationary model. The configuration shown

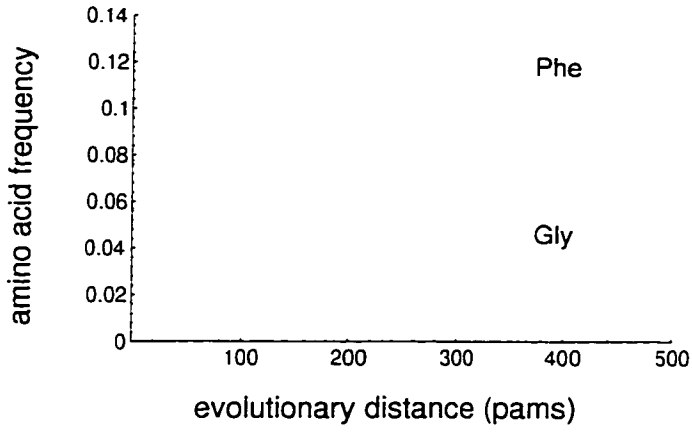


Figure 5.3: Amino acid frequencies of two amino acids, phenylalanine (Phe) and glycine (Gly), from original frequencies from $\boldsymbol{f}_{SCLFB}$, over evolutionary distance, using the mutation model $\boldsymbol{M}_{NH}$.

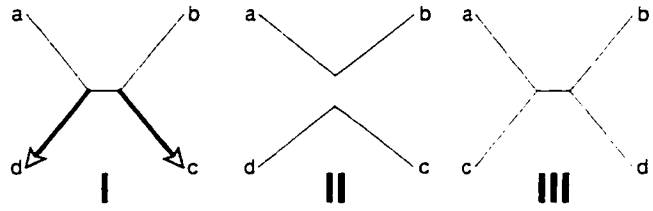


Figure 5.4: Tree topologies. Simulations were all made in the form of tree I. Control simulations were made using the stationary model on all five branches. When it was used, the directional model $\boldsymbol{M}_{NH}$ was used on two non-adjacent branches in the simulations, as shown in tree I. The stationary model $\boldsymbol{M}_{SCLFB}$ was used on the other three branches. The trees II and III are the other two possible trees.

in tree I in Figure 5.4 was used throughout the simulations, and so is the true tree against which the inferred trees are tested. Tree reconstruction from the sequences resulting from the simulations used the `protml` program with the JTT model.

The control simulation where all branches used the stationary model $\boldsymbol{M}_{SCLFB}$ is shown in Figure 5.5A. In these simulations, the correct tree, tree I, was recovered

most of the time even though the internal branch length was small. If however the two lower branches evolved using the model $\boldsymbol{M}_{NH}$ as shown in Figure 5.5B, tree II was recovered most of the time when the branch length was small. In this tree, the two biased branches group together. The correct tree was recovered most of the time only when the internal branch length was more than about 3 PAM.

To show that this effect is not peculiar to substitutions among mitochondrial sequences, an artificial mutation probability matrix was made (see the caption for Figure 5.6 for details of its construction) and the simulations duplicated using it as the directional mutation matrix, and the Dayhoff matrix as the stationary matrix. Results of the simulation are shown in Figure 5.6. Again we see that when all branches used the stationary model the correct tree (tree I) was obtained most of the time even when the internal branch length was small. However when the lower two branches used the directional model tree II was recovered most of the time when the internal branch length was small, and the correct tree was obtained most of the time only when the internal branch was greater than about 3 PAM.

5.5 Discussion

Simulations were made in the form of the unrooted 4-taxon tree shown in Figure 5.4, tree I. The length of the internal branch was varied. This branch provides information for subsequent analyses to distinguish among the three trees shown in Figure 5.4: when the internal branch length is small the tree is more star-like and it is more difficult for the analysis program to recover the correct tree. In the control simulations stationary models were used throughout, and the correct tree was usually recovered using `protml` even though in both examples (Figures 5.5 & 5.6) the simulations used the newly calculated stationary models and the analysis used the JTT stationary model. However, when two non-adjacent branches in the simulation were mutated using a non-stationary model, and the resulting sequences were subsequently analysed using the JTT stationary model, there was a tendency for the two compositionally biased branches to attract. It appears that we have a case here where an assumption

of the method is violated by the sequences under analysis, and because of that the analysis is compromised. This can be considered at least a partial explanation for the analysis described above where mitochondrial protein sequences of the nematode and honeybee grouped together in phylogenetic trees despite lack of close relatedness (Foster, 1997).

It is expected that this problem will be more serious in deep phylogenies, where a subtle trend can compound over long periods of time. The effect of compositional bias may also be important in analyses of star-like phylogenetic trees involving rapid radiation, where internal branches are small and subtle effects can bear on the result in a manner similar to that shown in the simulations in this study. Where the effects of compositional bias are not severe enough to affect the tree topology, the analysis may still be compromised by misestimated branch lengths or by lower confidence in the topology as shown in reduced bootstrap values.

We can speculate that phylogenetic reconstruction using more realistic models or combinations of models may be more successful than the current approach of using a single stationary model for the entire tree. At present, however, computational tools for analysis of phylogenetic trees using different models on different branches are not available. Additionally, sequence alignment is a crucial part of molecular phylogenetic analysis, but this consideration has not been addressed in this study. Sequence alignment bears not only on the phylogenetic analysis but the creation of the model as well. It may be worthwhile to develop alignment methodology which can use different models on different sequence combinations, and which can use directional models where appropriate.

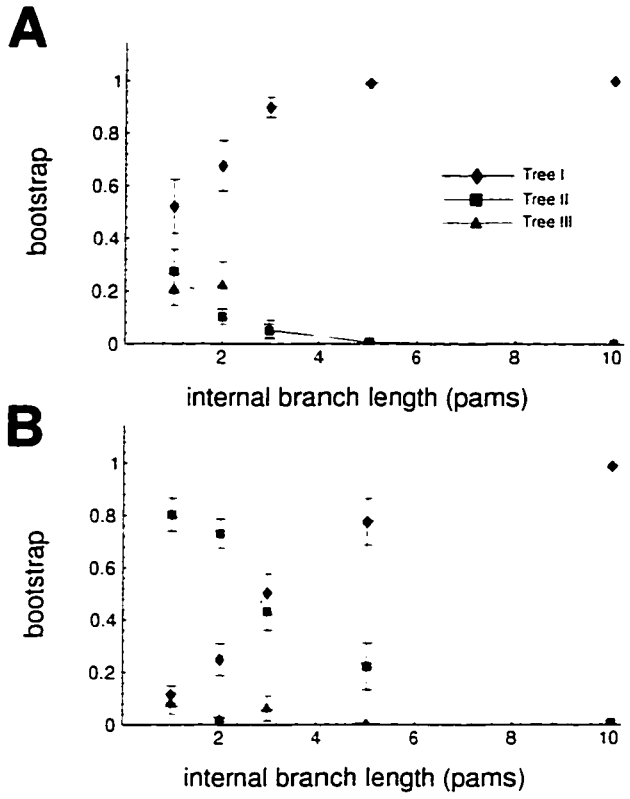


Figure 5.5: Simulation using M_{SCLFB} and M_{NH} and success of tree recovery. Simulations used sequences 20000 amino acids in length. The simulations were all of the form of tree I in Figure 5.4, with all external branches 100 PAM in length, and the internal branch length varied. At the end of the simulation, the bootstrap of the maximum likelihood of the three possible trees was estimated using `protml` version 2.2. In this analysis the JTT model with the f switch was used, taking the amino acid frequencies from the sequences. Ten replicates of each simulation were made, and the standard error of the mean of the bootstrap estimates is shown. **A.** Control simulation with all branches using M_{SCLFB}. **B.** Simulation as in **A** but with the lower two branches using the M_{NH} model.

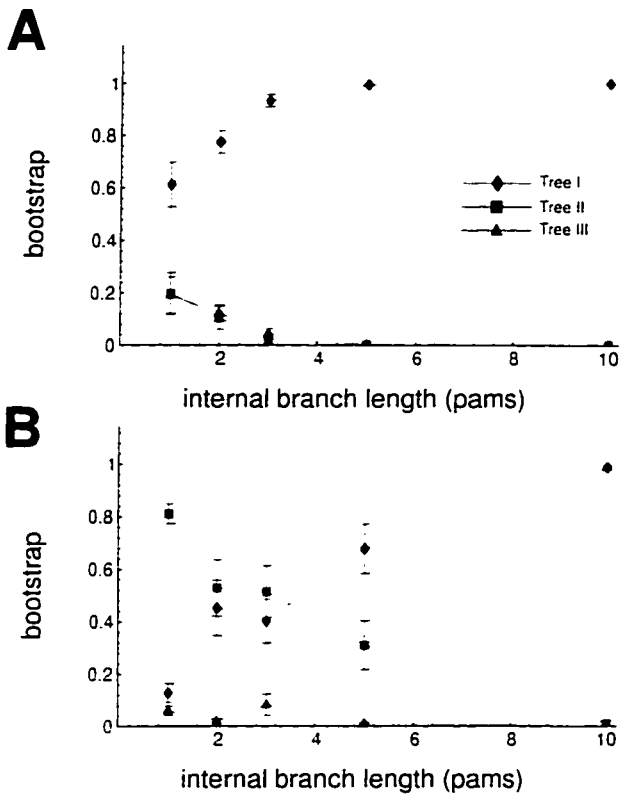


Figure 5.6: Simulation using the Dayhoff model $\boldsymbol{M}_{D78}$ and $\boldsymbol{M}_{bias}$. The $\boldsymbol{M}_{bias}$ model was made starting with a random amino acid sequence 10000 residues in length with frequencies taken from the Dayhoff model. The sequence was duplicated, and the duplicate mutated an arbitrary 30 PAM using the Dayhoff model. To then create an arbitrary compositional bias, 30% of the sequence was replaced with equal amounts of the amino acids ACDEFGHIKL, chosen because they were the first 10 in the alphabetical list of amino acids. $\boldsymbol{M}_{bias}$ was calculated using the matrix powering approach from these two sequences. Simulations and analysis were done as described in Figure 5.5. **A.** Control simulation with all branches using $\boldsymbol{M}_{D78}$. **B.** Simulation as in **A** but with the lower two branches using the $\boldsymbol{M}_{bias}$ model.

Chapter 6

Conclusions and future directions

6.1 Conclusions

6.1.1 Nucleotide AT/GC affects protein amino acid composition

In Chapter 1 I described the small body of previous research which showed a relationship between the AT/GC content of protein-coding genes and the amino acids of the encoded proteins. In Chapter 2, to further delineate this relationship between AT/GC content at the DNA level and composition at the protein level, the protein-coding sequences of two complete mitochondrial genomes were analysed. It was found that, relative to the AT/GC-neutral chicken genome, the AT-rich honeybee genome was rich in AT-rich codons, coding for the amino acids F, Y, M, I, N, and K, and poor in GC-rich codons, coding for the amino acids G, A, R, and P. This relationship was highly significant and applied to all AT- and GC-rich codons, and was shown not only in the entire genome but in all the genes individually as well.

In Chapter 3 I tested the generality of this observation by examining this relationship in several complete mitochondrial genomes. These genomes showed a strong correlation between AT/GC content at the DNA level and compositional bias at the protein level. In addition, I examined the genes of α-amylase and trypsin, which in eukaryotes are nuclear genes. It was found that these genes, and even the conserved

*hsp*70 and *ef*1α/*tu* genes, are susceptible to amino acid compositional bias correlated with AT/GC bias in the DNA.

6.1.2 Phylogenetic implications of DNA-driven amino acid composition bias

In Chapter 4 I showed that the relationship described above has phylogenetic implications. A striking example was presented using a set of animal mitochondrial protein sequences including two taxa with very biased amino acid compositions. Common phylogenetic tools failed to recover the correct tree from these sequences. This is an important observation because of a current trend in phylogenetics to favour use of protein sequences over compositionally-biased DNA sequences. That protein sequences can be compositionally biased and can also give erroneous phylogenetic results indicates that continued caution is warranted in phylogenetic reconstruction using protein sequences, especially those that are compositionally biased.

Since AT/GC bias in protein-coding genes is mostly absorbed in the synonymous sites of codons, the effect of compositional bias will not be as pronounced on the protein sequences as on the DNA. However, compositional bias is a problem in phylogenetic analysis, and is expected to be more of a problem in ancient or ambiguous phylogenies. In extreme cases, such as was presented in Chapter 4, the inferred tree topology was incorrect. In cases where there is less extreme compositional bias, while the inferred topology may be correct, it is expected that the results will be compromised as shown in decreased bootstrap values or misestimated branch lengths. We can speculate that compositional bias may have been a component in some phylogenetic studies with ambiguous results (Loomis and Smith, 1990; Liu and Beckenbach, 1992; Golding and Gupta, 1995; Nei, 1996).

The phylogenies described in Chapter 4 provided the additional observation that the two taxa with the most extreme compositional bias grouped together in the inferred trees, despite lack of recent common ancestry. I explored this theme in Chapter 5, using simulation. Some protein sequences, such as were described in Chapters 4, evolve toward a biased protein composition, and in order to model this I proposed in Chapter 5 directional (non-stationary) mutation probability matrices. Using these

models I was able to simulate evolution involving a change in amino acid composition. While the simulations used combinations of newly calculated stationary and directional mutation models, the subsequent phylogenetic analysis used only the stationary model built in to the maximum likelihood program. This served as a test of how robust the maximum likelihood method using a stationary model was to violation of stationarity. Results of the simulation showed that the method was indeed compromised by the biased sequences, and furthermore that biased branches tended to group together, in accord with the results of Chapter 4 using mitochondrial sequences.

6.2 Future directions

In molecular phylogenetics, we achieve our understanding of the evolutionary past by analysis of molecular sequences with our arsenal of phylogenetic tools. In that respect our understanding is only as good as our tools. Compositional bias of molecular sequences is a factor that has been largely left out of conventional phylogenetic methodology. As I have described above, there is now reason to believe that it is a significant factor.

Phylogenetic reconstruction methods often use Markov processes to model evolution using mutation probability matrices. These models are a statement of our conception of the probability that an amino acid in a protein will be replaced by another amino acid in a given period of evolution. The models should of course reflect as much as possible the reality of evolutionary processes in order to accurately reconstruct them. These models have various assumptions built-in, assumptions which may or may not be reflected in the real molecular sequences under analysis. The results presented above indicate that the assumptions of stationarity and homogeneity need to be addressed in phylogenetic reconstruction methodology.

An assumption of these models as they are currently formulated is that the amino acid composition is stationary, that it does not change over evolutionary time (Dayhoff et al., 1978). Another assumption built-in to phylogenetic methods is that evo-

lution represents a homogeneous Markov process, that the same model of evolution applies throughout the evolutionary tree (Zharkikh, 1994; Swofford et al., 1996). As described above, we can now appreciate that some branches of evolutionary trees evolve towards compositional bias, yet phylogenetic methods apply the same "all-purpose" model to these errant branches as to the rest of the tree. Using the same model for all parts of the tree might be quite inappropriate, yet we do so because the conceptual framework and the computational tools to do otherwise have not been available. I argue that the reality of evolution can be better approximated in our reconstructions by allowing some branches of phylogenetic trees to evolve in a qualitatively different manner, that is using a different model of evolution, than other branches on the same evolutionary tree. Phylogenetic software is needed which can use both stationary and non-stationary models, and which can use different models on different branches of the same phylogenetic tree.

Appendix A

Computation

Although much of the computation in this thesis used commonly available phylogenetic programs (`clustalw`, `Phylip`, `protml`), some of the computation required the use of novel software, which I will describe briefly here. A more extensive description is available at `http://bio02.bio.uottawa.ca/~peter`, which also has the complete API (application programmer's interface).

The software is a set of object-oriented tools to make sequence manipulation and analysis easier. I originally conceived these tools as a set of objects to help me write useful programs for manipulating and analysing molecular sequences of DNA and protein. Since then, with the introduction of Objective-Tcl with its shell (see below), it can now be used via scripts from a command-line interpreter, without a compiled program. In my role as a user of these tools, with this shell I can quickly make small scripted programs in a flexible, exploratory way.

Emphasis is placed on sequence analysis and algorithmic sequence manipulation. This is in contrast to graphic-user-interface (GUI) sequence editors, such as GDE (genetic data environment) or SeqPup, which emphasize sequence display and manual sequence manipulation. I knew when I wrote these tools that I could not completely predict what future questions I would want to ask, what manipulations I would want to do, or what analyses I would want to perform on sequences. For this reason I have designed these objects with power and flexibility in mind, and so use a

command-line interface (CLI).

I wrote the main body of the tools in Objective-C, a Smalltalk-like object-oriented superset of the C-language. Objective-C was chosen for various reasons, including clarity of the syntax, an excellent development environment, dynamic binding, computational speed, and the availability of an excellent public domain string class. Additionally, by virtue of its dynamic binding, Objective-C implements "categories", which allow functionality to be added to existing classes at runtime without re-compiling those classes. If the category is written in Objective-Tcl, it is not compiled, yet can still be added to existing objective-C classes at runtime. This allows very fast prototyping. The software therefore requires the system software and Objective-C runtime found in NeXTSTEP (`http://www.next.com`), the MiscString class (from the MiscKit `ftp://ftp.thoughtport.com/pub/next/misckit/`), and Objective-Tcl (from TipTop software `http://www.tiptop.com/`). The software interfaces with `readseq`, `clustalw`, and *Mathematica*, which also need to be available.

Extensive use is made of Objective-Tcl, an extension of the high-level scripting language Tcl which incorporates the syntax and object-orientation of Objective-C (`http://www.tiptop.com/`). There is an Objective-Tcl shell, which can load Objective-C object code (*.o files). With the shell one can instantiate and query "live" objects at the command line. Using the Objective-Tcl shell there is no need to compile Tcl code, and no need to link Objective-C code, which greatly speeds development time. With this configuration I can use Objective-C for the parts of the code that are permanent, and use Objective-Tcl code when a high-level language is desired, and use Tcl scripts for throw-away code and "one-offs".

The software is extensively documented, using a semi-automatic process. The source code contains embedded comments (as does any good source code), but some of those comments are specially positioned and formatted in the source code so that they can be recognized by `autodoc`, a `perl` script which is part of the MiscKit. This script takes the header and implementation files of the class and formats a document clearly describing the class, with all its variables and methods. This API, in rich

text format, is available at `http://bio02.bio.uottawa.ca/~peter/`, to which the interested reader is referred. Source code is available on request.

A.1 The main classes

Sequence Sequence objects can contain both DNA and protein sequences, or either separately. The sequences themselves are encapsulated in MiscString objects. In a Sequence object, I have MiscString objects for the DNA, the protein, the locus, and the name. Sequences have the ability to translate DNA using various genetic codes, and so can also check the given translation between the DNA and protein strings. Protein sequences can mutate themselves using a MutationMatrix as model.

SequenceList Sequences are handled by the group, by SequenceList objects, subclassed from the NeXT List class. It is here that the file reading and writing functionality resides. SequenceList objects can also make square plots (Chapter 2), and they interface with `clustalw` to align the sequences and return an Alignment object.

Alignment Alignment is a subclass of SequenceList. It contains functionality for manipulating, checking, comparing alignments, as well as "pretty-printing" highlighted alignments to encapsulated postscript files. It also interfaces with MutationMatrices and LogDet objects. Each Alignment object has its own MaskedSlice object for manipulation of alignment positions.

MaskedSlice MaskedSlice objects are vertical slices through Alignment objects. They are a means of indicating logical relationships, usually similarities, in the protein sequences of the alignment. You can use masks to decide how to highlight the printout of an alignment, or to decide whether to move columns from one Alignment to another.

LogDet LogDetDna and LogDetProtein objects are for calculating LogDet (equation 1 or 3) and paralinear distances between two Sequences in an Alignment

(Lockhart et al., 1994; Lake, 1994).

MutationMatrix MutationMatrix objects, under the direction of an Alignment object, collect the information necessary to calculate a mutation probability matrix, which it can then calculate. Only subclasses are used– the DayhoffMatrix and BennerMatrix classes. The Dayhoff78Matrix is a subclass of the DayhoffMatrix, and encapsulates the matrix as published (Dayhoff et al., 1978). Difficult calculations, such as non-integral matrix powering, use *Mathematica*, although in future such calculations should migrate to Objective-C code. Sequences use MutationMatrix objects to mutate themselves in simulations.

References

Adachi, J. and Hasegawa, M. (1992). *MOLPHY: programs for molecular phylogenetics, I.— PROTML: maximum likelihood inference of protein phylogeny.* Institute of Statistical Mathematics, Tokyo.

Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, 42:459–468.

Andersson, S. G. E. and Sharp, P. M. (1996). Codon usage and base composition in rickettsia prowazekii. *J Mol Evol*, 42:525–536.

Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng*, 7:1323–1332.

Berkhout, B. and van Hemert, F. J. (1994). The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. *Nucleic Acids Res*, 22:1705–11.

Collins, D. W. and Jukes, T. H. (1993). Relationship between G + C in silent sites of codons and amino acid composition of human proteins. *J Mol Evol*, 36:201–13.

Crozier, R. H. and Crozier, Y. C. (1993). The mitochondrial genome of the honeybee Apis mellifera: complete sequence and genome organization. *Genetics*, 133:97–117.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In Dayhoff, M. O., editor, *Atlas of protein sequences and structure*, volume 5 Suppl. 3, chapter 22, pages 345–352. Nat Biomed Res Found, Washington, D C.

D'Erchia, A. M., Gissi, C., Pesole, G., Saccone, C., and Arnason, U. (1996). The guinea-pig is not a rodent. *Nature*, 381:597–600.

Desjardins, P. and Morais, R. (1990). Sequence and gene organization of the chicken mitochondrial genome. A novel gene order in higher vertebrates. *J Mol Biol*, 212:599–634.

Diaconis, P. and Efron, B. (1983). Computer intensive methods in statistics. *Sci Am*, 248(5):116–130.

D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C., and Bernardi, G. (1991). Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J Mol Evol*, 32:504–10.

Doolittle, R. F., Feng, D. F., Tsang, S., Cho, G., and Little, E. (1996). Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science*, 271:470–7.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann Statist*, 7:1–26.

Efron, B. and Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science*, 253:390–395.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet*, 22:521–65.

Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package) version 3.5c.* Distributed by the author. Department of Genetics, University of Washington, Seattle.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol*, 266:418–27.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions. 2nd ed.* Wiley.

Foster, P. G. (1997). *Phylogenetic implications of the effect of nucleotide bias on amino acid composition.* PhD thesis, University of Ottawa.

Foster, P. G. and Hickey, D. A. (1997). Compositional bias can affect protein-based phylogenetic reconstruction. (In preparation).

Foster, P. G., Jermiin, L. S., and Hickey, D. A. (1997). Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol*, 44:282–8.

Golding, G. B. and Gupta, R. S. (1995). Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Mol Biol Evol*, 12:1–6.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science*, 256:1443–5.

Hasegawa, M. and Hashimoto, T. (1993). Ribosomal RNA trees misleading? *Nature*, 361:23.

Hasegawa, M., Hashimoto, T., Adachi, J., Iwabe, N., and Miyata, T. (1993). Early branchings in the evolution of eukaryotes: ancient divergence of Entamoeba that lacks mitochondria revealed by protein sequence data. *J Mol Evol*, 36:380–8.

Hashimoto, T., Nakamura, Y., Kamaishi, T., Nakamura, F., Adachi, J., Okamoto, K., and Hasegawa, M. (1995). Phylogenetic place of mitochondrion-lacking protozoan, Giardia lamblia, inferred from amino acid sequences of elongation factor 2. *Mol Biol Evol*, 12:782–793.

Hashimoto, T., Nakamura, Y., Nakamura, F., Shirakura, T., Adachi, J., Goto, N., Okamoto, K., and Hasegawa, M. (1994). Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, Giardia lamblia. *Mol Biol Evol*, 11:65–71.

He, M. and Haymer, D. S. (1995). Codon bias in actin multigene families and effects on the reconstruction of phylogenetic trees. *J Mol Evol*, 41:141–149.

Hillis, D. M., Allard, M. W., and Miyamoto, M. M. (1993). Analysis of DNA sequence data: phylogenetic inference. *Methods Enzymol*, 224:456–87.

Jermiin, L. S., Foster, P. G., Graur, D., Lowe, R. M., and Crozier, R. H. (1996). Unbiased estimation of symmetrical directional mutation pressure from protein-coding DNA. *J. Mol. Evol.*, 42:476–480.

Jermiin, L. S., Graur, D., Lowe, R. M., and Crozier, R. H. (1994). Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome b genes. *J Mol Evol*, 39:160–173.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8:275–82.

Jukes, T. H. and Bhushan, V. (1986). Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes [published erratum appears in J Mol Evol 1987;24(4):380]. *J Mol Evol*, 24:39–44.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217:624–6.

Kimura, M. (1983). *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge.

Lake, J. A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proc Natl Acad Sci U S A*, 91:1455–9.

Lee, K. Y., Wahl, R., and Barbu, E. (1956). Contenu en bases purique et pyrimidiques des acids desoxyribonucleiques des bacteries. *Ann Inst Past*, 91:212–224.

Liu, H. and Beckenbach, A. T. (1992). Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. *Mol Phylogenet Evol*, 1:41–52.

Lockhart, P. J., Howe, C. J., Bryant, D. A., Beanland, T. J., and Larkum, A. W. (1992). Substitutional bias confounds inference of cyanelle origins from sequence data. *J Mol Evol*, 34:153–62.

Lockhart, P. J., Larkum, A. W., Steel, M., Waddell, P. J., and Penny, D. (1996). Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc Natl Acad Sci U S A*, 93:1930–4.

Lockhart, P. J., Steel, M. J., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol*, 11:605–612.

Loomis, W. F. and Smith, D. W. (1990). Molecular phylogeny of Dictyostelium discoideum by protein sequence comparison. *Proc Natl Acad Sci U S A*, 87:9093–7.

Muto, A. and Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci USA*, 84:166–9.

Nei, M. (1996). Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet*, 30:371–403.

Osawa, S., Jukes, T. H., Watanabe, K., and Muto, A. (1992). Recent evidence for evolution of the genetic code. [Review]. *Microbiological Reviews*, 56:229–64.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*, 183:63–98.

Porter, T. D. (1995). Correlation between codon usage, regional genomic nucleotide composition, and amino acid composition in the cytochrome P-450 gene superfamily. *Biochim Biophys Acta*, 1261:394–400.

Rensing, S. A. and Maier, U. G. (1994). Phylogenetic analysis of the stress-70 protein family. *J Mol Evol*, 39:80–6.

Russo, C. A., Takezaki, N., and Nei, M. (1996). Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol Biol Evol*, 13:525–36.

Steel, M., Lockhart, P. J., and Penny, D. (1995). A frequency-dependent significance test for parsimony. *Mol Phylogenet Evol*, 4:64–71.

Steel, M. A., Lockhart, P. J., and Penny, D. (1993). Confidence in evolutionary trees from biological sequence data. *Nature*, 364:440–2.

Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol Biol Evol*, 13:964–969.

Sueoka, N. (1961). Compositional correlation between deoxyribonucleic acid and protein. *Cold Sp Harb Sym Quant Biol*, 26:35–43.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci USA*, 48:582–592.

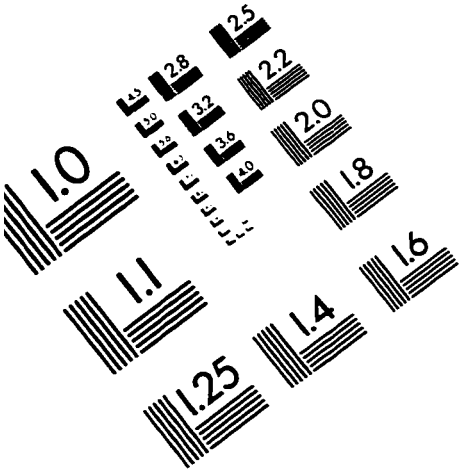
Swofford, D. L., Olson, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. In Hillis, D. M., Moritz, G., and Mable, B. K., editors, *Molecular systematics*. Sinauer, 2nd edition.

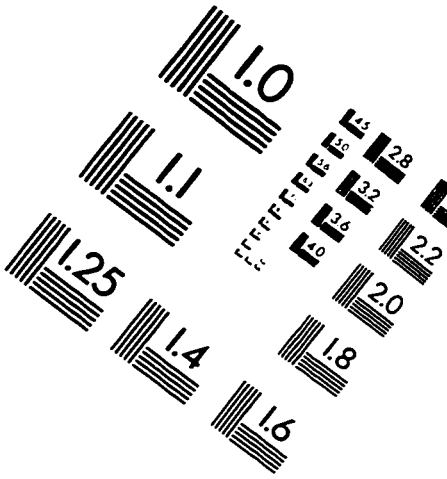
Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22:4673–4680.

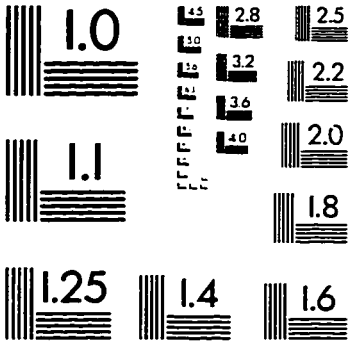
Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *J Mol Evol*, 39:315–29.

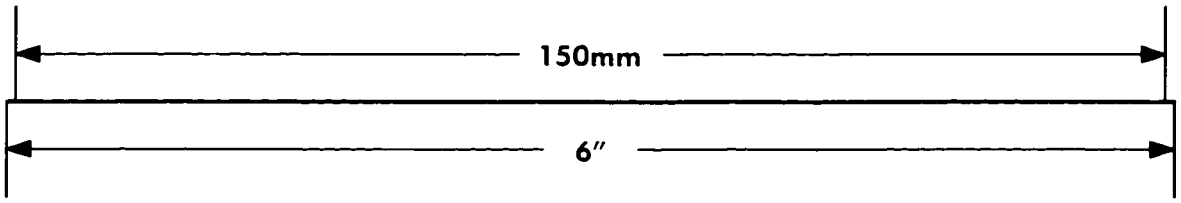
Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In Bryson, V. and Vogel, J., editors, *Evolving genes and proteins*, pages 97–166. Academic Press, New York.

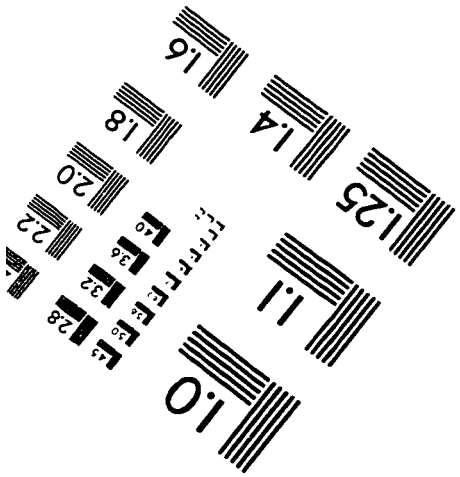
IMAGE EVALUATION
TEST TARGET (QA-3)

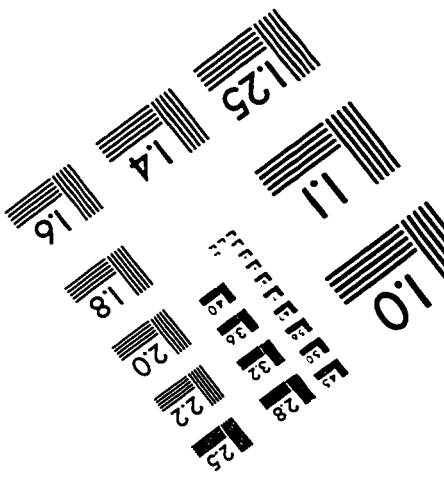












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989