Université d'Ottawa · University of Ottawa

Non-random neutral evolution

Gregory A. C. Singer

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

Thèse soumise à
Faculté des études supérieures et postdoctorales
Université d'Ottawa
en vue de l'obtention du doctorat ès sciences
L'Institut de biologie d'Ottawa-Carleton

9-Oct-02

Acknowledgements

I would like to thank my lab-mates throughout the years. Ada. David. Malgorzata. Huai-Chun. and the many summer students that I had the pleasure to work with. I would like to thank my two friends and colleagues. Rob Beiko and Mehrdad Hajibabaei. whom I've shared many cups of coffee and many slices of pizza with. I would like to thank the members of my advisory committee. Guy Drouin. Marcel Turcotte. George Carmody. and (formerly) Robert Charlebois. Finally—and most importantly—I would like to thank my supervisor. Donal Hickey. whom I've had the great fortune of working with since my undergraduate research project.

Abstract (English)

In this study, we investigate the causes and consequences of directional neutral evolution. We believe this is an evolutionary force that has not received the attention its importance to DNA and protein evolution warrants.

Traditional sequence analyses rely upon models that assume homogeneity of nucleotide and amino acid compositions across different lineages, but our examinations of modern sequences show that this assumption is often violated. Our findings imply that periods of directional changes to the nucleotide or amino acid compositions of biological sequences have occurred in the past. Moreover, we demonstrate that these directional changes are very common and can be quite extreme.

We show that modifications to the genetic code—a phenomenon that is common among mitochondrial genomes—can cause directional changes to the amino acid composition of proteins. Codon reassignments shift the relative number of codons assigned to each amino acid, and when coupled with neutral evolution over a period of time these changes will lead to a corresponding shift in the relative amino acid usages within the protein.

More common causes of directional changes to the amino acid composition of proteins are mutation biases in the DNA. In particular, we show that biases to the relative proportions of A+T and G+C are capable of bringing about amino acid composition changes to the proteins the DNA encodes, with AT-rich DNA favouring the encoding of the FYMINK amino acids, and GC-rich DNA favouring the encoding of GARP amino acids. We show that this effect is both pervasive across all kingdoms of life, and that it also affects every protein within the genome—whether highly or loosely conserved.

Finally, we show that directional neutral evolution in the DNA can operate in the presence of purifying selection in transmembrane domains. Since selection does not preclude neutral evolution if there are multiple favourable mutations possible, DNA mutation bias is able to influence the direction of evolution of the transmembrane domains without interfering with their fitness. We extend this same notion to include the possibility of mutation biases biasing the outcome of positive selection, as well.

Résumé

Dans cette étude, nous examinons les causes et les conséquences de l'évolution neutre directionnelle. Nous pensons qu'il s'agit d'une force évolutive qui n'a pas reçu l'attention qu'elle mérite compte tenu de son importance dans l'évolution de l'ADN et des protéines.

Les analyses conventionnelles de séquences s'appuient sur des modèles qui supposent l'homogénéité de la composition des nucléotides et des acides aminés chez différents lignages. L'examen de séquences modernes montre que dans de nombreux cas cette hypothèse ne tient pas. Cette découverte implique qu'il y a eu dans le passé des périodes de changements directionnels dans la composition des nucléotides ou des acides aminés des séquences biologiques. De plus, ces changements directionnels sont très fréquents et peuvent être assez extrêmes.

Nous montrons qu'il est possible que ces modifications au code génétique (phénomène courant parmi les génomes mitochondriaux) engendrent des changements directionnels dans la composition en acides aminés des protéines. Les redistributions de codons modifient le nombre relatif de codons attribués à chaque acide aminé. Associés à une évolution neutre pendant un certain temps, ces changements conduiront à un changement semblable dans les traitements relatifs des acides aminés au sein de la protéine.

Les changements directionnels dans la composition en acides aminés des protéines peuvent avoir des causes plus communes : les biais mutationnels dans l'ADN. Nous établissons notamment que des biais dans les proportions relatives de AT et de GC sont capables d'entraîner des changements dans la composition en acides aminés des protéines que l'ADN encode : l'ADN riche en AT facilite l'encodage des acides aminés FYMINK et l'ADN riche en GC celui des acides aminés GARP. Nous démontrons que cette action est omniprésente chez tous les êtres vivants et qu'elle a un effet sur chaque protéine au sein du génome, que celui-ci soit bien ou mal conservé.

Finalement, nous montrons que l'évolution neutre directionnelle dans l'ADN peut fonctionner en présence de la sélection purifiante dans les domaines transmembranaires. Comme la sélection n'empêche pas l'évolution neutre s'il est possible d'avoir de multiples mutations favorables, le biais mutationnel de l'ADN peut avoir un effet sur la direction de l'évolution des domaines transmembranaires sans pour autant nuire à leur santé. Nous élargissons la même notion pour y inclure la possibilité de biais mutationnels qui influencent le résultat de la sélection positive également.

Table of Contents

List of Figures

List of Tables

1 Introduction

In this study we describe situations in which protein evolution proceeds in a directional manner. Biological sequences that are subject to positive Darwinian selection fall into this category, but we instead focus on factors that cause non-Darwinian evolution to proceed directionally. Although these factors are not well known, we show that they may play a very significant role in the evolution of proteins.

Protein evolution is often assumed to follow a reversible Markov process. The condition of reversibility means that the model that describes the evolution from sequence O to sequence X can also describe the process of evolution in the reverse direction, X→O. This condition only holds if the relative frequencies of each of the characters in the sequence (nucleotides or amino acids) remain constant over time. Under this model of evolution the number of amino acid replacements in one direction must be equal to the number replacements in the opposite direction. For example, the number of isoleucine → valine mutations over time must be equal to the number of valine → isoleucine mutations, so there is no net increase of one amino acid relative to the other. It is clear that if the number of I→V mutations exceeded the number of V→I mutations from

sequence O to sequence X, then X would have more valine and less isoleucine than O. This model is not reversible, though, because we could not use it to describe evolution in the reverse direction, X→O, where isoleucine increases and valine decreases over time.

Under certain circumstances (discussed in the subsections below), the assumption of reversibility in protein evolution will be violated. In these cases, some (or all) of the substitutions between pairs of amino acids have lopsided frequencies, resulting in an increase of one amino acid in the sequence relative to the other. This sort of directional evolution should, in general, be a temporary phenomenon: after a period of time the amino acid proportions will reach a state of equilibrium and will no longer change (a proof of this statement appears in Section 1.2, below). Positive Darwinian selection is one force that is capable of bringing about such a directional change, and in fact, directional protein evolution is one criterion that is often used to detect natural selection acting on certain sites in the protein (Gu, 1999; Simon et al, 2002). However, it is also possible for neutral, or non-Darwinian evolutionary forces to bring about these directional changes. It is these latter forces that will be the focus of this thesis. We will show that two of these forces—changes to the genetic code and DNA mutation bias—are capable of causing a pattern of directional protein evolution. We also show that directional neutral evolution and natural selection are not mutually exclusive phenomena, and that both can influence the evolution of a protein sequence simultaneously. We provide both qualitative and quantitative models to describe this relationship.

1.1 The neutral theory of molecular evolution

Natural selection has a cost: greater selective intensity leads to fewer surviving individuals in each generation. Worse still, the more traits that are subject to intense selection, the more individuals in the population that must die (Haldane, 1957). Consider a population with two phenotypes, x and y, for example. If there is strong enough selection against x, we might expect as many as half of the x individuals to die in a single generation. Some elementary math shows that over

one million individuals will die in every generation if there are twenty traits that are subject to this level of selection ($\log_2(10^6) \approx 20$; $(½)^{20} \approx 10^{-6}$). Clearly, few populations could avoid extinction in such a situation. With these facts in mind, Haldane (1957) reasoned that a "safe" rate of evolution by natural selection could only produce one amino acid substitution per three hundred generations. It was several years after Haldane published this estimate before enough protein sequence data were available to make empirical measurements on the actual rate of amino acid replacements in proteins. Surprisingly, the measured rates far exceeded Haldane's estimate, and therefore, natural selection alone was deemed inadequate to describe the evolution of protein sequences (Kimura, 1968; King and Jukes, 1969). These results necessitated the development of the neutral theory of molecular evolution, which states that the majority of amino acid (and DNA) substitutions are selectively neutral and only a small proportion are the result of positive selection (Kimura, 1968). In other words, the theory proposes that genetic drift plays a much larger role than Darwinian selection in the evolution of protein and DNA sequences. There are a number of interesting implications that derive from this theory. Without selection, for example, the probability of a new mutation reaching fixation in a diploid population is $^1/_{(2N)}$ (where N is the number of individuals). The number of new mutations in a generation is $2N \times \mu$ (where μ is the mutation rate), so the number of new mutations that will become fixed in a population is $(2N \times \mu) \times {}^1/_{(2N)}$, or simply μ. In other words, the rate of evolution under the neutral hypothesis is equal to the mutation rate. Of course, this theory alone is also inadequate to describe the process of protein evolution, since some mutations are deleterious and therefore proteins do experience some degree of negative selection (Jukes and Kimura, 1984). Nevertheless, even conserved sites within DNA and proteins evolve in a fashion that can be reasonably approximated as neutral, but at a rate that is slower than the sites whose evolution is completely governed by genetic drift (Li, Wu, and Luo, 1985; Li, 1997).

1.2 Symmetrical substitution patterns

Models of molecular evolution generally assume symmetry in the pattern of substitutions, for both DNA (Kimura, 1980; Gojobori, Ishi, and Nei, 1982; Gu and Li, 1996) and proteins (Dayhoff et al, 1978; Henikoff and Henikoff, 1992; Jones, Taylor, and Thornton, 1992). Under these models, if there are two states that a particular site can take, x and y, then the number of $x \rightarrow y$ substitutions over time will be equal to the number of $y \rightarrow x$ mutations over time. Therefore, there is no net gain of x with respect to y or *vice versa*. This assumption is reasonable, since even if the mutation rate from $x \rightarrow y$ is much different than that for the $y \rightarrow x$ mutation, the system will eventually reach a state of equilibrium and symmetry will be restored. Note, however, that even though the number of $x \rightarrow y$ and $y \rightarrow x$ mutations are equal in a symmetrical evolutionary model, the proportions of x and y need not be so. Let us say that $p_{x \rightarrow y}$ is the probability of the $x \rightarrow y$ mutation, $p_{y \rightarrow x}$ is the probability of the $y \rightarrow x$ mutation, and π_x and π_y are the proportions of x and y in the sequence. At equilibrium, π_x is given by the formula:

$$\pi_x = p_{y \rightarrow x} \div (p_{x \rightarrow y} + p_{y \rightarrow x}) \qquad [1]$$

This formula can be derived in the following way. We know that the proportion of x at time t is equal to the amount of x that did not mutate, $((1 - p_{x \rightarrow y}) \times \pi_{x(t-1)})$, plus the amount of y that mutated to x, $(p_{y \rightarrow x} \times \pi_y)$. At equilibrium, though, the proportion of x in the sequence is not changing ($\pi_{x(t-1)} = \pi_{x(t)}$), so the equilibrium proportion of x is given by:

$$\pi_x = (1 - p_{x \rightarrow y}) \times \pi_x + p_{y \rightarrow x} \times \pi_y \qquad [2]$$

Since our hypothetical sequence is composed entirely of x and y, the proportion of y is one minus the proportion of x ($\pi_y = (1 - \pi_x)$). We can therefore substitute $(1 - \pi_x)$ for π_y in Equation 2, which then becomes:

$$\pi_x = (1 - p_{x \rightarrow y}) \times \pi_x + p_{y \rightarrow x} \times (1 - \pi_x) \qquad [3]$$

Some elementary algebra shows that Equation 3 is equivalent to Equation 1, above.

Using Equation 1, note that if the probability of an $x \rightarrow y$ mutation is equal to the probability of a $y \rightarrow x$ mutation, Equation 1 becomes 0.5; in other words, x and y occur in equal proportions within the sequence. If, on the other hand, the probability of an $x \rightarrow y$ mutation is three times that of a $y \rightarrow x$ mutation, Equation 1 becomes 0.25. At equilibrium, then, 75% of the sequence will be y and 25% will be x. When the sequence reaches this point, the relative proportions of x and y will cease to change, unless the mutation probabilities change again.

Given enough time, a symmetrical pattern of substitution will always be reached in a sequence with two characters. In Table 1.1 we show the frequencies of two characters, x and y, in a sequence before and after the relative mutation rates change. Note that there is a brief period of directional evolution in the sequence, after which a new equilibrium is found. At this point the relative proportions of x and y remain constant, and the number of $x \rightarrow y$ and $y \rightarrow x$ substitutions are the same in each generation. Interestingly, in sequences with more than two characters (like DNA and protein sequences), it is possible for non-reversibility to occur in sequences that are at equilibrium, although this is probably just a mathematical curiosity and is unlikely to have biological significance (Smith and Eyre-Walker, 2002).

Although directional evolution may be a short lived phenomenon within a given lineage, it is important to stress that once it has occurred, that lineage will forever have an asymmetrical substitution matrix with respect to other lineages. This is a problem, because many DNA sequence analyses involve models that assume a symmetrical pattern of substitution, since alternative models are mathematically complex (Kimura, 1981; Rodríguez et al, 1990; Gu and Li, 1998). We believe that the assumption of substitutional symmetry in DNA evolution is routinely violated, and that this leads to directional evolution in protein sequences. In the sections to follow we discuss some natural phenomena that can bring about directional changes to the proportions of amino acids in proteins.

1.3 Genetic code modifications

An interesting fact is that even if the DNA sequence of a gene has equal proportions of the four nucleotides, the encoded protein will not have equal proportions of the twenty amino acids. This is because the method by which the genetic code translates the information stored in genes into a protein sequence is biased. Since there are 61 codons (not counting stop codons) and 20 amino acids, there is redundancy in the genetic code. Although we might expect an average of three codons per amino acid, the actual assignment of these "synonymous"

Table 1.1: The frequencies of two states in a sequence, x and y will eventually return to a state of equilibrium after the relative mutation rates are changed. The original mutation rates are $p_{x\to y} = 0.2$ and $p_{y\to x} = 0.3$.

Generation	Proportion of x	Proportion of y	
1	0.600	0.400	
2	0.600	0.400	
3	0.500	0.500	← $p_{x\to y} = 0.3$; $p_{y\to x} = 0.2$
4	0.450	0.550	
5	0.425	0.575	
6	0.413	0.587	
7	0.406	0.594	
8	0.403	0.597	
9	0.402	0.598	
10	0.401	0.599	
11	0.400	0.600	
12	0.400	0.600	
...	...	...	

codons is uneven: some amino acids can be encoded by as many as six synonymous codons in the universal genetic code (arginine, serine, and leucine), while others can be encoded by only one (methionine and tryptophan). Other amino acids can each be encoded by two, three, or four different codons. A random DNA sequence, then, will encode a protein that is biased towards those amino acids with more codons. If we assume that actual DNA sequences evolve under the neutral model (where most mutations are selectively neutral), the biases in the genetic code should be realized as biases in the amino acid contents of actual proteins. Several studies have, in fact, confirmed this relationship (King and Jukes, 1969; Jukes et al, 1975; Dufton, 1983).

Despite being biased, the genetic code will still produce a pattern of protein evolution that is reversible (or, non-directional) once the amino acid composition of the proteins has reached equilibrium (see Section 1.2, above). Periods of directional evolution could, however, be caused by modifications to the genetic code. Since the proportion of an amino acid in a protein is directly related to the number of codons that can encode it, a reassignment of the relative proportions of codons to each amino acid will disrupt the previous state of equilibrium, and will cause directional evolution in the amino acid content of the proteins until a new equilibrium is reached, similar to the example given in Table 1.1, above. This is more than just a theoretical scenario, as modifications to the universal genetic code have been described in a number of different organisms. In particular, animal mitochondrial genomes show a remarkable plasticity in their genetic codes (Knight, Freeland, and Landweber, 2001). In Chapter 2, therefore, we look for evidence of directional evolution in response to genetic code modifications among the over two hundred completely sequenced animal mitochondrial genomes.

1.4 DNA mutation bias

According to the neutral theory of molecular evolution, the probability of fixation of a particular mutation is equal to the rate at which that mutation occurs (Kimura, 1983). If the rate of one mutation is greater than another, then, it has a greater

chance of being fixed in the population than the mutation with the lower rate of incidence. For this reason, mutation biases have the ability to cause a directional pattern of evolution.

In a random DNA sequence, it is expected that the four nucleotides will occur in approximately equal proportions. In order to maintain such a balance, the mutations from one nucleotide to any other must occur at exactly the same rates—a proposition that is unlikely due to the different chemical properties of the bases. In fact, it is commonly accepted that DNA mutations are biased (Lindahl, 1993). The most common form of base mutation in DNA occurs when cytosine, and to an even greater extent its homologue 5-methylcytosine, are deaminated. This process converts the original bases to uracil (in the case of cytosine) or thymine (in the case of 5-methylcytosine). Another common mutation is the oxidation of guanine by hydroxyl radicals, converting it to 8-hydroxyguanine. This new base preferentially pairs with adenosine instead of cytosine. After a round of DNA replication, these mutations collectively tend to increase the AT-richness of the DNA. For this reason, it is believed that the DNA repair process is biased towards G and C. In other words, in the absence of additional information the mismatch repair enzymes can do better than chance by assuming that the G or C nucleotides are "correct" while A and T nucleotides are the product of a mutation event. A G-T mismatch, for example, is far more likely to have originally been a G-C pair (because C→T mutations are common) than an A-T pair (since A→G mutations are relatively rare), so the repair machinery is more likely to replace the T nucleotide instead of the G. The biased mutation and repair mechanisms have been confirmed in animal cells (Bill et al, 1998), but the evidence is not as strong in the prokaryotes, where the process of DNA repair differs somewhat from that within the animals (Drummond and Bellacosa, 2001). Nevertheless, it is likely that similar mutation/repair biases exist in *E. coli* (Cox and Yanofsky, 1967; Nghiem et al, 1988; Au et al, 1989; Tsai-Wu et al, 1991).

Mutation and repair biases are not the only potential cause of nucleotide composition biases in DNA. There is some recent evidence for positive selection

on nucleotide composition skews within bacterial genomes, for example. Links between nucleotide composition and parasitism (Rocha and Danchin, 2002), thermophily (Wang and Hickey, 2002; Lynn, Singer, and Hickey, 2002; Singer and Hickey, 2002), and aerobiosis (Naya et al, 2002) have all been established. However, the current evidence suggests that mutation bias is still the chief factor affecting the base composition of genomes and that selection plays a smaller role (Lynn, Singer, and Hickey, 2002).

Since the most common type of DNA bias results in changes to the relative proportions of G+C and A+T, for simplicity we can reduce this system to two reaction rates, the mutation rate in one direction (G/C→A/T) and the rate of repair (A/T→G/C). If these two rates are perfectly balanced, the A+T and G+C contents of the DNA will be equal. If, however, the rate of mutation towards A+T is not perfectly matched to the rate of repair back to G+C, the DNA will become biased in one direction or the other (Lindhal, 1993). Recall that even in the event of such an imbalance of the mutation/correction rates, the DNA will eventually reach a state of equilibrium, in which the proportions of G+C and A+T remain constant over time (see Section 1.2). Should the relative rates of mutation and repair suddenly change, however, the DNA will mutate in a directional manner until new equilibrium nucleotide frequencies are achieved (see Table 1.1).

Early measurements of the genomic nucleotide contents of various bacteria found a wide range of G+C contents (Belozersky and Spirin, 1958). The ranges of G+C contents within other kingdoms are narrower, but a great deal of variability still exists (see Mooers and Holmes, 2000, for a review). Even if we assume that all modern DNA sequences are currently evolving in a reversible manner (i.e., the G+C contents are stable), the modern diversity in G+C contents combined with the fact that all of these organisms arose from a common ancestor implies that at some time in the past there were periods in which the nucleotide contents of the DNA were changing. Assuming the universal common ancestor had equal proportions of A+T and G+C, after successive speciation events certain lineages increased their A+T contents, while other lineages gained G+C. This process

necessarily involved a directional pattern of DNA evolution. Moreover, this process appears to be ongoing. There is abundant evidence for recent directional evolution in certain insect lineages, for example (Rodríguez-Trelles, Tarrío, and Ayala, 1999, 2000; Wirth, Le Guellec, and Veuille, 1999). These evolutionary trends can be problematic in DNA sequence analyses, since most models of DNA evolution assume a compositional homogeneity across lineages (Lockhart et al, 1992; Lockhart et al, 1994; Galtier and Gouy, 1994, 1995). Interestingly, since DNA encodes proteins, directional evolution in the DNA may cause a parallel pattern of directional evolution in the proteins.

1.5 DNA mutation bias and protein evolution

Although changes to the relative mutation rates of A/T→G/C and G/C→A/T can cause temporary directional evolution in the DNA, this pattern need not affect the evolution of the proteins that the DNA encodes. As long as all of the directional changes are confined to the synonymous codon positions, the amino acid sequence in the proteins will be unaffected (Loomis and Smith, 1990; Lockhart et al, 1992). On the other hand, if the nonsynonymous sites are also affected by directional mutation bias, a directional change in the amino acid composition the encoded proteins will result. Such a finding would be in agreement with the notion of neutral molecular evolution, since the changes are permitted to be fixed in the population (i.e., they are not selected against), yet they are also not the result of positive selection. These amino acid changes, then, would be both neutral and directional. In fact, directional DNA evolution has been shown to cause a corresponding pattern of directional evolution in proteins.

The first correlations between the nucleotide content of DNA and the amino acid composition of proteins were made in the early 1960's (Sueoka, 1961a and b), and subsequent studies have continued to support this relationship (D'Onofrio et al, 1991; Collins and Jukes, 1993; Berkhout and van Hermert, 1994; Hickey et al, 1994; Porter, 1995). In general, these studies found that GC-rich DNA encodes proteins that are rich in the G, A, R, and P amino acids (which are encoded by

GC-rich codons), and have low levels of the F, Y, M, I, N, and K amino acids (which are encoded by AT-rich codons). Conversely, AT-rich DNA encodes genes that are FYMINK-rich and GARP-poor. These results indicate that directional DNA evolution (as a result of mutation bias) causes a corresponding pattern of directional evolution in proteins. One weakness of these earlier studies was that they did not compare homologous proteins. A better method of inferring directional changes to the relative amino acid proportions within proteins is to measure the amino acid contents of the *same protein* in two or more different organisms. Since homologous proteins share a common ancestral sequence, any differences in the amino acid proportions among the modern sequences implies a period of directional evolution in the past. Thanks to the increasing amount of molecular data available, more recent investigations have been able to perform these studies on homologous sequences (Foster, Jermiin, and Hickey, 1997; Li, 1997, pp. 422-423; Lobry, 1997; Gu, Hewett-Emmett, and Li, 1998; D'Onofrio et al, 1999; Lafay et al, 1999; Rodriguez-Trelles, Tarrio, and Ayala, 1999; Wilquet and Van de Casteele, 1999). Without exception, they have found evidence of directional evolution in the proteins. Moreover, these patterns of evolution appear to be the result of directional evolution in the underlying DNA. This trend apparently transcends all phylogenetic borders, as it has been observed in a multitude of different organisms covering all of the domains of life. Despite these findings, the methods typically used in sequence analysis assume a stationary pattern of amino acid usage (Dayhoff et al, 1978; Henikoff and Henikoff, 1992; Jones, Taylor, and Thornton, 1992). Actual protein sequences routinely violate this assumption, which we believe may bias the outcome of these sequence analyses.

In Chapters 2 and 3 we present our own investigations into the effects of directional DNA evolution on the evolution of proteins. We have chosen to study the completely sequenced bacterial genomes, a very large and diverse data set. We first investigate the pervasiveness of DNA mutation bias and its affects on

amino acid usage among the genomes, and then follow up by examining the degree to which each gene within the genome is affected by DNA mutation bias.

1.6 Directional neutral evolution and negative selection

It is widely known that the rate of nonsynonymous nucleotide substitution is much lower than the rate of synonymous substitution (Li, Wu, and Luo, 1985; Li, 1997). Negative ("purifying") selection is responsible for this discrepancy, but the fact that proteins evolve at all indicates that not all mutations are rejected, and that neutral or positively selected mutations are allowed to change the protein sequence over time. The existence of positive selection on particular sites within a protein does not invalidate the neutral hypothesis of evolution, however (Stoltzfus, 1999). If half of the possible amino acids are selected against at a particular site in a protein, for example, neutral-like evolution can still occur among the remaining ten possibilities, which are neutral alternatives with respect to each other. Sites evolving under this model, however, would evolve at a slower rate than sites that were allowed to mutate in a completely neutral manner.

In the event of a new mutational pressure acting on the gene, the synonymous sites will change readily, and will quickly reach a state of equilibrium (i.e., they will return to a symmetrical pattern of substitution), with nucleotide proportions governed by a formula similar to Equation 1 (Section 1.2). Natural selection, on the other hand, will allow certain nonsynonymous mutations to be fixed in the population, but will reject other mutations that decrease the ability of the encoded protein to function properly. The nucleotide proportion at the nonsynonymous sites, then, may never reach the same nucleotide proportions of the synonymous sites. Certainly, the relative nucleotide proportions may reach a point of equilibrium, but it will be an equilibrium that is not determined by Equation 1 (Section 1.2), but by another equation that includes the conservative effects of natural selection. By extension to the protein sequence then, the equilibrium that is reached among the amino acids must also be a factor of both the mutation bias in the DNA and the degree and pattern of negative selection on the protein

sequence. This necessary interaction between directional mutations in the DNA and natural selection acting on the protein has not been studied in the past. In Chapter 5, however, we investigate the details of the relationship between these two forces and their combined effects on the evolution of transmembrane domains.

1.7 Directional neutral evolution and positive selection

The theory of neutral evolution does not deny the existence of positively selected amino acid replacements: it simply states that these substitutions are far less common than those that become fixed via genetic drift (Kimura, 1983). Recent preliminary findings, however, have suggested that as many as 45% of amino acid replacements may be the result of positive selection (Smith and Eyre-Walker, 2002; Fay et al, 2002). If we adopt a more liberal definition of "neutral mutation", however, then the role of neutral evolution can still play a highly important role in protein evolution, even in the presence of strong positive selection. For example, if there were positive selection for a particular residue in a protein to become acidic, both glutamate and aspartate would have positive selection coefficients. The choice between these two possibilities, then, would be selectively neutral—the final outcome would be guided by genetic drift. It may be more appropriate to refer to such mutations as "relatively neutral", since in this example aspartate and glutamate are selectively neutral relative to each other, but are not neutral with respect to the other amino acids. Since we already have cases in which purely neutral evolution can be directional, it is reasonable to assume that "relatively neutral" mutations can also be directional, and that they can bias the outcome of positive selection. In Chapter 6 we expand upon this idea, and revise Sewall Wright's adaptive landscape analogy to encompass the effects of directional neutral evolution (Wright, 1932).

2 Neutral evolution and the genetic code

2.1 Abstract

One hypothesis that might arise from the theory of neutral evolution is that the genetic code itself will affect the amino acid composition of proteins. Previous studies have borne out this relationship. Amino acids with six codons, for example, tend to occur more frequently in protein sequences than amino acids that are encoded by a single codon. We have updated these earlier results by studying the amino acid compositions of the proteomes from seventy completely sequenced bacteria. Our results confirm the positive relationship between the number of codons an amino acid is encoded by and the number of times that amino acid appears in each proteome. An analysis not performed previously is the effect that modifications to the genetic code have on this relationship. We have compared the amino acid frequencies of the proteins from over two hundred mitochondrial genomes, representing five different genetic codes. We found that modifications to the genetic code cause directional changes to the amino acid proportions of the proteins by changing the relative number of codons allotted to each amino acid.

2.2 Introduction

The theory of neutral molecular evolution states that many of the mutations in DNA and proteins are selectively neutral: that is, they do not increase or decrease the fitness of the organism (Kimura, 1968; Kimura, 1983). While positive selection is certainly not denied, it is believed to be of relatively small importance to the evolution of biological sequences compared to random genetic drift[1]. One of the implications of this hypothesis is that the genetic code itself will affect the amino acid composition of proteins. Since there are 61 codons (not including termination codons) but only twenty amino acids, each amino acid should be represented by an average of three codons in the genetic code. In actuality though, only one amino acid has exactly three codons (isoleucine), with the remainder being represented with as many as six codons (serine, arginine, and leucine) or as few as one (methionine and tryptophan). Even if the nucleotide composition of a gene is completely unbiased, the amino acid composition of the encoded protein will be skewed because of the biases within the genetic code. Early studies on a handful of protein sequences confirmed this relationship (King and Jukes, 1969; Dufton, 1983), although some amino acids are consistently over- or under-represented in proteins compared to what one would expect from the proportion of the genetic code assigned to them (Jukes et al, 1975). We decided that, given the large number of protein sequences presently available, it would be beneficial to revisit the link between the pattern of codon assignments in the genetic code and the amino acid usage in proteins.

The positive relationship between codon assignment and amino acid usage leaves open the question of causality. Has the genetic code evolved in such a way that amino acids that are generally "more useful" are assigned a larger proportion of the codons within the genetic code, or is it that the genetic code is a "frozen

---

[1] This view has been challenged by recent findings by Smith and Eyre-Walker (2002) and by Fay et al (2002), who have shown that nearly half of amino acid replacements may be the product of positive selection. However, these results are very preliminary and have only been demonstrated in *Drosophila*..

accident" (Crick, 1968), whose biased codon assignments influence the amino acid usage in proteins? Most research appears to support the former theory (Dufton, 1985; Xia and Li, 1998; Di Guilio and Medugno, 1999; Chiusano et al, 2000; Freeland et al, 2000), but this does not preclude a role for neutral evolution in the development of the genetic code (Maeshiro and Kimura, 1998; Judson and Haydon, 1999). To see the degree to which protein sequence evolution is guided by the genetic code, we have analyzed the effect of genetic code modifications on the amino acid usage in animal mitochondrial genomes. The neutral hypothesis predicts that changes to the relative proportion of codons assigned to each amino acid in the genetic code will have a corresponding directional effect on the relative proportions of those amino acids in the encoded proteins. Conversely, the selection hypothesis would predict no change in the amino acid proportions, since it is assumed that the current proportions are optimal and are maintained by purifying selection.

2.3 Methods

Seventy bacterial genomes[2] and 230 mitochondrial genomes[3] were downloaded from the GenBank FTP site (see Appendix 2.6.1 and Appendix 2.6.2, respectively, at the end of the chapter). Protein sequences (as identified in the GenBank flat files) were concatenated together for each genome and analysed as a whole. In the case of the mitochondrial genomes, the genetic code for each genome was determined from the GenBank flat files[4]. The amino acid proportions and other sequence manipulations were performed using programs written by GS using the Python programming language[5].

---

[2] ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/
[3] ftp://ftp.ncbi.nih.gov/genbank/genomes/MITOCHONDRIA/Metazoa/
[4] See http://www.ncbi.nih.gov/htbin-post/Taxonomy/wprintgc?mode=c
[5] http://www.python.org

Table 2.1: The PAM250 matrix. The matrix is symmetrical, with values representing the log odds of a mutation event between amino acids *i* and *j* versus no mutation event. Thus, negative values indicate that the two amino acids exchange less readily than chance would expect (based on the amino acid frequencies in the sequences), while positive values indicate exchanges that occur more often than is expected. For simplicity, we have removed the entries for gaps and ambiguous amino acids (*, B, X, and Z) that appeared in the original scoring matrix (Dayhoff et al. 1978).

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

2.4 Results

2.4.1 Even if DNA mutations are random, protein evolution is non-random because of biases in the genetic code

Since all protein mutations ultimately arise from mutations in the DNA, it makes sense that amino acids whose codons are very similar are able to inter-mutate more easily than amino acids whose codons are dissimilar, because fewer DNA mutations are necessary to convert one to the other. Glycine (GGN) and alanine (GCN), for example, are only one DNA mutation away from each other (a G→C mutation in the second codon position), so we would expect to see these amino acids mutate from one to the other far more often than phenylalanine (TTY) and glycine (GGN), which are separated by at least two DNA mutations (at the minimum, a T→G mutation in both the first and second codon positions). Some empirical studies of the pattern of amino acid substitution have already been carried out (Dayhoff et al. 1978; Henikoff and Henikoff, 1992; Jones, Taylor, and Thornton, 1992), and the results are tabulated in scoring matrices like the PAM250 matrix, reproduced in Table 2.1 (Dayhoff et al. 1978). The values in the PAM250 matrix correspond to the log odds of observing a substitution from one particular amino acid to another in two distantly related protein sequences, versus that amino acid not changing at all. In the matrix, we can see that glycine ↔ alanine substitutions (with a score of +1) indeed occur more often than glycine ↔ phenylalanine substitutions (with a score of −5), just as we had predicted based on the number of DNA mutations to convert one amino acid to the other. We have compared the PAM250 score for each possible amino acid substitution to the minimum number of DNA mutations required to make that substitution, and it is not surprising to find that amino acid substitutions that require fewer mutations in the DNA are far more likely to occur than those that require multiple substitutions (Figure 2.1). There is concordance, then, between our theoretical predictions based on the genetic code and the actual pattern of amino acid substitution. Indeed, many "mechanistic" evolutionary models that are based on theoretical

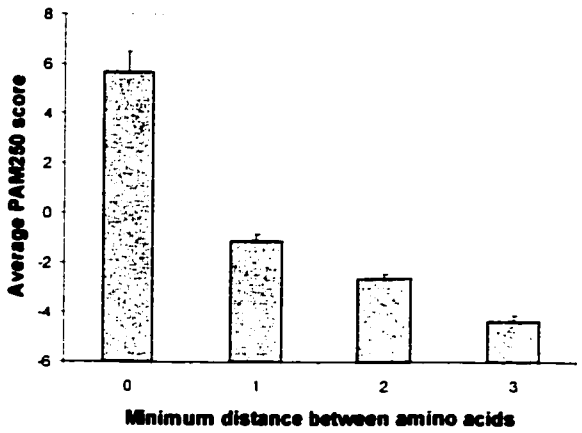


Figure 2.1: The probability of an amino acid substitution occurring is inversely related to the number of DNA mutations required for that substitution. A distance of zero indicates the absence of a mutation. Standard errors from the mean are shown.

predictions (as opposed to "empirical" models, like the PAM matrices, that rely on measurements of actual sequence data) do take into account the number of DNA substitutions required to convert one amino acid to another (Gojobori, 1983; Goldman and Yang, 1994; Yang, Nielsen, and Hasewaga, 1998; McClellan, 2000).

It makes intuitive sense that the more codons an amino acid has, the greater the number of amino acids that will be a single DNA mutation away. Our measurements of the standard genetic code confirm that this is the case (Figure 2.2). Eleven amino acids are a single mutational step away from leucine (which has six codons), for example, while tryptophan (with only a single codon) is a single mutation away from only five amino acids. The implications of this fact are that the more codons an amino acid has, the larger body of amino acids that are capable of mutating into it. Not only does the likelihood of being the endpoint of an amino acid substitution go up with increasing codons, the higher degeneracy of each of the codons will reduce the probability of the amino acid changing again once it exists in the sequence. In short, amino acids with many codons are easy to

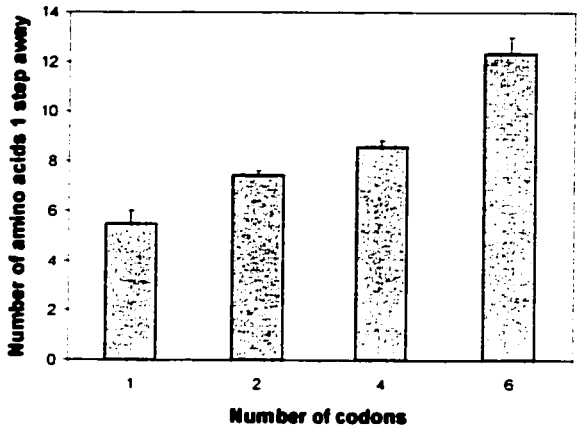


Figure 2.2: As the number of codons for an amino acid increases, so does the number of amino acids that are a single DNA mutation away from that amino acid. Standard errors from the mean are indicated.

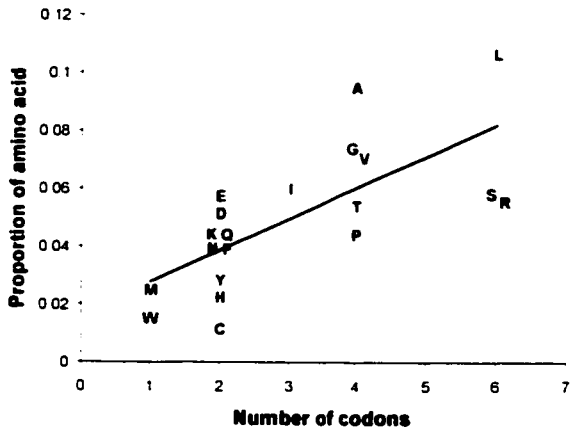


Figure 2.3: The proportion of an amino acid within the *E. coli* K12 proteome is positively related to the number of codons that can encode it ($r^2 = 0.51$). We have laterally shifted the positions of some amino acids for clarity.

be mutated to, but not easy to mutate away from. Under the neutral model of evolution then, we should expect to see amino acids with a large number of codons occurring with relatively high frequency in protein sequences.

2.4.2 The proportion of an amino acid within proteins is related to the number of codons for that amino acid

If one were to throw darts randomly at the genetic code table and record the number of times each amino acid had been hit, the scores for each amino acid would not all be equal. This is because some amino acids are allotted a larger proportion of codons in the genetic code table than others, so these are more likely targets. Similarly, in the DNA sequence of a gene it is far more likely that the codon will encode a leucine than a tryptophan, simply based on the fact that leucine has six codons while tryptophan has only one. When a gene is translated then, the average proportion of an amino acid within its protein will be correlated with the number of codons that can encode that amino acid. We tested this relationship in the entire *Escherichia coli* K12 proteome, and found that the codon count for each amino acid can indeed account for over fifty percent of the variation among the amino acids in the proteins (Figure 2.3). We repeated the analysis for an additional 69 bacterial proteomes, and found that twenty-seven additional genomes showed significant relationships, with an average r^2 of 0.53 (see Appendix 2.6.1 at the end of the chapter; the individual graphs for each genome are in Appendix 2.6.3). Interestingly, the genomes that do not show

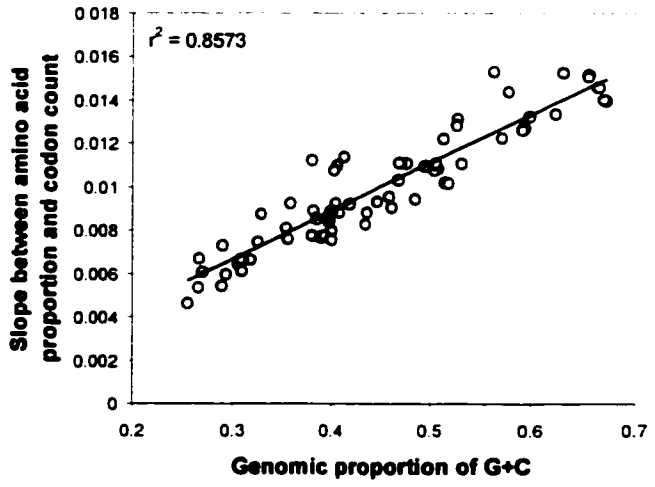


Figure 2.4: AT-rich genomes show a weaker relationship between amino acid content and codon count than GC-rich genomes.

significant relationships tend to have very low proportions of the G and C nucleotides. Indeed, there is a clear trend between GC-richness and the quality of the relationship between amino acid proportion and codon count (Figure 2.4). We will discuss the effect of nucleotide bias on amino acid composition in subsequent chapters, but for now it will suffice to note that our assumption thus far has been that each of the 61 amino acid encoding codons are equally likely to occur within a gene. When the nucleotide frequencies deviate from 25% each however, some codons become more likely to occur in the sequence than others. This effect causes AT-rich genomes to have high proportions of K and I, but low proportions of R and P. Because K and I have fewer codons than R and P, the slope of the regression line between amino acid proportion and codon count decreases. In any event, those genomes that do not violate our assumption (i.e., their nucleotide frequencies are approximately equal) all show a significant correlation between amino acid usage and codon count.

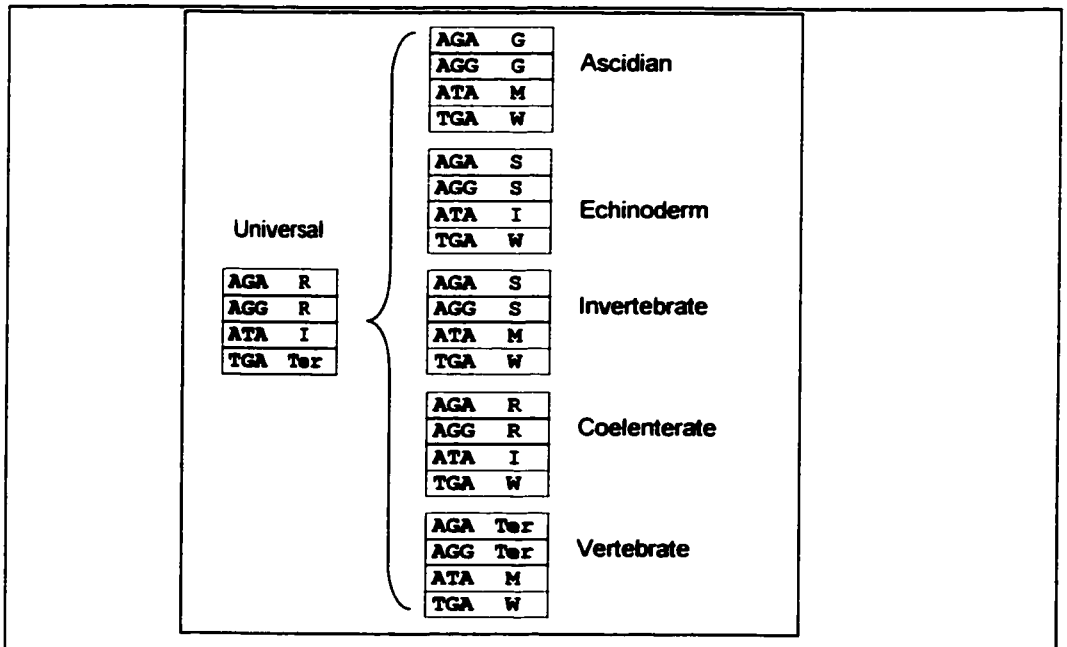


Figure 2.5: The five animal mitochondrial genetic codes are all slight variations of the standard code. Only the variable parts of each code are shown.

Here, though, we arrive at the age-old chicken-and-egg problem: has the genetic code evolved to match amino acid frequencies, or are the amino acid frequencies the result of the pattern of codon assignment in the genetic code? To provide an example: leucine occurs in high proportions in proteins, and also has a relatively high proportion of codons in the genetic code (six). In this case, has the genetic code evolved to provide a large number of codons for this amino acid because it is an important component of most proteins, or has leucine achieved its high proportions in proteins simply because it has a large number of codons? One way to distinguish between these two possibilities would be to modify the genetic code for a particular organism, then observe changes to the amino acid composition of the proteins over a long period of time (if any). If the amino acid proportions remained constant, this would be good evidence that the genetic code had evolved to match amino acid frequencies. On the other hand, if the changes to the genetic code table caused deviations to the relative amino acid proportions, this would suggest that the current amino acid proportions are simply the end result of a biased genetic code table combined with a long period of neutral evolution. Of course, this sort of experiment is impossible since it would take hundreds of thousands of years to accumulate a significant number of mutations to study. Fortunately though, Nature has already performed this experiment for us: animal mitochondrial genomes employ a variety of different genetic codes (Knight, Freeland, and Landweber, 2001). Have the amino acid contents of their proteins been affected by these genetic code modifications?

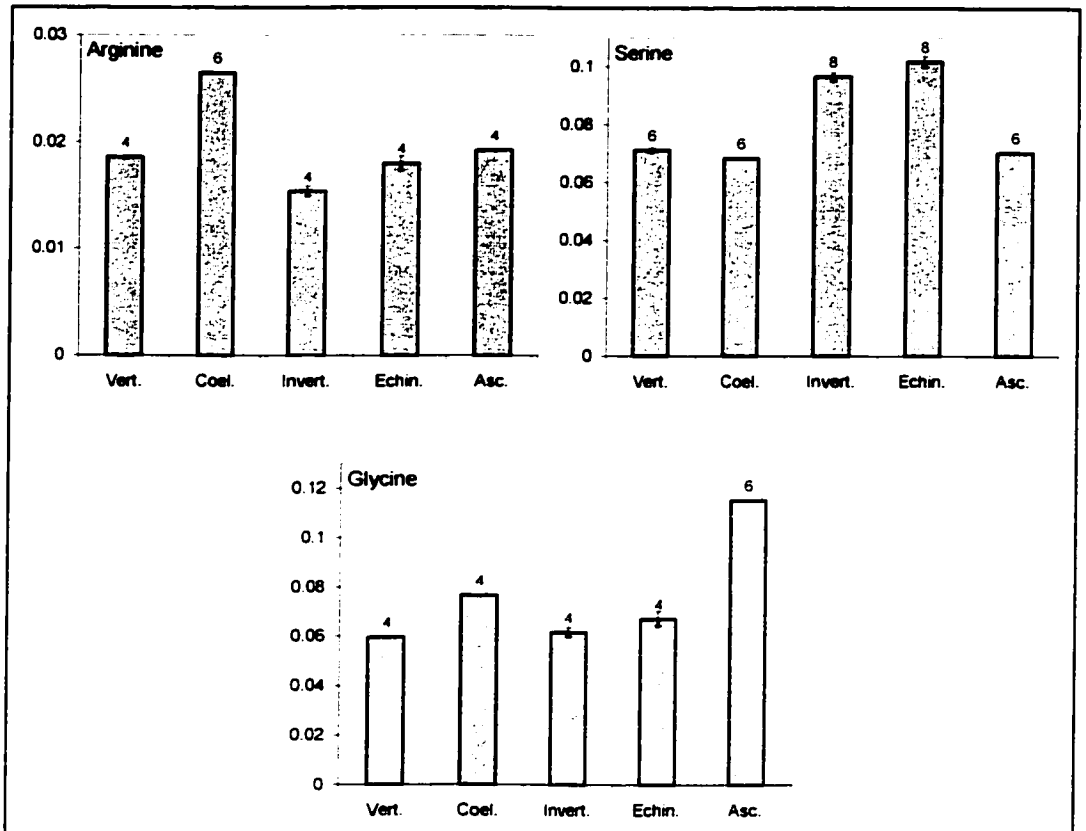


Figure 2.6: Modifications to the genetic code cause a corresponding change in the affected amino acid's proportion within mitochondrial genomes. The numbers on top of each bar indicate the number of codons that encode the amino acid in each genetic code. Error bars indicate standard errors from the mean where appropriate. (Top row, left) The coelenterate mitochondrial genetic code has two more codons than the other genetic codes, and a correspondingly large amount of arginine ($p < 4 \times 10^{-9}$ in a one-sample t-test, since there is only one coelenterate genome). (Top row, right) The invertebrate and echinoderm mitochondrial codes have two more serine codons than the other codes, and the proteins from these genomes therefore have more serine ($p < 10^{-36}$ in a two-sample t-test). (Bottom row, center) The ascidian mitochondrial genetic code has two more glycine codons than the other codes, and this genome has a significantly higher proportion of glycine than the other genomes ($p < 5 \times 10^{-13}$ in a one-sample t-test, since there is only one ascidian genome). Abbreviations for the mitochondrial genetic codes: "Vert"=Vertebrate, "Coel" = Coelenterate, "Invert" = Invertebrate, "Echin" = Echinoderm, "Asc" = Ascidian.

2.4.3 Genetic code modifications cause directional protein evolution

Genetic code modifications are a common phenomenon (O'Sullivan, Davenport, and Tuite, 2001), and this is especially true among the animal mitochondrial genomes (Knight, Freeland, and Landweber, 2001), probably as a result of chance codon reassignments (Knight, Landweber, and Yarus, 2001b; Yokobori, Suzuki, and Watanabe, 2001). Indeed, genetic code modifications have been used successfully as phylogenetic characters (Telford et al. 2000). There are five different genetic codes in use among the over two hundred completely sequenced animal mitochondrial genomes we analysed in this study. The universal genetic code is not among them, although all of the codes are simply minor variations of the standard one (Figure 2.5). Interestingly, the various genetic codes have tended to modify the same sets of codons. For example, most of the mitochondrial genetic codes have modified the meanings of the codons AGA and AGG, which encode arginine in the standard code. They are stop codons in the vertebrate mitochondria, glycine codons in the ascidians, and serine codons in the invertebrate and echinoderm codes. From this single example it is clear that the relative number of codons assigned to each amino acid varies from one genetic code to the next. The coelenterate mitochondrial code, for example, has not modified the AGA and AGG codons from their original meanings—they still encode arginine. This code, then, has two more arginine codons than the other codes. Do genomes that employ this code produce proteins that are more arginine-rich than homologous proteins from other mitochondrial genomes? Surprisingly, yes they do (Figure 2.6: upper row, left). In fact, in every case where the number of codons for a particular amino acid varies by two from one genome to the next (as they do for serine, arginine, and glycine), the plots in Figure 2.6 show that there is a corresponding shift in the proportion of the affected amino acid. Indeed, for some amino acids (methionine and lysine), even a single codon difference between codes is sufficient to cause a corresponding shift in the usage of the amino acids (Figure 2.7: upper row). The trends for isoleucine and asparagine are

less clear, but it is still possible to see that the ascidians, for example, have a significantly lower proportion than the other genomes (as predicted), even though the invertebrate and invertebrate genomes do not follow their predicted trends (Figure 2.7: lower row, left). Similarly, both the ascidian and coelenterate genomes show lower proportions of asparagine than the other genomes as predicted by their genetic codes, but the vertebrate and invertebrate genomes do not show decreased levels of this amino acid as we predicted they would (Figure 2.7: lower row, right). These few exceptions to the general rule appear to be related to the pattern of nucleotide usage in the genomes. Our analyses assume that each of the 61 amino acid encoding codons are equally likely to occur in a gene, but this is not the case when the four nucleotides are not in equal proportions. It is known that mitochondrial genomes typically violate the assumption of equal base frequencies, and have strong biases in their nucleotide compositions (Asakawa et al. 1991; Reyes et al. 1998; Boore, 1999; Saccone et al. 1999). This detail is discussed in more detail in Chapter 5.

All mitochondria have a common ancestor. Although we cannot know what the ancestral mitochondrion's genetic code or amino acid proportions were, we can infer from the variety of amino acid compositions among modern mitochondria that directional protein evolution has occurred in the past. Moreover, we have shown that these directional changes are correlated to the modifications to the genetic code that have occurred among these genomes.

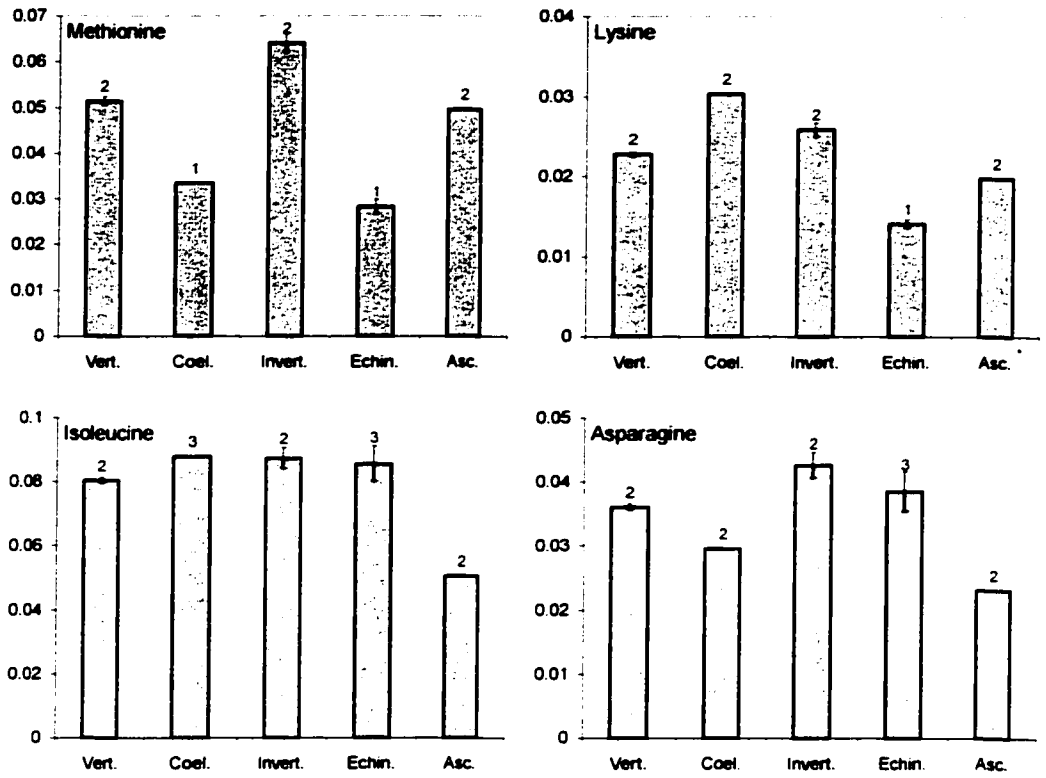


Figure 2.7: A single codon reassignment in the genetic code is sufficient to affect the amino acid proportions of encoded proteins. Symbols and abbreviations are as in Figure 2.6. (Top row, left) The coelenterate and echinoderm mitochondrial genetic codes have one less methionine codon than the other codes, resulting in a lower proportion of methionine in those genomes ($p < 6 \times 10^{-14}$ in two-sample t-tests). (Top row, right) The echinoderm mitochondrial genetic code as one less lysine codon than the other genetic codes, and it therefore has less lysine in its proteins ($p < 2 \times 10^{-10}$ in two-sample t-tests). (Bottom row, left) The ascidian mitochondrial genetic code has one less isoleucine codon than the other genetic codes, and it therefore the single ascidian genome has a lower proportion of isoleucine ($p < 2 \times 10^{-5}$ in a one-sample t-test). Note, however, that the vertebrate and invertebrate genomes do not follow their predicted trends. (Bottom row, right) The relationship between amino acid proportion and codon count is unclear for asparagine, although the echinoderms (with three codons) do show significantly higher proportions of asparagine then the coelenterate and ascidian genomes, which have only two asparagine codons ($p < 0.007$ in one-sample t-tests).

2.5 Discussion

To our knowledge, the number of codons assigned to each amino acid in the genetic code has never been viewed as a major force in protein evolution. However, our results have shown that within a single proteome, the proportion of each amino acid is correlated to the number of codons that are able to encode it (Figure 2.3 and Appendix 2.6.1). We have also shown that when there are modifications to the standard code, there is a corresponding shift in the amino acid usages in the proteins (Figure 2.6 and Figure 2.7). These findings are the logical conclusion of neutral evolution combined with a biased genetic code.

2.6 Appendices

Appendix 2.6.1: There is a positive relationship between the proportion of an amino acid and the number of codons that can encode it, in each of seventy bacterial proteomes. 28 of these relationships are significant at the $p<0.001$ level. The organisms are listed in order of decreasing slope. See Appendix 2.6.3 for accompanying figures.

Organism	Version	r^2	slope	p
Aeropyrum pernix	NC_000854.1	0.61	0.0153	5.03×10^{-5}
Deinococcus radiodurans	NC_001264.1	0.55	0.0153	1.88×10^{-4}
Mycobacterium tuberculosis CDC1551	NC_002755.1	0.53	0.0152	2.95×10^{-4}
Mycobacterium tuberculosis H37Rv	NC_000962.1	0.51	0.0151	3.81×10^{-4}
Pseudomonas aeruginosa	NC_002516.1	0.54	0.0146	2.17×10^{-4}
Mycobacterium leprae	NC_002677.1	0.56	0.0144	1.41×10^{-4}
Ralstonia solanacearum	NC_003296.1	0.51	0.0140	4.36×10^{-4}
Caulobacter vibrioides	NC_002696.2	0.46	0.0140	9.51×10^{-4}
Sinorhizobium meliloti	NC_003078.1	0.51	0.0134	3.76×10^{-4}
Mesorhizobium loti	NC_002682.1	0.51	0.0132	4.13×10^{-4}

Organism	Version	r^2	slope	p
Treponema pallidum	NC_000919.1	0.64	0.0131	2.55×10^{-5}
Xylella fastidiosa 9a5c	NC_002490.1	0.61	0.0128	4.86×10^{-5}
Agrobacterium tumefaciens str. C58 (U. Washington)	NC_003308.1	0.52	0.0127	3.65×10^{-4}
Agrobacterium tumefaciens str. C58 (Cereon)	NC_003065.1	0.52	0.0127	3.48×10^{-4}
Halobacterium sp. NRC-1	NC_002608.1	0.37	0.0126	4.40×10^{-3}
Brucella melitensis	NC_003318.1	0.50	0.0123	5.15×10^{-4}
Pyrobaculum aerophilum	NC_003364.1	0.45	0.0122	1.20×10^{-3}
Chlamydia trachomatis	NC_000117.1	0.54	0.0114	2.39×10^{-4}
Chlamydia muridarum	NC_002620.1	0.53	0.0112	2.95×10^{-4}
Yersinia pestis	NC_003143.1	0.52	0.0111	3.05×10^{-4}
Salmonella typhimurium LT2	NC_003277.1	0.51	0.0111	3.76×10^{-4}
Salmonella enterica enterica serovar Typhi	NC_003385.1	0.52	0.0110	3.33×10^{-4}
Synechocystis sp. PCC 6803	NC_000911.1	0.49	0.0110	5.97×10^{-4}
Chlamydophila pneumoniae J138	NC_002491.1	0.51	0.0110	4.15×10^{-4}
Chlamydophila pneumoniae CWL029	NC_000922.1	0.51	0.0110	4.42×10^{-4}
Methanothermobacter thermautotrophicus	NC_000916.1	0.45	0.0109	1.12×10^{-3}
Chlamydophila pneumoniae AR39	NC_002179.1	0.50	0.0109	4.46×10^{-4}
Escherichia coli K12	NC_000913.1	0.51	0.0108	4.45×10^{-4}
Escherichia coli O157:H7	NC_002695.1	0.52	0.0108	3.51×10^{-4}
Escherichia coli O157:H7 EDL933	NC_002655.2	0.51	0.0108	3.77×10^{-4}
Nostoc sp. PCC 7120	NC_003276.1	0.49	0.0108	5.83×10^{-4}

Organism	Version	r^2	slope	p
Vibrio cholerae	NC_002506.1	0.46	0.0103	1.02×10^{-3}
Neisseria meningitidis MC58	NC_003112.1	0.43	0.0102	1.71×10^{-3}
Neisseria meningitidis Z2491	NC_003116.1	0.42	0.0102	1.89×10^{-3}
Thermoplasma acidophilum	NC_002578.1	0.39	0.0095	3.22×10^{-3}
Archaeoglobus fulgidus	NC_000917.1	0.31	0.0094	1.07×10^{-2}
Pyrococcus abyssi	NC_000868.1	0.28	0.0093	1.76×10^{-2}
Sulfolobus solfataricus	NC_002754.1	0.30	0.0092	1.28×10^{-2}
Pasteurella multocida	NC_002663.1	0.38	0.0092	3.94×10^{-3}
Pyrococcus horikoshii	NC_000961.1	0.27	0.0092	1.80×10^{-2}
Thermotoga maritima	NC_000853.1	0.29	0.0090	1.52×10^{-2}
Haemophilus influenzae Rd	NC_000907.1	0.35	0.0089	5.65×10^{-3}
Thermoplasma volcanium	NC_002689.2	0.32	0.0089	9.49×10^{-3}
Pyrococcus furiosus DSM 3638	NC_003413.1	0.25	0.0088	2.46×10^{-2}
Bacillus halodurans	NC_002570.1	0.35	0.0088	5.93×10^{-3}
Sulfolobus tokodaii	NC_003106.1	0.26	0.0087	2.28×10^{-2}
Streptococcus pneumoniae TIGR4	NC_003028.1	0.31	0.0086	1.04×10^{-2}
Streptococcus pyogenes M1 GAS	NC_002737.1	0.31	0.0085	1.11×10^{-2}
Streptococcus pneumoniae R6	NC_003098.1	0.30	0.0084	1.26×10^{-2}
Bacillus subtilis	NC_000964.1	0.31	0.0083	1.05×10^{-2}
Lactococcus lactis lactis	NC_002662.1	0.27	0.0081	1.95×10^{-2}
Aquifex aeolicus	NC_001880.1	0.19	0.0080	5.29×10^{-2}
Helicobacter pylori J99	NC_000921.1	0.22	0.0077	3.52×10^{-2}
Listeria monocytogenes EGD-e	NC_003210.1	0.25	0.0077	2.58×10^{-2}

Organism	Version	r^2	slope	p
Helicobacter pylori 26695	NC_000915.1	0.22	0.0077	3.75×10^{-2}
Listeria innocua	NC_003383.1	0.24	0.0076	2.87×10^{-2}
Mycoplasma pneumoniae	NC_000912.1	0.23	0.0076	3.07×10^{-2}
Rickettsia conorii	NC_003103.1	0.20	0.0074	4.92×10^{-2}
Rickettsia prowazekii	NC_000963.1	0.18	0.0073	6.33×10^{-2}
Buchnera sp. APS	NC_002528.1	0.13	0.0067	1.12×10^{-1}
Mycoplasma genitalium	NC_000908.1	0.15	0.0067	9.42×10^{-2}
Staphylococcus aureus aureus N315	NC_003140.1	0.20	0.0066	4.83×10^{-2}
Staphylococcus aureus aureus Mu50	NC_002774.1	0.20	0.0066	4.94×10^{-2}
Campylobacter jejuni	NC_002163.1	0.13	0.0064	1.14×10^{-1}
Clostridium acetobutylicum	NC_003030.1	0.13	0.0061	1.15×10^{-1}
Clostridium perfringens	NC_003366.1	0.12	0.0060	1.38×10^{-1}
Borrelia burgdorferi	NC_001904.1	0.09	0.0059	1.96×10^{-1}
Methanococcus jannaschii	NC_001733.1	0.08	0.0054	2.18×10^{-1}
Mycoplasma pulmonis	NC_002771.1	0.08	0.0054	2.27×10^{-1}
Ureaplasma urealyticum	NC_002162.1	0.06	0.0046	2.93×10^{-1}

Appendix 2.6.2: The mitochondrial genomes analysed in this study and the genetic codes they employ.

Organism	Version	Genetic code
Halocynthia roretzi	NC_002177.1	Ascidian
Metridium senile	NC_000933.1	Coelenterate
Arbacia lixula	NC_001770.1	Echinoderm
Asterina pectinifera	NC_001627.1	Echinoderm
Balanoglossus carnosus	NC_001887.1	Echinoderm
Echinococcus multilocularis	NC_000928.2	Echinoderm
Fasciola hepatica	NC_002546.1	Echinoderm
Florometra serratissima	NC_001878.1	Echinoderm

Organism	Version	Genetic code
Hymenolepis diminuta	NC_002767.1	Echinoderm
Paracentrotus lividus	NC_001572.1	Echinoderm
Paragonimus westermani	NC_002354.1	Echinoderm
Schistosoma japonicum	NC_002544.1	Echinoderm
Schistosoma mansoni	NC_002545.1	Echinoderm
Schistosoma mekongi	NC_002529.1	Echinoderm
Strongylocentrotus purpuratus	NC_001453.1	Echinoderm
Taenia crassiceps	NC_002547.1	Echinoderm
Albinaria caerulea	NC_001761.1	Invertebrate
Anopheles gambiae	NC_002084.1	Invertebrate
Anopheles quadrimaculatus A	NC_000875.1	Invertebrate
Apis mellifera ligustica	NC_001566.1	Invertebrate
Artemia franciscana	NC_001620.1	Invertebrate
Ascaris suum	NC_001327.1	Invertebrate
Bombyx mori	NC_002355.1	Invertebrate
Branchiostoma floridae	NC_000834.1	Invertebrate
Branchiostoma lanceolatum	NC_001912.1	Invertebrate
Caenorhabditis elegans	NC_001328.1	Invertebrate
Cepaea nemoralis	NC_001816.1	Invertebrate
Ceratitis capitata	NC_000857.1	Invertebrate
Chrysomya chloropyga	NC_002697.1	Invertebrate
Cochliomyia hominivorax	NC_002660.1	Invertebrate
Crassostrea gigas	NC_001276.1	Invertebrate
Crioceris duodecimpunctata	NC_003372.1	Invertebrate
Daphnia pulex	NC_000844.1	Invertebrate
Drosophila melanogaster	NC_001709.1	Invertebrate
Drosophila yakuba	NC_001322.1	Invertebrate
Heterodoxus macropus	NC_002651.1	Invertebrate
Ixodes hexagonus	NC_002010.1	Invertebrate
Katharina tunicata	NC_001636.1	Invertebrate
Laqueus rubellus	NC_002322.1	Invertebrate
Limulus polyphemus	NC_003057.1	Invertebrate
Lithobius forficatus	NC_002629.1	Invertebrate
Locusta migratoria	NC_001712.1	Invertebrate
Loligo bleekeri	NC_002507.1	Invertebrate
Lumbricus terrestris	NC_001673.1	Invertebrate
Narceus annularus	NC_003343.1	Invertebrate
Onchocerca volvulus	NC_001861.1	Invertebrate
Ostrinia furnacalis	NC_003368.1	Invertebrate
Ostrinia nubilalis	NC_003367.1	Invertebrate
Pagurus longicarpus	NC_003058.1	Invertebrate
Penaeus monodon	NC_002184.1	Invertebrate
Platynereis dumerilii	NC_000931.1	Invertebrate
Pupa strigosa	NC_002176.1	Invertebrate
Rhipicephalus sanguineus	NC_002074.1	Invertebrate
Terebratalia transversa	NC_003086.1	Invertebrate
Terebratulina retusa	NC_000941.1	Invertebrate
Tetrodontophora bielanensis	NC_002735.1	Invertebrate
Thyropygus sp. DVL-2001	NC_003344.1	Invertebrate
Triatoma dimidiata	NC_002609.1	Invertebrate
Tribolium castaneum	NC_003081.1	Invertebrate

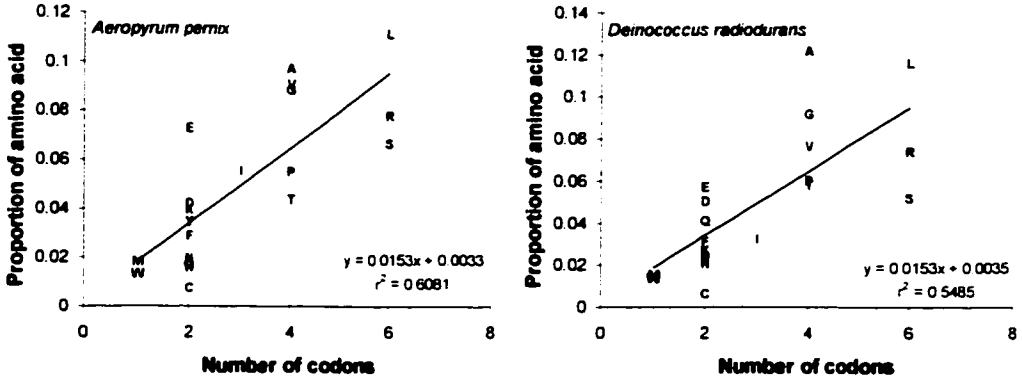
Organism	Version	Genetic code
Trichinella spiralis	NC_002681.1	Invertebrate
Alligator mississippiensis	NC_001922.1	Vertebrate
Anguilla japonica	NC_002707.1	Vertebrate
Anomalopteryx didiformis	NC_002779.1	Vertebrate
Antigonia capros	NC_003191.1	Vertebrate
Apteryx haastii	NC_002782.1	Vertebrate
Arctoscopus japonicus	NC_002812.1	Vertebrate
Artibeus jamaicensis	NC_002009.1	Vertebrate
Ateleopus japonicus	NC_003178.1	Vertebrate
Aulopus japonicus	NC_002674.1	Vertebrate
Aythya americana	NC_000877.1	Vertebrate
Balaenoptera musculus	NC_001601.1	Vertebrate
Balaenoptera physalus	NC_001321.1	Vertebrate
Beryx splendens	NC_003188.1	Vertebrate
Bos taurus	NC_001567.1	Vertebrate
Buteo buteo	NC_003128.3	Vertebrate
Caelorinchus kishinouyei	NC_003169.1	Vertebrate
Caiman crocodilus	NC_002744.2	Vertebrate
Canis familiaris	NC_002008.4	Vertebrate
Carassius auratus	NC_002079.1	Vertebrate
Casuarius casuarius	NC_002778.1	Vertebrate
Cavia porcellus	NC_000884.1	Vertebrate
Cebus albifrons	NC_002763.1	Vertebrate
Ceratotherium simum	NC_001808.1	Vertebrate
Chalinolobus tuberculatus	NC_002626.1	Vertebrate
Chauliodus sloani	NC_003159.1	Vertebrate
Chelonia mydas	NC_000886.1	Vertebrate
Chimaera monstrosa	NC_003136.1	Vertebrate
Chlorophthalmus agassizi	NC_003160.1	Vertebrate
Chrysemys picta	NC_002073.3	Vertebrate
Ciconia boyciana	NC_002196.1	Vertebrate
Ciconia ciconia	NC_002197.1	Vertebrate
Cololabis saira	NC_003183.1	Vertebrate
Conger myriaster	NC_002761.1	Vertebrate
Coregonus lavaretus	NC_002646.1	Vertebrate
Corvus frugilegus	NC_002069.1	Vertebrate
Crenimugil crenilabis	NC_003170.1	Vertebrate
Crossostoma lacustre	NC_001727.1	Vertebrate
Cyprinus carpio	NC_001606.1	Vertebrate
Dactyloptena peterseni	NC_003194.1	Vertebrate
Danaceticthys galathenus	NC_003185.1	Vertebrate
Danio rerio	NC_002333.2	Vertebrate
Dasypus novemcinctus	NC_001821.1	Vertebrate
Diaphus splendidus	NC_003164.1	Vertebrate
Didelphis virginiana	NC_001610.1	Vertebrate
Dinodon semicarinatus	NC_001945.1	Vertebrate
Dinornis giganteus	NC_002672.1	Vertebrate
Diplophos taenia	NC_002647.1	Vertebrate
Dogania subplana	NC_002780.1	Vertebrate
Dromaius novaehollandiae	NC_002784.1	Vertebrate
Dugong dugon	NC_003314.1	Vertebrate

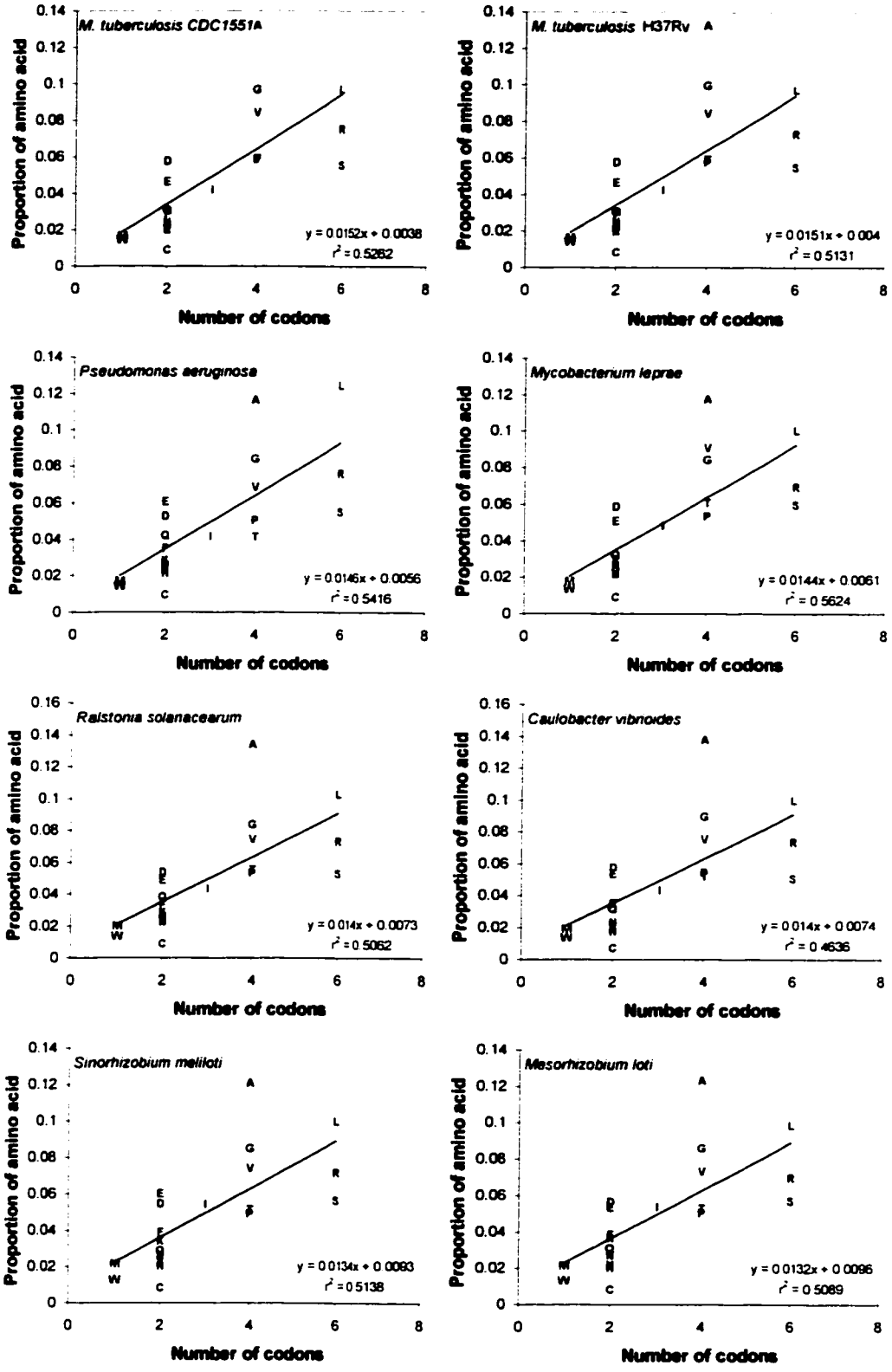
Organism	Version	Genetic code
Echinops telfairi	NC_002631.1	Vertebrate
Echinosorex gymnura	NC_002808.1	Vertebrate
Elassoma evergladei	NC_003175.1	Vertebrate
Emeus crassus	NC_002673.1	Vertebrate
Engraulis japonicus	NC_003097.1	Vertebrate
Eptatretus burgeri	NC_002807.1	Vertebrate
Equus asinus	NC_001788.1	Vertebrate
Equus caballus	NC_001640.1	Vertebrate
Erinaceus europaeus	NC_002080.1	Vertebrate
Eudromia elegans	NC_002772.1	Vertebrate
Eumeces egregius	NC_000888.1	Vertebrate
Exocoetus volitans	NC_003184.1	Vertebrate
Falco peregrinus	NC_000878.1	Vertebrate
Felis catus	NC_001700.1	Vertebrate
Gadus morhua	NC_002081.1	Vertebrate
Gallus gallus	NC_001323.1	Vertebrate
Gasterosteus aculeatus	NC_003174.1	Vertebrate
Gonostoma gracile	NC_002574.1	Vertebrate
Gorilla gorilla	NC_001645.1	Vertebrate
Halichoerus grypus	NC_001602.1	Vertebrate
Harpadon microchir	NC_003161.1	Vertebrate
Helicolenus hilgendorfi	NC_003195.1	Vertebrate
Heterodontus francisci	NC_003137.1	Vertebrate
Hippopotamus amphibius	NC_000889.1	Vertebrate
Homo sapiens	NC_001807.4	Vertebrate
Hoplostethus japonicus	NC_003187.1	Vertebrate
Hylobates lar	NC_002082.1	Vertebrate
Iguana iguana	NC_002793.1	Vertebrate
Ijimaia dofleini	NC_003179.1	Vertebrate
Isoodon macrourus	NC_002746.1	Vertebrate
Lama pacos	NC_002504.1	Vertebrate
Lampetra fluviatilis	NC_001131.1	Vertebrate
Lampris guttatus	NC_003165.1	Vertebrate
Latimeria chalumnae	NC_001804.1	Vertebrate
Lepidosiren paradoxa	NC_003342.1	Vertebrate
Loxodonta africana	NC_000934.1	Vertebrate
Macaca sylvanus	NC_002764.1	Vertebrate
Macropus robustus	NC_001794.1	Vertebrate
Mastacembelus favus	NC_003193.1	Vertebrate
Mertensiella luschani	NC_002756.1	Vertebrate
Monopterus albus	NC_003192.1	Vertebrate
Mugil cephalus	NC_003182.1	Vertebrate
Mus musculus	NC_001569.1	Vertebrate
Mustelus manazo	NC_000890.1	Vertebrate
Myctophum affine	NC_003163.1	Vertebrate
Myoxus glis	NC_001892.1	Vertebrate
Myripristis berndti	NC_003189.1	Vertebrate
Myxine glutinosa	NC_002639.1	Vertebrate
Neoceratodus forsteri	NC_003127.1	Vertebrate
Neoscopelus microchir	NC_003180.1	Vertebrate
Nycticebus coucang	NC_002765.1	Vertebrate

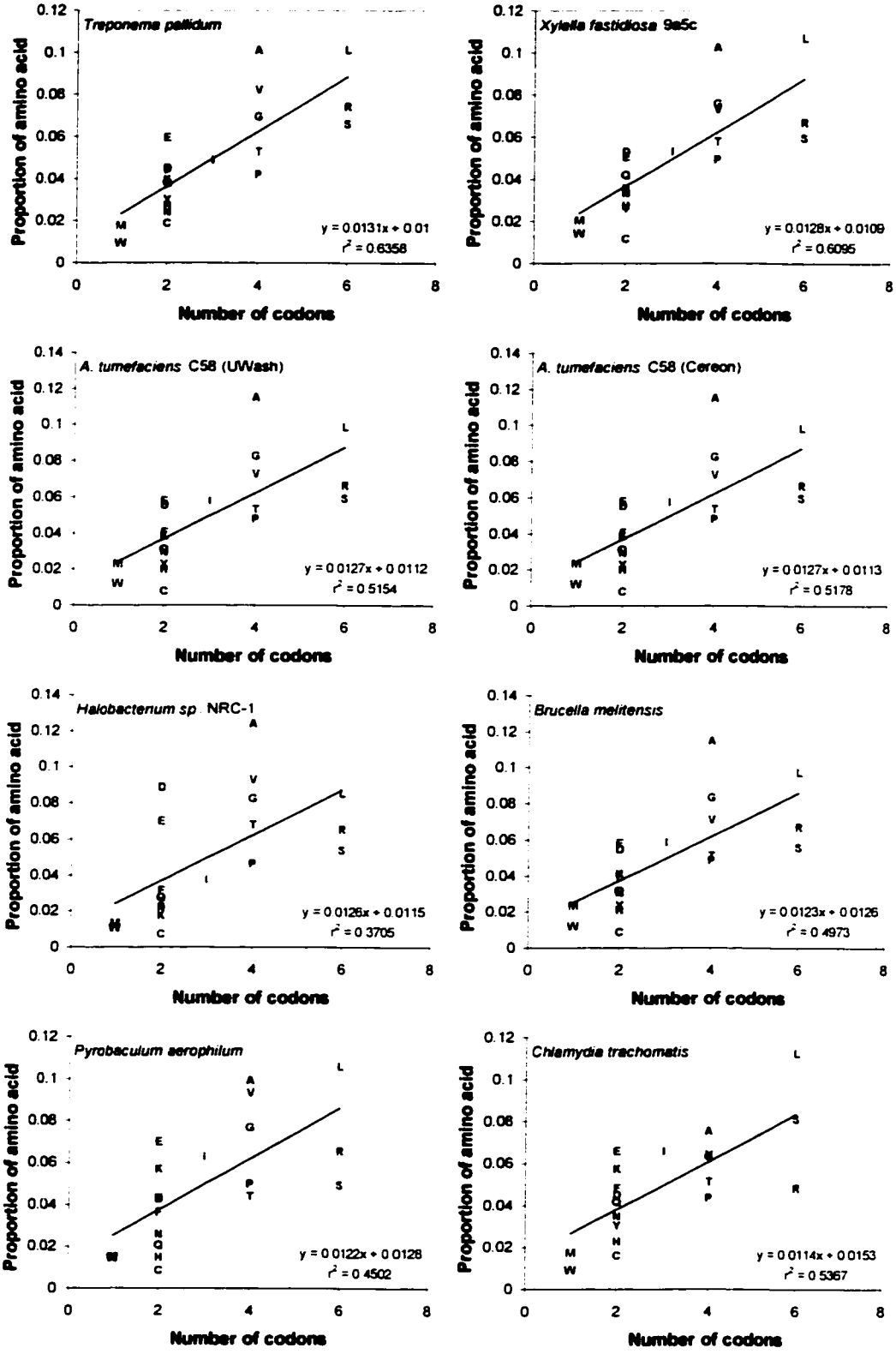
Organism	Version	Genetic code
Ochotona collaris	NC_003033.1	Vertebrate
Oncorhynchus mykiss	NC_001717.1	Vertebrate
Oncorhynchus tshawytscha	NC_002980.1	Vertebrate
Ornithorhynchus anatinus	NC_000891.1	Vertebrate
Orycteropus afer	NC_002078.1	Vertebrate
Oryctolagus cuniculus	NC_001913.1	Vertebrate
Osteoglossum bicirrhosum	NC_003095.1	Vertebrate
Ovis aries	NC_001941.1	Vertebrate
Pagrus major	NC_003196.1	Vertebrate
Pan paniscus	NC_001644.1	Vertebrate
Pan troglodytes	NC_001643.1	Vertebrate
Pantodon buchholzi	NC_003096.1	Vertebrate
Papio hamadryas	NC_001992.1	Vertebrate
Paralichthys olivaceus	NC_002386.1	Vertebrate
Pelomedusa subrufa	NC_001947.1	Vertebrate
Percopsis transmontana	NC_003168.1	Vertebrate
Petromyzon marinus	NC_001626.1	Vertebrate
Phoca vitulina	NC_001325.1	Vertebrate
Physeter catodon	NC_002503.1	Vertebrate
Platichthys bicoloratus	NC_003176.1	Vertebrate
Plecoglossus altivelis	NC_002734.1	Vertebrate
Polymixia japonica	NC_002648.1	Vertebrate
Polymixia lowei	NC_003181.1	Vertebrate
Polypterus ornatipinnis	NC_001778.1	Vertebrate
Pongo pygmaeus	NC_001646.1	Vertebrate
Pongo pygmaeus abelii	NC_002083.1	Vertebrate
Poromitra oscitans	NC_003172.1	Vertebrate
Protopterus dolloi	NC_001708.1	Vertebrate
Pterocnemia pennata	NC_002783.1	Vertebrate
Pteropus dasymallus	NC_002612.1	Vertebrate
Pteropus scapulatus	NC_002619.1	Vertebrate
Raja radiata	NC_000893.1	Vertebrate
Rana nigromaculata	NC_002805.1	Vertebrate
Rattus norvegicus	NC_001665.1	Vertebrate
Rhea americana	NC_000846.1	Vertebrate
Rhinoceros unicornis	NC_001779.1	Vertebrate
Rivulus marmoratus	NC_003290.1	Vertebrate
Rondeletia loricata	NC_003186.1	Vertebrate
Salmo salar	NC_001960.1	Vertebrate
Salvelinus alpinus	NC_000861.1	Vertebrate
Salvelinus fontinalis	NC_000860.1	Vertebrate
Sardinops melanostictus	NC_002616.1	Vertebrate
Saurida undosquamis	NC_003162.1	Vertebrate
Sciurus vulgaris	NC_002369.1	Vertebrate
Scopelogadus mizolepis	NC_003171.1	Vertebrate
Scyliorhinus canicula	NC_001950.1	Vertebrate
Smithornis sharpei	NC_000879.1	Vertebrate
Soriculus fumidus	NC_003040.1	Vertebrate
Squalus acanthias	NC_002012.1	Vertebrate
Stephanolepis cirrhifer	NC_003177.1	Vertebrate
Struthio camelus	NC_002785.1	Vertebrate

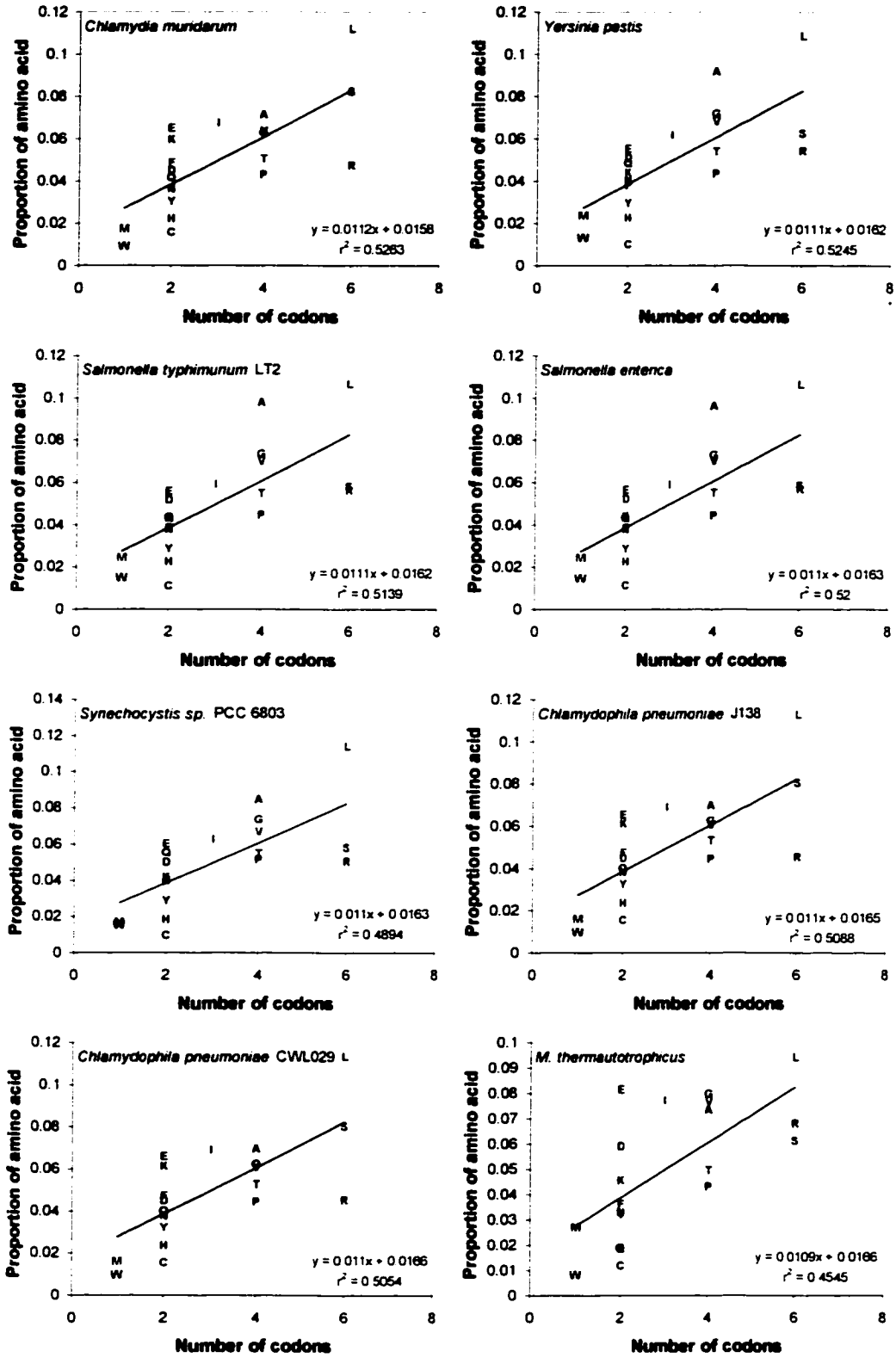
Organism	Version	Genetic code
Sus scrofa	NC_000845.1	Vertebrate
Tachyglossus aculeatus	NC_003321.1	Vertebrate
Talpa europaea	NC_002391.1	Vertebrate
Tarsius bancanus	NC_002811.1	Vertebrate
Thryonomys swinderianus	NC_002658.1	Vertebrate
Tinamus major	NC_002781.1	Vertebrate
Trachipterus trachypterus	NC_003166.1	Vertebrate
Trachurus japonicus	NC_002813.1	Vertebrate
Trichosurus vulpecula	NC_003039.1	Vertebrate
Tupaia belangeri	NC_002521.1	Vertebrate
Typhlonectes natans	NC_002471.1	Vertebrate
Vidua chalybeata	NC_000880.1	Vertebrate
Volemys kikuchii	NC_003041.1	Vertebrate
Vombatus ursinus	NC_003322.1	Vertebrate
Xenopus laevis	NC_001573.1	Vertebrate
Zenopsis nebulosus	NC_003173.1	Vertebrate
Zeus faber	NC_003190.1	Vertebrate
Zu cristatus	NC_003167.1	Vertebrate

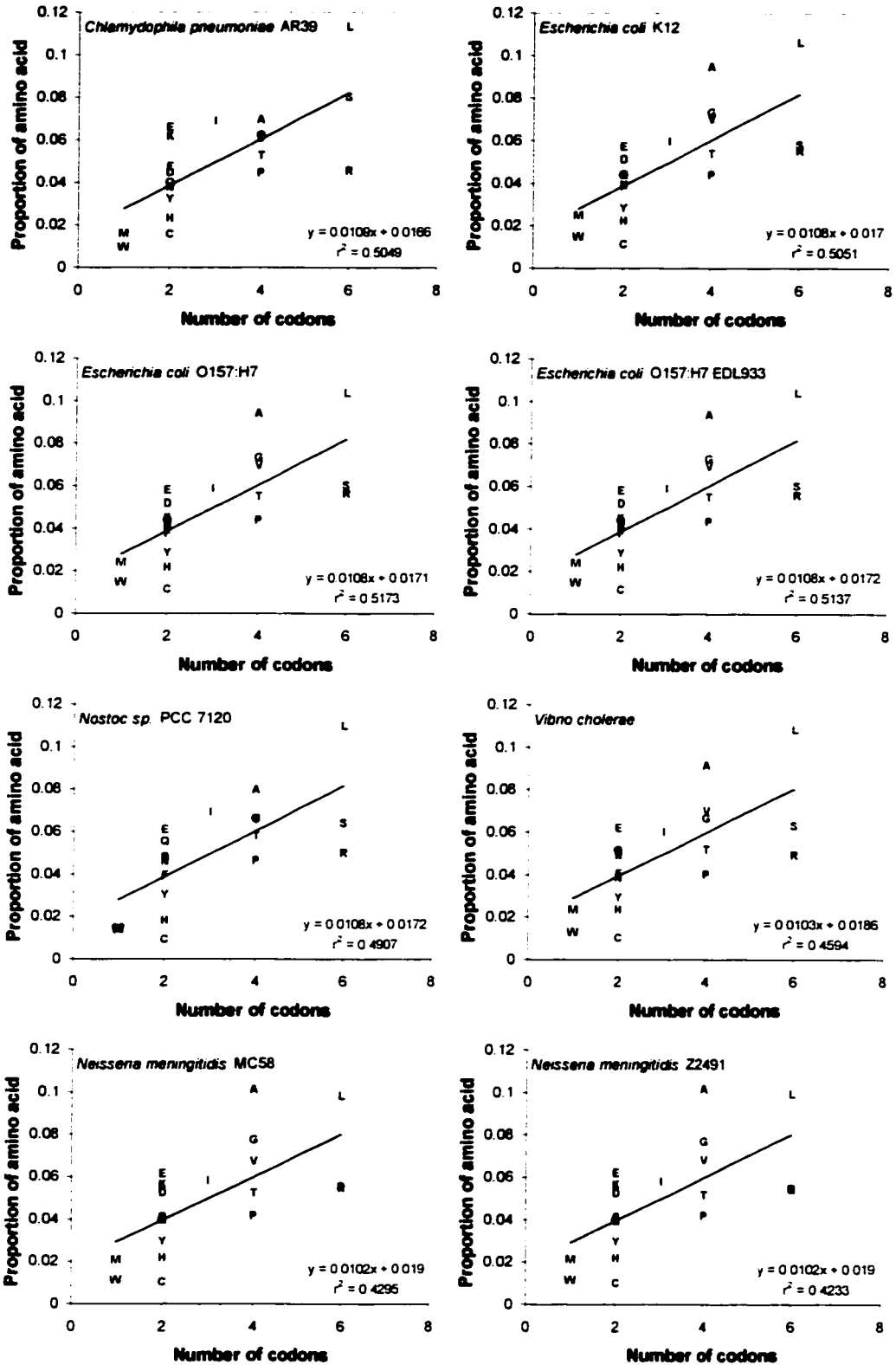
Appendix 2.6.3: Here we show the relationship between amino acid content and the number of codons for each amino acid in seventy complete bacterial proteomes. The organisms are listed in the same order as they appear in Appendix 2.6.1. from the steepest slope of the regression line to the shallowest.

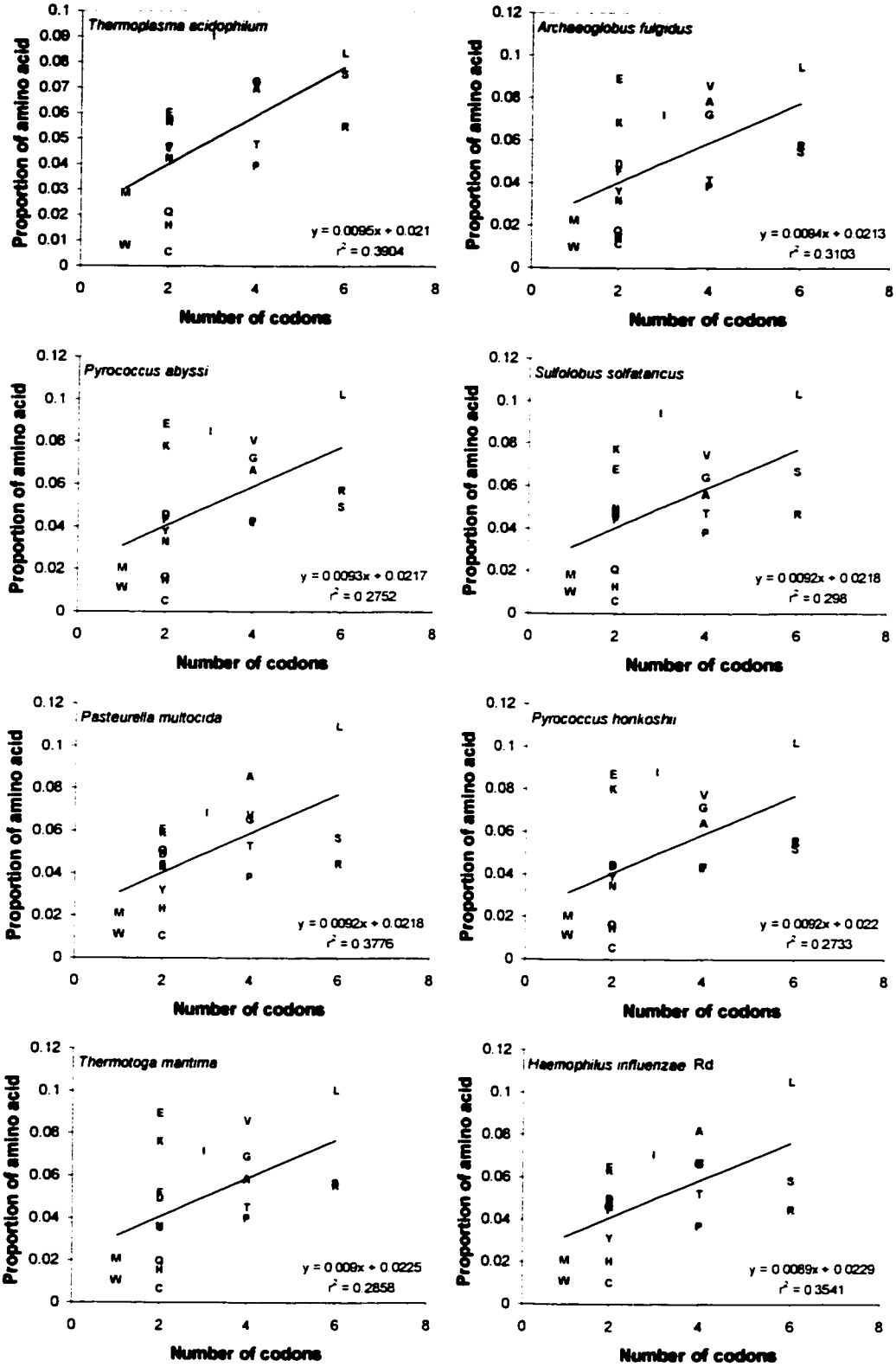


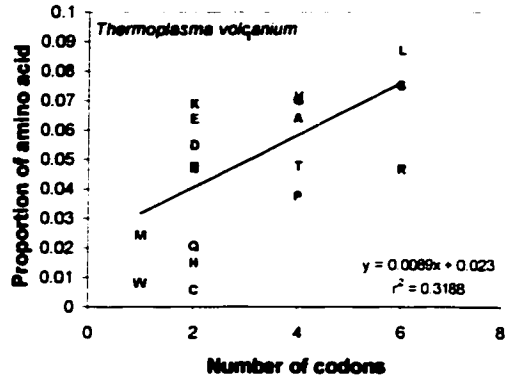

M. tuberculosis CDC1551A
y = 0.0152x + 0.0038
r² = 0.5282
M. tuberculosis H37Rv
y = 0.0151x + 0.004
r² = 0.5131
Pseudomonas aeruginosa
y = 0.0146x + 0.0056
r² = 0.5416
Mycobacterium leprae
y = 0.0144x + 0.0061
r² = 0.5624
Ralstonia solanacearum
y = 0.014x + 0.0073
r² = 0.5062
Caulobacter vibrioides
y = 0.014x + 0.0074
r² = 0.4636
Sinorhizobium meliloti
y = 0.0134x + 0.0093
r² = 0.5138
Mesorhizobium loti
y = 0.0132x + 0.0096
r² = 0.5089
Proportion of amino acid
Number of codons

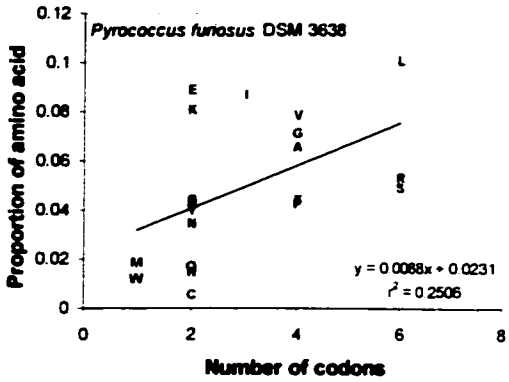

Treponema pallidum
y = 0.0131x + 0.01
r² = 0.6358
Xylella fastidiosa 9a5c
y = 0.0128x + 0.0109
r² = 0.6095
A. tumefaciens C58 (UWash)
y = 0.0127x + 0.0112
r² = 0.5154
A. tumefaciens C58 (Cereon)
y = 0.0127x + 0.0113
r² = 0.5178
Halobacterium sp. NRC-1
y = 0.0126x + 0.0115
r² = 0.3705
Brucella melitensis
y = 0.0123x + 0.0126
r² = 0.4973
Pyrobaculum aerophilum
y = 0.0122x + 0.0128
r² = 0.4502
Chlamydia trachomatis
y = 0.0114x + 0.0153
r² = 0.5367
Proportion of amino acid
Number of codons

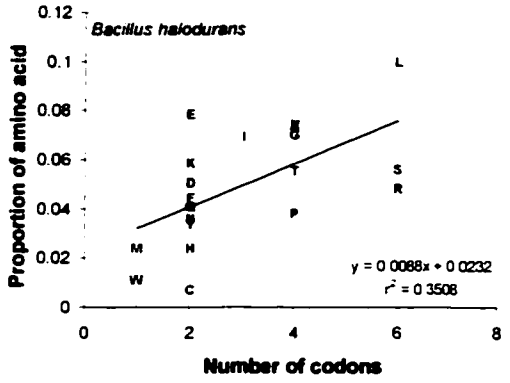

Chlamydia muridarum
y = 0.0112x + 0.0158
r² = 0.5263
Yersinia pestis
y = 0.0111x + 0.0162
r² = 0.5245
Salmonella typhimurium LT2
y = 0.0111x + 0.0162
r² = 0.5139
Salmonella enterica
y = 0.011x + 0.0163
r² = 0.52
Synechocystis sp. PCC 6803
y = 0.011x + 0.0163
r² = 0.4894
Chlamydophila pneumoniae J138
y = 0.011x + 0.0165
r² = 0.5088
Chlamydophila pneumoniae CWL029
y = 0.011x + 0.0166
r² = 0.5054
M. thermautotrophicus
y = 0.0109x + 0.0166
r² = 0.4545
Proportion of amino acid
Number of codons

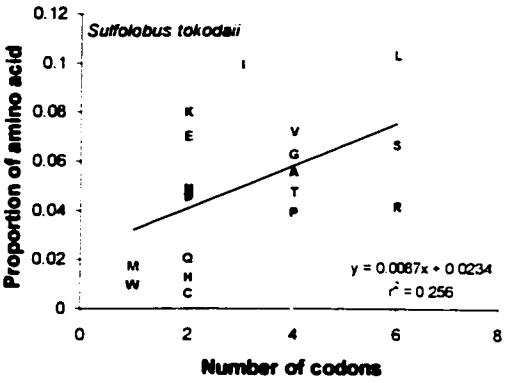

Chlamydophila pneumoniae AR39
y = 0.0109x + 0.0166
r² = 0.5049
Escherichia coli K12
y = 0.0108x + 0.017
r² = 0.5051
Escherichia coli O157:H7
y = 0.0108x + 0.0171
r² = 0.5173
Escherichia coli O157:H7 EDL933
y = 0.0108x + 0.0172
r² = 0.5137
Nostoc sp. PCC 7120
y = 0.0108x + 0.0172
r² = 0.4907
Vibrio cholerae
y = 0.0103x + 0.0186
r² = 0.4594
Neisseria meningitidis MC58
y = 0.0102x + 0.019
r² = 0.4296
Neisseria meningitidis Z2491
y = 0.0102x + 0.019
r² = 0.4233
Proportion of amino acid
Number of codons

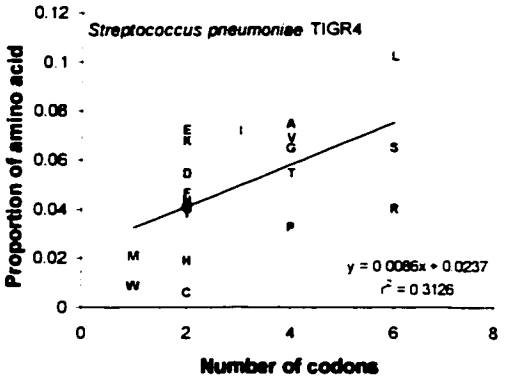

Thermoplasma acidophilum
y = 0.0095x + 0.021
r² = 0.3904
Archaeoglobus fulgidus
y = 0.0084x + 0.0213
r² = 0.3103
Pyrococcus abyssi
y = 0.0093x + 0.0217
r² = 0.2752
Sulfolobus solfataricus
y = 0.0092x + 0.0218
r² = 0.298
Pasteurella multocida
y = 0.0092x + 0.0218
r² = 0.3776
Pyrococcus horikoshii
y = 0.0092x + 0.022
r² = 0.2733
Thermotoga maritima
y = 0.009x + 0.0225
r² = 0.2858
Haemophilus influenzae Rd
y = 0.0089x + 0.0229
r² = 0.3541
Proportion of amino acid
Number of codons

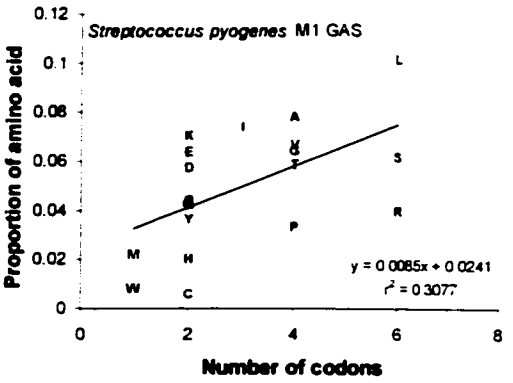

Thermoplasma volcanium
Proportion of amino acid
Number of codons
y = 0.0089x + 0.023
r² = 0.3188

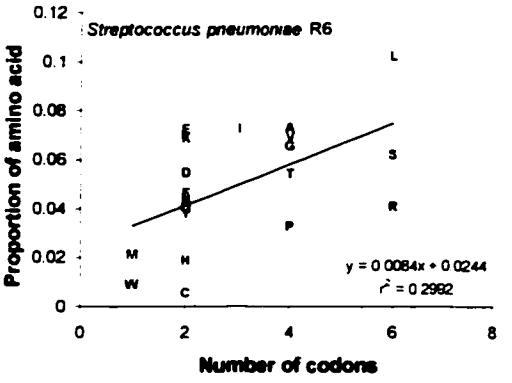

Pyrococcus furiosus DSM 3638
Proportion of amino acid
Number of codons
y = 0.0088x + 0.0231
r² = 0.2506

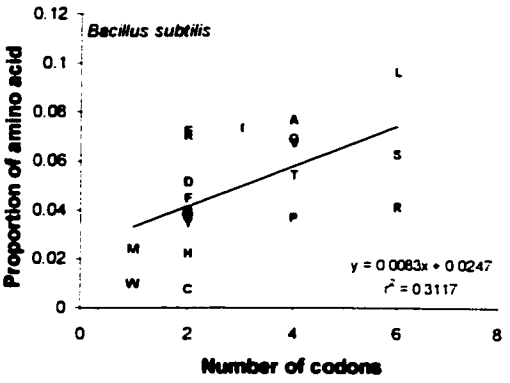

Bacillus halodurans
Proportion of amino acid
Number of codons
y = 0.0088x + 0.0232
r² = 0.3508

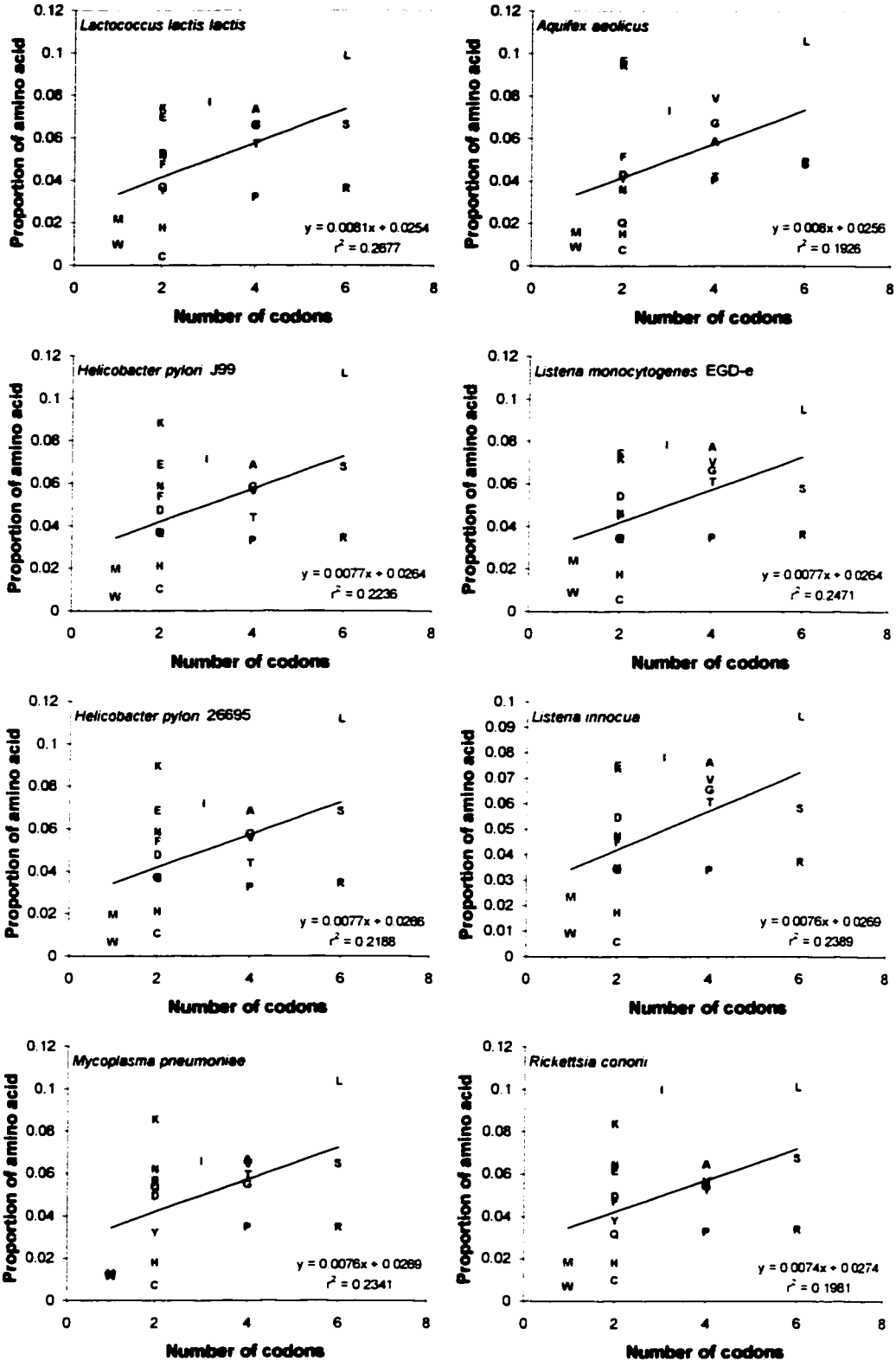

Sulfolobus tokodaii
Proportion of amino acid
Number of codons
y = 0.0087x + 0.0234
r² = 0.256

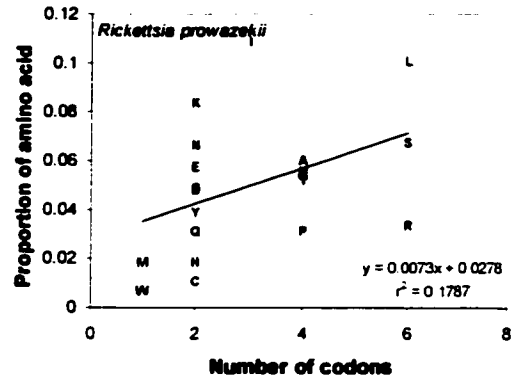

Streptococcus pneumoniae TIGR4
Proportion of amino acid
Number of codons
y = 0.0086x + 0.0237
r² = 0.3126

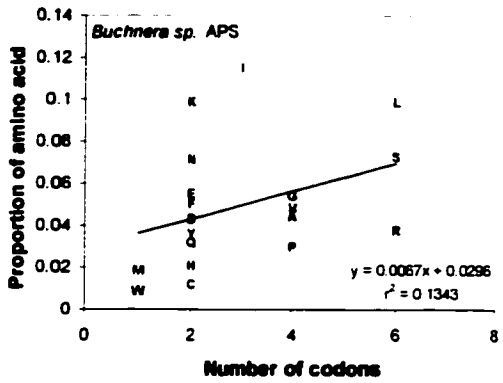

Streptococcus pyogenes M1 GAS
Proportion of amino acid
Number of codons
y = 0.0085x + 0.0241
r² = 0.3077

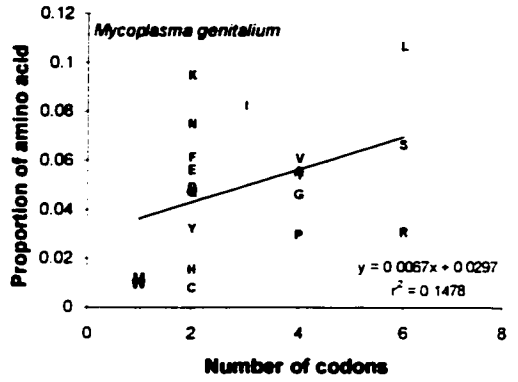

Streptococcus pneumoniae R6
Proportion of amino acid
Number of codons
y = 0.0084x + 0.0244
r² = 0.2992

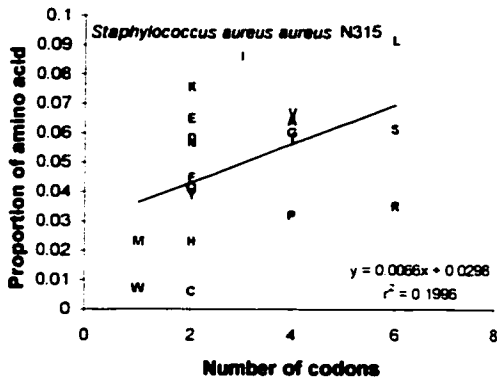

Bacillus subtilis
Proportion of amino acid
Number of codons
y = 0.0083x + 0.0247
r² = 0.3117

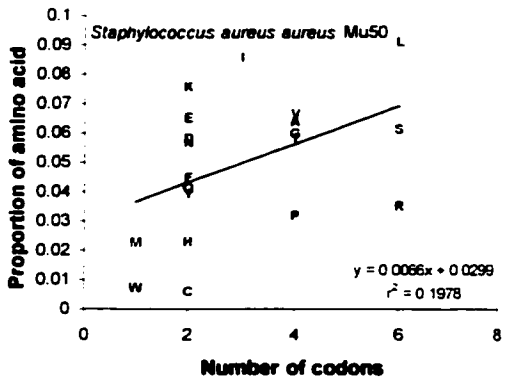

Lactococcus lactis lactis
y = 0.0081x + 0.0254
r² = 0.2677
Aquifex aeolicus
y = 0.008x + 0.0256
r² = 0.1926
Helicobacter pylori J99
y = 0.0077x + 0.0264
r² = 0.2236
Listeria monocytogenes EGD-e
y = 0.0077x + 0.0264
r² = 0.2471
Helicobacter pylori 26695
y = 0.0077x + 0.0266
r² = 0.2188
Listeria innocua
y = 0.0076x + 0.0269
r² = 0.2389
Mycoplasma pneumoniae
y = 0.0076x + 0.0269
r² = 0.2341
Rickettsia conorii
y = 0.0074x + 0.0274
r² = 0.1981
Proportion of amino acid
Number of codons

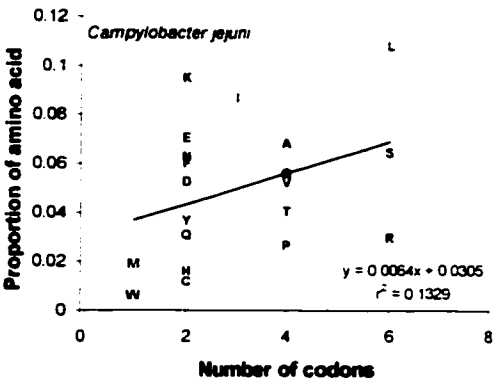

Rickettsia prowazekii
Proportion of amino acid
Number of codons
y = 0.0073x + 0.0278
r² = 0.1787

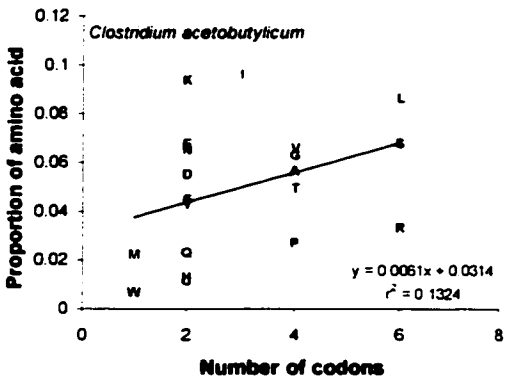

Buchnera sp. APS
Proportion of amino acid
Number of codons
y = 0.0067x + 0.0296
r² = 0.1343

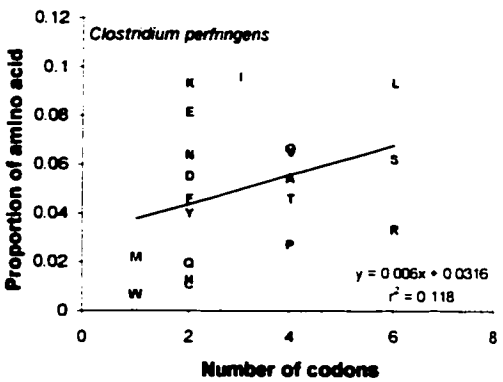

Mycoplasma genitalium
Proportion of amino acid
Number of codons
y = 0.0067x + 0.0297
r² = 0.1478

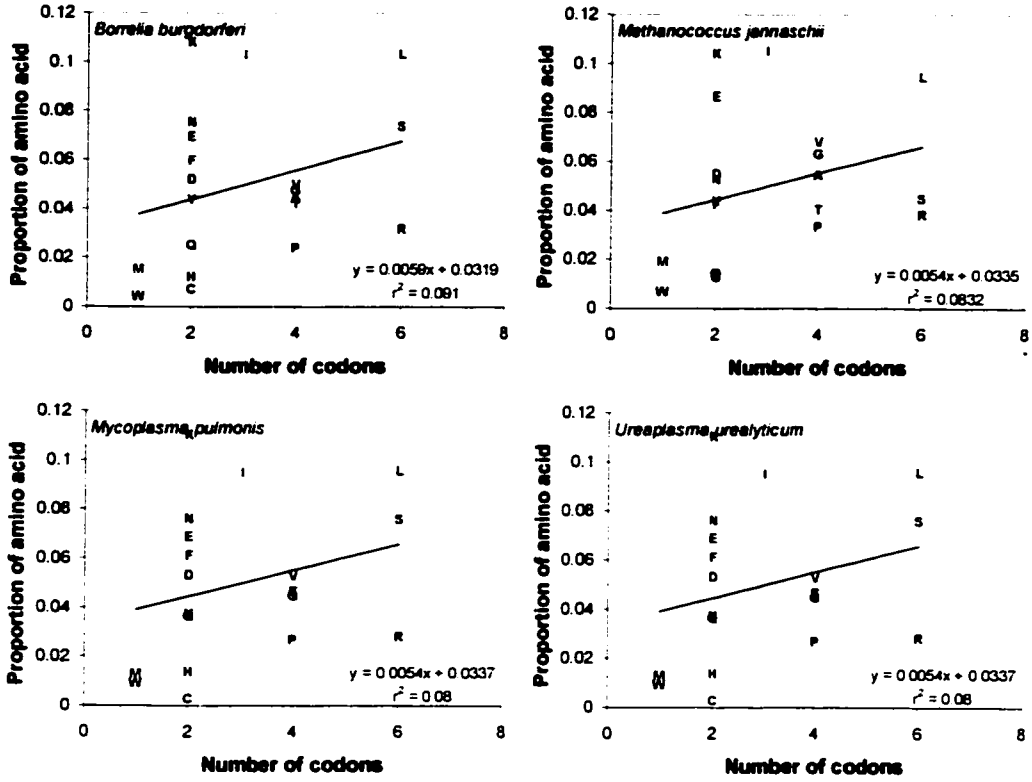

Staphylococcus aureus aureus N315
Proportion of amino acid
Number of codons
y = 0.0066x + 0.0298
r² = 0.1996


Staphylococcus aureus aureus Mu50
Proportion of amino acid
Number of codons
y = 0.0066x + 0.0299
r² = 0.1978


Campylobacter jejuni
Proportion of amino acid
Number of codons
y = 0.0064x + 0.0305
r² = 0.1329


Clostridium acetobutylicum
Proportion of amino acid
Number of codons
y = 0.0061x + 0.0314
r² = 0.1324


Clostridium perfringens
Proportion of amino acid
Number of codons
y = 0.006x + 0.0316
r² = 0.118


Borrelia burgdorferi
y = 0.0058x + 0.0319
r² = 0.091
Methanococcus jannaschii
y = 0.0054x + 0.0335
r² = 0.0832
Mycoplasma pulmonis
y = 0.0054x + 0.0337
r² = 0.08
Ureaplasma urealyticum
y = 0.0054x + 0.0337
r² = 0.08
Proportion of amino acid
Number of codons

3 Nucleotide bias causes a bias in the amino acid composition of all proteomes

3.1 Abstract

We have analysed the nucleotide contents from several completely sequenced bacterial genomes and we show that the pattern of nucleotide usage indicates a history of directional DNA evolution that appears to be independent of phylogenetic relationship. Moreover, this directional nucleotide bias appears to have had a dramatic effect on the amino acid composition of the encoded proteins. Based on the genetic code table, we predicted that AT-rich genes and genomes would have an increased frequency of certain amino acids (phenylalanine, tyrosine, methionine, isoleucine, asparagine, and lysine), and that homologous GC-rich genes would encode proteins with an increased proportion of a different subset of amino acids (glycine, alanine, arginine, and proline). By surveying the genes in twenty-one completely sequenced Eubacterial and Archaeal genomes, along with the entire *Saccharomyces cerevisiae* genome, we show that this prediction holds true. In short, we find that biased DNA encodes biased proteins.

3.2 Introduction

It is well known that both individual genes and entire genomes can vary significantly in their nucleotide composition (Muto and Osawa, 1987). Some organisms, for example, have genomes that are disproportionately rich in guanine and cytosine (G and C), while others have DNA that is rich in adenine and thymine (A and T). Variation in nucleotide composition is usually most pronounced at the synonymous codon positions of genes and, because of the redundancy in the genetic code, these variations in DNA content may have little effect on the amino acid content of the encoded proteins (Loomis and Smith, 1990; Lockhart et al., 1992). If, however, compositional bias at the DNA level affects both the synonymous and non-synonymous sites in protein-coding genes, then we expect proteins to change their amino acid composition over evolutionary time, and in a direction predicted by the underlying nucleotide bias. Several previous studies suggest that protein evolution may be affected by directional changes to the nucleotide composition of the encoding genes. The first evidence was provided many years ago by the finding that there was a correlation between the nucleotide content and amino acid content of bacterial cells, before the genetic code itself had been elucidated (Sueoka, 1961). Subsequently, a number of surveys of molecular sequences (D'Onofrio et al., 1991; Collins and Jukes, 1993; Berkhout and van Hermert, 1994; Hickey et al., 1994; Porter, 1995; Berkhout et al. 2002) identified correlations between the nucleotide composition of DNA and the amino acid content of the encoded proteins. Many of these surveys, however, involve intra-genomic comparisons of heterologous DNA and protein sequences. In general, more recent studies have focused primarily on inter-genomic comparisons of homologous gene sequences (Foster, Jermiin, and Hickey, 1997; Li, 1997, pp. 422-423; Lobry, 1997; Gu, Hewett-Emmett, and Li, 1998; D'Onofrio et al., 1999; Lafay et al., 1999; Rodriguez-Trelles, Tarrio, and Ayala, 1999; Wilquet and Van de Casteele, 1999; Knight et al, 2001; Schmitz et al, 2002). All of these studies, despite the fact that they are based on a wide variety of different genes and proteins, provide evidence for a significant correlation

between DNA composition and protein composition. We wished to test the generality of these correlations by doing a comprehensive analysis of completely sequenced genomes.

At the time of this study, there were over twenty completely sequenced prokaryotic genomes, and two completely sequenced eukaryotic genomes available in the public databases. One of the most striking features of these genomes is the range of nucleotide compositions represented (see the Methods). For example, *Borrelia burgdorferi*, has an overall GC content of only 25.5% in its coding DNA, while *Mycobacterium tuberculosis* has a GC content of 65.9%. With such a range in nucleotide composition, these genomes provide an excellent resource for studying the relationship between DNA bias and amino acid composition. We decided to perform a predictive analysis similar to that used by Foster, Jermiin, and Hickey (1997). By partitioning the universal codon table into GC-rich codons, AT-rich codons, and neutral codons (see Table 1), we could make the simple prediction that AT-rich coding sequences would encode proteins rich in the FYMINK amino acids (phenylalanine, tyrosine, methionine, isoleucine, asparagine, and lysine), while GC-rich coding sequences would produce proteins containing high levels of the GARP amino acids (glycine, alanine, arginine, and proline). In our comprehensive analysis, involving twenty-two completely sequenced genomes, we have shown that this prediction holds true.

Table 3.1: The universal codon table, rearranged to group codons according to their nucleotide composition.

First codon position	Second codon position: A or T				Second codon position: G or C			
A or T	**AAG**	**Lys**	**ATG**	**Met**	AGG	Arg	ACG	Thr
	AAA		**ATA**	**Ile**	AGA		ACA	
	AAC	**Asn**	**ATC**		AGC	Ser	ACC	
	AAT		**ATT**		AGT		ACT	
	TAG	**Ter**	**TTG**	**Leu**[a]	TGG	Trp[b]	TCG	Ser
	TAA		**TTA**		TGA	Ter	TCA	
	TAC	**Tyr**	**TTC**	**Phe**	TGC	Cys	TCC	
	TAT		**TTT**		TGT		TCT	
G or C	GAG	Glu	GTG	Val	**GGG**	**Gly**	**GCG**	**Ala**
	GAA		GTA		**GGA**		**GCA**	
	GAC	Asp	GTC		**GGC**		**GCC**	
	GAT		GTT		**GGT**		**GCT**	
	CAG	Gln	CTG	Leu	**CGG**	**Arg**[c]	**CCG**	**Pro**
	CAA		CTA		**CGA**		**CCA**	
	CAC	His	CTC		**CGC**		**CCC**	
	CAT		CTT		**CGT**		**CCT**	

[a] Because four of the six leucine codons are not AT-rich, leucine was not included amongst the AT-rich amino acids group

[b] While tryptophan has a GC-rich codon, the *Mycoplasma* genomes also use TGA to encode this amino acid, so its codons are not consistently GC-rich. We therefore did not include tryptophan in our group of GC-rich amino acids

[c] Arginine was included amongst our GC-rich amino acids group because the majority of its codons (4/6) are GC-rich.

3.3 Methods

3.3.1 Data sets

Twenty-one prokaryotic genomes, and the entire *Saccharomyces cerevisiae* genome were downloaded from the GenBank FTP site[1]: *Saccharomyces cerevisiae* (U00091, Y13134, X59720, Z71256, U00092, D50617, Y13135, U00093, Z47047, Y13136, Y13137, Y13138, Z71257, Y13139, Y13140, U00094 for chromosomes 1-16, respectively; 6274 genes; 40.3% GC in the coding sequences); *Aeropyrum pernix* K1 (Aero_p; 2694 genes; 57.5% GC); *Archaeoglobus fulgidus* (AE000782; 2407 genes; 49.4% GC); *Aquifex aeolicus* (AE000657; 1522 genes; 43.7% GC); *Borrelia burgdorferi* (AE000783 and seq_AE000784-seq_AE000794 for 11 plasmids; 1254 genes; 25.5% GC); *Bacillus subtilis* (AL009126; 4100 genes; 44.3% GC); *Chlamydia pneumoniae* (AE001363; 1052 genes; 41.3% GC); *Chlamydia trachomatis* (AE001273; 894 genes; 41.7% GC); *Escherichia coli* K-12 MG1655 (U00096; 4289 genes; 51.8% GC); *Haemophilus influenzae* Rd (L42023; 1709 genes; 38.8% GC); *Helicobacter pylori* 26695 (AE000511; 1566 genes; 39.6% GC); *Helicobacter pylori* J99 (AE001439; 1491 genes; 39.9% GC); *Mycoplasma genitalium* G37 (L43967; 480 genes; 31.6% GC); *Methanococcus jannaschii* (L77117; 1715 genes; 31.9% GC); *Mycoplasma pneumoniae* M129 (U00089; 677 genes; 40.7% GC); *Methanobacterium thermoautotrophicum* delta H (AE000666; 1869 genes; 50.6% GC); *Mycobacterium tuberculosis* (AL123456; 3918 genes; 65.9% GC); *Pyrococcus horikoshii* OT3 (Pyro_h; 2064 genes; 42.3% GC); *Rickettsia prowazekii* Madrid E (AJ235269; 834 genes; 30.4% GC); *Synechocystis* PCC6803 (AB001339; 3169 genes; 48.6% GC); *Thermotoga maritime* (AE000512; 1846 genes; 46.4% GC); *Treponema pallidum* (AE000520; 1031 genes; 52.6% GC). Both nucleotide and amino acid sequences were extracted directly from the GenBank flat files.

---

[1] ftp://ftp.ncbi.nih.gov/genbank/genomes/

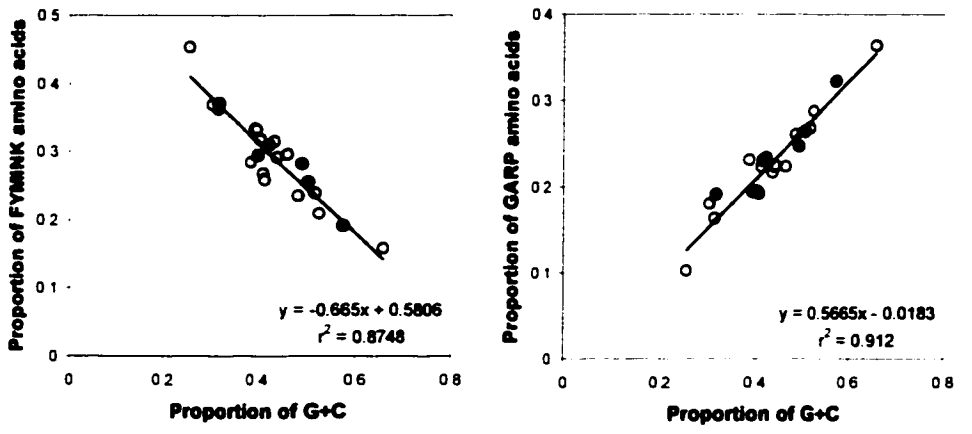


Figure 3.1: Correlation between mean G+C content and the amino acid content of 22 complete proteomes (see Methods, Section 3.3.1). Eubacterial genomes are shown as blue circles, Archaeal genomes as red circles, and the eukaryotic genome (*S. cerevisiae*) is shown as a green circle. Due to large sample sizes, the standard errors of the means are very small (usually less than 0.001); for this reason, standard error bars are not shown. (Left) The total genomic content of FYMINK amino acids (see Methods, section 3.3.2) decreases with increasing G+C content in the coding sequences, while (right) there is a corresponding increase in the proportion of GARP amino acids with increasing G+C content of the genomes.

3.3.2 FYMINK and GARP amino acids

We partitioned the codon table into three groups, in the manner introduced by Foster, Jermiin, and Hickey (1997): codons that are GC-rich at the first two codon positions, those that are AT-rich at the first two codon positions, and unbiased codons (Table 3.1). The AT-rich codons encode the FYMINK amino acids (phenylalanine, tyrosine, methionine, isoleucine, asparagine, and lysine). Although there are two AT-rich leucine codons, leucine was not included amongst this set because there are four other leucine codons that are not AT-rich. The GC-rich codons encode the GARP amino acids (glycine, alanine, arginine, and proline). In this set, arginine is included because four of its six codons are GC-rich.

3.3.3 Finding homologues

Homologous genes in common to all genomes were identified by performing gapped-BLASTP searches (Altschul et al., 1997) of each bacterial proteome against the *Bacillus subtilis* proteome, which is large and is relatively free from the effects of nucleotide bias. A cut-off expect score of 0.0001 was chosen since it appeared to give the largest data set whose hits seemed reliable. In addition, hits with less than 25% identities were ignored; this additional criterion resulted in the removal of many of the lower BLAST scores. Hits in common to all BLAST searches were assumed to be homologues (see Appendix 3.7.1).

3.4 Results

3.4.1 Comparison of complete genomes

In our first analysis, we included all of the genes, and all of the protein sequences from these genomes. We predicted that AT-rich (GC-poor) genomes would have a higher proportion of AT-rich codons in their genes, which encode the FYMINK amino acids. Likewise, GC-rich genomes should have proteins that are rich in the GARP amino acids. Our results show that this relationship holds true for the genomes analysed in this study (Figure 3.1), and these results are statistically highly significant ($p < 10^{-9}$).

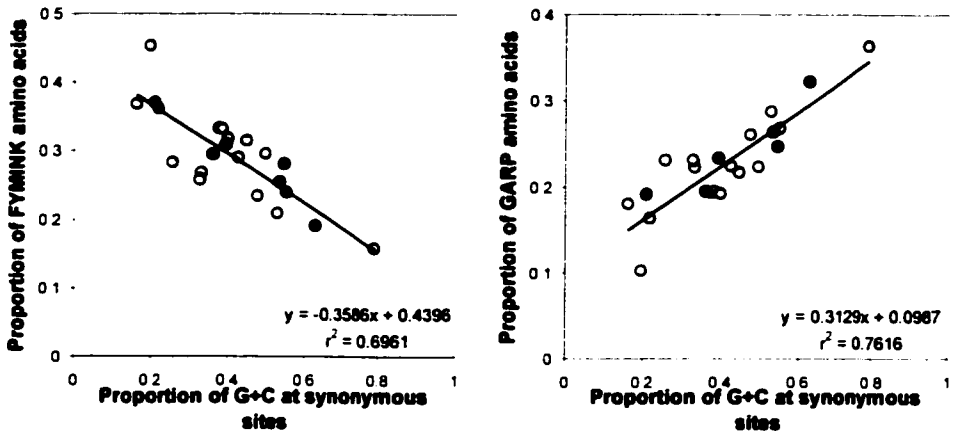


Figure 3.2: The (left) FYMINK and (right) GARP proportions of 22 proteomes are correlated to the synonymous G+C contents of the coding sequences. This Figure is similar to Figure 3.1, except that here we have plotted the amino acid contents against the synonymous G+C contents of the genes, rather than the overall G+C contents. Since directional changes to the amino acid contents of the proteins will have no effect on synonymous nucleotide proportions, the result from this Figure strongly suggests that it is DNA bias that is causing the shift in amino acid usage.

In order to demonstrate a causal relationship between nucleotide bias and amino acid composition, as opposed to a simple correlation between the two, we wished to eliminate the alternative possibility, i.e.,, that amino acid biases were affecting the nucleotide composition of genes rather than *vice versa*. This was accomplished by repeating the analysis using the nucleotide composition at the synonymous codon positions only. These results are shown in Figure 3.2, and show that there is still a very significant correlation between nucleotide composition and amino acid composition ($p < 10^{-6}$). Since the synonymous sites do not affect the protein sequence, this correlation reflects an underlying nucleotide bias, affecting both synonymous and non-synonymous sites. Additional supporting evidence for this conclusion stems from a comparison of the relative extent of the nucleotide bias at synonymous versus nonsynonymous codon positions (Figure 3.4). Nucleotide bias is much greater at the synonymous codon positions than at the nonsynonymous positions. If there were a primary

amino acid bias within proteins, which was then reflected as a secondary DNA bias, the nonsynonymous codon positions should be more biased than the synonymous codon positions, which are unaffected by amino acid content. Conversely, a primary bias at the DNA level is expected to be most pronounced at the synonymous codon positions, but less so at the nonsynonymous codon positions which are under stronger selective constraint. The data presented in Figure 3.4 strongly support the latter interpretation. Thus, these results are consistent with the view that the differences in amino acid composition are a consequence, rather than a cause, of the nucleotide bias.

Finally, we were concerned that our results might be skewed by the fact that not all genomes contain the same set of genes. For instance, some genomes might have a higher proportion of membrane-spanning proteins and this could result in differences in amino acid composition that are functionally based. In order to eliminate this possibility, we performed an additional comparison, using only the amino acid and nucleotide compositions of genes that are shared between these genomes. In this way, we could separate the effects of nucleotide bias from the effects of functional constraint. Homologous genes were identified by sequence similarity searches. Although we found hundreds of genes in common between any pair of genomes, we limited this analysis to a core set of forty-seven genes that were common to all twenty-two genomes. Again, we found a strong relationship ($p < 10^{-10}$) between the nucleotide content at synonymous sites and amino acid composition for this set of homologous genes (Figure 3.3A and B). It is important to note that most of the variability in amino acid frequencies between these sequences can be explained by nucleotide bias (r^2=0.64 for the FYMINK amino acids, and 0.67 for the GARP amino acids). Even more surprising is that because these genes were identified via sequence similarity searches, only the most conserved genes were included in the analysis. Because of the decreased flexibility in amino acid composition among these genes, the results based on this relatively conserved group gives an underestimate of the average effect of nucleotide bias on amino acid composition. In fact, this restriction of the analysis

to relatively conserved sequences explains why the slope of the regression lines in Figure 3.3 are less than in Figure 3.2.

Because phylogenetically related species are expected to have similar nucleotide and amino acid biases, it is plausible that the results shown in Figure 3.1, Figure 3.2, and Figure 3.3 reflect phylogenetic history rather than mutational bias. Pagel and Harvey (1988) have summarised the primary methods to distinguish evolutionary trends from phylogenetic artifact. All of these methods depend on some degree of replication of the trend within phylogenetically independent lineages. In our data set, we have an opportunity to look for repeated trends within two distinct lineages, the Eubacteria and Archaea. A simple inspection of the distribution of blue circles (Eubacteria) and red circles (Archaea) in Figure 3.1, Figure 3.2, and Figure 3.3 suggests that the trend is indeed repeated within both of these two major lineages, and that no significant difference exists between these two sets of data. In order to verify this impression, new regressions were calculated for each of these two kingdoms. We found that the regression was significant ($p<0.003$) within both the Eubacteria and the Archaea, indicating the trend is indeed repeated. We then used the analysis of covariance to test for differences between the slopes of the regression lines for the Eubacterial and Archaeal data sets. We found no significant difference between the two groups. In other words, not only is the trend repeated, but also the magnitude of the effect is very similar within both the Eubacteria and the Archaea. This is entirely consistent with the impression obtained from the simple visual inspection of the data points in Figure 3.3.

3.5 Discussion

Our main finding is not just that protein composition is affected by nucleotide bias, but that this effect is both very large and very widespread. We were surprised that our very simple predictive model based on the genetic code table could explain so much of the inter-genomic variation in protein composition. This result has implications for many studies that are based on the interpretation of

amino acid sequence data. For instance, it has already been shown that nucleotide bias can affect the functional properties of proteins (Naylor, Collins, and Brown, 1995; Gu, Hewett-Emmett, and Li, 1998, D'Onofrio et al., 1999) and that convergent amino acid composition can affect the construction of phylogenetic trees based on protein sequences (Gu, Hewett-Emmett, and Li, 1998; Foster and Hickey, 1999). For instance, Foster and Hickey (1999) showed that unrelated taxa are grouped together in a phylogenetic tree due to convergent amino acid sequences. Although this problem in phylogenetic reconstruction has been identified, a satisfactory method of dealing with the problem has not yet been found. Since the sequence of amino acids in a protein is linked to its structure, changes to the sequence driven by nucleotide bias may affect the evolution of the structure of the protein (Wood and Pearson, 1999).

The most parsimonious explanation of the observed patterns of amino acid composition in these genomes is an underlying mutational bias that varies between lineages (Sueoka, 1988). The resulting amino acid sequence changes are

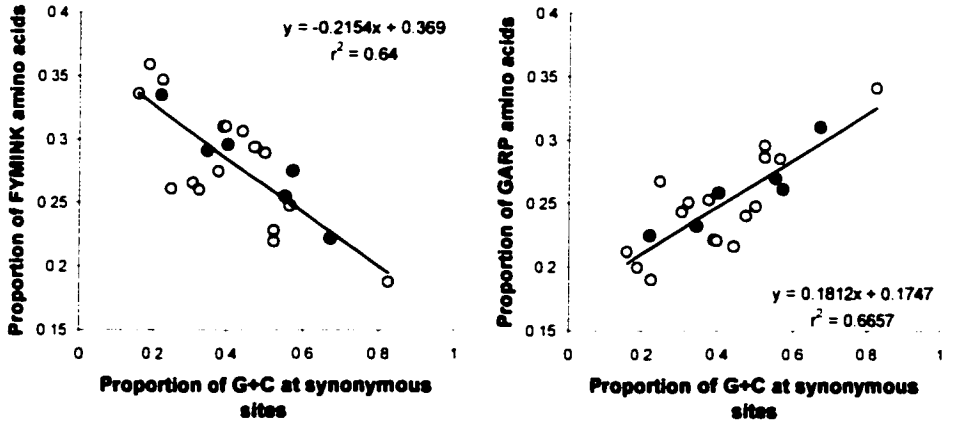


Figure 3.3: The proportions of the (left) FYMINK and (right) GARP amino acids in homologous proteins from 22 genomes are correlated to synonymous G+C content. Here we have expanded upon Figure 3.2 by only plotting the data for 47 genes that these organisms have in common with each other. In other words, the exact same protein has evolved to become GARP-rich in GC-biased DNA, while simultaneously evolving to become more FYMINK-rich in AT-biased DNA.

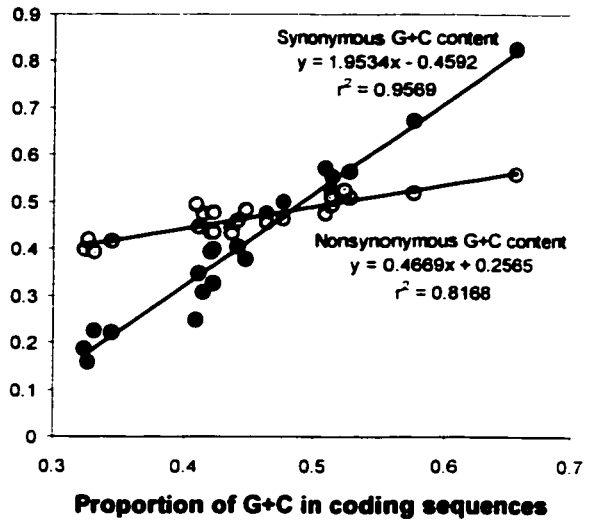


Figure 3.4: Comparison of nucleotide compositions at synonymous and non-synonymous sites. For each of the genomes shown in Figure 3.3, the G+C content at synonymous sites (dark circles) and non-synonymous sites (light circles) were plotted on the y-axis against the total G+C content of the coding sequences.

non-random, since the mutational bias is strongly directional, and yet they are not caused by natural selection acting directly on protein function. Consequently, their evolutionary dynamics cannot be described either in terms of Darwinian selection or random genetic drift.

In conclusion, we recognise that other factors such as selective constraint, adaptive change and genetic drift all play important roles in protein sequence evolution. The results presented here, however, demonstrate that mutational pressure on DNA composition can also be a very powerful and pervasive force in long-term protein evolution.

3.6 Note

The results from this Chapter were published in Singer and Hickey (2000).

3.7 Appendices

Appendix 3.7.1: The 47 homologous genes from the 22 genomes analysed, with descriptions and GenBank GI numbers.

Description (from Bsub)	Aero	Aful	Aquae	Bbur	Bsub	Cpneu	Ctra	Ecoli	Hinf	Hpyl	Hpyl99	Mgen	Mjan	Mpneu	Mthe	Mtub	Pyro	Rpxx	Synecho	Tmar	Tpal	Yeast
seryl-tRNA synthetase	5105674	2648500	2982982	2688129	2632280	4377187	3329186	1787120	1573062	2314654	4155982	1045673	1591725	1673809	2622225	1781123	3257118	3861310	1652304	4981940	3322945	642295
methionyl-tRNA synthetase	5104799	2649118	2983692	2688510	2632305	4376385	3328422	1788432	1574731	2313521	4155547	3844630	1591898	1673790	2621666	2052144	3257407	3861220	1001640	4981630	3323103	1323302
dimethyladenosine transferase	5104065	2648786	2984285	2688073	2632306	4377097	3329038	1787342	1571428	2314758	4156101	1045677	1592198	1673805	2621281	2052145	3257625	3861219	1652444	4981188	3323301	536084
dimethyladenosine transferase	5104206	2648762	2984099	2688507	2632109	4377394	3328775	1786236	1573534	2314603	4155935	3845058	1591684	1673824	2622432	2052147	3258259	3861209	1651740	4982002	3322613	1370549
cell-division protein and general stress protein (class III heat-shock)	5105710	2648566	2983448	2688712	2632336	4377127	3329313	1789568	1574793	2314219	4154882	3845051	1591785	1671833	2621817	2113966	3256587	3860613	1652085	4981098	3323067	807972
glutamyl-tRNA synthetase	5106018	2650372	2983664	2688265	2632359	4376852	3328877	1788743	1573240	2313587	4154965	3845057	1592022	1671825	2621087	2791589	3258115	3861166	1001692	4981911	3322975	1419825
ribosomal protein L11 (BL11)	5105873	2650083	2984196	2688304	2632369	4376335	3328736	2367334	1573500	2314164	4155722	3844668	1591080	1674314	2622809	1877378	3256386	3860703	1652498	4980960	3322507	1302170
RNA polymerase (beta' subunit)	5105544	2648656	2984193	2688294	2632375	4376340	3328731	2367335	1573495	2314357	4155718	3844922	2826350	1674001	2622150	2143292	3257973	3860708	1653352	4980968	3322512	1420309
ribosomal protein S12 (BS12)	5105540	2648653	2983320	2688295	2632377	4376843	3328870	1789740	1573569	2314356	4155709	3844674	2826360	1674308	2622154	2009655	3257969	3860698	1653410	4982071	3322515	1302537
ribosomal protein S7 (BS7)	5105537	2648652	2983319	2688296	2632378	4376842	3328869	1789739	1573568	2314355	4155708	3844675	1591701	1674307	2622155	2009656	3257968	3860699	1653409	4982070	3322516	1015830
elongation factor G	5105115	2648651	2982776	2688449	2632379	4376841	3328868	1789738	1573567	2314354	4155707	3844676	1591702	1674306	2622156	3261535	3258339	3860700	1653408	4982069	3323075	1360422
elongation factor Tu	5105535	2649659	2982777	2688415	2632380	4376331	3328740	1789737	1573560	2314366	4155731	3845045	1591042	1673840	2622158	1806151	3257908	3861201	1653407	4982068	3322451	1420449
ribosomal protein S10 (BS13)	5105533	2649669	2982773	2688414	2632382	4376840	3328867	1789717	1573786	2314482	4155817	3844744	1591041	1674372	2622159	1806168	3257907	3861200	1653406	4982065	3322452	976131
ribosomal protein L3 (BL3)	5103618	2648621	2982767	2688413	2632383	4376945	3328965	1789716	1573787	2314481	4155816	3844745	1590928	1674371	2621048	1806169	3258212	3861199	1652427	4982064	3322453	1323396
ribosomal protein L2 (BL2)	5103609	2648624	2982766	2688410	2632386	4376942	3328962	1789713	1573790	2314478	4155813	3844748	2826257	1674368	2621052	1806172	3258208	3861196	1652424	4982061	3322456	603629
ribosomal protein S19 (BS19)	5104006	2648642	2982775	2688409	2632387	4376941	3328961	1789712	1573791	2314477	4155812	3844749	1592272	1674367	2621053	1806173	3258207	3861195	1652423	4982060	3322457	1302539
ribosomal protein L22 (BL17)	5104004	2648643	2982770	2688408	2632388	4376940	3328960	1789711	1573792	2314476	4155826	3844750	1591162	1674366	2621054	1806174	3258206	3861194	1652422	4982059	3322458	1008376
ribosomal protein S3 (BS3)	5104002	2648625	2982772	2688407	2632389	4376939	3328959	1789710	1573793	2314475	4155825	3844751	1591163	1674365	2621055	1806175	3258205	3861193	1652421	4982058	3322459	1302158
ribosomal protein S17 (BS16)	5103999	2648627	2982774	2688404	2632392	4376936	3328956	1789707	1571796	2314473	4155822	3844754	1591167	1674362	2621040	1806178	3258202	3861190	1652418	4982055	3322462	854458
ribosomal protein L14	5103998	2648628	2983985	2688403	2632393	4376935	3328955	1789706	1571798	2314472	4155821	3844755	1591168	1674361	2621041	1806182	3258200	3861189	1652417	4982054	3322463	486297
ribosomal protein L5 (BL6)	5103993	2648645	2983983	2688401	2632395	4376933	3328953	1789704	1573800	2314470	4155819	3844757	1591171	1674359	2621044	1806184	3258197	3861187	1652415	4982052	3322465	817823
ribosomal protein S8 (BS8)	5103991	2648631	2983992	2688399	2632397	4376932	3328952	1789702	1573802	2314469	4155815	3844759	1591172	1674357	2621046	1806186	3258195	3861185	1652414	4982050	3322466	1008400
ribosomal protein L6 (BL8)	5103989	2648648	2983984	2688398	2632398	4376931	3328951	1789701	1573803	2314468	4155834	3844760	1591173	1674356	2621047	1806187	3258194	3861184	1652413	4982049	3322467	500660
ribosomal protein S5	5103985	2648634	2983990	2688396	2632400	4376929	3328949	1789699	1571805	2314466	4155832	3844762	1591177	1674354	2621060	1806189	3258188	3861182	1652411	4982047	3322469	536669
ribosomal protein L15	5105434	2648636	2983986	2688394	2632402	4376928	3328948	1789697	1573807	2314465	4155831	3844763	2826259	1674353	2621062	1806191	3258065	3861180	1652410	4982045	3322477	1302361
preprotein translocase subunit	5104652	2648637	2982812	2688393	2632403	4376927	3328947	1789696	1573808	2314464	4155830	3844764	2826280	1674352	2621063	2911006	3258185	3861179	1652409	4982044	3322478	609428
ribosomal protein S13	5105425	2648638	2982822	2688391	2632408	4376926	3328946	1789693	1573810	2314462	4155844	3844771	2826258	1674347	2621070	2104383	3258070	3861177	1652405	4982041	3322479	1301975
ribosomal protein L13	5105435	2649459	2984157	2688238	2632416	4376522	3328524	1789626	1574283	2313165	4154587	3845010	2826260	1673891	2621075	2104364	3258063	3860796	1652400	4982019	3323353	1420377
class I heat-shock protein (molecular chaperonin)	5104805	2649475	2983164	2688702	2632907	4376465	3328603	1789445	1573514	2314767	4156114	3844655	2826367	1671750	2622538	1449168	3258431	3860607	1653193	4980638	3322977	1431146
class I heat-shock protein (molecular chaperonin)	5105771	2649120	2984179	2688576	2632916	4376199	3328508	1790586	1573528	2313084	4154502	3844983	1591659	1671939	2621883	2909515	3256402	3861167	1652489	4981018	3322286	577181
tryptophanyl-tRNA synthetase	5106165	2648859	2983495	2687887	2633496	4377113	3329028	1789786	1573633	2314417	4155771	3844717	1592065	1674268	2621301	2894246	3257425	3861023	1652781	4981003	3322931	1230641
signal recognition particle	5105423	2650007	2984037	2688634	2633970	4376178	3328415	1788963	1571059	2314305	4155673	1045721	1590878	1673748	2622427	1405951	3258124	3860739	1653392	4982134	3322699	1230693
DNA topoisomerase I	5105059	2648740	2983253	2688756	2633984	4377077	3329092	1787529	1574197	2313198	4154624	3844712	1590401	1674272	2622750	2105048	3257028	3860886	1652946	4980754	3322672	609366
ribosomal protein S2	5105440	2649456	2984239	2688009	2634021	4376990	3329132	1786365	1571934	2314737	4156073	1045746	2826355	1674326	2621070	1403411	3258058	3860655	1653232	4981290	3322900	500819
initiation factor IF-2	5106076	2649835	2984268	2688747	2634035	4376594	3328492	1789559	1574742	2314195	4154899	3844735	1590888	1674382	2621309	1648881	3257511	3861101	1651769	4981303	3323208	1419805

Description (from Bsub)	Aero	Aful	Aquae	Bbor	Bsub	Cpneu	Ctra	Ecoli	Hinf	Hpyl	Hpyl99	Mgen	Mjan	Mpneu	Mthe	Mtub	Pyro	Rpra	Synecho	Tmar	Tpal	Yeast
ribosomal protein S15 (BS18)	5104809	2649805	2982947	2688744	2634040	4377129	3329115	1789550	1574788	2314181	4154915	1845014	2826217	1671886	2622536	2624107	3256443	3861055	1651956	4981903	3323204	1230664
multifunctional SOS repair regulator	5101500	2648416	2984348	2688005	2634066	4377070	3329089	1789051	1571588	2311215	4154654	1844021	2691451	1674029	2622824	2624258	3256505	3861290	1653947	4982444	3322905	1420786
alanyl-tRNA synthetase	5105865	2648270	2981727	2688110	2635186	4377211	3329299	1789048	1573826	2314404	4155760	1844871	2691270	1674107	2622813	2621815	3256687	3861381	1653611	4981959	3323345	1420729
GTPase activity	5104150	2648391	2984292	2688719	2635257	4376835	3328847	1789574	1573894	2311401	4154802	1844073	2691967	1673950	2622747	2666144	3257622	3861368	1652022	4980588	3323049	458902
valyl-tRNA synthetase	5105495	2648303	2983819	2688666	2635274	4376354	3328717	1790708	1574225	2314306	4155674	1844911	2691666	1674039	2621857	2791488	3256704	3861224	1001391	4982399	3323365	1321141
phenylalanyl-tRNA synthetase (beta subunit)	5106006	2649146	2984040	2688420	2635128	4376889	3328910	1788806	1574770	2311508	4155564	1844792	2691750	1673698	2621860	1839002	3257065	3860975	1653476	4981353	3322665	1360406
phenylalanyl-tRNA synthetase (alpha subunit)	5106003	2648587	2981471	2688421	2635120	4377321	3329107	1788007	1574760	2311590	4155565	1844701	2692593	1673699	2621831	1839001	3257066	3860974	1001185	4981352	3323297	836732
ribosomal protein S4 (BS4)	5105427	2648246	2982810	2688529	2635450	4377018	3329071	1789691	1571812	2314460	4155842	1046011	2498964	1674077	2621071	2104381	3258669	3860905	1001162	4982039	3322583	1302082
thioredoxin reductase	5104731	2649006	2981154	2688441	2635092	4376591	3328496	1787114	1574715	2313960	4155129	1844690	2692167	1674293	2621705	2808698	3257840	3861002	1653956	4981404	3323124	849175
serine hydroxymethyltransferase	5105660	2649750	2981131	2688518	2636215	4376810	3328862	1788902	1571908	2313271	4154686	1844985	2692207	1673936	2622490	2896730	3258083	3861272	1652200	4981245	3322607	536692
arginyl-tRNA synthetase	5105446	2649707	2983452	2688523	2636270	4376861	3328887	1788184	1574424	2311410	4154811	1844966	2690971	1671957	2622560	1132416	3257902	3860635	1001346	4981037	3323142	1230668
antagonist of Sey	5104835	2649213	2981235	2688121	2636010	4376598	3328488	1787454	1573365	2311680	4155071	1844631	2691073	1671787	2622634	1256023	3256910	3861148	1653033	4981704	3323387	536234

4 Nucleotide bias can affect the amino acid composition within all proteins in the proteome

4.1 Abstract

Our results from the previous chapter showed that the effects of nucleotide bias on amino acid usage are widespread among the bacteria. Here we expand upon those results by performing pair-wise comparisons between the most oppositely biased organisms for each of the three major domains of life: Archaea, Eubacteria, and Eukaryota. Our results show that directional nucleotide bias in the DNA affects the amino acid usage of proteins in all three kingdoms, and that nearly every gene appears to be affected, although to varying degrees: highly conserved genes are affected by DNA mutation bias to a lesser extent than loosely conserved genes. Finally, we present evidence that the directional changes in amino acid composition within proteins may cause changes in the metabolic requirements for protein synthesis. Therefore, directional mutation bias in the DNA may have significant effects on the energy requirements of the cell.

4.2 Introduction

In the previous chapter we showed that directional nucleotide bias is very common among the completely sequenced bacterial genomes, and that this bias transcends phylogenetic boundaries (the Archaea, for example, showed a range in G+C bias that nearly spanned the same range as the Eubacteria). Although the nucleotide biases in these organisms are the greatest at the synonymous sites within the genes, we showed that the nonsynonymous nucleotide positions are also affected by DNA mutation bias. In terms of the encoded proteins, we showed that AT-rich genomes tend to encode proteins that are rich in the FYMINK amino acids (which are encoded by AT-rich codons), but that have low proportions of the GARP amino acids (which are encoded by GC-rich codons). We showed that the GC-rich genomes, on the other hand, encode proteins that show the opposite trend in amino acid composition.

In this chapter, we examine the relationship between nucleotide bias and amino acid content on a finer scale. In each of the three main domains of life (Archaea, Eubacteria, and Eukaryota) we have identified two organisms whose nucleotide biases are extremely polarized. We have identified homologous genes and proteins for each pair of organisms, and have measured their nucleotide and amino acid frequencies. By performing these gene-by-gene pair-wise analyses, we are able to determine the extent to which each gene within the genome is affected by DNA mutation bias, and how this bias affects the amino acid content of the encoded protein. Moreover, we show that these changes to the amino acid composition of the proteins may affect the metabolic requirements of protein synthesis.

Table 4.1: The GenBank versions and G+C contents of the organisms analysed in this study.

Organism	Kingdom	GenBank Accession	Synonymous %GC	Nonsynonymous %GC
Streptomyces coelicolor	Eubacteria	N/A[a]	94.0%	56.7%
Borrelia burgdorferi	Eubacteria	AE000783	19.4%	37.0%
Aeropyrum pernix	Archaea	Aero_p	67.8%	53.2%
Methanococcus jannaschii	Archaea	L77117	22.2%	40.1%
Drosophila melanogaster	Eukaryota	N/A[a]	63.8%	48.3%
Plasmodium falciparum	Eukaryota	AE001362, MAL3 (chr 2 and 3)	14.9%	30.5%

[a] At the time of this study, the *S. coelicolor* and *D. melanogaster* genomes were not complete, so protein and nucleotide sequences from these organisms were downloaded from GenBank using Entrez searches (http://www.ncbi.nih.gov/Entrez).

4.3 Methods

Pairs of organisms from the Eubacteria, the Archaea, and the Eukaryota were downloaded from GenBank (see Table 4.1). The pairs represent highly (but approximately equally) divergent organisms within their respective kingdoms (Nei et al. 2001; Clarke et al. 2002). Homologous sequences were identified using BLASTP searches (Altschul et al. 1997). We selected only those homologous pairs that shared 25% or more identical residues and had an alignment expectation score less than 10^{-4}. Although this expect score is high, the 25% identity cut-off caused the vast majority of alignments to have very good expect scores. The median expect score in among the *S. coelicolor* – *B. burgdorferi* homologues was

7×10^{-31}, for example. The amino acid frequencies of these proteins were measured using programs written by GS using the Python programming language[1]. The gene sequences corresponding to these proteins were then extracted from GenBank flat files, and the nucleotide proportions were measured.

4.4 Results

4.4.1 Nucleotide bias and phylogenetic relationships

Table 4.1 lists the organisms that we have chosen to study. The existence of both AT-rich and GC-rich organisms within each kingdom shows the pliability of the nucleotide content over the course of evolution. Even the most parsimonious explanation for these patterns of nucleotide bias requires the independent onset of a new direction of bias in at least three lineages. If the universal ancestor to all of these organisms was AT-rich, for instance, we could only explain the present distribution of nucleotide biases by positing a shift in the direction of bias towards G+C in each of the *S. coelicolor*, *A. pernix*, and *D. melanogaster* lineages. In addition, there is variation in the extent of the nucleotide biases among both the GC- and AT-rich organisms. The GC-richness of *A. pernix*, for example (67.8% at synonymous sites) pales in comparison to the GC-richness of *S. coelicolor* (94% at synonymous sites).

The range in nucleotide biases in this small data set are truly remarkable, from *S. coelicolor*'s synonymous GC content of 94.0% down to *P. falciparum*'s synonymous GC content of only 14.9% (Table 4.1). The nucleotide proportions at the nonsynonymous sites are not so divergent, although they tend to follow the same pattern as the synonymous sites—a result that is consistent with our findings from the previous chapter, and with the predictions of the neutral theory of molecular evolution. Although the Eubacteria show the greatest degree of polarity in terms of nucleotide bias, this is likely a sampling artifact since there are many

[1] http://www.python.org

completely sequenced Eubacterial genomes and relatively few Archaeal and Eukaryote genomes. Presumably, as additional genomes are sequenced the range in nucleotide biases among the three kingdoms will even out.

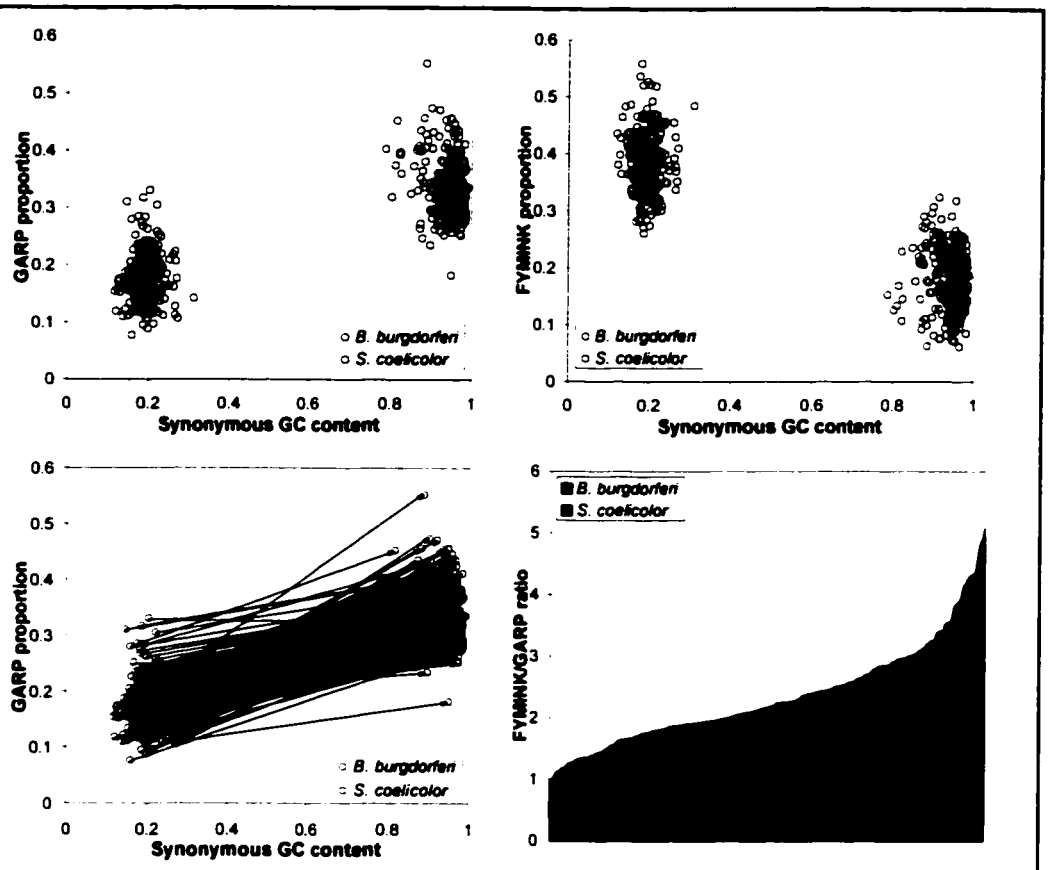


Figure 4.1: The GARP (upper left) and FYMINK (upper right) amino acid proportions in 255 homologous proteins from two Eubacteria: *B. burgdorferi*, which is AT-rich, and *S. coelicolor*, which is GC-rich. Although there appears to be overlap in the distributions on the y-axis, pair-wise comparisons show that nearly all *S. coelicolor* proteins have more GARP than their homologues in *B. burgdorferi* (lower left). In the Figure on the lower right we have presented this same trend in a different way. We have ordered the *B. burgdorferi* genes in increasing order of FYMINK/GARP ratio, and have plotted the FYMINK/GARP ratios for the homologous *S. coelicolor* genes at the same x-axis positions. For all homologues, the FYMINK/GARP ratio is greater in the AT-rich *Borrelia* than in the GC-rich *Streptomyces*.

4.4.2 Gene-by-gene analyses of directional evolution

Since GC-rich codons encode the GARP amino acids while the AT-rich codons encode the FYMINK amino acids, we expected the AT-biased Eubacterium *B. burdorferi* to encode proteins that have more FYMINK and less GARP than the proteins from its GC-biased cousin, *S. coelicolor*. Indeed, this is exactly what we found (Figure 4.1, top two plots). Note that we have plotted only homologous proteins: in other words, after the divergence since their most recent common ancestor, the proteins in *B. burgdorferi* became increasingly FYMINK-rich and GARP-poor, while the *same proteins* in *S. coelicolor* became biased in the exact opposite direction. There is no overlap in the two genomes' distributions along the x-axis, meaning the synonymous nucleotide contents of the genes from these genomes are highly polarized. Indeed, there is a flattening effect in the *S. coelicolor* distribution as it "bunches up" against the 100% G+C limit. There is, however, overlap along the y-axes of the two plots. It should be noted, though, that those genes towards the top of each distribution are paired together, as are the genes towards the bottom of each distribution, so for each gene there is an increase in GARP and a decrease in FYMINK from *B. burgdorferi* to *S. coelicolor*. This can be seen in Figure 4.1 (lower left), where we have drawn arrows to connect each *B. burgdorferi* protein to its homologue in *S. coelicolor*. Note that in nearly every case the line is sloping upward: that is, each *S. coelicolor* protein has more GARP than the same protein *B. burgdorferi*, as we would predict from the nucleotide biases in each genome. We have performed the same analysis on the FYMINK proportions with identical results (Figure not shown). This gene-by-gene trend is more clear in the lower right plot in Figure 4.1, where we have ordered *B. burgdorferi*'s genes from lowest FYMINK/GARP ratio to highest, and have plotted the same ratio for each homologous gene from *S. coelicolor* at the corresponding x-axis position. For every gene that these two organisms have in common, the *B. burgdorferi* gene has a higher FYMINK/GARP ratio than its *S. coelicolor* homologue—again, this is exactly the

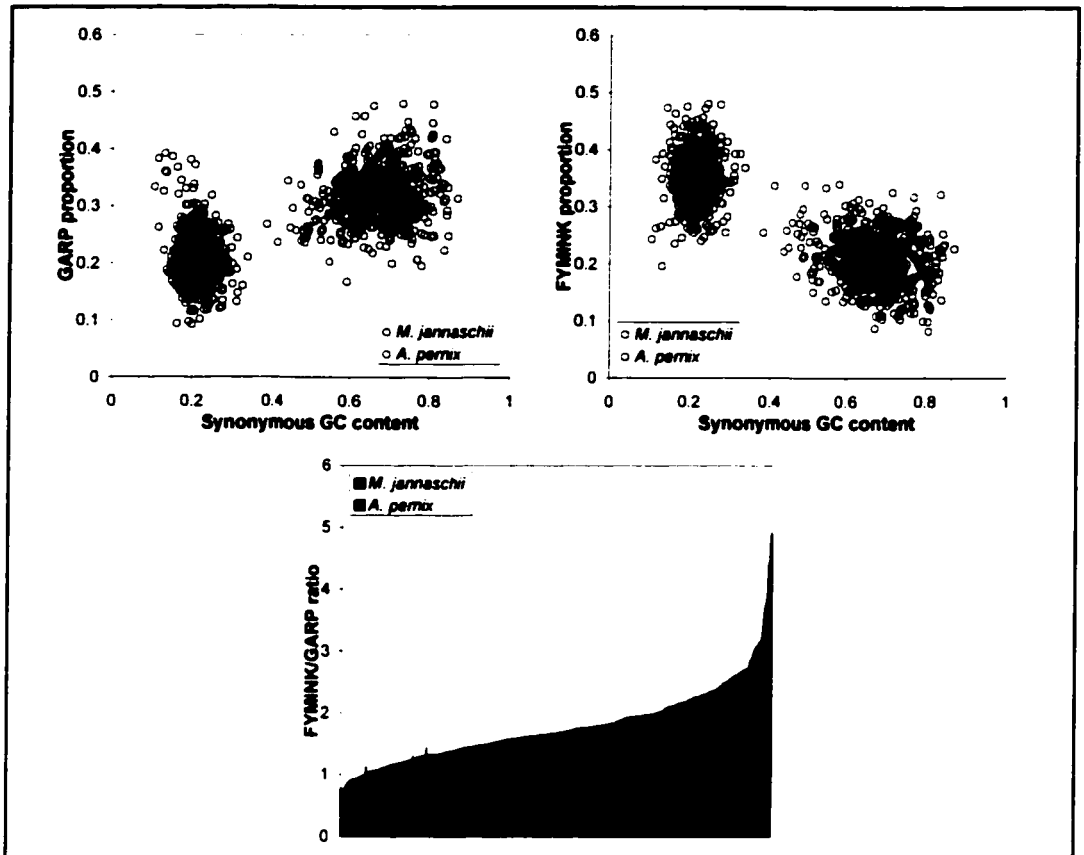


Figure 4.2: These Figures are analogous to those from Figure 4.1, except that in this case we have plotted 509 genes in common between two Archaea: *M. jannaschii*, which is AT-rich, and *A. pernix*, which is GC-rich. Seven *M. jannaschii* genes have a lower FYMINK/GARP ratio than their homologues in *A. pernix*. From left to right, their descriptions (from the *M. jannaschii* GenBank entry) are: LSU ribosomal protein L12A, LSU ribosomal protein L15E, H^+-transporting ATP synthase subunit B, LSU ribosomal protein L14E, a conserved hypothetical protein, DNA topoisomerase VI subunit A, and transcription initiation factor IID.

trend that we would predict based on the direction of nucleotide bias in each genome.

We know that *B. burgdorferi* and *S. coelicolor* share a common ancestor from deep within the Eubacterial tree (Clarke et al, 2002), and we have identified 255 genes that these two modern organisms share with each other through that ancestor. Figure 4.1 shows that sometime after the speciation event that

eventually led to these two modern organisms, the synonymous G+C contents of the two genomes became strikingly different: GC-rich for *S. coelicolor* and AT-rich for *B. burgdorferi*. These nucleotide usage patterns show that there must have been periods of directional evolution in the DNA—there is no other possible explanation, since under a stationary model of DNA evolution these two organisms would have G+C contents that were identical to each other and their common ancestor. Figure 4.1 also shows, however, that this trend is not limited to the synonymous sites in the DNA. The amino acid compositions of the proteins are also very different: *B. burgdorferi* proteins have more FYMINK and less GARP amino acids than the corresponding proteins in *S. coelicolor*. Again, the only possible conclusion we can make is that there has been a directional pattern of evolution in these proteins, and the direction of evolution has been the same in every protein. We believe that the directional evolution of both the DNA and the proteins is not mere coincidence; we prefer the more parsimonious explanation that a change in the relative mutation rates in the DNA caused a period of directional DNA evolution—in opposite directions in each organism—that affected the amino acid contents of their proteins through directional (but neutral) nucleotide substitutions in nonsynonymous codon positions within the genes.

Figure 4.2 shows analogous plots for the pair-wise comparison between two Archaea, the AT-rich *M. jannaschii* and the GC-rich *A. pernix*. Although these two organisms are not as polarized in terms of their nucleotide biases as the two Eubacteria are, we see very similar trends in their amino acid compositions. In this case only seven of the 509 protein sequences that these two organisms have in common have a FYMINK/GARP ratio that goes against the predicted trend (although not by a large margin); all of the other proteins follow the direction of amino acid bias we would predict based on the nucleotide contents of the genomes (Figure 4.2, right). We see these same results once again in the comparison of the two Eukaryota (Figure 4.3). Among the 161 genes that the AT-rich *P. falciparum* and the GC-rich *D. melanogaster* have in common, only one has a FYMINK/GARP ratio that goes against the predicted trend, and only to a

very small degree (Figure 4.3, right). Of interest here is the scale along the y-axis, which is fully three times larger than that for the Eubacteria and Archaea. These extremely biased amino acid compositions correspond to the extreme AT-richness of the *P. falciparum* genome (see Table 4.1).

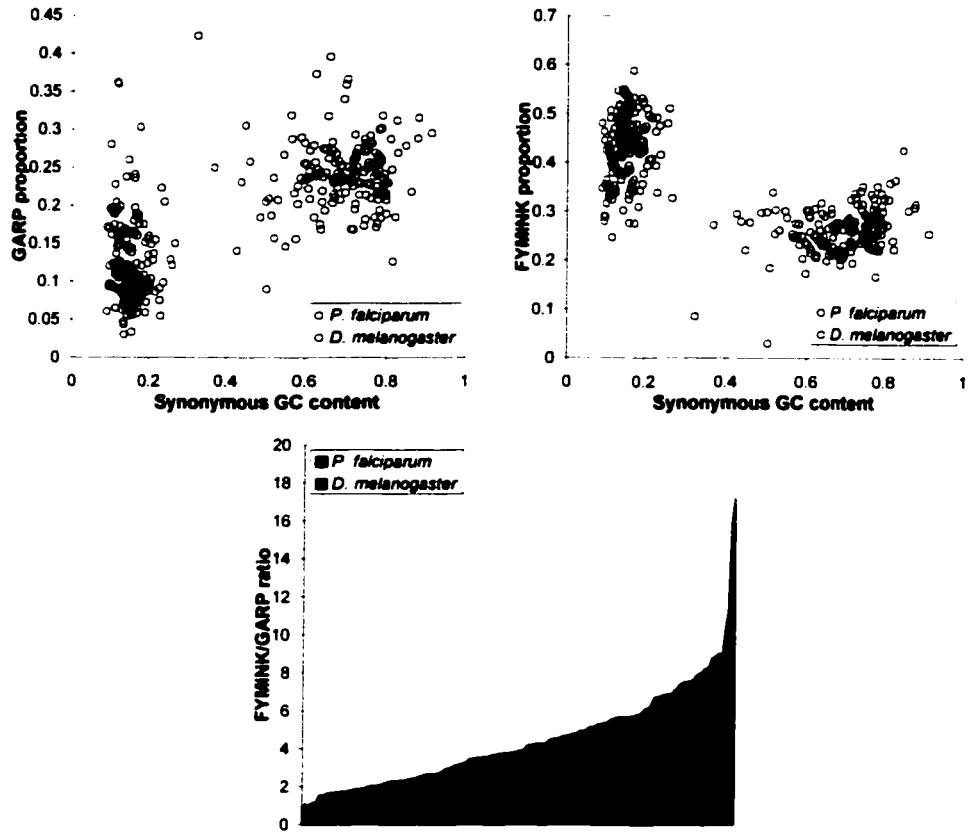


Figure 4.3: These Figures are analogous to those from Figure 4.1, except that in this case we have plotted 161 genes in common between *P. falciparum*, which is extremely AT-rich, and *D. melanogaster*, which is GC-rich. In only one case does a *D. melanogaster* gene have a higher FYMINK/GARP ratio than its homologue in *P. falciparum* (40S ribosomal protein S23, according to the *P. falciparum* GenBank entry).

4.4.3 Gene conservation and the polarization of amino acid substitutions

The lower plots in Figure 4.1, Figure 4.2, and Figure 4.3 show that the vast majority of proteins within a genome have an amino acid composition that is affected by nucleotide bias. At the same time, however, it is clear that the degree to which each protein is affected is different. Those to the left of the plots, for example show little difference in FYMINK/GARP ratio between the two organisms, while those to the right show a greater divergence. We analysed the relationship between the degree of divergence versus the degree of amino acid usage bias in the homologous proteins from the two Eubacteria in our study, *B. burgdorferi* and *S. coelicolor*. As the results in Figure 4.4 (left plot) show, highly conserved genes (those with a low degree of divergence between the two sequences) show the least amount of amino acid usage bias, while loosely conserved genes (those with a high degree of divergence) show a larger amount of polarized amino acid bias. We infer from this finding that when variation is allowed within a protein sequence, the DNA bias determines which direction the amino acid substitutions will take. Those proteins that conserve only small portions of the sequence, then, show a large degree of amino acid bias. On the other hand, proteins that are more highly conserved have fewer numbers of variable sites, and therefore the amount of amino acid bias is small. These findings are similar to those of Andersson and Sharp (1996), who found a positive relationship between degree of bias and sequence divergence. We have gone a step further, though, and have measured the FYMINK and GARP proportions at the variable sites (i.e., those sites in which the two sequences differ) within the proteins (Figure 4.4, right plot). Note that the difference between conserved and variable proteins disappears, indicating that all sites that are free to mutate—whether within conserved proteins or not—are affected by DNA mutation bias. In Table 4.2 we show the GARP and FYMINK proportions in detail for several of these proteins. The *tlyA* and *rnhB* genes are known to be highly variable, and this freedom has allowed DNA bias to modify the amino acid contents of these

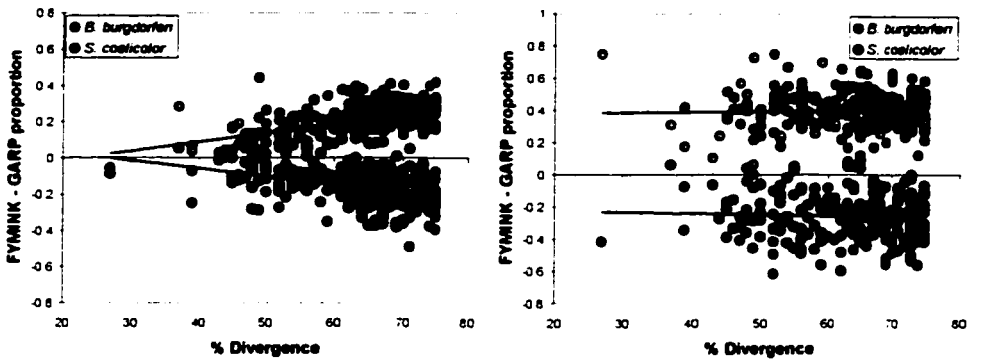


Figure 4.4: The relative amounts of the FYMINK and GARP amino acids in the AT-rich *B. burgdorferi* and GC-rich *S. coelicolor* genomes are correlated with the degree of divergence of the proteins. We have only plotted homologues with greater than 25% sequence identity. (Left plot) The more loosely conserved genes (low sequence identity) tend to show the greatest divergence in terms of FYMINK and GARP proportions between the two organisms, while the more highly conserved proteins (high sequence identity) show FYMINK and GARP proportions that are more similar. (Right plot) When the FYMINK and GARP proportions are measured at variable sites only, the difference between conserved and variable proteins disappears.

proteins in opposite directions in the two bacteria. Conversely, the *gap* and *tuf* genes are highly conserved, so there is little sequence variation between the organisms. Therefore, there is a correspondingly small amount of amino acid composition bias in these proteins.

4.4.4 Examination of individual amino acid changes

Our original prediction was that GC-rich coding DNA would produce proteins rich in the GARP amino acids, while AT-rich DNA would produce FYMINK-rich proteins. In this and the previous chapter we have shown that this is the case. Nevertheless, we decided that it was important to determine whether certain amino acids within each group were more apt to change than others. We scored the proportions of individual amino acids within the 255 gene pairs of the AT-rich *B. burgdorferi* and the GC-rich *S. coelicolor*. The results are shown in Figure 4.5. Each amino acid changed its proportion in the predicted direction, although

methionine changed only slightly. We had predicted that the proportion of methionine residues would increase in the AT-rich genome, but our results did not support this contention. On closer inspection of the data, we observed that while methionines replace other more GC-rich codons in this genome, the methionine codon, ATG, is itself replaced by the even more AT-rich isoleucine codons, ATA and ATT. In other words, in AT-biased genomes the 3rd position G in the methionine codon is quickly mutated to an A or T, resulting in an isoleucine codon. Thus, there is no net increase of methionine in the very AT-rich genomes.

Table 4.2: A comparison of the proportions of FYMINK and GARP amino acids in four homologous gene pairs from the genomes of *B. burgdorferi* and *S. coelicolor*. *tlyA* and *rnhB* are loosely conserved genes, while *gap* and *tuf* are very highly conserved.

Gene	**Borrelia burgdorferi**		**Streptomyces coelicolor**	
	GARP	**FYMINK**	**GARP**	**FYMINK**
Hemolysin (*tlyA*)	31 (11.8%)	109 (41.4%)	110 (40.6%)	27 (10.0%)
Ribonuclease H (*rnhB*)	30 (16.6%)	81 (44.8%)	84 (36.1%)	34 (14.6%)
Glyceraldehyde 3-phosphate dehydrogenase (*gap*)	77 (23.0%)	91 (27.2%)	85 (25.3%)	86 (25.6%)
Translation elongation factor TU (*tuf*)	101 (25.2%)	113 (28.2%)	103 (25.9%)	98 (24.7%)

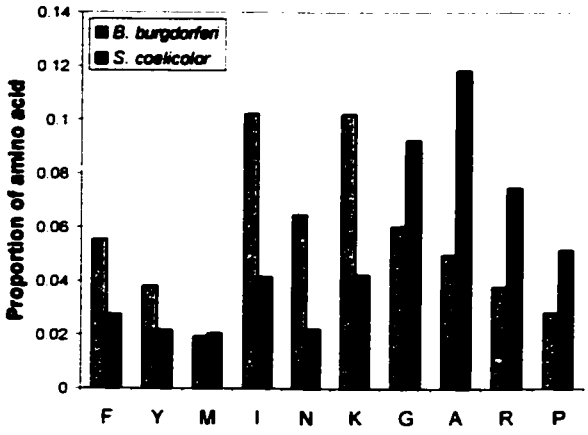


Figure 4.5: An analysis of the individual amino acid contents within the 255 homologous proteins in the two Eubacteria *B. burgdorferi* (AT-rich) and *S. coelicolor* (GC-rich). Because of the large amount of data sampled, standard error bars are too small to plot on the Figure. Note that all of the amino acids in the FYMINK and GARP categories change in the predicted directions, with the exception of methionine.

4.4.5 The metabolic costs of DNA bias

Recently, it was reported that there is positive selection on amino acid usage in proteins to reduce the energy required for amino acid synthesis (Akashi and Gojobori, 2002). We list the "energetic costs" for the synthesis of each amino acid in Table 4.3. Given that we have observed great flexibility in the amino acid contents of proteins in response to directional nucleotide bias, we wondered if changes to the amino acid composition of proteins had any impact on the metabolic requirements of protein synthesis. Presumably, selection would discourage changes that increased the energy requirements for protein synthesis. However, consultation with Table 4.3 shows that the average metabolic cost required to synthesize one of the FYMINK amino acids is 35.60, while it is only 17.75 for the GARP amino acids. We have analysed the homologous pairs of proteins from *B. burgdorferi* and *S. coelicolor*, and have confirmed that the average metabolic cost per amino acid in each set of homologous proteins is different: the AT-rich *B. burgdorferi* has an average metabolic cost per amino

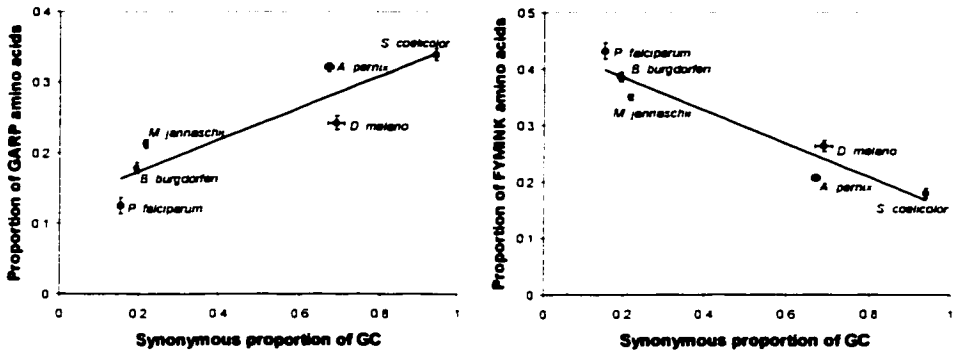


Figure 4.6: The GARP (left plot) and FYMINK (right plot) amino acid proportions in the proteins from all six organisms and their relationship to synonymous GC content. The bars indicate 99% confidence intervals from each genomic mean. Note that FYMINK and GARP proportions are guided by nucleotide bias and not by phylogenetic relationship.

acid of 24.69, while the average among the homologous proteins in *S. coelicolor* is only 22.02—a highly significant difference ($p < 7 \times 10^{-95}$ in a pair-wise t-test). Similarly, the average metabolic cost per protein is 8908 in *B. burgdorferi*, but only 8470 in *S. coelicolor*, which again is a significant difference ($p < 0.007$). We presume that either these differences have no biological significance or that the *B. burgdorferi* genome has altered its protein expression pattern to compensate for the additional energy requirements of the synthesis of its proteins.

4.5 Discussion

We have shown that the direction of nucleotide bias is remarkably plastic, and directional changes have occurred independently in many different lineages. The degree of the bias, too, can be quite striking as our representative organisms from the Archaea, the Eubacteria, and the Eukaryota show. Although the pairs of organisms we analysed are all highly divergent from each other, this does not appear to be a prerequisite for significant changes in nucleotide content. For example, within the bacterial family Mycoplasmatacea, there is a striking amount of variation in genomic G+C content (*Mycoplasma genitalium*: 31.7%; *Mycoplasma pneumoniae*: 40.0%; *Mycoplasma pulmonis*: 26.6%; *Ureaplasma*

Table 4.3: The energy requirements for the synthesis of the 20 amino acids (from Akashi and Gojobori, 2002).

Amino acid	Metabolic cost[a]		
A	11.7	M	34.3
C	24.7	N	14.7
D	12.7	P	20.3
E	15.3	Q	16.3
F	52	R	27.3
G	11.7	S	11.7
H	38.3	T	18.7
I	32.3	V	23.3
K	30.3	W	74.3
L	27.3	Y	50

[a] The metabolic cost is given in approximate units of the number of high energy phosphate bonds that need to be broken (ATP → ADP, for example) to fuel the reaction.

urealyticum: 25.5%). Note that three of these organisms are from the same genus, yet the G+C content ranges from 26.6% to 40.0%.

Interestingly, our findings show that rather than being confined to the synonymous sites within the genes, nucleotide bias also affects the nonsynonymous codon positions, forcing a change in the amino acid content of the encoded proteins. Bias in the DNA, then, is responsible for shifts in the relative proportions of the FYMINK and GARP amino acids, which are encoded by AT-rich and GC-rich DNA, respectively. We have shown not only that these amino acid shifts are present in the organisms from the three kingdoms, but that they appear to affect nearly every gene within the genome, to varying degrees. This trend leads to an interesting convergence in amino acid composition between

evolutionarily distant taxa. All of the AT-rich genomes are FYMINK-rich and GARP-poor, regardless of the fact that they are from different kingdoms. These same organisms show a very distinct pattern of amino acid usage compared to the GC-rich organisms from their own kingdoms. We have plotted these trends in Figure 4.6, where it is clear that the organisms group together in terms of their nucleotide and amino acid biases rather than by phylogenetic relatedness.

Our results suggest that all sites within proteins that have the freedom to mutate will do so in a biased manner, as directed by the bias in the underlying DNA. Conserved proteins, having fewer mutable sites, show the least degree of amino acid bias for this reason. Conversely, loosely conserved proteins show a very great degree of amino acid bias in the directions predicted by the nucleotide biases in the DNA. If we confine our analysis to only the variable sites within proteins, the difference between highly and loosely conserved proteins disappears. In other words, any site that is free to mutate—whether within a conserved protein or not—is affected by DNA mutation bias. Sueoka (1992) showed that the G+C content of the DNA represented a balance between mutation bias and negative selection. We have shown that the same principle applies at the level of protein sequences. In the next chapter we will further investigate this relationship between negative selection in proteins and mutation bias in the DNA.

It is possible that the amino acid changes in response to nucleotide bias affect the higher-order properties of the encoded protein. Oliver and Marin (1996) have shown that proteins encoded by AT-rich genes tend to be shorter than those encoded by GC-rich genes, for example. The explanation for this relationship was that since stop codons are AT-rich, the likelihood of a nonsense mutation (a mutation from an amino acid-encoding codon to a stop codon) is greatly increased in AT-biased DNA than in GC-biased DNA. Similarly, the likelihood of a nonsynonymous mutation in a stop codon is increased in GC-rich DNA. AT-rich genes have a tendency to be truncated, then, while GC-rich genes have a tendency to be elongated—provided such shortenings do not negatively impact on the fitness of the encoded protein. Our findings suggest, though, that AT-rich DNA

also produces proteins that have a higher metabolic cost than their counterparts in GC-rich DNA. Perhaps gene-shortening in AT-rich genomes is a response to help minimize this increased energy cost. An examination of protein alignments would help solve this mystery: the stop-codon hypothesis predicts that the GC-rich genes will be extended at the 3′ end relative to homologous AT-rich genes, while the energy minimization hypothesis makes no prediction as to the location of gene shortening in the AT-rich genes, since a deletion in any region of the gene lowers the energy requirements of assembling the encoded protein. This is a potentially interesting area for future study.

4.6 Note

Some of the results from this Chapter were published in Singer and Hickey (2000).

5 Both DNA mutation bias and natural selection shape the amino acid composition of membrane-spanning domains

5.1 Abstract

Mutation bias in DNA is known to affect the amino acid compositions of the encoded proteins. This effect is necessarily entwined with the effects of negative selection, which will tend to limit the changes that mutation bias introduces. The nature of this relationship has not been studied previously. We have analysed the helical transmembrane domains in proteins from over 150 metazoan mitochondrial genomes and have verified that the variation in amino acid usage between genomes is a function of both the pattern of nucleotide bias in the genes and the force of natural selection. We have quantified this relationship in a mathematical model that is capable of accounting for nearly 90% of amino acid variation in transmembrane domains. Our findings show that

natural selection is conserving these domains at the biochemical level (as opposed to the amino acid sequence level), leaving some flexibility in their amino acid compositions. It is within these constraints that DNA mutation bias is able to control the amino acid usage.

5.2 Introduction

In the previous chapter, directional DNA evolution was shown to influence the amino acid usage of all proteins studied. The degree to which each protein was affected, however, depended on the degree to which natural selection conserved its sequence: highly conserved proteins were affected to a lesser degree than loosely conserved proteins. We can infer from this finding that DNA mutation bias at the level of the gene must interact with natural selection acting on the protein. This process is not fundamentally different from the relationship purifying selection has with the normal process of mutation in the DNA: mutations always occur at some rate, and most are rejected. Even in very highly conserved proteins, however, some nonsynonymous mutations are selectively neutral and can become fixed in the population. The introduction of a biased pattern of DNA mutation simply produces a directional pattern of amino acid substitution. In this chapter we investigate the details of this relationship in the evolution of helical transmembrane domains.

Alpha-helical membrane-spanning protein domains are highly conserved, since they must maintain certain biochemical properties to be able to pass through the lipid bilayer. These requirements are supplied by amino acids that are strongly hydrophobic, and which have a propensity to form tightly packed α-helical structures (Kyte and Doolittle, 1982; Cornette et al, 1987; von Heijne, 1995; Eilers et al, 2000; Deber et al, 2001). These characteristics are highly consistent from one protein to the next and even from one organism to the next, allowing transmembrane domains to be predicted with very good accuracy (Rost et al, 1996; Tusnády and Simon, 1998). Although selection clearly constrains the course of evolution within these domains, the conservation is not complete—at least not at the sequence level. It seems, though, that protein hydropathy can be

maintained throughout a range of nonsynonymous nucleotide biases (Mrázek and Kypr, 1992). Particularly relevant to our study is that it is known that mutational biases in the DNA are capable of affecting the amino acid composition within transmembrane domains, in a manner that is very similar to the trends we have shown in the previous chapters (Stevens and Arkin, 2000). It would seem, therefore, that both natural selection and DNA mutation bias are important factors in the overall amino acid usage within transmembrane domains.

To study the interaction between DNA bias and natural selection, we focused on the completely sequenced metazoan mitochondrial genomes. In addition to being a large and diverse data set (there are over 200 completely sequenced genomes at the time of writing, from ten different phyla), metazoan mitochondrial genomes have a very consistent set of genes, allowing for direct comparison between different genomes. Finally, mitochondrial proteins are known to have multiple membrane-spanning regions (making up fully half of the total protein content); yet, previous research has shown that these same proteins are subject to amino acid substitution biases caused by the underlying DNA mutation bias in the genes that encode them (Foster et al, 1997). For these reasons, the mitochondrial genomes appeared to be ideal for our study.

5.3 Methods

5.3.1 Data selection

At the time of writing, over 200 metazoan mitochondrial genomes are available for downloading from GenBank[1]. Since many of the genomes show a significant strand-specific nucleotide bias, it was necessary to consider the strand on which each gene was encoded (Asakawa et al, 1991; Reyes et al, 1998; Boore, 1999; Saccone et al, 1999). Fortunately, most genomes encode the majority of their genes on one strand of the chromosome, leaving 157 genomes from ten phyla for

[1]ftp://ftp.ncbi.nih.gov/genbank/genomes/MITOCHONDRIA/Metazoa/

our analysis[2]. These genomes shared eleven genes with each other: ATP6, CYTB, COX1, COX2, COX3, ND1, ND2, ND3, ND4, ND4L, and ND5.

We found that the phyla were not equally represented among the remaining genomes: the vast majority of them are from chordates, with smaller proportions coming from the other taxa. Because of this sequencing bias, we decided to randomly select only two genomes from each phylum (or one, if there was only a single representative), and perform our analyses on this reduced set of data. Although this procedure left us with a mere 17 genomes (Table 5.1), we show later in the Results, Section 5.5.5, that these genomes represent the whole data set very well.

5.3.2 Identification of transmembrane domains

The genes listed in the previous paragraph were extracted from each genome's GenBank file, and were then submitted to the TMHMM 2.0 server[3] (Krogh et al. 2001), which is believed to be the most accurate method of alpha-helical transmembrane domain prediction available (Moller et al. 2001). The predicted transmembrane domains were then removed from each protein sequence, and their corresponding DNA sequences were removed from each gene sequence. All of the non-transmembrane and transmembrane sequences (protein and DNA) for each genome were separately concatenated together to form two long sequences (per genome) that were analysed as a whole. Thus, for each genome there were two protein sequences (and their corresponding DNA sequences), corresponding to the transmembrane and non-transmembrane regions of the proteome. All analyses were performed using programs written by GS using the Python programming language[4].

---

[2]Although the extremely AT-rich genomes do encode their genes on both strands of the chromosome, they also lack a strand-specific nucleotide bias, so they were included in our analysis (Perna and Kocher, 1995; Saccone et al, 1999).

[3] http://www.cbs.dtu.dk/services/TMHMM/

[4] http://www.python.org

Table 5.1: The seventeen randomly selected mitochondrial genomes that were analysed in this study, including at most two representatives from each of ten phyla. We have also included the GenBank version for each genome, the percentage of each nucleotide, and the number of nucleotides in the transmembrane and non-transmembrane portions of the protein-encoding genes we analyzed.

Organism	Phylum	Version	A	C	G	T	TM	Non-TM
Platynereis dumerilii	Annelida	NC_000931.1	30	21	16	33	5625	4754
Ceratitis capitata	Arthropoda	NC_000857.1	31	12	13	44	5784	4803
Chrysomya chloropyga	Arthropoda	NC_002697.1	31	12	13	44	5667	4830
Terebratalia transversa	Brachiopoda	NC_003086.1	17	13	28	41	5712	4560
Terebratulina retusa	Brachiopoda	NC_000941.1	27	30	14	29	5598	4890
Dogania subplana	Chordata	NC_002780.1	34	28	11	26	5379	5253
Poromitra oscitans	Chordata	NC_003172.1	25	33	15	26	5685	5055
Metridium senile	Cnidaria	NC_000933.1	25	17	20	38	6033	5067
Arbacia lixula	Echinodermata	NC_001770.1	27	21	17	35	6054	4785
Florometra serratissima	Echinodermata	NC_001878.1	27	12	16	46	6252	4527
Balanoglossus carnosus	Hemichordata	NC_001887.1	23	33	17	28	5847	5007
Pupa strigosa	Mollusca	NC_002176.1	25	18	21	36	5526	4641
Venerupis (Ruditapes) philippinarum	Mollusca	NC_003354.1	27	10	21	42	5796	6381
Ancylostoma duodenale	Nematoda	NC_003415.1	26	07	18	50	5631	4218
Ascaris suum	Nematoda	NC_001327.1	19	08	22	52	5654	4305
Fasciola hepatica	Platyhelminthes	NC_002546.1	14	10	27	49	5853	3798
Paragonimus westermani	Platyhelminthes	NC_002354.1	13	18	30	38	5192	3701

5.3.3 Quantification of nucleotide bias

To quantify the effect that DNA mutation bias has on the amino acid composition of proteins, we decided to calculate the "expected" amino acid frequencies we would see if DNA bias were the only factor governing protein evolution. By comparing these expected frequencies to the actual amino acid frequencies in the proteins, we could estimate the degree to which DNA bias influences amino acid content. We measured the nucleotide frequencies at 4-fold degenerate codon positions, and used these nucleotide frequencies to calculate our expected codon frequencies, which could then be translated into amino acid frequencies (see Figure 5.1). Since synonymous nucleotide positions in the gene are essentially free of selective constraint, the expected amino acid frequencies that we calculate are a function of the DNA bias alone. The degree to which the actual amino acid frequencies of the protein correlate with these expected values is, in our opinion, a good estimator of the degree to which DNA mutation bias governs the amino acid usage within the protein.

Table 5.2: The Kuhn and Leigh (1985) hydrophobicity scale. Negative integers indicate hydrophilic residues; positive integers indicate hydrophobic residues.

Amino Acid	Hydrophobicity		
H	-2.80	S	-0.87
K	-2.50	Y	-0.79
N	-2.40	M	-0.21
Q	-2.40	T	0.04
R	-2.40	G	0.95
E	-2.30	F	1.05
D	-2.10	A	2.20
C	-2.00	V	2.90
P	-1.60	I	3.30
W	-1.10	L	5.00

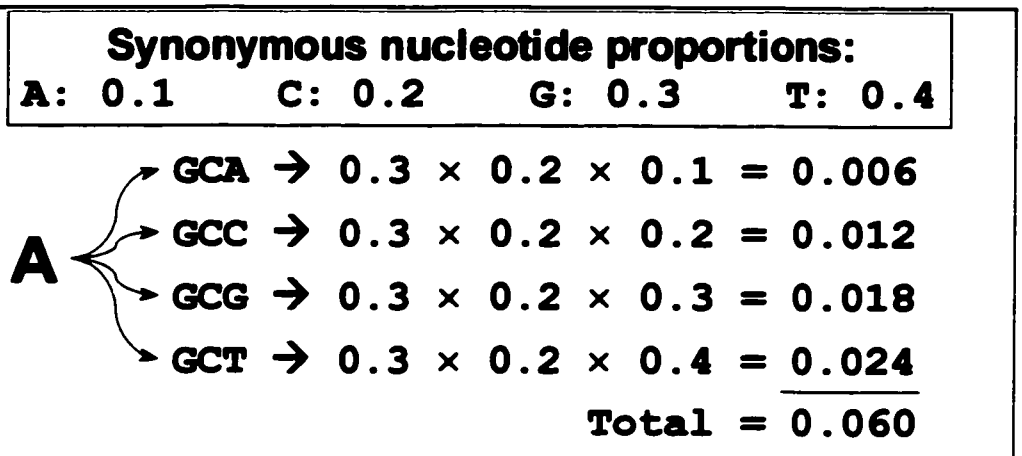


Figure 5.1: Measurement of the "expected" proportion of alanine encoded by a gene evolving under a completely neutral model with the nucleotide frequencies shown. Here, the expected proportion of alanine is 0.060, based on the expected proportions of its four codons. The synonymous nucleotide proportions are measured from an actual gene, and are assumed to be the same as what the overall base proportions would be if the gene were unconstrained by natural selection. Correspondence between this predicted value and the actual proportion of each amino acid in the protein gives an indication of the degree to which DNA bias in the gene is governing amino acid usage in the protein.

5.4 Qualitative Results

5.4.1 Transmembrane domains are more hydrophobic than non-transmembrane regions in proteins

Previous research has shown that there is strong selective pressure for hydrophobic residues in transmembrane domains (Kyte and Doolittle, 1982; Cornette et al, 1987; von Heijne, 1995; Eilers et al, 2000; Tourasse and Li, 2000; Deber et al, 2001). Therefore, we decided to begin our investigation by confirming the relationship between amino acid usage and hydrophobicity in the transmembrane domains from mitochondrial proteins. Figure 5.2 shows that the transmembrane domains within the proteins we studied are indeed much more hydrophobic than the non-transmembrane portions of the same proteins, as

measured using the Kuhn and Leigh (1985) hydropathy scale (Table 5.2). The increased hydrophobicity of the transmembrane domains is being maintained by a pattern of negative selection that is clearly distinct from that within the non-transmembrane protein regions. We were curious to determine how this different pattern of selection on transmembrane domains affected the relationship between nucleotide bias and amino acid usage that we have discussed in the previous chapters.

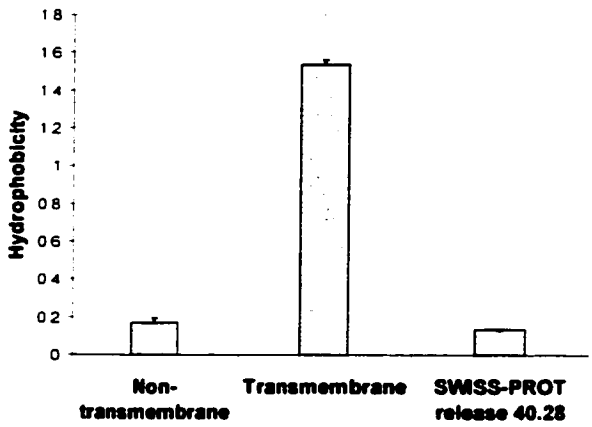


Figure 5.2: The average amino acid within transmembrane domains is much more hydrophobic than the average amino acid in the non-transmembrane regions from the same proteins. Note that if these values are compared to the average hydrophobicity of an amino acid in all of SWISS-PROT (release 40.28), it is apparent that the difference in hydrophobicity between transmembrane and non-transmembrane protein segments is completely due to positive selection for increased hydrophobicity in the transmembrane domains. The non-transmembrane domains do not differ significantly from the "average" protein (in SWISS-PROT) in terms of hydrophobicity ($p>0.1$ in a one-sample t-test). The hydrophobicities were measured using the Kuhn and Leigh (1985) scale.

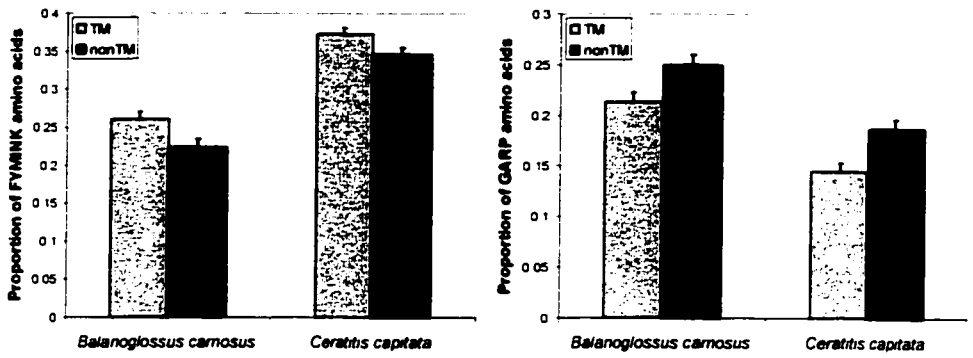


Figure 5.3: The AT-rich *C. capitata* genome encodes proteins that are more FYMINK-rich (left) and GARP-poor (right) than proteins from the GC-rich *B. carnosus* genome. Similar trends are observed both within and outside of transmembrane domains.

5.4.2 Comparison of AT-rich and GC-rich mitochondrial genomes shows that FYMINK and GARP change as expected, but are still subject to purifying selection acting on the transmembrane domains

Foster et al (1997) showed that the GARP and FYMINK amino acid proportions within mitochondrial proteins are affected by DNA composition biases—the same trend that we have reported among bacterial and eukaryotic genomes in the previous chapters. In light of the high degree of selection acting upon transmembrane domains though, we wondered whether the Foster et al results were entirely confined to the non-transmembrane portions of the mitochondrial proteins, or if the amino acid composition within the transmembrane domains was also affected by DNA bias. Since it is known that transmembrane domains evolve slower than the other parts of the protein (Tourasse and Li, 2000), it is logical to believe that transmembrane domains are relatively unaffected by DNA mutation biases. We have chosen the most AT-rich mitochondrial genome (*Ceratitis capitata*, 75.2% AT in the coding sequences) and the most GC-rich

(*Balanoglossus carnosus*, 49.5% GC in the coding sequences) and have compared their homologous proteins. Unlike the Foster et al analysis however, we have separated the transmembrane and non-transmembrane portions of the proteins and have analysed these two regions separately. The results of these analyses are presented in Figure 5.3. In both the transmembrane and non-transmembrane protein regions, the relatively GC-rich *B. carnosus* has a higher proportion of the GARP amino acids and a lower proportion of the FYMINK amino acids than the corresponding regions in the AT-rich *C. capitata*. Interestingly, these two groups of amino acids change in proportionately similar ways in both protein regions. At first glance then, the transmembrane domains appear to be able to change their amino acid compositions as easily as the non-transmembrane protein segments. This finding is intriguing, since the amino acid compositions of these two regions must be very different to account for their distinct hydrophobicities (Figure 5.2). Indeed, the Figure does show an amino acid composition difference between the two protein regions. Note that the transmembrane domains have a slightly lower proportion of the GARP amino acids and a higher proportion of the FYMINK amino acids than the non-transmembrane protein regions in each organism ($p < 0.05$ for FYMINK in *C. capitata*; $p < 0.004$ for other comparisons). This trend is likely due to selection for hydrophobicity, since the two hydrophobic amino acids in the GARP group—glycine and alanine—have an average hydrophobicity of 1.6 in the Kuhn and Leigh (1985) scale, while the two hydrophobic amino acids in the FYMINK category—phenylalanine and isoleucine—have an average hydrophobicity of 2.15. This difference in amino acid composition between the protein regions is dwarfed by the difference between the two organisms, however. The transmembrane regions in *B. carnosus*, for example, have less FYMINK than the non-transmembrane regions in *C. capitata*. This simple result indicates that intra-protein amino acid variation is governed by selection for hydrophobicity (FYMINK proportions are higher and GARP proportions are lower in transmembrane regions than in non-transmembrane regions within proteins), while nucleotide bias appears to affect the inter-genome amino acid variation

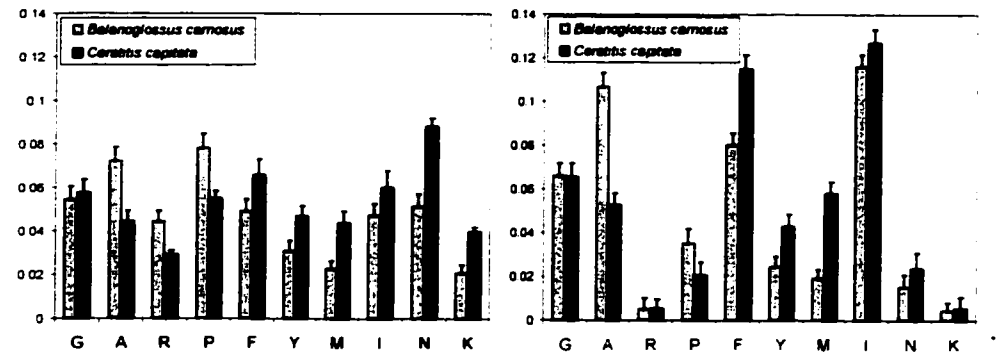


Figure 5.4: Comparison of individual amino acid proportions in proteins from the AT-rich *C. capitata* and the relatively GC-rich *B. carnosus* show that the GARP and FYMINK amino acids have changed in predictable ways. The pattern of change is quite different, however, between the non-transmembrane (left) and transmembrane (right) protein regions.

(FYMINK proportions are higher and GARP proportions are lower in the AT-rich *C. capitata* than in the relatively GC-rich *B. carnosus*).

The effects of natural selection on the amino acid composition of transmembrane domains are clearer when we examine the proportions of individual amino acids. While the FYMINK and GARP groups of amino acids show similar interspecies variation within both transmembrane and non-transmembrane protein regions (Figure 5.3), the amino acids within those groups show very different patterns of variation (Figure 5.4). Although each of the amino acids in the GARP and FYMINK categories varies in the predicted direction within both the transmembrane and non-transmembrane protein segments, the effect is evenly spread across all of the amino acids in the non-transmembrane segments, but is concentrated in only a few amino acids within the transmembrane domains. Within the transmembrane domains, hydrophilic amino acids like arginine, proline, asparagine, and lysine occur in very low proportions overall, and show little difference between the two organisms. On the other hand, the hydrophobic amino acids like alanine, phenylalanine, and isoleucine occur in high proportions overall, but tend to show greater levels of variation between the two organisms. In

both the non-transmembrane and the transmembrane protein regions, then, there is pressure from biases in the underlying DNA to change the relative proportions of the GARP and FYMINK amino acids. Since the non-transmembrane domains have no specific amino acid requirements, these amino acid changes occur roughly equally across the amino acids in the GARP and FYMINK groups. Conversely, the transmembrane domains do have a specific amino acid requirement that favours hydrophobic amino acids and disfavours the hydrophilic amino acids. Therefore, regardless of the underlying pressure from the DNA to increase R and P proportions in the *B. carnosus* transmembrane domains, the strong selection against these amino acids limits the degree to which this change can occur. Alanine, on the other hand, is favoured by selection and is therefore able to increase greatly. In fact, even though the magnitude of the GARP increase is the same between transmembrane domains and non-transmembrane protein regions (Figure 5.3), nearly all of this change is confined to alanine within the transmembrane domains (Figure 5.4). Similarly, phenylalanine and methionine absorb nearly all of the change among the FYMINK amino acids. These interesting patterns of amino acid usage show elements of both selection for hydrophobicity and GC/AT DNA mutation bias. In the next section we proceed to measure these two factors more accurately and generate a quantitative model of the effects of both selection and mutation bias on amino acid usage within the transmembrane domains.

5.5 Quantitative Results

5.5.1 Intra-genomic amino acid variation in transmembrane domains is the result of selection for hydrophobicity

Although the previous section demonstrated the relationship between negative selection and DNA mutation bias in a qualitative way, we decided to expand these results into a quantitative model. To begin with, we have already shown that the average amino acid within transmembrane domains has a much higher

hydrophobicity than the average amino acid within non-transmembrane protein regions, or within an average amino acid within all of SWISS-PROT (Figure 5.2). The increased hydrophobicity within transmembrane domains is likely the result of positive selection for hydrophobic amino acids coupled with selection against hydrophilic amino acids. We confirmed this trend by measuring the amino acid contents in the transmembrane domains from seventeen completely sequenced mitochondrial genomes (the details of the selection of these genomes is given in Section 5.3.1). Figure 5.5 shows that there is indeed a positive relationship between the hydrophobicity of an amino acid (as measured using the Kuhn and Leigh (1985) scale) and its proportion within transmembrane domains. Indeed, hydrophobicity alone can explain 75% of the variation in usage between amino acids. Nevertheless, this still leaves 25% of the variation in amino acid usage unexplained.

Note that although there is no between-genome variation on the x-axis position of each amino acid in Figure 5.5 (since each amino acid's hydrophobicity is constant and cannot vary from one genome to the next), there is a great deal of variation

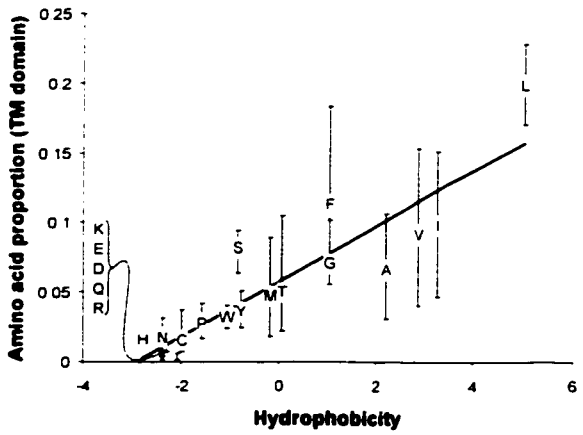


Figure 5.5: The proportion of an amino acid in transmembrane domains is correlated to its hydrophobicity (r^2 = 0.75). Shown are the means across the seventeen genomes analysed, with the bars indicating the range in values (rather than the standard deviation, to present the actual spread in the data).

along the y-axis. In other words, the amino acid content of transmembrane domains varies from genome-to-genome (a trend that we also observed in the two genomes analysed in Section 5.5.2). Interestingly, the inter-genomic variability appears to be greater for the hydrophobic amino acids than for the hydrophilic amino acids. This is most apparent among the most hydrophilic amino acids, where inter-genomic variation is virtually nonexistent. In Figure 5.6 we show that there is indeed a weak but positive relationship between the inter-genomic variability of a particular amino acid and its hydrophobicity. We have confirmed that this is true even if we measure the variance proportionally (i.e., the hydrophobic amino acids have a disproportionately higher variance than the hydrophilic amino acids). Since hydrophilic amino acids are strongly selected against in transmembrane domains, there is little room for variability: they must occur at very low levels in all genomes. Among the hydrophobic amino acids however, a great deal of variability is possible so long as the domain as a whole remains sufficiently hydrophobic. A clear exception to this trend is the amino acid leucine—the most hydrophobic amino acid—which shows a surprisingly low degree of inter-genomic variability. We have identified two possible explanations for this deviation. First, natural selection may be keeping leucine in high proportion within transmembrane domains. In this respect, then, its pattern of variation would be similar to that of the hydrophilic amino acids, since selection would allow little room for variability from one genome to the next. Alternatively, the factors that cause the inter-genomic variability of some amino acids (like nucleotide bias) may have no effect on leucine. In Section 5.4.2 for example, we showed that the variation in AT/GC contents of two mitochondrial genomes affected the GARP and FYMINK groups of amino acids—leucine proportions, therefore, are not expected to change in response to changes in the GC content of the underlying DNA.

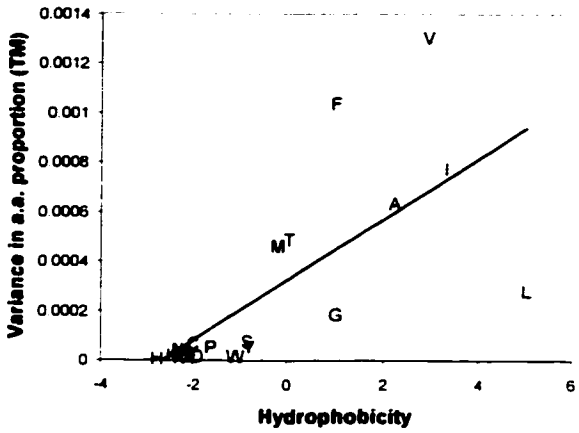


Figure 5.6: The inter-genomic variability in usage for a particular amino acid goes up with increasing hydrophobicity (r^2 = 0.52; 0.75 without leucine). In other words, amino acids that are hydrophilic have proportions that do not vary from one genome to the next, while hydrophilic amino acids can be quite variable.

5.5.2 Inter-genomic amino acid variation affects both transmembrane and non-transmembrane protein regions

25% of amino acid variation within helical transmembrane domains cannot be explained by selection for hydrophobicity (Figure 5.5). The remaining variation occurs between genomes: that is, the amino acid usage within transmembrane domains varies from one genome to the next. Our preliminary results from Section 5.4 indicated that GC/AT biases in the DNA were responsible for some amino acid usage differences between genomes. Nevertheless, we cannot rule out other factors, like positive natural selection, that may be causing a directional change in the amino acid compositions of these transmembrane domains. We can test for positive selection within transmembrane domains using the non-transmembrane protein segments as a control. If positive selection is acting upon the transmembrane domains, it should have no effect on the amino acid composition of the non-transmembrane regions of the same proteins unless the entire protein is subject to the same selection pressures. Conversely, if directional nucleotide bias in the DNA were affecting the transmembrane amino acid

proportions, we would expect it to also affect the amino acid composition of the non-transmembrane amino acid proportions. In Figure 5.7 we have plotted the transmembrane proportion of each amino acid versus its proportion in the non-transmembrane protein regions for each genome. We have plotted each amino acid on a different scale to ensure that the variation between genomes is clearly visible. Although some amino acids show a weaker correlation than others, in every case there is a positive relationship between the transmembrane and non-transmembrane proportions of the amino acid. This means that the directional changes to the amino acid composition within the transmembrane domains also occur in the non-transmembrane portions of the same proteins. This is a result that is consistent with the notion that a bias in the underlying DNA is causing directional changes to the amino acid sequence in the encoded proteins.

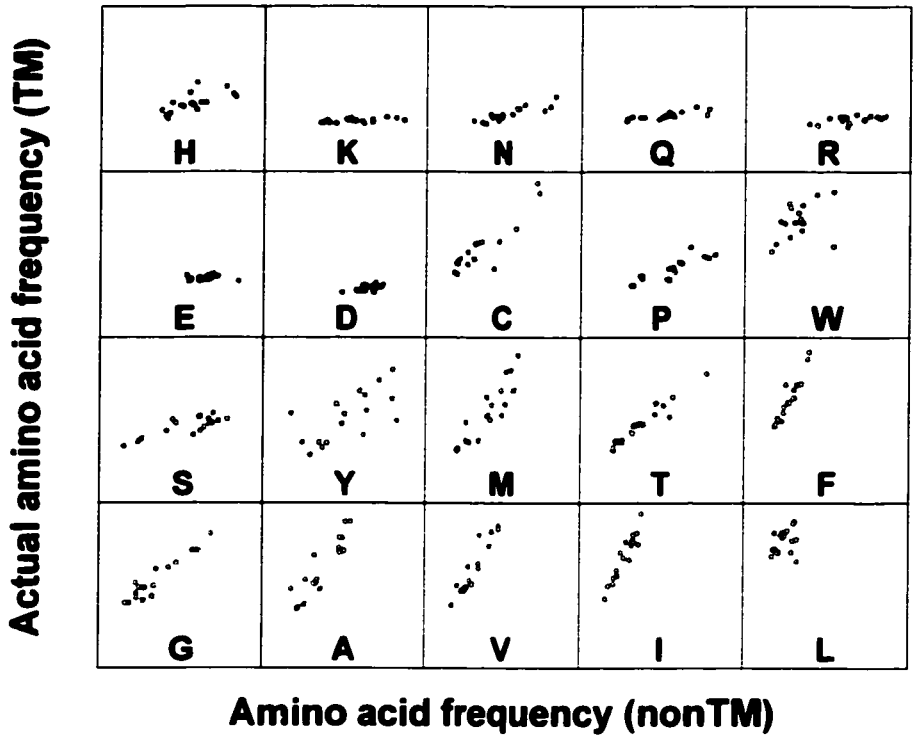


Figure 5.7: The proportion of each amino acid in transmembrane domains is positively correlated to its proportion in the non-transmembrane protein regions. In each plot, the line $x=y$ is indicated: if the points tend to fall above the line, there is a higher proportion of that amino acid in the transmembrane domains than in the non-transmembrane portions of the proteins. Conversely, points falling below the line indicate a higher proportion in the non-transmembrane regions. The amino acids are shown in order of increasing hydrophobicity to make the relationship between slope and hydrophobicity more apparent: hydrophobic amino acids tend to have a slope greater than one, while hydrophilic amino acids tend to have a slope less than one. An interesting exception to this trend is W, which is a hydrophilic amino acid but appears to occur in the transmembrane domains with a higher frequency than in the non-transmembrane domains. This could be related to other positively selected biochemical properties, like helix stabilization, that we are not considering.

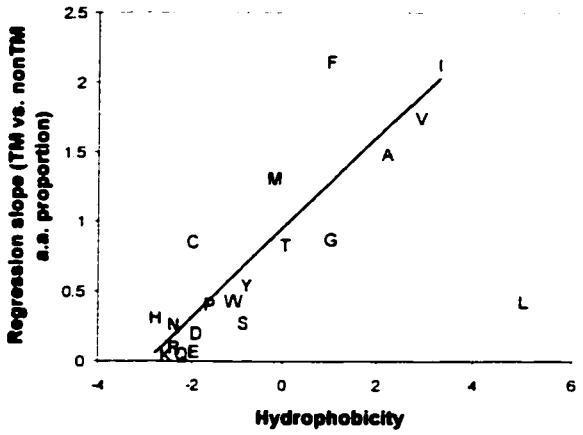


Figure 5.8: The regression slope between transmembrane and non-transmembrane amino acid proportions increases with amino acid hydrophobicity (r^2 = 0.79 without leucine; r^2 = 0.46 when leucine is included: the regression line in the figure does not include leucine). In other words, hydrophobic amino acids are acquired more readily in transmembrane domains than they are in non-transmembrane protein regions. On the other hand, hydrophilic amino acids are acquired more readily in non-transmembrane regions.

5.5.3 Using non-transmembrane amino acid frequencies to predict the frequencies within transmembrane domains

Directional changes to the amino acid composition are not confined to the transmembrane domains within proteins; similar changes occur simultaneously within the non-transmembrane protein regions (Figure 5.7). Within the transmembrane domains however, we know that there is negative selection to maintain a high level of hydrophobicity (Figure 5.5). The effects of natural selection on directional protein evolution are actually apparent in Figure 5.7 If amino acids are gained and lost at the same rate in both the transmembrane and non-transmembrane portions of the proteins, the points will fall along the dotted line in the middle of each plot. On the other hand, if the points tend to fall below the line, the amino acid is changing at a slower rate in the transmembrane domains, while if the points fall above the line the amino acid is changing at a faster rate within transmembrane domains relative to the non-transmembrane

protein portions. We note that the amino acids whose points fall below the line tend to be hydrophilic—amino acids like aspartate, glutamate, lysine, and arginine, which are selected against within the transmembrane domains. Conversely, hydrophobic amino acids like alanine, phenylalanine, isoleucine, and valine all have points that fall above the line, since they are being positively selected for in transmembrane domains. We have plotted the slope of the regression line for each amino acid versus its hydrophobicity in Figure 5.8. Here it is quite clear that those amino acids with a shallow slope in Figure 5.7 are hydrophilic, while those with a steep slope are hydrophobic. Curiously, leucine does not follow this pattern, and in fact shows little correlation between the transmembrane and non-transmembrane regions of the proteins. This is because leucine shows very little inter-genomic variation, which as we discussed in Section 5.5.1, may either indicate strong selection for a high leucine proportion in the mitochondrial proteins, or it may simply be that the factor that is causing the inter-genomic amino acid variation does not affect leucine.

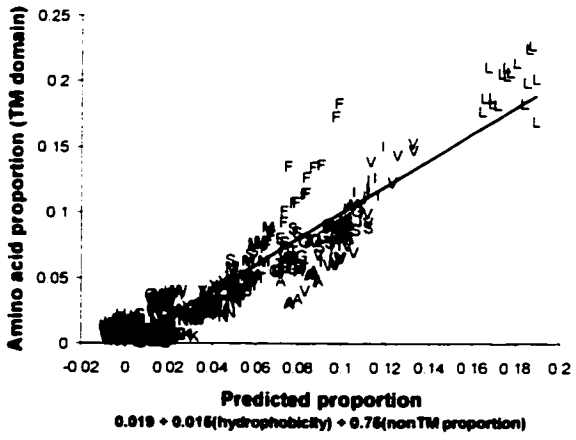


Figure 5.9: A regression model based on the hydrophobicity and the non-transmembrane proportion of an amino acid is an excellent predictor of the transmembrane proportion of that amino acid (r^2 = 0.87). The model is an improvement over the one that only takes the hydrophobicity of an amino acid into account (Figure 5.5), because inter-genomic variation has been accounted for in this model, by way of the non-transmembrane amino acid proportions.

We have already shown that amino acid hydrophobicity is an independent measure of the effect of selection acting upon transmembrane domains (Figure 5.5). We have also shown that the non-transmembrane portions of the same proteins may be a good independent measure of the inter-genomic variation in amino acid usage within transmembrane domains, since the non-transmembrane protein regions appear to be free of the compositional constraints that affect transmembrane domains. Combining these two factors—hydrophobicity and non-transmembrane amino acid proportions—we built a model that can predict the amino acid frequencies within transmembrane domains (Figure 5.9). Note that by considering the proportion of the amino acid in the non-transmembrane regions of the proteins (as an independent measure of inter-genomic amino acid variation) we have greatly improved upon the regression model that considers hydrophobicity alone (Figure 5.5). The success of this model indicates that the amino acid usage *outside* of the transmembrane domains is a good predictor of what the amino acid usage *inside* of the transmembrane domains will be. Again, this is a finding that is consistent with the notion that these proteins have evolved in a directional manner in response to DNA mutation bias.

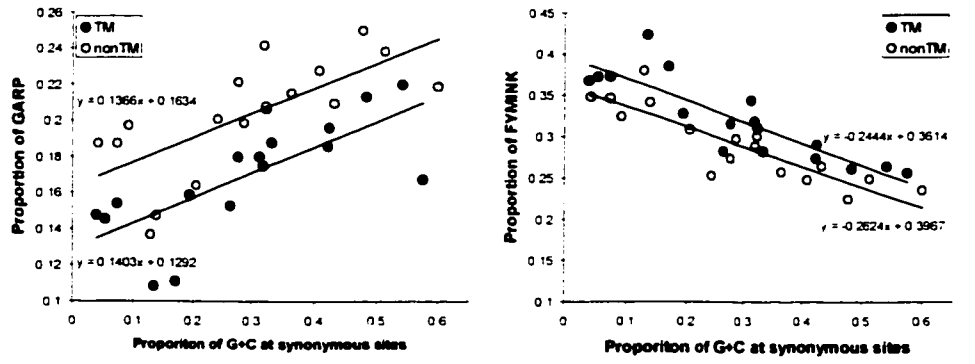


Figure 5.10: The proportions of GARP (left) and FYMINK (right) in relation to synonymous GC content in transmembrane domains and non-transmembrane protein regions. The slopes of the regression lines are virtually identical (0.137 vs. 0.140 for GARP; and -0.244 vs. –0.262 for FYMINK). The y-intercepts of the lines are different, though: transmembrane domains have slightly more FYMINK and less GARP than the non-transmembrane regions of the proteins.

5.5.4 The cause of inter-genomic amino acid variation is nucleotide bias, not selection

Although our results thus far have been consistent with DNA mutation bias acting upon the amino acid composition of these mitochondrial proteins, we cannot completely rule out positive selection. It is plausible that there is positive selection on whole-protein amino acid composition changes, which is why the inter-genomic variability in amino acid usage is correlated between the non-transmembrane and transmembrane regions of the proteins. To demonstrate that this latter scenario is unlikely however, we have plotted the GARP and FYMINK amino acid proportions against the synonymous G+C contents for both the transmembrane and non-transmembrane protein regions (Figure 5.10). Just as we discovered in Section 5.4.2, GC/AT mutation biases in the DNA appear to be affecting the amino acid compositions of the proteins that the DNA encodes. Moreover, the slopes of these regression lines are virtually identical between the different protein domains.

Although GC/AT mutation biases can explain the inter-genomic variation in the GARP and FYMINK amino acids, it is incapable of explaining the inter-genomic variation observed in other amino acids, like valine and threonine for example (Figure 5.5). A closer examination of the pattern of nucleotide bias in the mitochondrial genomes reveals the reason for this discrepancy: in addition to AT/GC mutation biases, animal mitochondrial genomes also have a strand-specific nucleotide bias in which one strand becomes rich in the G and T nucleotides while the other strand becomes rich in A and C (Asakawa et al. 1991; Reyes et al. 1998; Boore, 1999; Saccone et al. 1999). These two biases are related to each other: very AT-rich genomes tend to show little or no strand bias at all (since it is essentially "washed out" by the AT-richness), but the more GC-rich genomes tend to have a very strong skew in the nucleotide composition from one strand to the other (Perna and Kocher, 1995; Saccone et al. 1999). Within these latter genomes, genes encoded on one strand will be very AC-rich, while those on the other strand are very GT-rich (this complicated the selection of our data set, as

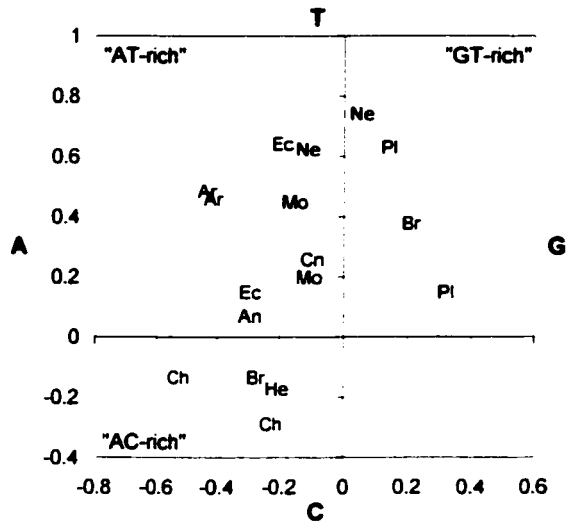


Figure 5.11: The pattern of synonymous nucleotide bias in the coding sequences from seventeen mitochondrial genomes analysed. Each genome is indicated by its abbreviated phylum name (An=Annelida, Ar=Arthropoda, Br=Brachiopoda, Ch=Chordata, Cn=Cnidaria, Ec=Echinodermata, He=Hemichordata, Mo=Mollusca, Ne=Nematoda, and Pl=Platyhelminthes). The sides of the plot to which the genome is closest indicates the direction of nucleotide bias, as labeled in the corners of the plot. Interestingly, no "GC-biased" genomes exist.

described in the Methods, Section 5.3.1). These combined factors together allow us to group mitochondrial genes (and genomes) into three categories: AT-rich, GT-rich, and AC-rich (Figure 5.11). Interestingly, no mitochondrial genomes are "GC-rich" in their synonymous nucleotide positions.

Previous studies have shown that strand-specific DNA biases like those in the mitochondria are capable of affecting the amino acid composition of the encoded proteins (Lafay et al, 1999; Mackiwicz et al, 1999). However, just as AT/GC measurements do not fully encompass the variation in nucleotide usage in mitochondrial genomes, measurements of the FYMINK and GARP amino acid proportions do not fully encompass all of the variation in the inter-genomic amino acid usage that is caused by strand-specific nucleotide biases. For this reason we used a more complicated analysis to measure the relationship between DNA mutation bias and amino acid usage. We used the synonymous nucleotide

frequencies of the genes to calculate "expected" codon frequencies, which were then translated into "expected" amino acid frequencies (this procedure is described in detail in the Methods, Section 5.3.3). The amino acid frequencies that we calculated are the values we would expect the actual protein to have if nucleotide bias were the only factor controlling its evolution (i.e., the gene is evolving under a completely neutral model of evolution). The correlation between these two values indicates the degree to which the evolution of the actual protein sequence is governed by nucleotide bias. The results of this analysis are presented in Figure 5.12, and show that nucleotide bias does play a significant role in determining the amino acid usage of transmembrane domains. We have presented these same results in a different manner in Figure 5.13, where the relationship between the actual and predicted amino acid frequencies is plotted separately for each amino acid. Although the predicted values are quite different from the actual values, the trend is usually predicted correctly (note that most of the slopes are positive). In other words, the nucleotide bias-based predictions can fairly accurately predict which genomes will have higher levels of a particular amino acid and which genomes will have a lower proportion. This result offers strong support to the notion that it is DNA mutation bias that is affecting the amino acid usage in the transmembrane domains, rather than *vice versa*, since selection for directional changes to the amino acid composition of these proteins would have no effect on the nucleotide composition of the synonymous sites within the genes. Although positive selection is unlikely to be affecting the transmembrane domains, the effects of negative selection are quite apparent in this Figure. We note that the relationship between the actual amino acid frequencies and the values we predicted seem to improve with increasing amino acid hydrophobicity. Put another way, hydrophilic amino acids do not vary the way we would expect them to, given the underlying nucleotide composition of the genes. Presumably, this is because there is negative selection against the accumulation of these amino acids. Conversely, hydrophobic amino acids are not selected against, and therefore their proportions are allowed to vary in response to nucleotide bias.

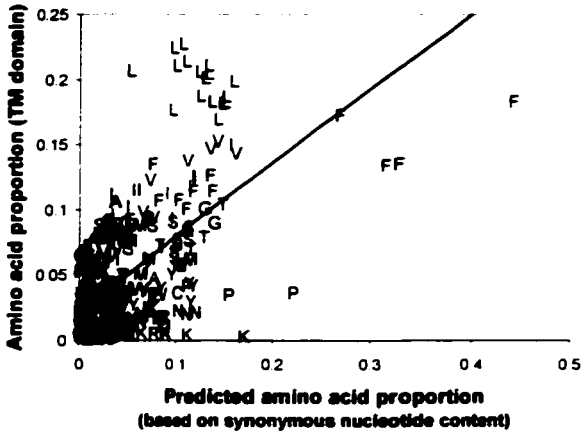


Figure 5.12: The true amino acid proportions within transmembrane domains resembles the predictions we made under the assumption that protein evolution is governed solely by DNA mutation bias. Although the relationship is not strong (r^2 = 0.37), it is highly significant (p = 0); thus, nucleotide bias plays a small but significant role in the evolution of transmembrane domains.

5.5.5 Both nucleotide bias and selection affect the amino acid usage within transmembrane domains

In Figure 5.7 we established that the amino acid compositions within the non-transmembrane protein regions are good predictors of the amino acid compositions within the transmembrane domains. This relationship shows that the directional amino acid substitutions occur both within and between the transmembrane domains within a protein. In the previous section we showed that the directional amino acid changes appear to be caused by mutation biases in the DNA (Figure 5.13). Thus, synonymous nucleotide composition within genes is another independent estimator of the amino acid usage within the transmembrane domains of their encoded proteins. Just as we were able to accurately predict the amino acid composition within transmembrane domains by combining effects of selection (as measured by amino acid hydrophobicity) and inter-genomic amino acid variation (as measured by the non-transmembrane amino acid proportions), we can now build a new model in which hydrophobicity and synonymous

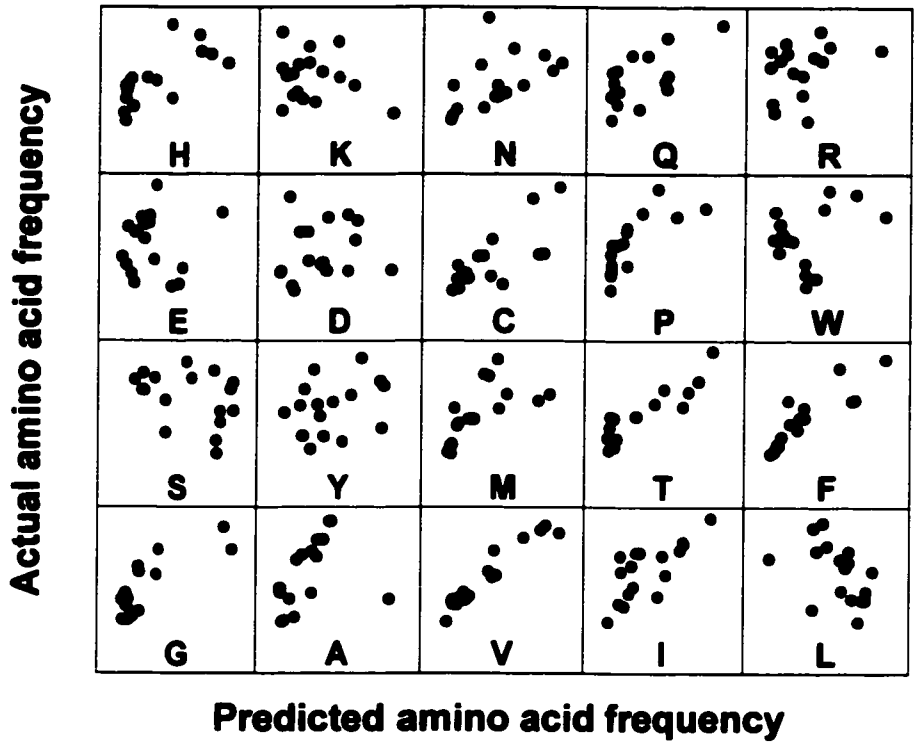


Figure 5.13: The relationship between the actual amino acid proportions in transmembrane domains and the expected proportion, based on the synonymous nucleotide bias of the genes. Although the values of the axes have been removed for the sake of simplicity, the important feature to note is that, in general, there is a positive relationship between the actual and predicted amino acid frequencies. The amino acids are listed in order of increasing hydrophobicity to illustrate that the relationship between the actual and predicted amino acid frequencies improves with hydrophobicity, in general. Red = AT-rich genomes, green = GT-rich genomes, and blue = AC-rich genomes.

nucleotide contents are combined. Figure 5.14 shows how well this model performs.

In Figure 5.4, we could see the effects of both selection and nucleotide bias on the amino acid composition of the transmembrane domains. We have now quantified these two factors and have combined them into a model that is capable of explaining about 85% of the amino acid variation in transmembrane domains. Naturally, we wondered whether this regression model is generally applicable to

all transmembrane domains, or whether it is specific to the data set that we used to build it. Figure 5.15, however, shows that the same model can explain the vast majority of the amino acid variation among all the available mitochondrial genomes. Indeed, we have even applied the same formula to transmembrane domains from bacterial genomes, and the model is able to accurately predict the amino acid proportions in them, as well (Figure 5.16). Particularly interesting is that one of the bacteria, *Mycobacterium tuberculosis*, has GC-rich DNA—a pattern of nucleotide bias that was not included in the dataset used to create the model (Figure 5.17). Nevertheless, our model still correctly predicts the proportions of its amino acids.

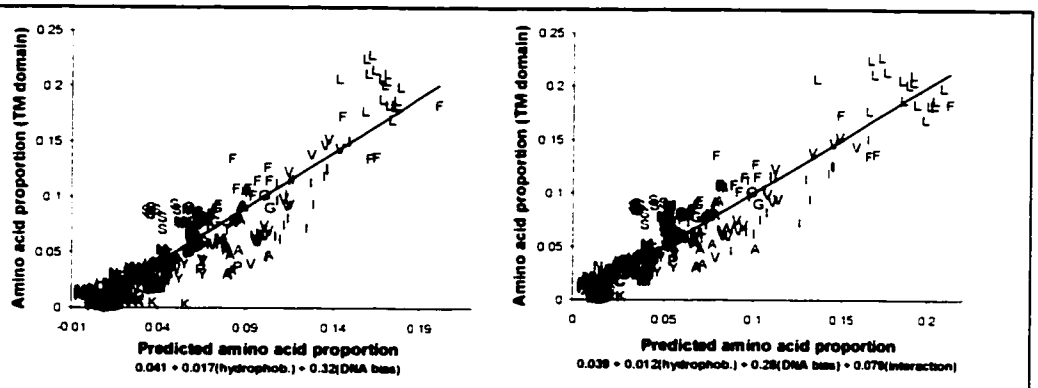


Figure 5.14: (Left) A regression model that takes both hydrophobicity and nucleotide bias into account predicts amino acid usage in transmembrane domains better than simply considering hydrophobicity alone (r^2 = 0.85, vs. 0.75 when only hydrophobicity is considered, as shown in Figure 5.5). (Right) When a term is included in the model to account for interaction between nucleotide bias and hydrophobicity, the r^2 increases slightly to 0.88.

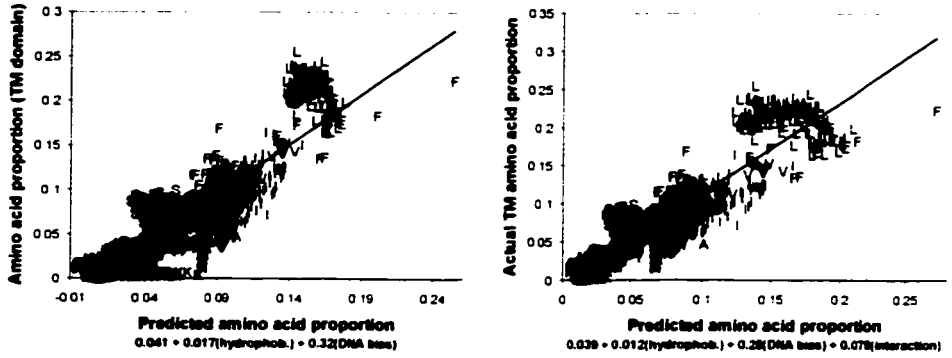


Figure 5.15: (Left) The model built using seventeen mitochondrial genomes works very well when applied to all of the mitochondrial genomes (r^2 = 0.82). (Right) Again, allowing for interaction between hydrophobicity and nucleotide bias allows for slightly better predictions (r^2 = 0.87).

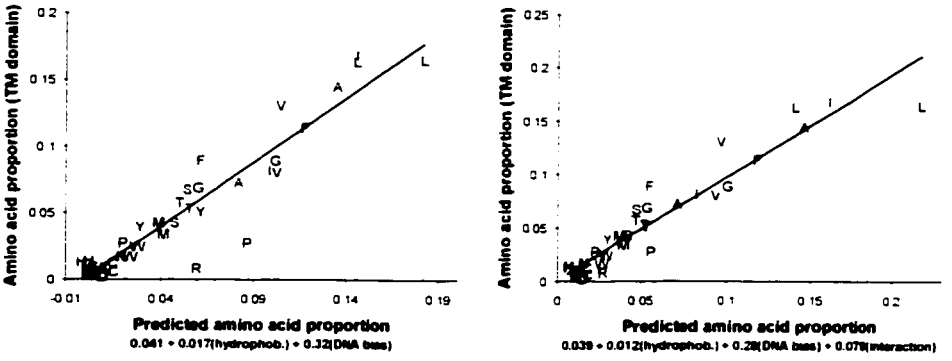


Figure 5.16: The amino acid usages from bacterial transmembrane domains follow the model built from mitochondrial proteins. (Left) r^2 = 0.89; (right) r^2 = 0.91 when interaction between hydrophobicity and nucleotide bias is taken into account.

5.5.6 The effect of nucleotide bias on protein hydropathy

The regression equation used to make the amino acid predictions in the rightmost plots from Figure 5.14, Figure 5.15, and Figure 5.16 includes a term describing the interdependence between the effect of nucleotide bias and hydrophobicity. In all three cases, including this term in the regression model increased its explanatory power, suggesting that nucleotide bias and amino acid hydrophobicity

are related to each other in some way. Indeed, previous studies have found a link between nucleotide bias in genes and the hydrophobicity of the encoded proteins, although these results have not been consistent. Gu et al (1998), for example, found a positive correlation between GC content and hydrophobicity, while other researchers have found the opposite trend (D'Onofrio et al. 1999; Cruveiller et al. 1999). A more commonly accepted trend is the positive relationship between T-richness and hydrophobicity (Kypr and Mrázek. 1987; Naylor et al. 1995; Chiusano et al. 2000). Our own results both support and contrast these findings. Figure 5.18 (left) shows that the transmembrane domains do, in fact, have a higher nonsynonymous T proportion than the non-transmembrane protein regions, which is consistent with the link between hydrophobicity and T proportion. However, among the transmembrane domains from different organisms, there is actually a negative correlation between hydrophobicity and nonsynonymous T content. Conversely, there is a positive relationship between nonsynonymous T proportion and hydrophobicity in the non-transmembrane protein regions. In other words, the relationship between nucleotide bias and hydrophobicity depends on the region of the protein being investigated. A similar relationship exists when we compare hydrophobicity and synonymous T proportion (Figure 5.18, right). These findings suggest that DNA mutation bias does have the capacity to affect the biochemical properties of a protein, but the exact nature of this relationship is not clear, and appears to be different for each region of the protein.

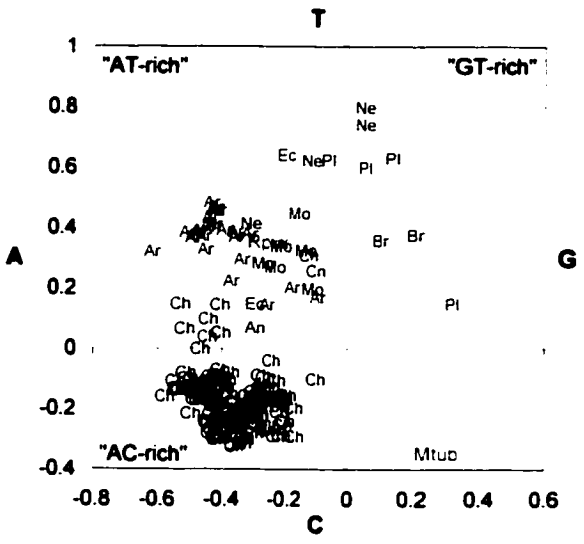


Figure 5.17: The nucleotide bias patterns in the genomes studied. Abbreviations are as described in the legend for Figure 5.11, with the addition of the two bacterial genomes, Mtub (*Mycobacterium tuberculosis*) and Rpxx (*Rickettsia prowazekii*). The original seventeen genomes used in our analyses are coloured red, and the two bacterial genomes are coloured green. Mtub is the only organism studied that has a "GC-rich" pattern of nucleotide bias.

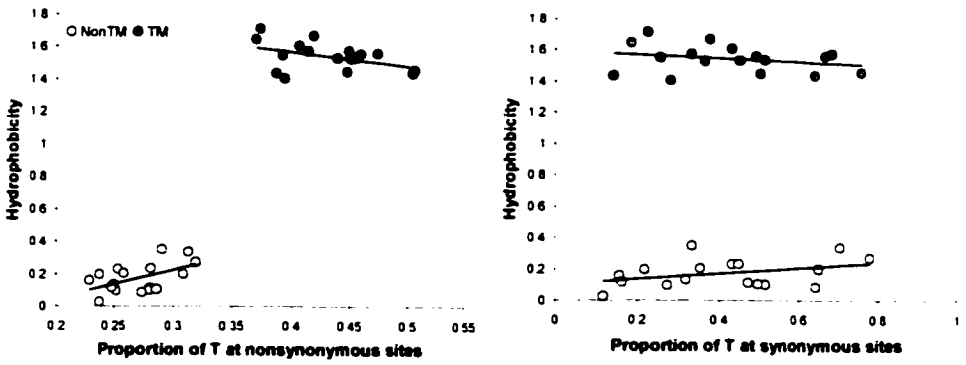


Figure 5.18: (Left) A positive relationship between the hydrophobicity of a peptide sequence and the nonsynonymous T proportion in the underlying DNA is found in non-transmembrane protein regions (r^2 = 0.28), but the opposite relationship occurs in the transmembrane domains (r^2 = 0.19). (Right) Similar, but weaker trends are observed when hydrophobicity is compared to the synonymous T proportion (r^2 are 0.066 and 0.15 for TM and nonTM, respectively).

5.6 Discussion

In this study we have not only confirmed that both selection and mutation bias affect the evolution of transmembrane domains; we have also quantified this relationship. In general, we found that natural selection accounts for variation in amino acid proportions relative to other amino acids *within* individual genomes (Figure 5.5), while the *between-genome* variation is controlled by DNA mutation bias (Figure 5.13). Since hydrophilic amino acids (like D, E, and R) decrease the ability of the transmembrane domain to pass through the lipid bilayer, for example, they are selected against in all genomes, regardless of the pattern of nucleotide bias in the DNA. Among the hydrophobic amino acids, selection appears to be more tolerant as to which amino acids are used, allowing for a large degree of variation between genomes. It is here that the patterns of DNA mutation bias in each genome are able to exert control over the amino acid usage. Genomes whose coding sequences are GT-biased, for example, will tend to encode amino acids whose codons are rich in the G and T nucleotides (like valine), while genomes with AC-biased coding sequences will encode amino acids whose codons are rich in A and C (isoleucine, for example). In a way this is a simple corollary to our findings from previous chapters. In transmembrane domains encoded by GC-rich DNA, for example, there is pressure for increased usage of the G, A, R, and P amino acids. Due to the constraints of natural selection, however, only the hydrophobic amino acids (G and especially A) can increase significantly. Stevens and Arkin (2000) found empirical evidence for this trend among bacterial transmembrane domains, and discovered that almost all of the amino acid variation between the transmembrane domains from GC- and AT-rich bacteria occurred between the amino acids V+A (which were preferred in the GC-rich organisms) and F+I (which were preferred in the AT-rich organisms).

Although this study has focused on the evolution of transmembrane domains, it is likely that similar interactions between natural selection and DNA bias occur in other protein domains as well, and indeed in entire proteins. In the previous chapter we showed that even highly conserved proteins are affected by DNA

mutation bias. In that case, it appeared as if all of the amino acid changes were occurring in the loosely conserved sites within the proteins. The results from the present study, however, are a bit different. The amino acids within transmembrane domains are conserved, but at the biochemical level instead of the amino acid level. This flexibility is enough for DNA mutation bias to exert some control over the amino acid compositions of the domains. Selection at the biochemical level is probably very common: many proteins have features like hydrophobic binding pockets or charged surfaces, in which a number of different amino acids could be substituted while still maintaining the biochemical properties of these regions. We believe it is highly probable that DNA mutation bias affects the amino acid usage of these protein regions in much the same manner as we observed its influence on transmembrane domains here.

It has been suggested that mutation bias is an "orienting factor" in the direction that evolution proceeds in (Yampolsky and Stoltzfus, 2001), and our results fit well into this model. In transmembrane domains, the choice between hydrophobic amino acids appears to be arbitrary, since they all suit the requirements of the domain equally well. In the presence of a mutation bias, however, the choice of one amino acid over another is no longer random. Rather than being directionless, the progress of evolution is now squarely oriented in one direction or another, depending on the direction of the nucleotide bias in the DNA. We expand upon this idea in Chapter 6.

6 Conclusions, and how directional neutral evolution and positive selection interact

6.1 Abstract

We have shown that neutral evolution can bring about directional, asymmetrical changes to the amino acid composition of proteins. These changes can come about via modifications to the genetic code, or by mutation biases in the DNA. We have shown that this is a pervasive trend that appears to transcend all boundaries of phylogenetic relatedness. We have also shown that directional neutral evolution is not incompatible with negative selection, and that both factors interact to shape the direction of protein evolution. In this chapter we expand upon this idea to encompass the interaction between directional neutral evolution and positive natural selection. We expand upon Sewall Wright's adaptive landscapes as an analogy for this mode of evolution.

6.2 Conclusions

6.2.1 The evolution of proteins is affected by the genetic code

The number of codons assigned to each amino acid in the universal genetic code ranges from one to six (and even eight in some mitochondrial genetic codes). In Chapter 2 we showed that this bias in the genetic code has an effect on the amino acid usage in proteins: amino acids with many codons are more common than those with few codons. We also showed that modifications to the genetic code cause periods of asymmetrical protein evolution, where the amino acid frequencies change to match the new codon frequencies in the code. The relative proportions of seven amino acids were shown to vary in response to five different modifications to the genetic code among the metazoan mitochondrial genomes. These results seem to support the notion of a neutrally-evolved genetic code rather than one optimized by selection.

6.2.2 The evolution of proteins is affected by DNA mutation bias

In Chapter 3 we showed that mutation bias in the DNA—causing an accumulation of G and C nucleotides with respect to A and T nucleotides or *vice versa*—was capable of bringing about massive changes to the relative proportions of amino acids in the proteins. A change towards GC-richness in the DNA, for example, will tend to produce proteins that have increased proportions of the G, A, R, and P amino acids lower proportions of the F, Y, M, I, N, and K amino acids. AT-biased DNA, on the other hand, tends to produce proteins that show the opposite trend in their amino acid compositions. In Chapter 4 we demonstrated that these changes are widespread, affecting every protein studied—even those that are subject to strong purifying selection. These same patterns were observed in the three primary domains of life: the Archaea, the Eubacteria, and the Eukaryota, with nearly identical results in each case.

Our results from Chapter 4 showed that the degree to which directional protein evolution occurs in response to directional DNA evolution depends on how well

negative selection preserves the amino acid sequence of the protein. In Chapter 5 we developed both qualitative and quantitative models describing the interaction between directional mutation biases in the DNA and negative selection acting on the protein to shape the overall amino acid usage within transmembrane domains. Our results showed that negative selection acts at the biochemical level within these domains rather than at the amino acid level. This allows some flexibility in the amino acid sequence, allowing directional evolution to occur in response to the directional evolution occurring in the underlying DNA. Since many regions within proteins are known to be conserved at the biochemical level instead of the amino acid level, this phenomenon is probably very widespread.

6.2.3 Adaptive landscapes

The concept of neutral evolution can be modeled using a modified version of Sew Sewall Wright's classic analogy of the adaptive landscape (Wright, 1932). Wright's original concept modeled the fitness of an organism as the height it had attained on a landscape of peaks and valleys, while the x and y positions on the landscapes indicated allele frequencies in the population. We now know that—certainly at the level of proteins and DNA—the true fitness landscape is more akin to a series of plateaus, since many mutations are selectively neutral. In other words, a protein can withstand changes to its amino acid sequence without negatively impacting on its performance, as we saw in Chapter 5. This is even truer at the DNA level, where mutations to the synonymous codon positions of the gene may have no effect on the fitness of the protein whatsoever[1]. Even a plateau does not adequately model protein evolution though, since not all mutations at the top are neutral: some mutations do negatively impact on the protein's fitness. For this reason, we envision the landscape being composed of ridges or mesas.

---

[1] Strictly speaking this is not true, since synonymous codons may be optimized to match the tRNA population within the cell. Mutations to synonymous positions, therefore, may affect the efficiency with which the mRNA is translated, thus affecting the expression of the protein (Ikemura, 1981a,b; Ikemura, 1982; Shields and Sharp, 1987; Shields et al, 1988; Stenico, Lloyd, and Sharp, 1994; McInerney, 1998). For simplicity, however, we will consider only the fitness of the encoded protein, which is not affected by synonymous changes to the DNA.

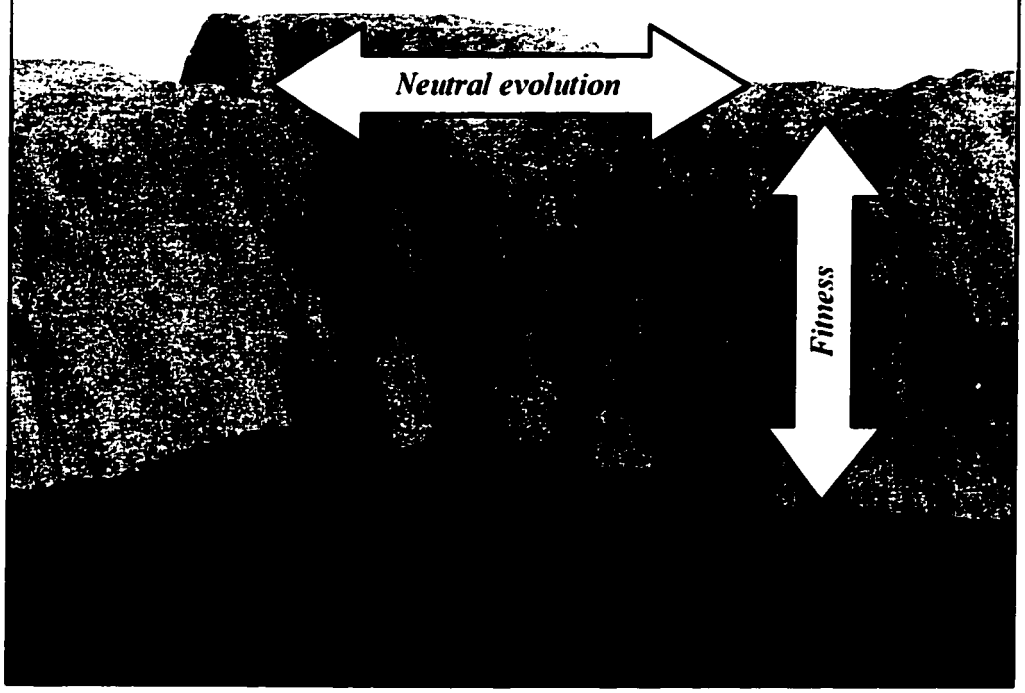


Figure 6.1: A mesa-type adaptive landscape describes conservative evolution. Vertical height corresponds to the fitness of the protein, so movement along the apex is neutral, while movement down the slope is selected against. Under a symmetrical model of evolution, movement along the top is random, whereas under an asymmetrical model of evolution, the movement along the top is directional.

Mutations that maintain a certain level of fitness of the protein can be interpreted as movement along the top of the mesa, while deleterious mutations correspond to lateral movements that send the protein toppling down the steep cliffs on either side.

If we use transmembrane domains as an example, any movement along the top of the ridge corresponds to mutations from one hydrophobic residue to another, which do not negatively affect the fitness of the domains. A mutation from a hydrophobic amino acid to a hydrophilic one, however, corresponds to movement

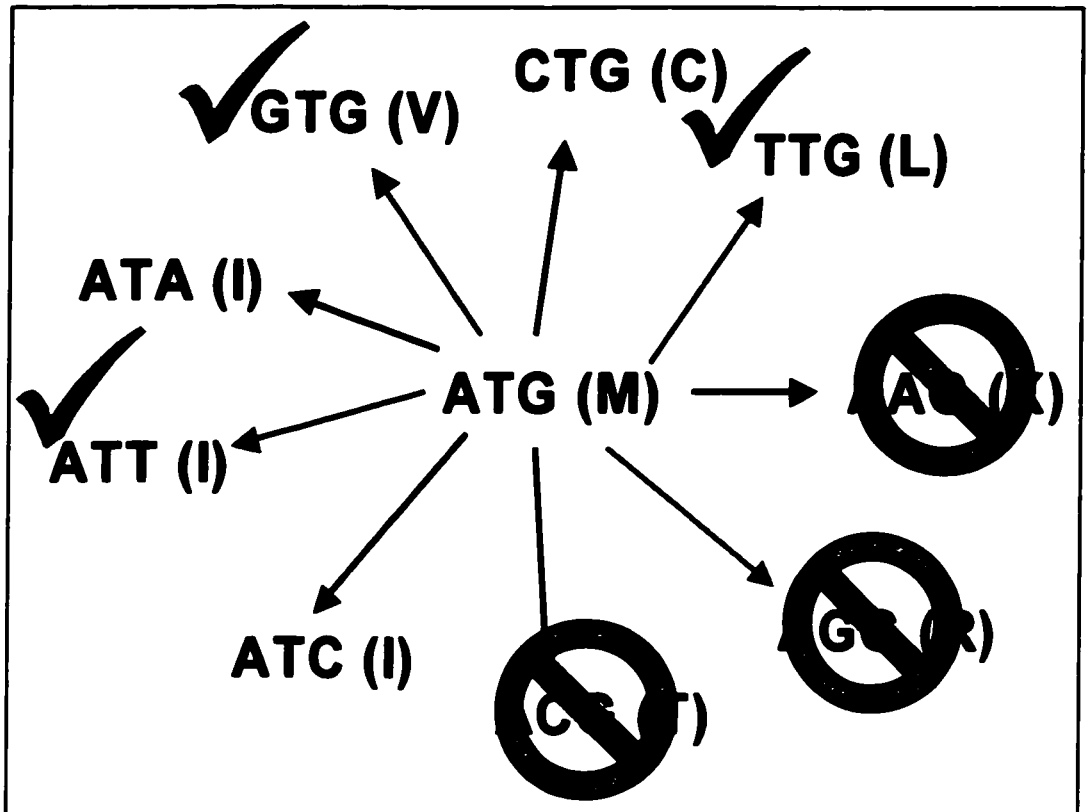


Figure 6.2: The interaction between mutation bias and negative selection on the mutation of a methionine codon. In DNA that is GT-biased, the arrows coloured green represent likely paths, while those coloured red represent unlikely paths. Meanwhile, negative selection forbids the fixation of mutations that lead to hydrophilic residues (T, R, and K). These two factors combined, then, will tend to favour evolution along those paths indicated by the checks.

down the side of the ridge (see Figure 6.1). Thus, non-random neutral evolution is possible even when it is constrained by purifying selection. Within transmembrane domains for example, the hydrophobic amino acids are selectively neutral relative to each other, but are favoured by positive selection relative to the hydrophilic amino acids.

Under a symmetrical model of evolution, movement along the top of an evolutionary ridge will be random. As our results from Chapters 3 through 5 show, however, asymmetrical evolution in proteins can be caused by directional

evolution in the underlying DNA. AC-biased mitochondrial genes, for example, showed a pattern of amino acid usage that is quite different than the pattern within the GT-biased genes. Going back to the analogy of the evolutionary ridge, this corresponds to movement in one direction over the other along the top (Figure 6.1).

We illustrate this point at the codon level in Figure 6.2. Here we show a methionine codon and the nine other codons that are a single mutational step away from it. If the DNA sequence was evolving in a completely neutral, symmetrical manner, each of these possible mutations would have an equal chance of reaching fixation in the population. If the DNA was evolving in a directional manner towards the G and T nucleotides however, some of these possibilities would have a greater chance of being fixed than others. The other factor affecting the fate of this codon in transmembrane domains is selection against hydrophilic residues. From the remaining possibilities in which the ATG codon can evolve then, AGG (arginine, a hydrophilic amino acid) will be selected against and will not be allowed to be fixed in the population, while the other three codons encode hydrophobic amino acids, so evolution can proceed in any of those three directions.

6.2.4 Mutation bias and positive selection

Just as we have shown that neutral evolution and negative selection are not mutually exclusive concepts (see Section 6.2.3), we argue that neutral evolution and positive Darwinian selection can work simultaneously upon one site in a protein. If we imagine an emerging transmembrane domain, the same principles apply to the selective pressures on amino acid usage as did when we were examining transmembrane domains that already exist: hydrophilic amino acids are selected against, while hydrophobic amino acids are positively selected for. Since there are multiple hydrophobic amino acids, there are various directions in which evolution can proceed while still increasing the fitness of the domain. In other words, the choice between two sufficiently hydrophobic amino acids is

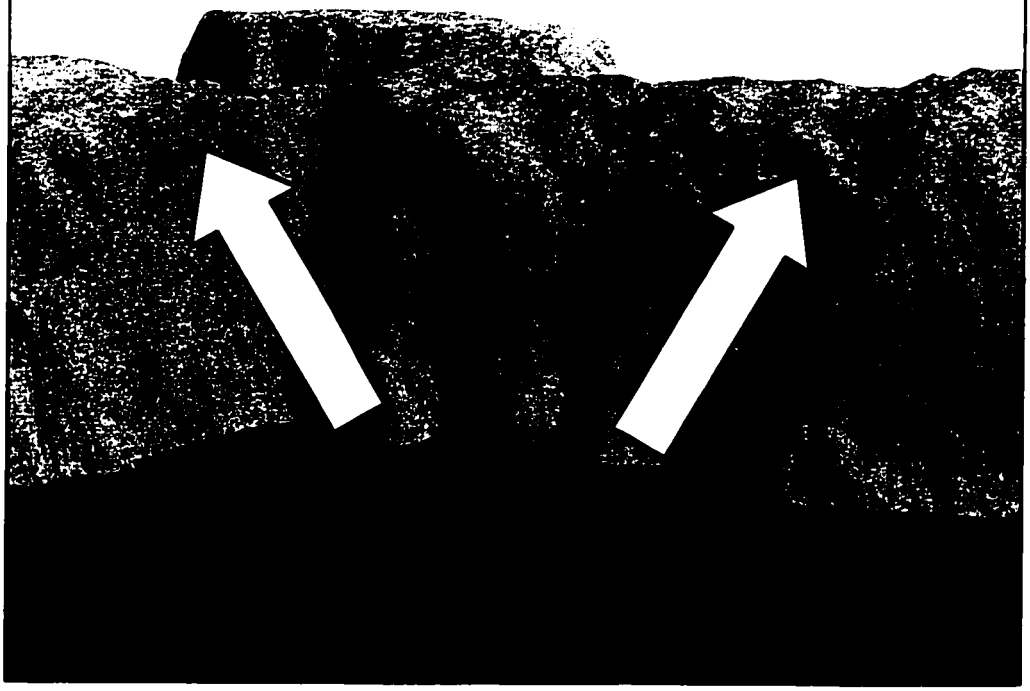


Figure 6.3: The mesa-type landscape can also model the evolution of a protein during positive selection. As the protein ascends the mesa, directional mutation pressure can cause the protein to move towards the right or the left, but the end result is the same in this case.

selectively neutral. In Chapters 2 through 4 we showed that certain factors could bring about a directional pattern of neutral evolution, so we believe this is possible when positive selection is involved, as well.

Yampolsky and Stoltzfus (2001) have provided a mathematical proof of this effect, and have referred to directional mutation biases as "orienting factors" in evolution. That is, if positive selection can occur in more than one direction, biases in the mutation process will decide the path that the selection process takes. This concept can also be modeled well on an adaptive landscape, if we consider a protein that is starting at a low point in the landscape instead of on top of a peak. In Figure 6.3 we show how positive selection in combination with a directional mutational pressure can influence the path along which evolution proceeds. In this

case, the end result is the same, since both paths lead to the top of the same peak—the protein is destined to become a transmembrane domain, but the particular amino acid sequence within that domain will be determined by the pattern of nucleotide bias in the underlying gene.

Figure 6.4 shows a two-peaked positive selection scenario. Note that the two peaks are distict and that the choice of which one to ascend is arbitrary. Indeed, the peak that the protein ultimately climbs would be random if the sequence was evolving in a symmetrical manner. However, in the presence of a directional mutation bias, the ascension of one peak will become more likely than the other. In essence, then, the ultimate fate of natural selection can be ordained by directional "neutral" evolution.

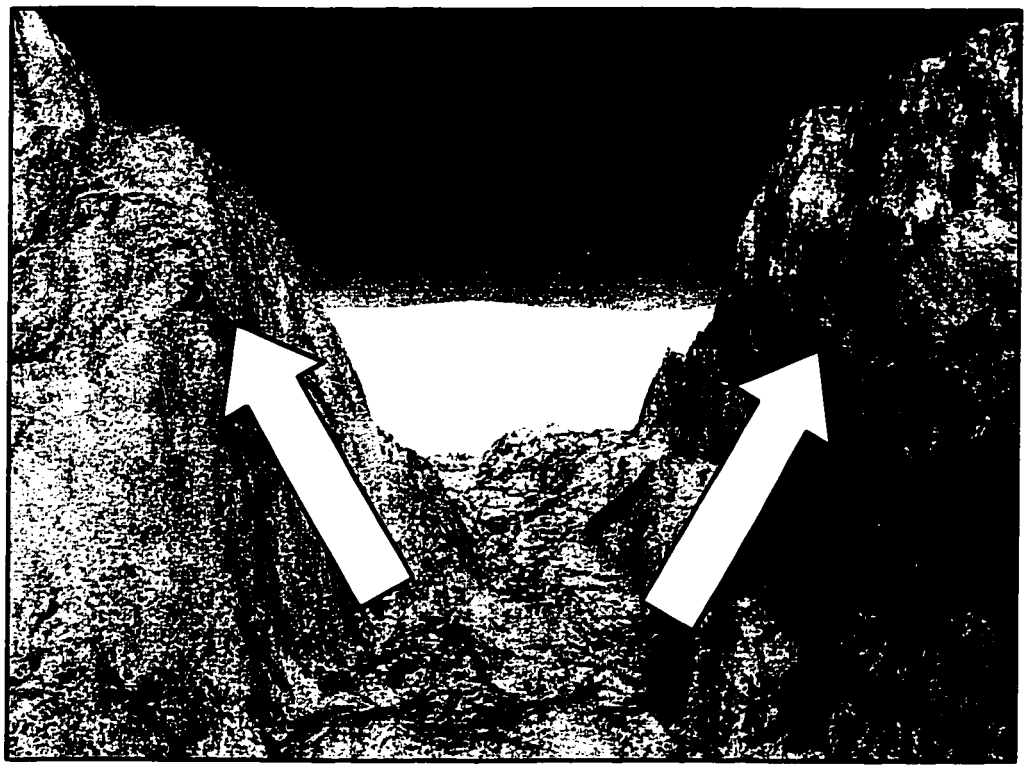


Figure 6.4: Here we present a landscape in which two routes are equally favoured by selection. In this case, directional mutation bias will decide which peak is ascended, and the end result is different: that is, directional neutral evolution has determined the ultimate fate of the protein.

Although we present the two peaks in Figure 6.4 as being equally favourable, Yampolsky and Stoltzfus showed that directional mutation bias can also affect the outcome of positive selection when the two peaks are not equal. We have presented this situation in Figure 6.5, where the slope of one peak is significantly greater than the other. In this case the differences in slope are analogous to differences in the selection coefficients of the two peaks: every step to the right brings about a greater increase in the fitness of the protein than a step to the left does. All other things being equal, therefore, we would expect the probability of fixation of the rightward mutation to be much greater than the one to the left. In the presence of a directional mutation bias, though, this may not be the case. Indeed, a strong enough mutation bias to the left can meet and even exceed the selective pressure to go to the right. Mutation bias, then, may cause protein evolution to proceed in a direction that is ultimately less advantageous than it might have achieved in the absence of bias. This isn't necessarily the case, however, as the bias may be in the "correct" direction, which would actually increase the probability of the better evolutionary path is chosen by the protein.

6.3 Implications to sequence analysis

Heterogeneity of nucleotide compositions across lineages is known be problematic in DNA-based phylogenetic studies (Lockhart et al. 1992; Hashimoto et al. 1994; Lockhart et al. 1994; Galtier and Gouy, 1994, 1995; He and Haymer, 1995; Hashimoto et al. 1995). For this reason, a number of more complicated models have been devised to counteract these problems (Steel, Lockhart, and Penny, 1993; Lockhart, 1994; Ferretti et al. 1994; Galtier and Gouy, 1995, 1998). Even when these new models are employed, however, directional changes in the DNA can confound phylogenetic analyses (Chang and Campbell, 2000). Since we have shown that directional changes in the DNA lead to directional changes in protein evolution, these same problems are likely to be applicable to protein-based sequence analyses. Indeed, Foster and Hickey (1999) showed that during phylogenetic reconstruction, branches with similar amino acid biases tend to

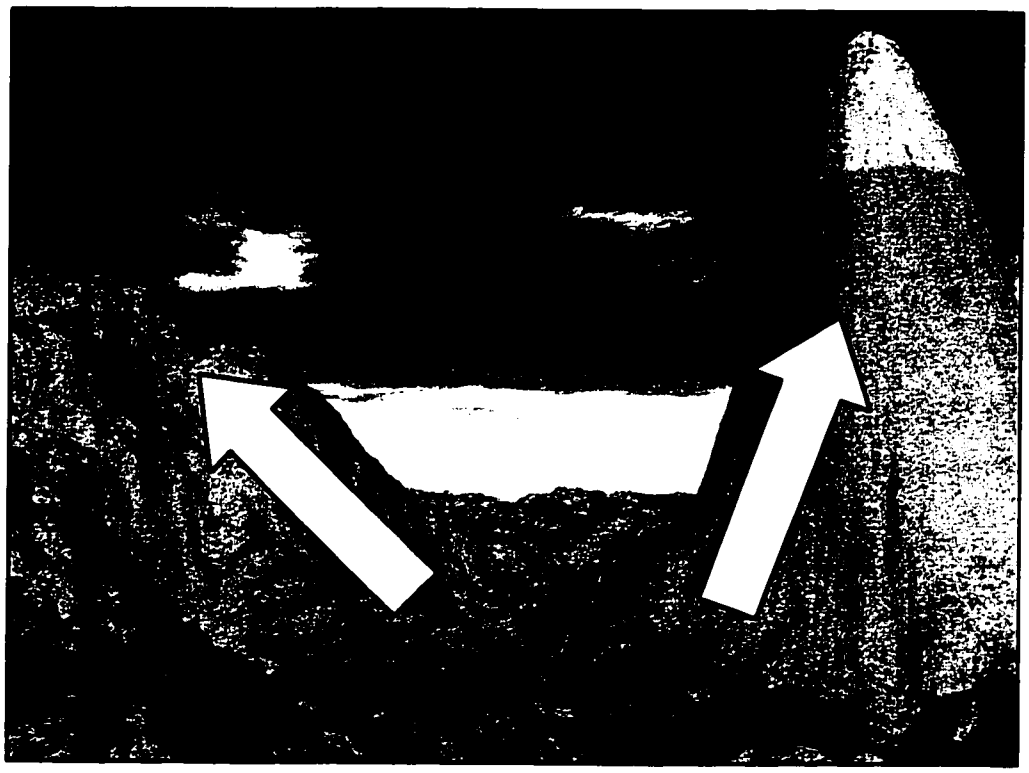


Figure 6.5: Directional neutral evolution can overcome the selective advantage to climb one peak over another. This Figure represents a situation similar to that in Figure 6.4, except that the two routes are not equal: the peak to the right is much steeper (i.e., has a higher selection coefficient) than the peak on the left. Regardless of the advantage to moving rightward, a sufficient level of directional bias can cause the protein to evolve toward the left.

group together. Other sequence analyses like the inference of ancestral nodes in a tree are also likely to be affected by directional changes to the amino acid composition of proteins (Perna and Kocher, 1995; Eyre-Walker, 1998).

Since we have shown that directional evolution at both the DNA and protein levels is a pervasive phenomenon, we believe traditional sequence analyses should be used with some caution, especially when it is clear that the assumptions of the underlying model have been violated.

6.4 Other types of non-random neutral evolution

This thesis has concentrated almost exclusively on non-random patterns of point mutations in the DNA, and their effects on protein evolution. However, point mutations are not the only type of DNA mutation. DNA sequences are also modified by deletions, duplications, translocations, inversions, gene conversion, and unequal crossing, to name only a few of the known mutation processes. All have the potential to be non-random in nature. Thus, there are probably forms of non-random, non-Darwinian evolution in addition to those that have been discussed in this thesis. Detecting and measuring these interesting mechanisms of evolution will pose an interesting problem for future studies in this area.

7 References

Akashi H and Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci USA 99:3695-3700

Altschul SF. Madden TL. Schaffer AA. Zhang J. Zhang Z. Miller W. and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402

Andersson SGE and Sharp PM. 1996. Codon Usage and Base Composition in *Rickettsia prowazekii*. J Mol Evol 42:525-536

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

Au KG. Clark S. Miller JH. and Modrich P. 1989. *Escherichia coli* mutY gene encodes an adenine glycosylase active on G/A mispairs. Proc Natl Acad Sci USA 86:8877-8881

Belozersky AN and Spirin AS. 1958. A Correlation between the Compositions of Deoxyribonucleic and Ribonucleic Acids. Nature 182:111-112.

Berkhout B and van Hemert FJ. 1994. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res 22:1705-1711

Berkhout B, Grigoriev A, Bakker M, and Lukashov VV. 2002. Codon and Amino Acid Usage in Retroviral Genomes is Consistent with Virus-Specific Nucleotide Pressure. AIDS Res Hum Retroviruses 18:133-141

Bill CA, Duran WA, Miselis NR, and Nickoloff JA. (1998). Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells. Competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. Genetics 149:1935-1943

Boore JL. 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27:1767-1780.

Chang BSW and Campbell DL. 2000. Bias in Phylogenetic Reconstruction of Vertebrate Rhodopsin Sequences. Mol Biol Evol 17:1220-1231

Chiusano ML, Alvarez-Valin F, Di Giulio M, D'Onofrio G, Ammirato G, Colonna G, and Bernardi G. 2000. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. Gene 261:63-69

Clarke GD, Beiko RG, Ragan MA, and Charlebois RL. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J Bacteriol 184:2072-2080

Collins DW and Jukes TH. 1993. Relationship between G + C in silent sites of codons and amino acid composition of human proteins. J Mol Evol 36:201-213

Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, and DeLisi C. 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J Mol Biol 195:659-685

Cox EC and Yanofsky C. 1967. Altered base ratios in the DNA of an *Escherichia coli* mutator strain. Proc Natl Acad Sci USA 58:1895-1902

Crick FHC. 1968. The origin of the genetic code. J Mol Biol 38:367-379

Cruveiller S, Jabbari K, D'Onofrio G, and Bernardi G. 1999. Different hydrophobicities of orthologous proteins from *Xenopus* and human. Gene 238:15-21

Dayhoff MO (Ed.), Hunt LT, Barker WC, Schwartz RM, Orcutt BC, and Young CL. 1978. Atlas of Protein Sequence and Structure, Vol 5 Supplement 3. Nat Biomed Res Found, Washington D.C.

Deber CM, Wang C, Liu LP, Prior AS, Agrawal S, Muskat BL, and Cuticchia AJ. 2001. TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. Protein Sci 10:212-219.

Di Giulio M and Medugno M. 1999. Physicochemical optimization in the genetic code origin as the number of codified amino acids increases. J Mol Evol 49:1-10

D'Onofrio G, Jabbari K, Musto H, and Bernardi G. 1999. The correlation of protein hydropathy with the base composition of coding sequences. Gene 238:3-14

Drummond JT and Bellacosa A. 2001. Human DNA mismatch repair *in vitro* operates independently of methylation status at CpG sites. Nucleic Acids Res 29:2234-2243

Dufton MJ. 1983. The significance of redundancy in the genetic code. J Theor Biol 102:521-526

Dufton MJ. 1985. Genetic Code Redundancy and the Evolutionary Stability of Protein Secondary Structure. J Theor Biol 116:343-348

Duret L and Hurst LD. 2001. The Elevated GC Content at Exonic Third Sites Is Not Evidence Against Neutralist Models of Isochore Evolution. Mol Biol Evol 18:757-762

Eilers M, Shekar SC, Shieh T, Smith SO, and Fleming PJ. 2000. Internal packing of helical membrane proteins. Proc Natl Acad Sci USA 97:5796-5801.

Eyre-Walker A. 1998. Problems with Parsimony in Sequences of Biased Base Composition. J Mol Evol 47:686-690

Fay JC, Wyckoff GJ, and We CI. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415:1024-1026

Ferretti V, Lang BF, and Sankoff D. 1994. Skewed Base Compositions, Asymmetric Transition Matrices, and Phylogenetic Invariants. J Comput Biol 1:77-92

Foster PG, Jermiin LS, and Hickey DA. 1997. Nucleotide Composition Bias Affects Amino Acid Content in Proteins Coded by Animal Mitochondria. J Mol Evol 44:282-288

Foster PG and Hickey DA. 1999. Compositional Bias May Affect Both DNA-Based and Protein-Based Phylogenetics Reconstructions. J Mol Evol 48:284-290

Freeland SJ, Knight RD, Landweber LF and Hurst LD. 2000. Early fixation of an optimal genetic code. Mol Biol Evol 17:511-518

Galtier N, Gouy M. 1994. Molecular phylogeny of Eubacteria: a new multiple tree analysis method applied to 15 sequence data sets questions the monophyly of gram-positive bacteria. Res Microbiol 145:531-541

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci USA 92:11317-11321

Gojobori T, Ishi K, and Nei M. 1982. Estimation of Average Number of Nucleotide Substitutions When the Rate of Substitution Varies with Nucleotide. J Mol Evol 18:414-423

Gojobori T. 1983. Codon substitution in evolution and the "saturation" of synonymous changes. Genetics 105:1011-1027

Goldman N and Yang Z. 1994. A Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences. Mol Biol Evol 11:725-736

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664-1674.

Gu X, Hewett-Emmett D, and Li W-H. 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica 102/103:383-391

Gu X and Li W-H. 1996. A general additive distance within time-reversibility and rate variation among nucleotide sites. Proc Natl Acad Sci USA 93:4671-4676

Gu X and Li W-H. 1998. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proc Natl Acad Sci USA 95:5899-5905

Haldane. JBS. 1957. The cost of natural selection. J Genet 55:511-524

Hashimoto T. Nakamura Y. Nakamura F. Shirakura T. Adachi J. Goto N. Okamoto K-I and Hasewaga 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

Hashimoto T. Nakamura Y. Kamaishi T. Nakamura F. Adachi J. Okamoto K-I and Hasewaga M. 1995. Phylogenetic Place of Mitochondriaon-Lacking Protozoan. *Giardia lamblia*. Inferred from Amino Acid Sequences of Elongation Factor 2. Mol Biol Evol 12:782-793

He M and Haymer DS. 1995. Codon Bias in Actin Multigene Families and Effects on the Reconstruction of Phylogenetic Relationships. J Mol Evol 41:141-149

Henikoff. S and Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915-10919

Hickey DA. Wang S. and Magoulas C. 1994. Gene duplication. gene conversion and codon bias. Pp. 199-207 in B. Golding. ed. Non-Neutral Evolution. Chapman and Hall. New York.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146:1-21

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151:389-409

Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158:573-597

Jones DT, Taylor WR, and Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-282

Judson OP and Haydon D. 1999. The Genetic Code: What Is It Good For? An Analysis of the Effects of Selection Pressures on Genetic Codes. J Mol Evol 49:539-550

Jukes TH, Holmquist R, and Moise H. 1975. Amino acid composition of proteins: Selection against the genetic code. Science 189:50-51

Kimura M. 1968. Evolutionary Rate at the Molecular Level. Nature 217:624-626

Kimura M. 1980. A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences. J Mol Evol 16:111-120

Kimura M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454-458

Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

King JL and Jukes TH. 1969. Non-Darwinian Evolution. Science 164:788-798

Knight RD, Landweber LF, and Yarus M. 2001. How mitochondria redefine the code. J Mol Evol 53:299-313

Knight RD, Freeland SJ, and Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol 2:RESEARCH0010

Krogh A, Larsson B, von Heijne G, and Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol 305:567-580

Kuhn LA and Leigh JS Jr. 1985. A statistical technique for predicting membrane protein structure. Biochim Biophys Acta 828:351-61

Kypr J and Mrázek J. 1987. Occurrence of nucleotide triplets in genes and secondary structure of the coded proteins. Int J Biol Macromol 9:49-53

Kyte J and Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105-132.

Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, and Wolfe KH. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 27:1642-1649

Li W-H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, Mass.

Li W-H, Wu C-I, and Luo C-C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150-174

Lindahl T. 1993. Instability and decay of the primary structure of DNA. Nature 362:709-715

Lobry JR. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205:309-316

Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ and Larkum AW. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol 34:153-162

Lockhart PJ, Steel MA, Hendy MD, and Penny D. 1994. Recovering Evolutionary Trees under a More Realistic Model of Sequence Evolution. Mol Biol Evol 11:605-612

Loomis WF and Smith DW. 1990. Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. Proc Natl Acad Sci USA 87:9093-9097

Lynn DJ, Singer GAC, and Hickey DA. 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res 30:4272-4277

Mackiewicz P, Gierlik A, Kowalczuk M, Dudek MR, and Cebrat S. 1999. How Does Replication-Associated Mutation Pressure Influence Amino Acid Composition of Proteins? Genome Res 9:409-416

Maeshiro I and Kimura M. 1998. The role of robustness and changeability on the origin and evolution of genetic codes. Proc Natl Acad Sci USA 95:5088-5093

McClellan DA. 2000. The Codon-Degeneracy Model of Molecular Evolution. J Mol Evol 50:131-140

McInerney JO. 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc Natl Acad Sci USA 95:10698-10703

Möller S. Croning MDR. and Apweiler R. 2001. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17:646-653

Mooers AØ and Holmes EC. 2000. The evolution of base composition and phylogenetic inference. Tree 15:365-369

Mrázek J and Kypr J. 1992. Nucleotide composition of genes and hydrophobicity of the encoded proteins. FEBS Lett 305:163-165 [Note: the first author's name was incorrectly spelled as "Mazek" in the original article, and this is how the entry is listed in PubMed]

Muto A and Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84:166-169

Naya H. Romero H. Zavala A. Alvarez B. and Musto H. 2002. Aerobiosis Increases the Genomic Guanine Plus Cytosine Content (GC%) in Prokaryotes. J Mol Evol 55:260-264

Naylor GJP. Collins TM. and Brown WM. 1995. Hydrophobicity and phylogeny. Nature 373:565-566

Nei M. Xu P. and Glazko G. 2001. Estimation of divergence times from multiprotein sequences from a few mammalian species and several distantly related organisms. Proc Natl Acad Sci USA 98:2497-2502

Nghiem Y. Cabrera M. Cupples CG. and Miller JH. 1988. The MutY gene: a mutator locus in *Escherichia coli* that generates G:C to T:A transversions. Proc Natl Acad Sci USA 85:2709-2713

Oliver JL and Marín A. 1996. A Relationship Between GC Content and Coding-Sequence Length. J Mol Evol 43:216-223

O'Sullivan JM. Davenport JB. and Tuite MF. 2001. Codon reassignment and the evolving genetic code: problems and pitfalls in post-genome analysis. Trends Genet 17:20-22

Pagel MD and Harvey PH. 1988. Recent Developments in the Analysis of Comparative Data. Q Rev Biol 63:413-440

Perna NT and Kocher TD. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J Mol Evol 41:353-358

Porter TD. 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

Reyes A, Gissi C, Pesole G, and Saccone C. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol Biol Evol 15:957-966

Rocha EPC and Danchin A. 2002. Base composition bias might result form competition for metabolic resources. Trends Genet 18:291-294

Rodríguez F, Oliver JL, Marin A, and Medina JR. 1990. The General Stochastic Model of Nucleotide Substitution. J Theor Biol 142:485-501

Rodríguez-Trelles F, Tarrío R, and Ayala FJ. 1999. Switch in Codon Bias and increased Rates of Amino Acid Substitution in the *Drosophila saltans* Species Group. Genetics 153:339-350

Rodríguez-Trelles FR, Tarrío R, and Ayala FJ. 2000. Fluctuating Mutation Bias and the Evolution of Base Composition in *Drosphila*. J Mol Evol 50:1-10

Rost B, Fariselli P, and Casadio R. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 5:1704-1718

Saccone C, De Giorgi C, Gissi C, Pesole G, and Reyes A. 1999. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. Gene 238:195-209

Schmitz J, Ohme M, and Zischler H. 2002. The Complete Mitochondrial Sequence of *Tarsius bancanus*: Evidence for an Extensive Nucleotide Compositional Plasticity of Primate Mitochondrial DNA. Mol Biol Evol 19:544-553

Shields DC and Sharp PM. 1987. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res 15:8023-8040

Shields DC, Sharp PM, Higgins DG, and Wright F. 1988. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704-716

Simon AL, Stone EA, and Sidow A. 2002. Inference of functional regions in proteins by quantification of evolutionary constraints. Proc Natl Acad Sci USA 99:2912-2917

Singer GAC and Hickey DA. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol 17:1581-1588

Singer GAC and Hickey DA. 2002. Patterns of Codon Usage and Amino Acid Composition in Thermophiles: Evidence for Natural Selection or Nucleotide Bias? Gene [Submitted]

Smith NGC and Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature 415:1022-1024

Steel MA. Lockhart PJ and Penny D. 1993. Confidence in evolutionary trees from biological sequence data. Nature 364:440-442

Stenico M. Lloyd AT. and Sharp PM. 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res 22:2437-2446

Stevens TJ and Arkin IT. 2000. The effect of nucleotide bias upon the composition and prediction of transmembrane helices. Protein Sci 9:505-511

Stoltzfus A. 1999. On the Possibility of Constructive Neutral Evolution. J Mol Evol 49:169-181

Sueoka N. 1961a. Compositional Correlation between Deoxyribonucleic Acid and Protein. Cold Spring Harb Symp Quant Biol 26:35-43

Sueoka N. 1961b. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. Proc Nat Acad Sci USA 47:1141-1149

Sueoka N. 1988. Directional Mutation Pressure and Neutral Molecular Evolution. Proc Natl Acad Sci USA 85:2653-2657

Sueoka N. 1992. Directional Mutation Pressure. Selective Constraints. and Genetic Equilibria. J Mol Evol 34:95-114

Sueoka N and Kawanashi Y. 2000. DNA G+C content of the third codon position and codon usage biases of human genes. Gene 261:53-62

Telford MJ. Herniou EA. Russell RB. and Littlewood DTJ. 2000. Changes in mitochondrial genetic codes as phylogenetic characters: Two examples from the flatworms. Proc Natl Acad Sci USA 97:11359-11364

Tourasse NJ and Li W-H. 2000. Selective Constraints, Amino Acid Composition, and the Rate of Protein Evolution. Mol Biol Evol 17:656-664

Tsai-Wu JJ, Radicells LP, and Lu A-L. 1991. Nucleotide sequence of the *Escherichia coli* micA gene required for A/G-specific mismatch repair: identity of MicA and MutY. J Bacteriol 173:1902-1910

Tusnády GE and Simon I. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283:489-506

von Heijne G. 1995. Membrane protein assembly: rules of the game. Bioessays 17:25-30

Wang H-C and Hickey DA. 2002. Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes. Nucleic Acids Res 30:2501-2507

Wilquet V and Van de Casteele M. 1999. The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. Res Microbiol 150:21-32

Wirth T, Le Guellec R, and Veuille M. 1999. Directional Substitution and Evolution of Nucleotide Content in the *Cytochrome Oxidase* II Gene in Earwigs (Dermapteran Insects). Mol Biol Evol 16:1645-1653

Wood TC and Pearson WR. 1999. Evolution of Protein Sequences and Structures. J Mol Biol 291:977-995

Wright S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proceedings of the 6th International Congress of Genetics 1:356-366

Xia X and Li W-H. 1998. What Amino Acid Properties Affect Protein Evolution? J Mol Evol 47:557-564

Yampolsky LY and Stoltzfus A. 2001. Bias in the introduction of variation as an orienting factor in evolution. Evol Dev 3:73-83

Yang Z, Nielsen R, and Hasewaga M. 1998. Models of Amino Acid Substitution and Applications to Mitochondrial Protein Evolution. Mol Biol Evol 15:1600-1611

Yokobori S, Suzuki T, and Watanabe K. 2001. Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. J Mol Evol 53:314-26