Genetica **102/103:** 383–391, 1998.

Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria

Xun Gu*, David Hewett-Emmett & Wen-Hsiung Li**
*Human Genetics Center, SPH, University of Texas, P.O. Box 20334 Houston, TX 77225, USA; *Present address: The Institute of Molecular Evolutionary Genetics and Department of Biology, 328 Mueller Laboratory, Pennsylvania State University, University Park, PA 16802, USA; **Author for correspondence (Phone: 713-500-9814; Fax: 713-500-0900; e-mail: li@hgc.sph.uth.tmc.edu.)*

Key words: directional mutational pressure, genomic GC content, amino acid composition, hydrophobicity, phylogeny

Abstract

The relationship between change in genomic GC content and protein evolution in bacteria was studied by simple correlational analysis (at the genus level) and by Felsenstein's (1985) independent contrast test. We first used the *dnaA* gene in bacteria as an example to show (1) that the amino acid composition of a protein can be dramatically affected by mutational pressure (the genomic GC content), (2) that surprisingly, deleting relatively closely-related genera may increase rather than decrease the correlation between genomic GC content and amino acid composition, and (3) that most unexpectedly, as the genomic GC content increases, both strongly hydrophobic and strongly hydrophilic amino acids tend to change to ambivalent amino acids, suggesting that the majority of these amino acid substitutions are not caused by positive Darwinian selection.

These patterns were then also shown to hold for the 14 other genes studied, indicating their generality for the evolution of bacterial proteins. As directional mutation pressure can affect the amino acid composition of proteins, it may mislead phylogenetic inference, even if protein instead of DNA sequences are used.

Introduction

The GC content (G+C%) of bacterial genomes varies greatly (25% $\sim$ 75%) and this variation is thought to be largely due to differences in mutational pattern among species (Sueoka, 1962, 1988; Muto & Osawa, 1987). The neutralist view of molecular evolution predicts a correlation between mutational pattern and amino acid composition in proteins (Sueoka, 1961; Jukes & Bhushan, 1986; Nei, 1987), because it holds that protein evolution is mutation driven (Kimura, 1968, 1983; King & Jukes, 1969; Ohta, 1973). In contrast, the selectionist view would predict no correlation because it considers mutation unimportant in protein evolution. A correlation has indeed been found in several studies (Sueoka, 1961; Jukes & Bhushan, 1986; Collins & Jukes, 1993; Bronson & Anderson, 1994; Moran, 1996), though some authors (e.g., Schachtel et al., 1991; D'Onofrio et al., 1991) argued that it supports the selectionist view.

This problem is also important to molecular systematics and bioinformatics. It is known that phylogenetic analysis of ribosomal DNA sequences could be misled by variation in GC content among sequences, because sequences with similar GC contents may be grouped together, irrespective of their true evolutionary relationships (e.g., Hasegawa & Hashimoto, 1993; Sogin, Hinkle & Leipe, 1993; Lockhart et al., 1994). Therefore, some authors suggested the use of protein sequences when the GC content varies considerably among species. This suggestion was based on the common belief that the amino acid composition of a protein would not be seriously affected by variation in genomic GC content, but no solid evidence in favor of this view has been provided. In bioinformatic studies, amino acid composition and biochemical proper-

Table 1. Protein sequences of *dnaA* genes and genomic GC content of species

Acc.#	Species	GC_4 (%)
U01667	*Caulobacter crescentus*	90.6
M82836	*Streptomyces coelicolor*	90.5
U17833	*Mycobacterium smegmatis*	85.0
L25439	*Rhizobium meliloti*	80.5
X14791	*Pseudomonas putida*	78.2
M17353	*Serratia marcescens*	77.9
M17352	*Salmonella typhimurium*	68.0
L36958	*Synechocystis* sp.	67.1
J01602	*Escherichia coli*	61.0
L47617	*Vibrio harveyi*	36.6
M58352	*Proteus mirabilis*	32.2
U39734	*Mycoplasma genitalium*	17.4
M80817	*Buchnera aphidicola*	14.8
D90426	*Mycoplasma capricolum*	11.2
Z19108	*Spiroplasma citri*	11.9

ties (e.g., hydrophobicity) have been used frequently in predicting the presence of functional or structural domains. However, current methods do not consider the possibility that either amino acid composition or hydrophobicity can be affected by genomic GC contents, which may cause some bias in prediction of function or homology searches, especially in the study of microbial genomes.

In this paper, we first will use *dnaA* protein sequences as an example of a detailed analysis to show the correlation between mutational pattern and amino acid composition. Then, the co-evolution between the genomic GC content and amino acid frequency among bacterial organisms will be examined by Felsenstein's (1985) independent contrast test. Second, we will show that the hydrophobicity of a protein can be affected by mutational pressure. Finally, we will show that these findings hold in general for the evolution of bacterial proteins.

Data and methods

Protein sequences

We collected the DNA sequence data of 15 genes in bacteria from the Genbank.

The *dnaA* gene will be used as an example to show the detailed analysis; this gene encodes a 438 amino acid protein in *E. coli* that is required for the initiation of DNA replication. For each *dnaA* sequence, the Genbank accession number, the species name, and the genomic GC content at the four-fold degenerate sites (see below) are given in Table 1. The information for the other genes is available upon request to Xun Gu.

Table 2. Classification of amino acids [a]

Amino acids	Genetic codons
Group A (high-GC)	
proline (P)	CCU, CCC, CCA, CCG
alanine (A)	GCU, GCC, GCA, GCG
glycine (G)	GGU, GGC, GGA, GGG
tryptophan (W)	UGG [b]
Group B (intermediate-GC)	
valine (V)	GUU, GUC, GUA, GUG
threonine (T)	ACU, ACC, ACA, ACG
histidine (H)	CAU, CAC
glutamic acid (E)	GAA, GAG
aspartic acid (D)	GAU, GAC
glutamine (Q)	CAA, CAG
cysteine (C)	UGU, UGC
serine (S)	UCU, UCC, UCA, UCG, AGU, AGC
Group C (low-GC)	
phenylalanine (F)	UUU, UUC
tyrosine (Y)	UAU, UAC
asparagine (N)	AAU, AAC
lysine (K)	AAA, AAG
isoleucine (I)	AUU, AUC, AUA
methionine (M)	AUG

(*a*) Two amino acids, arginine (R) and leucine (L), are not included, because both of them are encoded by two synonymous codon families that belong to different groups (for detail, see the text). (*b*) The genetic codes for tryptophan (W) in *Mycoplasma* are UGA and UGG, which are different from the universal codes. In our computation, this effect has been considered.

Classification of amino acids

Amino acids can be classified into three groups according to the GC content of their codons (Table 2, see also Jukes & Bhushan, 1986): Group A consists of proline (P), alanine (A), glycine (G), and tryptophan (W), which are encoded by high-GC codons; e.g., alanine is encoded by GCU, GCC, GCA, or GCG. Group B consists of valine (V), threonine (T), histidine (H), glutamic acid (E), aspartic acid (D), glutamine (Q), cysteine (C), and serine (S), which are encoded by codons with an intermediate-GC content; e.g., aspartic acid is encoded by either GAU or GAC. And group C consists of phenylalanine (F), tyrosine (Y), asparagine

(N), lysine (K), isoleucine (I), and methionine (M), which are encoded by low-GC codons; e.g., phenylalanine is encoded by either UUU or UUC. Arginine (R) and leucine (L) are not included in these groups, because R is encoded by an intermediate-GC (AGA, AGG) and a high-GC codon family (CGU, CGC, CGA, CGG), and L is encoded by a low-GC (UUA, UUG) and an intermediate-GC (CUU, CUC, CUA, CUG) codon family. We denote the frequencies of groups A, B, and C by f_A, f_B, and f_C, respectively. Clearly, if amino acid frequencies are affected by the nucleotide composition of the genome, f_A will increase, f_C will decrease, but f_B will change little as GC% increases.

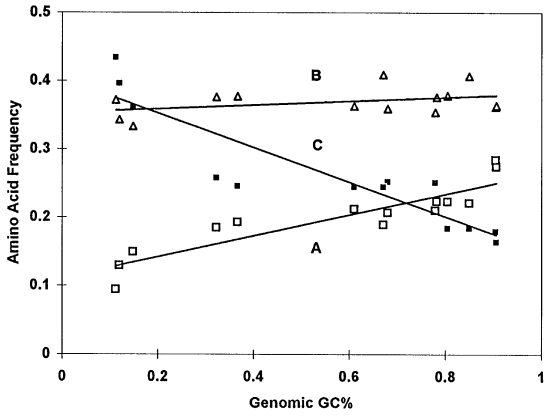

Figure 1. Frequencies of amino acid groups A, B, and C plotted against the genomic GC_4 in *dnaA* gene.

Estimation of the genomic GC content

We use the genomic GC content at the four-fold degenerate sites of genes (GC_4) as a measure of the intensity of the GC mutational pressure, because all nucleotide changes at these sites are synonymous and do not change amino acids. We estimated the GC_4 of a species by the average GC_4 of 100 genes randomly chosen from Genbank, or of all available genes in Genbank, if the number of available genes for any of the species under study is less than 100. Although GC_4 is affected by codon usage bias, it is better than the average GC content of genome, which is strongly affected by the functional constraints of proteins. However, for the species we studied, the average GC content of genome and GC_4 are highly correlated, and using either of them gives essentially the same results (data not shown).

Statistical inferrences

For a given protein, when the genomic GC content (GC_4) and the amino acid frequencies (f_A, f_B, or f_C) are estimated for each species, a straightforward approach to studying the correlation between them is the (cross-species) linear regression. However, this simple analysis may be problematic, because some of the species under study may be closely related.

This non-independence problem may become serious at the species level because of overestimation in the true degree of freedom. Previous studies (e.g., Sueoka, 1961; Muto & Osawa, 1987) did not take this problem into consideration.

Because species generally cannot be considered independent, it was suggested that simple correlational analysis should be conducted at the genus or even family level (e.g., Harvey & Mace, 1982). Although this criterion is empirical, most of the dependency may be taken away by this approach (Harvey & Pagel, 1991). Therefore, in our study, only one sequence from the same genus is used and the coefficient of correlation is denoted as r_g.

Furthermore, we will show that deletion of relatively closely related genera has little effect on or may even increase the correlation.

As another analysis, we use Felsenstein's (1985) independent contrast test to study the above correlation. Given a rooted phylogenetic tree with n species, a total of $n-1$ pairwise contrasts can be obtained for each pair of characters X (the genomic GC%) and Y (the amino acid frequency), respectively. Felsenstein (1985) showed that the pairwise contrasts of X and Y are independent of each other, and the coefficient of correlation between X and Y can be estimated from the contrasts of X and Y. Note that under the gradual (Brownian) model, the variance of X or Y, and the covariance between X and Y, are proportional to the evolutionary time t (in practice, t can be converted into branch length if the molecular clock holds approximately). Therefore, the pairwise contrasts should be standardized, i.e., divided by the square root of the branch length. However, if the change of X (or Y) is punctuational (i.e., branch length independent), the original contrast may be more appropriate. Indeed, our preliminary result has suggested that there is little correlation between the difference of genomic GC% and the divergence time, because closely related bacteria may have very different genomic GC%.

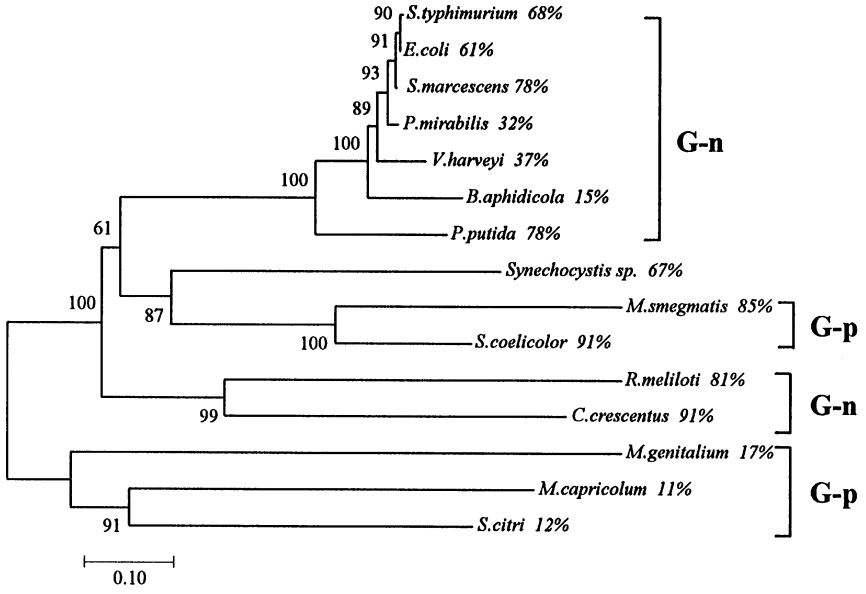

Figure 2. The neighbor-joining tree of *dnaA* protein sequences; the pairwise distance is estimated by $d = -\ln(1-p)$, where p is the proportion of amino acid differences between two sequences. The numbers for interior branches refer to the bootstrapping values for 500 replications. The percentage noted after the species name is the genomic GC content at the four-fold degenerate sites. Gp is short for Gram positive bacteria, and Gn for Gram negative bacteria.

Results

Correlation between GC content and amino acid frequencies

Figure 1 shows that the amino acid composition of the *dnaA* protein varies greatly among the 15 bacterial genera listed in Table 1. The frequency of group A (f_A) is 9% in the low-GC bacterium *Mycoplasma capricolum* ($GC_4 = 11\%$ and the average genomic GC = 27%), but it increases to 28% in the high-GC bacterium *Streptomyces coelicolor* ($GC_4 = 91\%$ and the average genomic GC = 67%). The correlation between f_A and GC_4 at the genus level is highly significant ($r_g = 0.91$, $p < 0.001$). In contrast, f_C is 43% in *M. capricolum* but is reduced to only 18% in *S. coelicolor* ($r_g = -0.92$, $p < 0.001$). There is no correlation between f_B and GC_4. It clearly shows that the amino acid usage of a protein can be very different between high-GC and low-GC bacteria.

Although the above analysis was conducted at the genus level, there is still the possibility that the correlations have been overestimated. To check this point, let us consider the deletion of some relatively closely related genera. We note from Figure 2, which is constructed from protein sequences by the neighbor-joining method (Saitou & Nei, 1987), that *Salmonella typhimurium* and *Serratia marcescens* are rather closely related to *E. coli*, but deletion of these two genera has little effect: the correlation between f_A and GC% reduces from 0.91 to 0.90, and the correlation between f_C and GC% reduces from -0.92 to -0.91. Furthermore, deletion of *S. typhimurium*, *S. marcescens*, *P.mirabilis*, and *V. harveyi* (the four genera relatively closely-related to *E. coli*), actually increases the two correlation coefficients ($f_A \sim GC_4$ and $f_C \sim GC_4$) to 0.93 and -0.98, respectively. This is also true for some other deletions. For example, deletion of *S. typhimurium* and *V. harveyi* also increases the correlations (to 0.92 and -0.94, respectively). A possible reason for the increases is as follows. As is clear from Figure 2 (see the top 6 genera), GC% can change rapidly. On the other hand, amino acid substitution occurs much more slowly because of stronger functional constraints at the protein level. In other words, the effect of change in GC% on amino acid accumulates only gradually.

Next, let us use Felsenstein's (1985) contrast test. Based on the phylogenetic tree given in Figure 2, no correlation was observed between the (absolute) contrast value of GC_4 and the branch length ($r = 0.02, p > 0.8$), which implies that the change of genomic GC content is not strongly associated with the

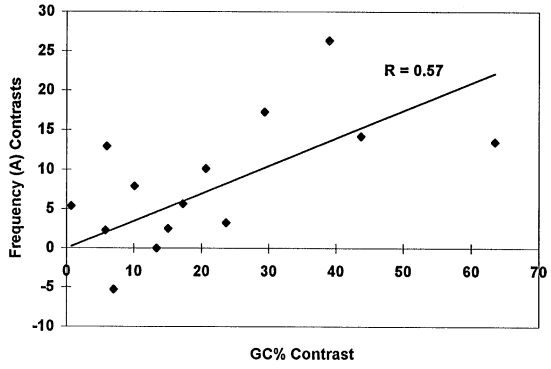

Figure 3. The correlation between the contrast of GC_4 and the (standardized) contrast of group A (amino acid) frequency in *dnaA* gene.

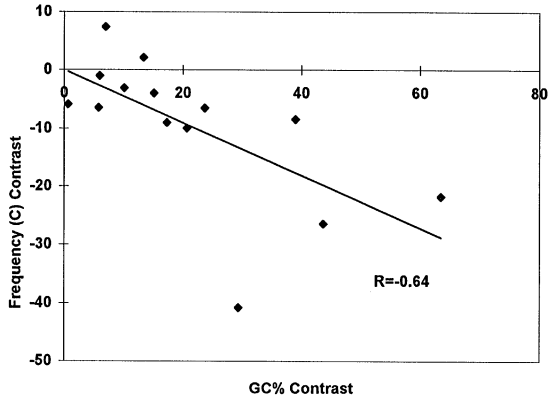

Figure 4. The correlation between the contrast of GC_4 and the (standardized) contrast of group C (amino acid) frequency in *dnaA* gene.

divergence time. Figure 3 shows a significant positive correlation between GC_4 contrast and (standardized) f_A contrast ($r = 0.57, p < 0.05$). If the nonstandardized f_A contrast is used, the coefficient of correlation is $r = 0.573$ ($p < 0.05$). A similar result is obtained for the correlation between f_C and GC_4; $r = -0.64$ ($p < 0.02$) for the standardized contrasts of f_C (Figure 4) and $r = -0.71$ ($p < 0.01$) for the nonstandardized contrast, respectively.

The topology in Figure 2 is largely consistent with previous studies (Woese, 1987; Eisen, 1995; e.g., the rRNA tree), except that the gram-positive group and gram-negative group are not monophyletic; they are separated according to the genomic GC content. Another problem is about the uncertainty of the root. However, the correlations from the independent contrast test were always significant when alternative topologies and roots were used, indicating that the effect of the above problems is not serious.

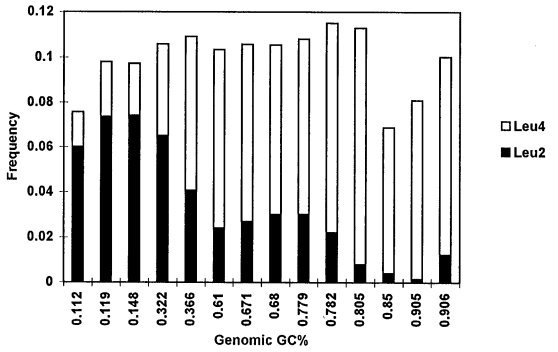

Figure 5. Frequency of amino acid leucine in *dnaA* gene among bacteria. Leu2: the frequency of codon family (UUA, UUG); Leu4: the frequency of codon family (CUU, CUC, CUA, CUG).

Frequencies of leucine and arginine

The frequency of leucine (L) in *dnaA* shows no correlation with GC_4 ($r_g = -0.06, p > 0.80$), because the frequency changes in the two codon families for leucine (Table 2) largely cancel each other, that is, low-GC bacteria use codons UUA and UUG more frequently than codons CUU, CUC, CUA, and CUG, and *vice versa* in high-GC bacteria (Figure 5). A similar pattern was observed in arginine, except that the frequency of arginine increases with increasing GC_4 ($r_g = 0.82, p < 0.001$). This is because in bacteria, high-GC codons (CGT, CGC, CGA, and CGG) for arginine are more abundant than intermediate-GC codons (AGA and AGG).

Effects of GC% on biochemical properties of proteins

An interesting question is whether some of the biochemical properties of a protein are affected by GC mutational pressure. Our analysis shows that the frequencies of positively charged (K, R, H), negatively charged (D, E), and noncharged amino acids (all others) are not significantly correlated with GC_4 in *dnaA*. This can be explained as follows. The two negatively charged amino acids, D and E, are encoded by GC-intermediate codons (Table 2), so that their usage in a protein is little affected by biased GC mutational pressure. For the positively charged amino acids, R is encoded by high-GC and intermediate-GC codons, K is encoded by low-GC codons, and H is encoded by GC-intermediate codons (Table 2), so that the fre-

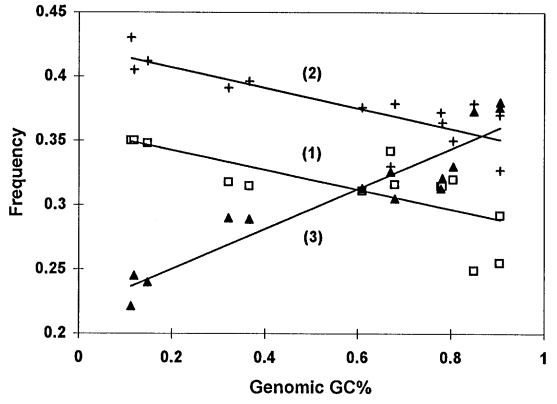

Figure 6. Frequencies of internal amino acids (curve 1), external amino acids (curve 2), and ambivalent amino acids (curve 3) plotted against the genomic GC_4 in *dnaA* gene, respectively.

quency changes of R, K, and H largely cancel each other. For example, in *dnaA*, R is more frequently used than K in *S. coelicolor* ($GC_4 = 91\%$) but the reverse is true in *M. capricolum* ($GC_4 = 11\%$). In short, the total charge of a protein is barely affected by mutation pressure.

Amino acids can also be classified into external, internal, and ambivalent (Dickerson & Geis, 1983). External amino acids (D, E, K, R, H, N, and Q) are polar or charged and found frequently in an aqueous environment (hydrophilic) on the outside of a protein; internal amino acids (F, L, I, M, V, Y, and W) are hydrophobic and tend to be buried on the inside of a protein; and ambivalent amino acids (all others) are neither strongly hydrophobic nor strongly hydrophilic so that they can be either external or internal. In *dnaA* gene, we have shown that both the frequency of external amino acids (f_e) and that of internal amino aicds (f_i) are negatively correlated with GC_4; and the frequency of ambivalent amino acids (f_a is positively correlated with GC_4 (Figure 6). For example, the frequencies of external, internal, and ambivalent amino acids are 43%, 35%, and 22%, respectively, in *M. capricolum* ($GC_4 = 11\%$), but are 37%, 25%, and 38%, respectively, in *S. coelicolor* ($GC_4 = 91\%$). By the simple correlational analysis, the coefficient of correlation between f_e and GC_4 is $r_g = -0.76, p < 0.001$; that between f_i and GC_4 is $r_g = -0.92, p < 0.001$; and that between f_a and GC_4 is $r_g = 0.62, p < 0.02$, respectively. These conclusions have been confirmed by the contrast test: for f_e and GC_4, $r = -0.46, p < 0.05$; for f_i and GC_4, $r = -0.62, p < 0.01$; and for f_a and GC_4, $r = 0.53, p < 0.05$, respectively.

Table 3. Correlations between amino acid frequencies and GC_4 for multiple genes

Genes	n	Correlation coefficient (r_g)		
		f_A	f_B	f_C
dnaA	14	0.91**	0.37	-0.92**
dnaN	12	0.90**	0.46	-0.82**
trpC	12	0.92**	0.25	-0.85**
grpE	12	0.90**	0.01	-0.78**
aroA	16	0.82**	0.00	-0.85**
rpoB	13	0.64**	0.66*	-0.76**
trpE	16	0.86**	0.02	-0.86**
trpD	12	0.78**	0.40	-0.82**
gyrB	13	0.81**	0.62*	-0.84**
gyrA	13	0.91**	0.02	-0.77**
dnaJ	13	0.83**	0.06	-0.78**
trpG	12	0.79**	0.60*	-0.97**
ftsZ	11	0.70*	0.10	-0.80**
nusG	10	0.86**	0.15	-0.80**
hsp70	14	0.68**	0.47	-0.75**

n is the number of genera studied. f_A, f_B, and f_C are frequencies of amino acid groups A, B, and C, which are classified according to the GC content of codons (see text).
*: $p < 0.05$ and **: $p < 0.01$.

General pattern in bacterial protein evolution

For *dnaA* and the other 14 bacterial proteins we have examined, a strikingly general pattern has emerged (Tables 3, 4). First, f_A is always positively correlated with GC_4: all proteins show a significant correlation ($r_g = 0.64 \sim 0.93$). Second, f_C is always negatively correlated with GC_4 ($r_p = 0.75 \sim 0.97$). Third, only a weak correlation between f_B and GC_4 is observed: 12 of the 15 proteins show no significant correlation. Fourth, the frequency of leucine (L) in all 15 proteins shows no correlation with GC_4. Fifth, all the proteins show a similar pattern for the correlation between biochemical properties of proteins and the genomic GC content. Therefore, although the simple correlational analysis at the genus level may overestimate the statistical confidence, it seems clear that the correlation between amino acid composition (or hydrophobicity) and GC_4 is a general feature in bacterial genomes.

Table 4. Correlations between hydrophobicity and GC_4 for multiple genes

Genes	n	Correlation coefficient (r)		
		f_i	f_e	f_a
dnaA	14	-0.76**	-0.62*	0.92**
dnaN	12	-0.74**	-0.83**	0.86**
trpC	12	-0.77**	-0.81**	0.83**
grpE	12	-0.68*	-0.82**	0.89**
aroA	16	-0.72**	-0.72**	0.80**
rpoB	13	-0.44	-0.46	0.57*
trpE	16	-0.53*	-0.73**	0.73**
trpD	12	-0.71**	-0.68*	0.77**
gyrB	13	-0.84**	-0.24	0.84**
gyrA	13	-0.63*	-0.53	0.83**
dnaJ	13	-0.55*	-0.73**	0.76**
trpG	12	-0.88**	-0.56	0.83**
ftsZ	11	-0.29	-0.68**	0.59*
nusG	10	-0.71*	-0.09	0.79*
hsp70	14	-0.55*	0.22	0.14

n is the number of genera studied. f_e, f_i, and f_a are frequencies of external, internal, or ambivalent amino acids, respectively. Note that the classification of external, internal, or ambivalent amino acids is not unambigous, but our result is robust against this variation in classification.
$*$: $p < 0.05$ and $**$: $p < 0.01$.

Discussion

Neutralist vs. selectionist view

To explain the above observations under the selectionist view one has to assume that amino acids encoded by GC-rich codons are advantageous in species with a GC-rich genome, whereas amino acids encoded by GC-poor codons are advantageous in species with a GC-poor genome. It is doubtful that this assumption holds for so many genes in so many genera. Some authors argued that, in thermophilic bacteria, strong preferential usage of thermally stable amino acids encoded by high-GC codons (e.g., alanine and arginine) is adaptive, because these bacteria inhabit very hot niches (Argos et al., 1979; Kagawa et al., 1984). However, the 50 genera included in our study inhabit very different environments but show similar relationships between genomic GC content and amino acid compositions. On the other hand, our results can be easily explained under the neutralist view by assuming that either most of these amino acid changes are nearly neutral or their selective disadvantages are not large, so that substitutions can still occur when mutation pressure is strong.

It is also remarkable that directional mutation pressure can lead to so many exchanges between external or internal and ambivalent amino acids. It is difficult to argue that such exchanges are selectively advantageous, because they may affect to some extent the biochemical properties of the protein. Rather, they are more likely to be disadvantageous. However, the disadvantage may be very small, because ambivalent amino acids can be either external or internal so that exchanges between external (internal) and ambivalent amino acids can occur by mutation pressure. At any rate, the observation is much more difficult to explain by the selectionist view than by the neutralist view.

The pairwise contrast test

We realize that accurate estimation of the correlation between genomic GC content and amino acid composition is not a simple problem, because use of phylogenetically related species can lead to overestimation. In order to reduce the effect of phylogenetic correlation, we used different genera instead of species in the correlation analysis and also used Felsenstein's (1985) contrast test. As indicated by the *dnaA* data, the results were generally consistent, though the contrast test usually gave weaker correlations. Of course, the pairwise contrast test requires several assumptions that may not be realistic, and there is a possibility that the phylogenetic trees used for this test are not correct. Fortunately, after examining several alternative topologies, we found the result robust.

Protein phylogeny, time estimation, and structure/function prediction

The effect of genomic GC content on amino acid composition established from the extensive protein sequence data indicates that the pattern of protein sequence evolution is complex if the mutational pressure varies considerably among species. For example, convergent evolution in amino acids may arise between two lineages if they have similar GC mutational pressures. In this case, a protein phylogeny can be misleading, contrary to the conclusion of Hasegawa et al. (1993). Because the mutational effect is significant even for such highly conservative proteins as Hsp70 and rpoB, the effect of mutation pressure on protein phylogeny and divergence time estimation (Doolittle et al., 1996) needs to be re-examined. For example, in the

phylogenetic tree inferred from *dnaA* sequences (Figure 2), the gram-positive (as well as gram-negative) bacteria are clustered into two groups according to the genomic GC content: the high-GC group of gram-positive bacteria is more closely related to the high-GC group of gram-negative bacteria, and so is the case for the low-GC bacteria. The same problem was also observed in Hsp70 protein sequences, which have been used widely for studying the universal tree of life (Gu, Golding & Li, unpublished data). The effect of genomic GC% on protein phylogeny should be examined in detail in the future.

In studying the structure-function relationships of a protein, particularly when the homology search cannot provide clues on protein function, predictive methods based on amino acid composition have been proposed (e.g., Zhang & Chou, 1992; Nakashima & Nishikawa, 1994; Cedano et al., 1997). These methods assume that a group of proteins with similar function must share a similar amino acid composition and do not consider the effect of mutation pressure on the amino acid composition and biochemical properties of proteins. Thus, it may be problematic when proteins in *Mycoplasma genitalium* ($GC_4 = 27\%$) are treated in the same way as *Streptomyces coelicolor* ($GC_4 = 91\%$). As many bacterial genomes will soon be sequenced, it is obviously important to develop more accurate methods for structure/function prediction.

Acknowledgment

This study was supported by NIH grants to W.H. Li.

References

Argos, P., M.G. Rossmann, U.M. Grau, A. Zuber, G. Frank & J.D. Tratschin, 1979. Thermal stability and protein structure. Biochemistry 18: 5698–5703.

Bronson, E. C. & J. N. Anderson, 1994. Nucleotide composition as a driving force in the evolution of retroviruses. J. Mol. Evol. 38: 506–532.

Cedano, J., A. Patrick, J. Perez-Pons & E. Querol, 1997. Relation between amino acid composition and cellular location of proteins. J. Mol. Biol. 266: 594–600.

Collins, D.W. & T.H. Jukes, 1993. Relationship between G+C in silent sites of codons and amino acid compositions of human proteins. J. Mol. Evol. 36: 201–203.

Dickerson, R.E. & I. Geis, 1983. Hemoglobins: Structure, Function, Evolution, and Pathology. The Benjamin/Cummings Publishing Company, Inc. Menlo Park, CA.

D'Onofrio, G., D. Mouchiroud, B. Aissani, C. Gautier & G. Bernardi, 1991. Correlations between the compositional properties of human genes, codon usage and amino acid composition of proteins. J. Mol. Evol. 32: 504–510.

Doolittle, R.F., D.-F Feng, S. Tsang, G. Cho & E. Little, 1996. Determining divergence times of the major kindoms of living organisms with a protein clock. Science 271: 470–477.

Eisen, J.A., 1995. The RecA proteins as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. J. Mol. Evol. 41: 1105–1123.

Felsenstein, J., 1985. Phylogenies and the comparative method. American Naturalist. 125: 1–15.

Fitch, W.M., 1977. Phylogenies constrained by the crossover process as illustrated by human hemoglobines and a thirteen cycle, eleven-amino-acid repeat in human apolipoprotein A-1. Genetics 86: 623–644.

Harvey, P.H. & G.M. Mace, 1982. Comparisons between taxa and adpative trends: problems of methodology. Current Problems in Sociobiology (ed. King's College Sociobiology group), pp.343–361. Cambridge University Press.

Harvey, P.H. & M.D. Pagel., 1991. The Comparative Methods in Evolutionary Biology. Oxford university press.

Hasegawa, M., T. Hashimota, J. Adachi, N. Iwabe & T. Miyata, 1993. Early branchings in the evolution of eukaryotes: ancient divergence of entamoeba that lacks mitochondria revealed by protein sequence data. J. Mol. Evol. 36: 380–388.

Jukes T.H. & V. Bhushan, 1986. Silent nucleotide substitutions and G+C contents of some mitochondrial and bacterial genes. J. Mol. Evol. 24: 39–44.

Kagawa, Y., N. Nojima, N. Nukiwa, M. Ishizuka, T. Nakajima, T. Yasuhara, T. Tanaka & T. Oshima, 1984. High guanine plus cytocine content in the third letter of codons of an extreme thermophile. J. Biol. Chem. 259: 2956–2960.

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature. 217: 624–626.

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

King, J.L. & T.H. Jukes, 1969. Non-Darwinian evolution. Science 164: 788–798.

Kushiro, A., M. Shimizu & K. I. Tomita, 1987. Molecular cloning and sequence determination of the *tuf* gene coding for the elongation factor Tu of *Thermus thermophilus*. Eur. J. Biochem. 170: 93–98.

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

Moran, N.A., 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA. 93: 2873–2878.

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

Nakashima, H. & K. Nishikawa, 1994. Discrimination of intracellular and extracellular proteins using amino acid compositions and residue-pair frequencies. J. Mol. Biol. 238: 54–61.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Ohta, T., 1973. Slightly deleterious mutant substitutions in evolution. Nature 246: 96–98.

Saitou, N. & M. Nei, 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406–425.

Schachtel, G.A., P. Bucher, E. Mocarski, B.E. Blaisdel & S. Karlin, 1991. Evidence for selective evolution in codon usage in con-

served amino acid segments of human alphaherpesvirus proteins. J. Mol. Evol. 33: 483–494.
Sogin, M.L., G. Hinkle & D.D. Leipe, 1993. Universal tree of life. Nature 362: 795.
Sueoka, N., 1961. Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harbor Symp. Quant. Biol. 26: 35–43.
Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA. 48: 582–592.
Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA. 85: 2653–2657.
Woese, C.R., 1987. Bacterial evolution. Microbiol. Rev. 51: 221–271.
Zhang, Ch. & K. Chou, 1992. A correlation-coefficient method to predicting protein structural classes from amino acid compositions. Eur. J. Biochem. 207: 429–433.