Protein Structure, Neighbor Effect, and a New Index of Amino Acid Dissimilarities

*Xuhua Xia**† *and Zheng Xie*‡

*Bioinformatics Laboratory, HKU-Pasteur Research Center, Hong Kong; †Department of Microbiology, University of Hong Kong; and ‡Institute of Environmental Protection, Hunan University, Changsha, People's Republic of China

Amino acids interact with each other, especially with neighboring amino acids, to generate protein structures. We studied the pattern of association and repulsion of amino acids based on 24,748 protein-coding genes from human, 11,321 from mouse, and 15,028 from *Escherichia coli*, and documented the pattern of neighbor preference of amino acids. All amino acids have different preferences for neighbors. We have also analyzed 7,342 proteins with known secondary structure and estimated the propensity of the 20 amino acids occurring in three of the major secondary structures, i.e., helices, sheets, and turns. Much of the neighbor preference can be explained by the propensity of the amino acids in forming different secondary structures, but there are also a number of intriguing association and repulsion patterns. The similarity in neighbor preference among amino acids is significantly correlated with the number of amino acid substitutions in both mitochondrial and nuclear genes, with amino acids having similar sets of neighbors replacing each other more frequently than those having very different sets of neighbors. This similarity in neighbor preference is incorporated into a new index of amino acid dissimilarities that can predict nonsynonymous codon substitutions better than the two existing indices of amino acid dissimilarities, i.e., Grantham's and Miyata's distances.

Introduction

The genetic variation of protein-coding genes represents a major component in genetic biodiversity, and much effort has been spent in understanding how proteins evolve and diversify by amino acid substitutions. Two approaches have been taken to study the pattern of amino acid substitutions. The first is empirical (Dayhoff and Barker 1972; Dayhoff, Schwartz, and Orcutt 1978; Dayhoff, Barker, and Hunt 1983), based on a comparative analysis of amino acid sequences. The second is parametric, initialized by studies on the relationship between amino acid dissimilarities and substitution patterns (Zuckerkandl and Pauling 1965; Sneath 1966; Epstein 1967; Clarke 1970; Grantham 1974; Miyata, Miyazawa, and Yasunaga 1979; Kimura 1983; Xia and Li 1998), with the objective of building a realistic model of amino acid substitutions. Some of these findings have been incorporated into codon-based or amino acid–based models for phylogenetic analysis (Goldman and Yang 1994; Yang, Kumar, and Nei 1995; Yang, Nielsen, and Hasegawa 1998).

These two approaches assume that amino acid substitutions at different amino acid sites are independent of each other. In other words, the amino acid substitution occurring at site i is irrelevant to what amino acid is found at sites $i - 1$, $i + 1$, or any other site. This assumption is problematic for the following reason. The normal functioning of proteins depends on its three-dimensional conformation that, in the micro scale, depends on the angles of the peptide chain, especially the angles of the $N–C_\alpha$ (ϕ) and the $C_\alpha–C$ (ψ) bonds. According to the Ramachandran plot (Ramachandran and Sasisekharan 1968) and subsequent empirical studies (Morris et al. 1992), only particular combinations of these two angles can give rise to, or maintain, the basic secondary structures such as α-helices or β-sheets. In other words, only particular combinations of amino acids can cooperate to form particular secondary structures. For example, a stretch of Glu, Ala, or Met tends to form an α-helix, but the insertion of Gly or Pro would tend to break the α-helix (Chou and Fasman 1978*a*). This implies that an amino acid substitution at site i may depend on what the neighboring amino acids are.

Different amino acids have different preferences either for or against being in certain secondary structures. For example, Ala and Glu are good α-helix formers, whereas some others such as Gly and Pro tend to disrupt the α-helix structure. Similarly, Ile and Val are good, whereas Glu and Pro are poor β-sheet formers (Chou and Fasman 1974*a*, 1978*b*; Branden and Tooze 1998). This kind of empirical evidence has led to the derivation of the Chou-Fasman conformational parameters that can be used to predict secondary structures of protein molecules (Chou and Fasman 1978*a*). A corollary of this is that α-helix formers should be found more frequently as neighbors, as should β-sheet formers.

The existing conformational parameters (Chou and Fasman 1978*a*) were based on a small data set and should be revised. Proteins with known structures now have accumulated to about 15,000 in the PDB database (Berman et al. 2000). One of the objectives of this paper is to obtain a more updated estimate of the propensity of the 20 amino acids occurring in the three major secondary structures, i.e., helices, sheets, and turns. This will not only complement a previous study on interactions of non-neighbor amino acids (Singh and Thornton. 1992), but will also help us to better interpret the neighbor preference of amino acids.

A study on neighboring amino acids can also shed light on amino acid dissimilarities. Two indices of amino acid dissimilarities have been proposed (Grantham

Key words: protein structure, neighbor preference, amino acid, amino acid distance, phylogenetics.

Address for correspondence and reprints: Xuhua Xia, Bioinformatics Laboratory, HKU-Pasteur Research Center, Dexter H.C. Man Building, 8 Sassoon Road, Pokfulam, Hong Kong. E-mail: xxia@hkusua.hku.hk.

Mol. Biol. Evol. 19(1):58–67. 2002
 ISSN: 0737-4038

1974; Miyata, Miyazawa, and Yasunaga 1979), with Grantham's distance based on the volume, the polarity, and the chemical property of the side chain, and Miyata's distance based on the first two amino acid properties. Amino acids can differ in many ways, and Sneath (1966) has indeed listed 134 properties. Of the 10 properties studied in detail, all exhibit significant relationship with substitution rates (Xia and Li 1998), suggesting that they are all important properties related to the normal functioning of proteins. It is difficult to agree upon which amino acid properties should be used to construct an index of amino acid dissimilarities, and the choice of three properties to build Grantham's distance and two properties to build Miyata's distance is, to a large extent, arbitrary.

The arbitrary choice of amino acid properties and potentially false formulation of the amino acid dissimilarities may be responsible for some of the old controversies between Kimura (1983, p. 159) and Gillespie (1991, p. 43). Kimura, being a neutralist, argued that the most frequent nonsynonymous substitutions were those involving similar amino acids and the substitution rate would decrease monotonously with increasing dissimilarity between involved amino acids (fig. 7.1 in Kimura 1983). This is of course what one would expect from the neutral theory of molecular evolution, in which positive selection plays a negligible role in molecular evolution and purifying (negative) selection eliminates those mutations with major effects. Gillespie, on the other hand, argued that the most frequent nonsynonymous substitutions were not between the chemically most similar amino acids, but instead were between amino acids with a Miyata's distance near 1 (fig. 1.12 in Gillespie 1991). It is difficult to appreciate or interpret the latter finding, and we are inclined to think that the finding may be an artifact because of inappropriate formulation of amino acid dissimilarities. For example, those amino acid pairs with a Miyata's distance near 1 may actually be more similar to each other than what Miyata's distances would let us believe. The peak of substitutions at Miyata's distance near 1 may disappear when better indices are formulated.

An entirely different approach to study amino acid dissimilarities is to look at whether two amino acids have similar sets of neighbors. We know that an amino acid in a protein needs to interact with neighbors in certain ways to maintain the normal functional structure of the protein. If an amino acid has no preference for its neighbors, then the probability of having one particular amino acid as its neighbor is simply the proportion of the amino acid among all 20 amino acids. The deviation from this random expectation represents the degree of preference for its neighbors. If two amino acids have strong but identical preferences for the same set of amino acids as their neighbors, then we can say that the two amino acids are functionally equivalent, no matter how they differ in their amino acid properties. This would seem to be a more objective way of obtaining amino acid dissimilarities, objective in the sense that we do not need to choose arbitrarily two or three out of many amino acid properties to build a dissimilarity index.

This paper has three objectives. The first is to estimate the propensity of the 20 amino acids occurring in the three major categories of secondary structures, i.e., helices, sheets, and turns, by using the large number of proteins now available with known structures. The second is to document the genomic pattern of neighbor preference for the 20 amino acids by taking advantage of the huge amount of available protein data and interpret the neighbor preference with reference to protein secondary structures. The third is to incorporate the differences in neighbor preference between amino acids into a new formulation of amino acid dissimilarity index.

Materials and Methods

Propensity of Amino Acids Occurring in Helices, Sheets, and Turns

We retrieved 7,342 proteins with known structures from the PDB database (Berman et al. 2000), extracted helices, sheets, and turns according to the PDB Format Description, Version 2.2, and counted the frequency distribution of amino acids in each of the three structure categories. A total of 935 files did not conform to the format description and were discarded.

The propensity of an amino acid occurring in one of the three structure categories is calculated as follows. Let N_{Tot} be the total number of amino acids in the three structure categories; N_i (where $i = 1, 2, \ldots, 20$ corresponding to the 20 amino acids) be the number of amino acid i found in all three structure categories; N_h, N_s, and N_t be the number of amino acids found in helices, sheets, and turns, respectively; and $N_{h,i}$, $N_{s,i}$, and $N_{t,i}$ be the number of amino acids in helices, sheets, and turns, respectively. If amino acids occur equally likely in the three secondary structures, then the expected numbers of $N_{h,i}$, $N_{s,i}$, and $N_{t,i}$ are, respectively,

$$E(N_{h,i}) = \frac{N_h N_i}{N_{Tot}}; \qquad E(N_{s,i}) = \frac{N_s N_i}{N_{Tot}};$$

$$E(N_{t,i}) = \frac{N_t N_i}{N_{Tot}} \tag{1}$$

The propensity of amino acid i occurring in helices is defined as

$$P_{h,i} = \frac{N_{h,i} - E(N_{h,i})}{N_i} \tag{2}$$

$P_{h,i}$ measures how strongly an amino acid is associated with one particular secondary structure and is independent of sample size. $P_{s,i}$ and $P_{t,i}$ are calculated in the same way. We retrieved only 7,342 proteins instead of all proteins in the PDB database, because the $P_{h,i}$, $P_{s,i}$, and $P_{t,i}$ values are stabilized after analyzing just 3,000 protein structures. Using more data will not change the $P_{h,i}$, $P_{s,i}$, and $P_{t,i}$ values.

Neighbor Preference in Amino Acids

A total of 25,467 protein-coding sequences (CDS) from human (*Homo sapiens*), 11,490 CDS from mouse (*Mus musculus*), and 15,028 CDS from *Escherichia coli* were retrieved and translated into protein sequences by using the ACNUC retrieval system (Gouy et al. 1985). We excluded from further analysis 719 human CDS, 169 mouse CDS, and 20 E. coli CDS, which contain embedded stop codons. These sequences are likely pseudogenes and are irrelevant to this study.

Some genes have been sequenced and deposited in GenBank multiple times, and this may bias the result in the way that the observed pattern of neighbor preference may not reflect the genomic pattern but instead may reflect the pattern of those over-represented genes. For this reason, we have also analyzed all 4,289 CDS in the complete genome of *E. coli* K-12. The *E. coli* genomic data set will be referred to as *E. coli*^{G} hereafter.

With 20 amino acids, there are 400 possible amino acid doublets (i.e., neighbors). Let N_{ij} (where i and j = 1, 2, . . . , 20 corresponding to the 20 amino acids) be the number of amino acid pairs, with amino acid j following amino acid i. For example, $N_{Ala,Arg}$ is the number of Ala-Arg pairs in all sequences; $N_{Arg,Ala}$ is the number of Arg-Ala pairs in all sequences, and so on. The counting is from the N-terminal to the C-terminal of the amino acid sequences. The first methionine is not counted. Data extraction is done with DAMBE (Xia 2000).

For amino acid i, all 20 N_{ij} values, with j = 1, 2, . . . , 20 corresponding to the 20 amino acids, make up a profile of neighbor preference for amino acids found after amino acid i, and all 20 N_{ji} values makes another profile for amino acids found before amino acid i along the amino acid sequence. The former set will be referred hereafter as the $Profile_a$ of amino acid i, with the subscript a meaning after. The latter set of 20 N_{ji} values will be referred hereafter as the $Profile_b$ of amino acid i, with the subscript b meaning before.

The N_{ij} values apparently depend on amino acid usage. If amino acid j is very abundant, then obviously N_{ij} and N_{ji} will be large, too. If amino acid i does not have any neighbor preference, then the expected value for N_{ij} is

$$E(N_{ij}) = P_j \sum_{j=1}^{20} N_{ij} \tag{3}$$

where P_j is the frequency of amino acid j. For example, if $P_{ala} = 0.1$, and the sum of the 20 $N_{Gly,j}$ values is 10,000 (i.e., Gly has 10,000 downstream neighbors), then the expected value of $N_{Gly,Ala}$ is 1,000 ($=0.1 \times 10{,}000$). Given our reasoning (see *Introduction*) we expect certain amino acids to be neighbors more likely than expected from random association. For example, good α-helix formers should be more likely to be neighbors, as should β-sheet formers.

Whether the 20 N_{ij} values for amino acid i deviate significantly from the expectation of random association can be tested by a chi-square goodness-of-fit test with

$$\chi_i^2 = \sum_{j=1}^{20} \frac{[N_{ij} - E(N_{ij})]^2}{E(N_{ij})}. \tag{4}$$

The degree of freedom associated with χ^2 is 19 rather than 18 because P_j is not calculated from the 20 N_{ij} values.

The strength of the neighbor preference for amino acid i (SP_i) can be simply measured by

$$SP_i = \sqrt{\frac{\chi_i^2}{\sum_{j=1}^{20} N_{ij}}}. \tag{5}$$

Note that we should not use χ^2 directly to measure the strength of preference because the χ^2 value depends on the sample size, i.e., a more abundant amino acid tends to yield a large χ^2 value than a less abundant amino acid, everything else being equal. In contrast, SP_i is independent of sample size and can therefore facilitate comparisons among amino acids. As SP_i can only take positive values and therefore cannot indicate which amino acid is favored or disfavored by amino acid i, we also use the following index (I_{ij}) to measure the preference of amino acid i for amino acid j:

$$I_{ij} = \frac{N_{ij} - E(N_{ij})}{E(N_{ij})}. \tag{6}$$

Apparently, I_{ij} will be positive if amino acid i has amino acid j as its neighbor more frequently than expected, and negative if amino acid i has amino acid j as its neighbor less frequently than expected.

N_{ij} may differ from N_{ji}, i.e., amino acid i may have different preferences for amino acids that go before it and those that go after it. This difference, or similarity, between these two profiles can be measured by the Pearson correlation coefficient between the 20 N_{ij} values and the 20 N_{ji} values (where j = 1, 2, . . . , 20). Note that such correlation coefficients measure only the similarity between $Profile_a$ and $Profile_b$. They do not measure the strength of preferences. For example, if there is no preference at all, then $Profile_a$ and $Profile_b$ will both be expected to approach the relative abundance of the 20 amino acids, and will have a correlation coefficient near 1 given the large data set.

If two amino acids, x and y, have similar neighbor preference, then N_{xj} and N_{yj} will be highly correlated, and we can use the correlation coefficient to measure similarity in neighbor preference between the two amino acids. Alternatively, we can treat the 20 N_{ij} values as allele frequencies for one locus, and calculate a pairwise genetic distance between amino acids by using genetic distances based on allele frequencies (e.g., Cavalli-Sforza and Edwards 1967; Nei 1972; Reynolds, Weir, and Cockerham 1983). The amino acid distance based on similarity in neighbor preference will be referred to hereafter as D_{np}, with np standing for neighbor preference.

To test whether D_{np} is related to the rate of amino acid substitutions, we compiled substitution data from two sets of protein-coding sequences. One set consists of 58 presumably orthologous genes from the human,

Table 1
Frequency Distribution of Amino Acids in Helices, Sheets, and Turns, Together with Calculated Propensity (P) of the Amino Acids to Occur in These Secondary Structures. P_h and P_s Are Strongly and Negatively Correlated

AA	Helix	Sheet	Turn	P_h	P_s	P_t
Ala	679,844	223,981	200,056	0.1046	−0.1001	−0.0044
Arg	338,964	133,987	8,505	0.0727	−0.0643	−0.0085
Asu	234,507	106,612	15,800	0.0257	−0.0439	0.0181
Asp	328,812	130,958	20,586	0.0532	−0.0699	0.0167
Cys	76,692	64,848	4,168	−0.1050	0.1025	0.0025
Gln	263,627	103,004	7,817	0.0727	−0.0675	−0.0053
Glu	466,583	158,786	14,058	0.0984	−0.0942	−0.0041
Gly	302,707	202,319	32,948	−0.0686	0.0335	0.0351
His	128,750	80,222	5,673	−0.0315	0.0312	0.0003
Ile	317,142	272,940	7,490	−0.1006	0.1142	−0.0136
Leu	639,508	301,347	14,699	0.0380	−0.0272	−0.0108
Lys	399,975	161,685	13,806	0.0637	−0.0616	−0.0021
Met	220,677	99,733	7,026	0.0427	−0.0380	−0.0047
Phe	231,278	164,420	5,924	−0.0554	0.0668	−0.0114
Pro	159,634	93,287	16,138	−0.0380	0.0041	0.0338
Ser	306,728	187,261	18,716	−0.0330	0.0227	0.0104
Thr	289,276	225,337	16,455	−0.0866	0.0817	0.0049
Trp	87,666	60,463	2,834	−0.0506	0.0579	−0.0074
Tyr	198,432	150,680	6,982	−0.0741	0.0806	−0.0065
Val	393,229	368,733	11,344	−0.1228	0.1343	−0.0115

the mouse, and the cow, and the other is made of the 13 protein-coding genes from each of the 19 completely sequenced mitochondrial sequences used in Xia (1998). The ancestral sequences were reconstructed using the CODEML program in the PAML package (Yang 2000), with jones.dat for the nuclear genes and mtmam.dat for the mitochondrial genes. Pair-wise comparisons were made between neighboring nodes along the tree. The tree for the first data set with only three operational taxonomic units (OTUs) is simply a trifurcating tree with one internal node, and the tree for the second data set is the same as in Xia and Li (1998). The number of substitutions involving amino acids i and j is designated as NS_{ij}.

We expect NS_{ij} to be large between similar amino acids and small between different amino acids. However, NS_{ij} values depend not only on the amino acid dissimilarities, but also on the frequencies of the amino acids involved. For example, $NS_{i,j}$ (where i, j = 1, 2, . . . , 20 corresponding to the 20 amino acids and $i \neq j$) will necessarily be zero if the sequences contain no amino acid i or amino acid j. Thus, NS_{ij} should be adjusted for amino acid frequencies before it is used to evaluate indices of amino acid distances.

Let P_i (where i = 1, 2, . . . , 20 corresponding to the 20 amino acids) be the frequency of amino acid i in the set of amino acid sequences, and N_s be the total number of amino acid substitutions. The expected value of $NS_{i,j}$, when amino acids replace each other randomly, is

$$E(NS_{ij}) = \frac{N_s P_i P_j}{\sum_{i=1}^{20} P_i \sum_{j=2, j>i}^{20} P_j}. \quad (7)$$

The quantity

$$R_{ij} = NS_{ij} - E(NS_{ij}) \quad (8)$$

can then be taken as a measure of substitution rate for evaluating the indices of amino acid dissimilarities. Whether one amino acid distance is better than others depends on whether it can predict the R_{ij} values better than others.

Another method for evaluating the relative performance of different amino acid distances is to apply them in a likelihood-based phylogenetic analysis (Yang, Nielsen, and Hasegawa 1998). The best distance should generate larger likelihood values than other distances. For this purpose, we have used the 13 protein-coding genes from six OTUs, with two chimpanzees (GenBank LOCUS names: CHPMTB and CHPMTE), one gorilla (GGMTG), one human (HSMITG), one orangutan (ORAMTD), and one gibbon (HLMITCSEQ).

Results and Discussion

Propensity of Amino Acids to Occur in Helices, Sheets, and Turns

Different amino acids have strong association with particular secondary structures, with Ala and Glu found most frequently in helices, Val, Cys, and Ile found most frequently in sheets and Gly and Pro found most frequently in turns (table 1). A dendrogram of amino acids, based on the average linkage clustering method on the P_h, P_s, and P_t values, grouped the helix-forming amino acids in one cluster, the sheet-forming amino acids in another, and the turn-forming amino acids (Gly, Pro) in the third (fig. 1).

The P_h and P_s values are positively correlated with the helix- and sheet-forming propensities (Chou and Fasman 1978*a*), with r being 0.5742 and 0.7229, respectively. P_t is correlated with the turn-forming propensity (Fasman and Chou 1974), with r = 0.7977. It has long been known that Pro and Gly often occur together in reverse turns (Schulz and Schirmer 1979, p. 111; Thornton 1992; Creighton 1993, Pp. 225–226;).

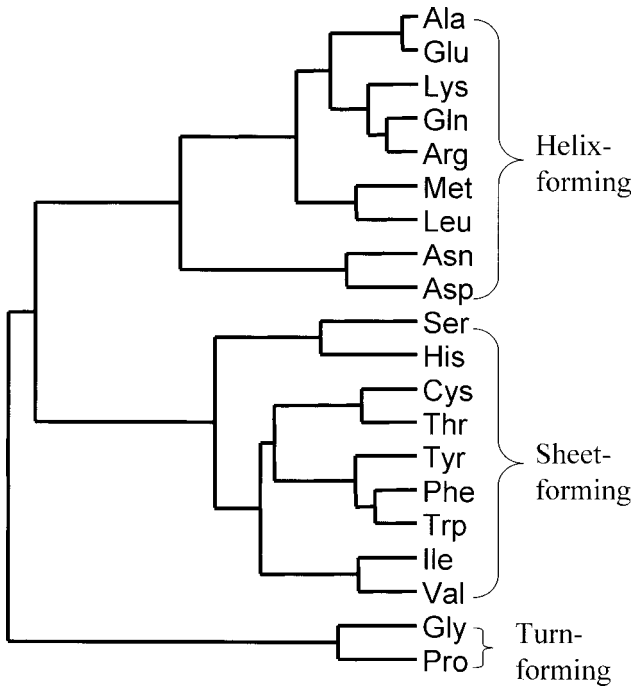

FIG. 1.—Dendrogram of the 20 amino acids based on their propensity to occur in helices, sheets, and turns.

For example, Pro occurs frequently in such turns typically at position i + 1. The turn requires a residue with a positive φ angle and Gly, having no side chain to constrain the angle, is one of the few amino acids that can take such a conformation.

The three aromatic amino acids (Tyr, Phe, and Trp) are clustered together and tend to occur in sheets (table 1). Aromaticity can affect the rate of amino acid substitutions (Xia and Li 1998), and our observation that they are all sheet-formers suggests that the replacement

Table 2
Amino Acid Usage (%) Based on 6,407 Proteins in the PDB Database (from PDB), on the 4289 CDS in *E. coli* K-12 genome (*E. coli*^{G}), and on the Sequences in GenBank for the Three Species

AA	From PDB	*E. coli*^{G}	*E. coli*	Human	Mouse
Ala	8.7233	9.5156	9.4163	7.0431	6.9145
Arg	4.7054	5.5572	5.5660	5.5667	5.5122
Asn	4.3958	3.9653	4.0672	3.7398	3.7131
Asp	5.7834	5.1534	5.1965	4.9099	4.9074
Cys	1.4881	1.1743	1.1415	2.2247	2.3452
Gln	3.7046	4.4436	4.3766	4.6207	4.6353
Glu	6.1897	5.7645	5.8062	6.9512	6.7225
Gly	7.7763	7.3926	7.3609	6.7576	6.8129
His	2.2690	2.2812	2.2320	2.4969	2.5188
Ile	5.3053	6.0236	5.9990	4.5217	4.4387
Leu	8.4923	10.6731	10.3972	9.7560	9.7087
Lys	5.9917	4.4208	4.6156	5.6947	5.6302
Met	3.0119	2.5431	2.7162	2.2326	2.2435
Phe	3.8491	3.9114	3.8536	3.7514	3.7621
Pro	4.6081	4.4416	4.3379	6.0910	6.1794
Ser	6.0741	5.8438	6.0202	7.9078	8.1370
Thr	5.8517	5.4239	5.4725	5.3887	5.4710
Trp	1.4446	1.5322	1.4435	1.2933	1.2924
Tyr	3.4318	2.8580	2.9124	2.8634	2.8827
Val	6.9040	7.0810	7.0685	6.1887	6.1724

Table 3
Neighbor Preference Data (N_{ij} values in rows and N_{ji} values in columns) from *E. coli*^{G}. Each Row Represents one Profile of Neighbor Preference. For Example, the Row Headed by Ala Is the Profile$_a$ of Alanine

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	12,813	7,040	4,336	6,599	1,550	5,816	7,700	9,669	2,502	7,952	14,894	5,482	3,566	4,744	4,678	7,049	6,411	2,144	2,601	8,881
Arg	5,850	4,670	2,830	3,977	845	3,923	5,091	4,435	2,105	4,497	8,177	3,346	1,732	3,303	2,777	3,677	3,356	1,268	2,880	5,008
Asn	4,898	2,728	2,442	2,857	558	2,268	2,803	4,436	1,175	3,463	4,879	2,359	1,202	1,928	2,871	2,866	2,769	843	1,699	3,618
Asp	6,697	3,341	3,104	3,899	744	2,162	4,735	4,967	1,320	4,692	6,174	3,516	1,537	2,773	2,925	3,719	3,365	1,085	2,574	5,081
Cys	1,395	911	501	864	294	663	897	1,492	471	851	1,468	480	288	649	768	1,016	724	281	516	1,077
Gln	5,837	3,904	2,112	2,323	520	4,332	3,064	3,984	1,690	3,135	6,671	2,548	1,414	2,031	2,881	3,087	2,971	979	1,677	3,826
Glu	6,847	4,790	3,417	3,229	715	4,607	4,894	4,844	1,925	4,725	8,059	4,602	2,200	2,487	2,750	4,007	3,954	1,108	2,127	5,070
Gly	8,181	4,873	3,901	4,908	1,435	3,693	6,109	7,238	2,184	6,850	9,804	5,500	2,931	4,388	2,672	5,383	5,049	1,703	3,558	7,825
His	2,518	1,673	1,210	1,559	480	1,552	1,544	2,310	1,006	1,862	2,970	1,080	588	1,388	1,784	1,710	1,436	538	1,235	1,787
Ile	8,782	4,082	3,865	4,774	1,048	2,504	4,426	6,289	1,669	4,898	6,845	3,396	1,799	3,004	3,891	5,394	4,909	962	2,153	5,462
Leu	14,843	8,071	5,905	6,938	1,771	5,498	7,361	9,368	3,009	7,977	15,538	6,166	3,730	5,703	7,567	9,127	8,496	1,937	3,409	9,389
Lys	5,761	3,516	2,589	2,874	447	2,734	3,647	4,024	1,221	3,427	5,968	3,126	1,409	1,663	2,978	3,152	3,553	636	1,574	4,142
Met	3,492	1,807	1,265	1,508	305	1,533	1,578	2,523	646	1,922	4,089	1,417	1,152	1,134	1,728	2,011	2,144	354	705	2,533
Phe	5,066	2,516	2,427	3,078	778	1,403	2,358	4,263	1,104	3,529	4,482	1,838	1,234	2,313	2,076	4,200	3,323	858	1,649	3,458
Pro	5,630	2,651	1,720	3,607	559	3,125	4,569	4,705	1,404	2,599	6,824	2,112	1,377	2,494	2,196	2,873	2,813	1,001	1,612	5,173
Ser	7,107	4,554	2,678	3,992	889	3,353	4,294	6,992	1,904	4,115	8,849	2,834	1,812	3,060	3,429	4,663	4,026	1,306	2,179	5,551
Thr	6,787	4,031	2,261	3,547	756	2,924	3,518	6,091	1,679	4,154	9,610	2,119	1,622	2,872	4,271	3,890	4,048	1,141	1,600	5,280
Trp	1,368	1,467	668	839	282	1,710	775	1,255	604	1,034	3,094	734	575	937	873	1,132	792	331	616	1,259
Tyr	3,278	2,445	1,447	2,119	541	2,122	1,761	2,992	1,003	2,010	4,013	1,381	753	1,631	1,860	2,421	2,045	590	1,310	2,234
Val	9,275	4,725	3,879	5,052	1,130	3,101	5,475	6,469	1,701	6,345	9,509	4,353	2,828	3,447	4,005	5,933	5,744	1,319	2,350	7,567

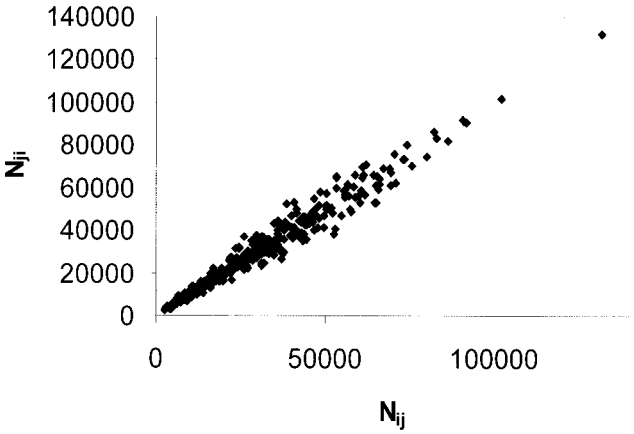

FIG. 2.—N_{ij} and N_{ji} values are similar, as illustrated here with human data. The same pattern is also true for the mouse and *E. coli*.

of a helix-forming amino acid (which are all nonaromatic) by one of these aromatic amino acids may destabilize the secondary structure. Consequently, purifying selection should act against such replacements. The similarity in sheet-forming among these amino acids represents a new dimension of similarity that is ignored by previous formulation of amino acid distances.

Amino Acid Usage

Amino acid usage for the human and mouse sequences are very similar (table 2), with a Pearson correlation coefficient equal to 0.999, suggesting that amino acid usage is conserved among distantly related mammalian species. The correlation coefficient is 0.903 between *E. coli* and human and 0.894 between *E. coli* and mouse. The correlation coefficient between *E coli* and *E coli*^{G} is 0.9991, suggesting that, at the amino acid usage level, the potential bias caused by differential representation of genes in GenBank is not obvious. The amino acid usage from the PDB database is closer to *E. coli* than to the two mammalian species, with the correlation coefficient being 0.9607, 0.9519, 0.9005, and 0.8920, respectively, for *E coli*^{G}, *E coli*, human, and mouse.

Neighbor Preference in Amino Acids

The N_{ij} and N_{ji} values are generally very similar (table 3 and fig. 2) which is true for all the three species studied. However, for each of the 20 amino acids, the N_{ij} values deviate highly significantly from $E(N_{ij})$ for all the three species, with $P = 0.0000$ for all the species and for all individual χ^2 tests, of which one is illustrated in table 4, for amino acid alanine. Different amino acids exhibit different degrees of neighbor preference, with Glu and Pro consistently having strong preference in all the three species (table 5). Glu is the best α-helix former and Pro is the ultimate α-helix breaker (Chou and Fasman 1974*a*, 1974*b*, 1978*a*). It is understandable that they should occur mostly in particular combinations of amino acids. The amino acids with the least preference are Leu and Val, which happen to be two of the three most typical amino acids (Sneath 1966). The typicalness of an amino acid in Sneath (1966) is measured by the

Table 4
χ^2-Test of Goodness-of-Fit for Profile$_b$ of Alanine, Based on *E. coli*^{G}. The Column Headed by χ^2 Shows the Individual Terms of the χ^2 Statistic, i.e., $(N_{ij} - E[N_{ij}]^2/E(N_{ij})$, and the Last Column $I_{ala,j}$, Is Calculated According to Equation (6)

AA	$N_{ala,j}$	$E(N_{ala,j})$	χ^2	$I_{ala,j}$
Ala....	12,813	12030.3	50.9	0.0651
Arg....	7,040	7025.8	0.0	0.0020
Asn ...	4,336	5013.2	91.5	−0.1351
Asp ...	6,599	6515.3	1.1	0.0128
Cys....	1,550	1484.6	2.9	0.0440
Gln....	5,816	5617.9	7.0	0.0353
Glu....	7,700	7287.9	23.3	0.0565
Gly....	9,669	9346.2	11.1	0.0345
His....	2,502	2884.0	50.6	−0.1325
Ile.....	7,952	7615.5	14.9	0.0442
Leu....	14,894	13493.7	145.3	0.1038
Lys....	5,482	5589.0	2.0	−0.0192
Met ...	3,566	3215.1	38.3	0.1091
Phe....	4,744	4945.1	8.2	−0.0407
Pro....	4,678	5615.4	156.5	−0.1669
Ser....	7,049	7388.2	15.6	−0.0459
Thr....	6,411	6857.2	29.0	−0.0651
Trp....	2,144	1937.1	22.1	0.1068
Tyr....	2,601	3613.3	283.6	−0.2802
Val....	8,881	8952.3	0.6	−0.0080
χ^2 =			954.5	
DF =			19	
P =			0.0000	

average differences between the amino acid and all the other amino acids, with the most typical amino acid having the smallest mean difference. The weak preference of these two amino acids suggests that they are general-purpose amino acids that can perhaps be put anywhere in protein molecules.

One particular neighbor preference in amino acids that is consistent in the three species is the preference of its own kind, with the only exception of Pro in *E. coli* (table 6). There are several possible explanations for the self preference. One reviewer suggested that many proteins are transmembrane, and similar amino acids would cluster in the intramembrane hydrophobic and cytoplasmic hydrophilic regions. An alternative explanation is replication slippage leading to stretches of identical codons.

Pro-Pro doublets are common in mammalian proteins, but rare in *E. coli* proteins. In general, the self preference is weaker in *E. coli* than in the two mammalian species. This corroborates recent studies (Nishizawa and Nishizawa 1999; Nishizawa, Nishizawa, and Kim 1999) showing that modern proteins have a tendency for repetitive use of the same amino acid at a local scale, whereas this local repetitiveness is weak in ancient proteins, e.g., human homologues of *E. coli* proteins.

Aside from the self preference, different amino acids also exhibit association and repulsion with other amino acids. A subset of these association and repulsion patterns, with I_{ij} values either greater than 0.2 or lesser than −0.2 is shown in table 7. All of these associations and repulsions can be easily explained with reference to figure 1. In general, those amino acids with a high propensity for occurring in the same secondary structure

Table 5
Strength of Neighbor Preference, i.e., SP_i Values in Equation (5). Values Greater than 0.14 Are in Bold

	Human		Mouse		*E. coli*		*E. coli*^{G}	
AA	Before	After	Before	After	Before	After	Before	After
Ala	0.1383	**0.1496**	0.1323	**0.1495**	0.0895	0.0875	0.0869	0.0913
Arg	0.1018	0.1153	0.1088	0.1141	0.1393	0.1004	**0.1446**	0.1066
Asn	0.1072	0.1179	0.1100	0.1049	0.0986	0.1373	0.0976	0.1338
Asp	0.0933	0.1139	0.0981	0.1190	0.1334	0.1033	0.1345	0.1053
Cys	0.1104	0.1165	0.0934	0.1154	**0.1689**	**0.1796**	**0.1648**	**0.1804**
Gln	0.1261	0.1186	0.1363	0.1362	**0.1710**	**0.2466**	**0.1776**	**0.2627**
Glu	**0.2214**	**0.1950**	**0.2173**	**0.1933**	**0.1450**	**0.1396**	**0.1492**	**0.1421**
Gly	0.0811	0.1282	0.0765	0.1260	0.1360	0.1147	0.1463	0.1123
His	**0.1650**	0.1180	**0.1627**	0.1136	**0.1634**	**0.1530**	**0.1669**	**0.1536**
Ile	0.1255	0.1304	0.1200	0.1356	0.1252	0.1070	0.1365	0.1064
Leu	0.0805	0.0861	0.0755	0.0841	0.0913	0.1151	0.0901	0.1232
Lys	**0.1577**	**0.1656**	**0.1556**	**0.1622**	0.1043	**0.1699**	0.1107	**0.1762**
Met	0.1339	0.1068	**0.1478**	0.1104	0.1331	0.1329	0.1324	0.1330
Phe	0.1284	0.1263	0.1307	0.1237	0.1766	0.1075	**0.1836**	0.1105
Pro	**0.2015**	**0.1743**	**0.1972**	**0.1658**	**0.1805**	**0.1868**	**0.1758**	**0.1916**
Ser	0.1320	**0.1573**	0.1253	**0.1476**	0.0922	0.1189	0.0893	0.1201
Thr	0.0908	0.0752	0.0889	0.0792	**0.1467**	0.1064	0.1551	0.1073
Trp	0.1201	0.1094	**0.1416**	0.1248	**0.2906**	0.1193	**0.3039**	0.1305
Tyr	0.1249	0.1199	0.1263	0.1196	0.1293	**0.1876**	0.1316	**0.2022**
Val	0.0831	0.0811	0.0859	0.0738	0.0971	0.0928	0.1025	0.0912

are associated and those with a high propensity for occurring in different secondary structures are repulsive.

One might wonder why Pro and Trp are not associated because both occur very frequently in reverse turns (typically made of four amino acid residues indexed, i, i + 1, i + 2, and i + 3). This is because Pro is almost exclusively found in position i + 1 and Trp appears to occur only in position i + 3, i.e., they do not occur as immediate neighbors (Schulz and Schirmer 1979, p. 111; Thornton 1992; Creighton 1993, pp. 225–226).

Amino Acid Distance Based on Neighbor Preference

We have so far focused only on the neighbor preference of individual amino acids, but have not yet studied the similarity in neighbor preference between amino acids. We could measure the similarity in neighbor preference between amino acids x and y by calculating the Pearson correlation coefficient between the N_{ij} values for x and the N_{ij} values for y. However, the correlation coefficient measuring the similarity between amino acids is not convenient for comparison with other indices such as Grantham's and Miyata's distances that measure the dissimilarity but not the similarity between amino acids. An alternative measure of amino acid dissimilarities in neighbor preference is to treat the profile for each amino acid as one locus with 20 alleles, i.e., 20 N_{ij} values. We can then calculate a genetic distance by using available formulation of genetic distances (e.g., Cavalli-Sforza and Edwards 1967; Nei 1972; Reynolds, Weir, and Cockerham 1983). In this study, we used Nei's method and the *E. coli*^{G} data to obtain D_{np}, with the subscript np standing for neighbor preference.

The reason for deriving D_{np} values from the *E. coli*^{G} data is that modern proteins tend to make repetitive use of the same amino acids, whereas ancient proteins

Table 6
Preference of Its Own Kind Measured by I_{ij}Values According to Equation (6). Those I_{ii} Values Larger than 0.4 Are in Bold. Ala Means Ala-Ala Doublet, Arg Means Arg-Arg Doublet and So on

AA	Human	Mouse	*E. coli*	*E. coli*^{G}
Ala	**0.4053**	**0.4114**	0.0943	0.0651
Arg	0.3500	0.3792	0.1209	0.1391
Asn	0.1371	0.0985	0.1826	0.1706
Asp	0.0987	0.1031	0.0887	0.1049
Cys	0.3238	0.3010	**0.6599**	**0.6022**
Gln	0.3589	**0.4713**	**0.6174**	**0.6522**
Glu	**0.5755**	**0.5772**	0.0944	0.1101
Gly	0.2060	0.1762	0.0187	−0.0036
His	**0.5085**	**0.4800**	**0.4603**	**0.4566**
Ile	0.2401	0.2317	0.0138	0.0152
Leu	0.1456	0.1548	0.0264	0.0262
Lys	**0.4388**	**0.4288**	0.2307	0.2105
Met	0.2651	0.2561	0.3386	0.3403
Phe	0.1814	0.1540	0.1567	0.1383
Pro	**0.5533**	**0.5155**	−0.1392	−0.1622
Ser	0.3493	0.3328	0.0569	0.0303
Thr	0.0535	0.0427	0.0599	0.0356
Trp	0.3620	0.3644	0.0730	0.0608
Tyr	0.3070	0.3293	0.2038	0.2065
Val	0.1153	0.0923	0.1236	0.1343

Table 7
Association (defined as having an $I_{ij} > 0.2$) and Repulsion (with an $I_{ij} < -0.2$) Between Amino Acids. Leu and Gln Do Not Have Association with or Repulsion Against Other Amino Acids According to This Definition and Are Not Listed. Based on Human Data Only

AA Association	Repulsion
Ala	Tyr
Cys-His	
Glu-Lys	His, Phe
Gly-Pro	
His-Cys	Glu
Lys-Glu	
Met	Pro
Phe-Tyr	Glu
Pro-Gly	Met
Tyr-Phe	Ala

Table 8
Substitution Data from the 58 Nuclear Protein-coding Genes from Human, Mouse, and Cow, and 13 Protein-coding Mitochondrial Genes from 19 Mammalian Species. NS_{Nuc}: Observed Number of Substitutions from the 58 Nuclear Genes; NS_{MT}: Observed Number of Substitutions from the Mitochondrial Genes

AA	Grantham	Miyata	D_{np}	NS_{Nuc}	NS_{MT}
Arg-Ala . . .	111	2.915	102.1	14	7
Asn-Ala . . .	110	1.773	116.0	6	10
Asn-Arg . . .	85	2.036	67.1	16	0
Asp-Ala . . .	126	2.361	93.6	35	41
Asp-Arg . . .	96	2.335	64.5	4	0
Asp-Asn . . .	23	0.65	22.4	112	69
Cys-Ala . . .	195	1.389	127.8	2	0
Cys-Arg . . .	180	3.057	54.7	6	3
—.	—	—	—	—	—
Val-Phe . . .	50	1.429	47.6	18	15
Val-Pro	68	1.781	169.0	5	0
Val-Ser	123	2.145	71.9	11	5
Val-Thr. . . .	69	1.415	16.8	24	48
Val-Trp	88	2.504	55.9	1	0
Val-Tyr	55	1.514	51.4	0	2

(e.g., *E. coli* proteins) do not (Nishizawa and Nishizawa 1999; Nishizawa, Nishizawa, and Kim 1999). Thus, the local repetitiveness may be a derived character caused by factors such as replication slippage. The resulting repetitiveness may distort the similarity in neighbor preference between amino acids. For this reason, we used the ancient proteins in *E. coli* instead.

To test whether D_{np} is related to the rate of amino acid substitutions, we compiled substitution data from two sets of sequences. One set consists of 58 presumably orthologous protein-coding genes from the human, the mouse, and the cow, and the other is made of the 13 protein-coding genes from each of the 19 completely sequenced mitochondrial sequences used in Xia (1998). The number of substitutions involving each amino acid pairs, obtained by comparing neighboring nodes along a phylogenetic tree, is partially shown in table 8 (NS_{Nuc} and NS_{MT}). For example, there are 14 amino acid substitutions involving Arg and Ala for the nuclear genes (table 8).

R_{ij} values were computed, according to equation (8), for both the nuclear gene and for the mitochondrial gene. The resulting R_{Nuc} and R_{MT} values, however, are highly correlated with NS_{Nuc} and NS_{MT}, respectively ($r > 0.95$

Table 9
Regression of the Pair-wise Amino Acid Substitution Rate on the Three Measures of Amino Acid Dissimilarities

		Estimate	SE	T	*P*
Nuc. . . .	Intercept	5.0567	0.2795	18.0947	0.0000
	Grantham	−0.0124	0.0031	−4.0233	0.0001
	Miyata	−0.0841	0.1443	−0.5828	0.5607
	D_{np}	−0.0168	0.0021	−7.8559	0.0000
Mito. . .	Intercept	5.0603	0.3323	15.2268	0.0000
	Grantham	−0.0098	0.0037	−2.6580	0.0085
	Miyata	−0.1831	0.1716	−1.0669	0.2874
	D_{np}	−0.0188	0.0025	−7.4044	0.0000

NOTE.—Nuc. - Nuclear substitution data; Mito. - mitochondrial substitution data; Grantham - Grantham's; Miyata - Miyata's distance; SE - Standard error.

Table 10
Amino Acid Dissimilarities Based on Neighbor Preference from *E. coli* genomic DNA and Miyata's Distance

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr
Arg . . .	2.36																		
Asn . . .	1.64	2.01																	
Asp . . .	2.18	2.11	0.8																
Cys . . .	1.81	2.68	2	2.95															
Gln . . .	1.84	1.41	1.41	2.11	2.36														
Glu . . .	2.05	1.46	1.48	1.62	3.3	1.04													
Gly . . .	1.24	2.55	1.93	1.72	2.47	2.73	2.41												
His. . . .	2.34	1.27	1.24	2.01	1.82	1.25	2	2.74											
Ile	2.39	2.86	2.19	2.65	1.92	2.75	3.25	3.06	2.49										
Leu . . .	1.9	2.53	2.29	2.98	1.95	2.16	2.82	3.02	2.31	0.47									
Lys . . .	2.3	1.33	1.46	1.9	2.87	1.31	1.12	3.1	1.73	2.24	2								
Met . . .	1.85	2.68	2.38	3.22	2.1	2.05	2.65	3.2	2.55	1.12	0.47	1.77							
Phe . . .	2.95	3.03	2.79	3.2	2.23	3.22	3.73	3.3	2.62	0.71	1.23	2.9	1.89						
Pro. . . .	0.85	2.39	2.01	2.51	1.74	2.23	2.64	1.77	2.58	2.9	2.69	2.68	2.48	3.6					
Ser. . . .	0.6	2.23	1.09	1.96	1.21	1.54	2.18	1.4	1.77	2.41	2.1	2.18	2.04	2.78	0.99				
Thr. . . .	1.23	2.45	1.58	2.62	1.67	1.52	2.59	2.57	1.8	2.45	1.79	2.09	1.46	2.93	1.46	0.82			
Trp. . . .	4.01	3	4.58	5.38	3.99	2.94	3.79	5.05	3.49	3.9	3.05	3.89	3.05	3.69	4.33	4.07	3.6		
Tyr. . . .	2.41	1.84	2.38	3.35	1.7	1.99	2.79	3.43	1.36	1.53	1.17	2.26	1.43	1.47	2.88	2.06	2.06	2.04	
Val. . . .	1.39	2.47	1.99	2.26	1.78	2.32	2.58	2.06	2.4	0.82	0.75	1.99	0.94	1.47	2.21	1.88	1.87	3.92	1.95

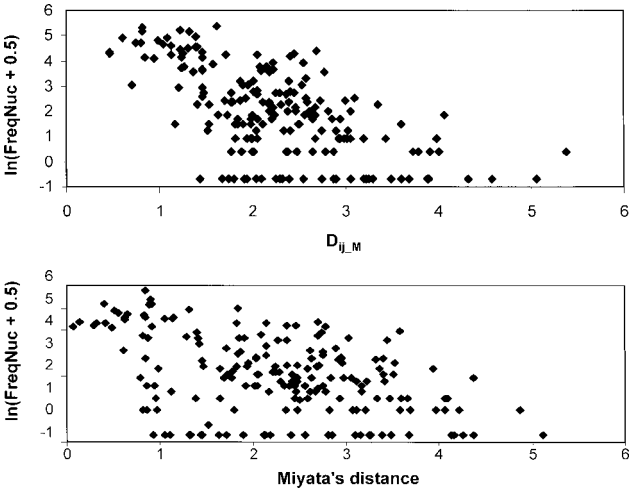

FIG. 3.—Frequency distribution of amino acid substitutions over D_{ij_M} (top panel) and Miyata's distance (bottom panel). Data from 58 presumably orthologous protein-coding genes from human, mouse, and cow.

for both), suggesting that the adjustment of amino acid frequencies does not really matter much. We regressed R_{Nuc} and R_{MT} separately on the three amino acid distances in table 8, after log-transforming R_{Nuc} and R_{MT}. The transformation is done after adding a constant value to R_{Nuc} and R_{MT} so that the minimum value of R_{Nuc} and R_{MT} is 0.5. The reason for the log-transformation follows from the simple formulation in Kimura (1983):

$$R = Ae^{BD_{ij}} \quad (9)$$

where R is equivalent to R_{Nuc} and R_{MT}, and D_{ij} is the distance between amino acids i and j. The equation implies that ln(R) is linearly related to D_{ij}, and hence the transformation.

A multiple regression (table 9) shows that all three amino acid dissimilarities are negatively correlated with R_{Nuc} and R_{MT}. The model accounts for 43.35% of the total variation in R_{Nuc} and 37.76% of the total variation in R_{MT}. The nonsignificant *P* value for Miyata's distance suggests that the distance does not add much to improve the model once Grantham's distance and D_{np} are already in the model.

The result suggests that D_{np} should be incorporated into the index of amino acid dissimilarities. We have taken a simple approach by rescaling D_{np} to have the same mean and variance as Grantham's distance, and obtain a new index as

$$D_{ij\_G} = \frac{G_{ij} + D_{np,ij}}{2}. \quad (10)$$

where the subscript G indicates the fact that D_{ij} results from a combination of D_{np} with Grantham's distance. Similarly, we also obtained D_{ij_M} (table 10), where the subscript M indicates the fact that D_{ij_M} results from a combination of D_{np} with Miyata's distance. A plot of D_{ij_M} versus log-transformed NS_{Nuc} is shown in figure 3. The number of substitutions seems to decrease monotonously with increasing D_{ij_M}, consistent with Kimura's (1983) observation but not with Gillespie's (1991). However, this may not be because of an improvement of D_{ij_M} over Miyata's distance, because the monotonous decrease in the number of substitutions is also visible with Miyata's distance. Thus, the pattern observed by Gillespie, that the substitution rate increases first with amino acid distance and then decreases with amino acid distance, may simply be caused by a less representative data set.

To evaluate the performance of these two new distance indices, we have used them together with Grantham's and Miyata's distances in a codon-based phylogenetic reconstruction involving the 13 protein-coding genes from six ape species. The maximum likelihood values (table 11) show that setting all amino acid distances as equal is the worst, followed by Grantham's and Miyata's distances. This result is consistent with a previous study (Yang, Nielsen, and Hasegawa 1998). D_{ij_G} is better than all preceding distances, but D_{ij_M} is the best of all (table 11). Because D_{np} was derived solely from the neighbor preference data of the 4289 CDS from the genome of *E. coli* K-12, not from mitochondrial genes, the better performance of D_{ij_M} involving mitochondrial genes suggests that D_{ij_M} may be generally applicable to other genes.

Table 11
Comparison of the Performance of Four Indices of Amino Acid Dissimilarity in Tree Estimation

Locus	Length	Equal	Grantham	Miyata	D_{ij_G}	D_{ij_M}
AtPase 6	681	−1963.98	−1960.49	−1957.72	−1957.19	−1955.38
ATPase 8 . . .	207	−629.23	−628.61	−625.86	−625.98	−624.02
COI	1542	−3900.56	−3894.92	−3895.49	−3893.76	−3894.02
COII	684	−1720.63	−1717.93	−1714.51	−1715.96	−1712.90
COIII.	784	−2079.52	−2074.85	−2072.97	−2070.53	−2068.74
Cyt-*b*	1141	−3219.01	−3206.96	−3203.93	−3203.36	−3201.70
NADH1.	957	−2635.90	−2634.50	−2628.91	−2629.22	−2624.30
NADH2.	1044	−3013.57	−2988.25	−2991.49	−2981.33	−2984.96
NADH3.	346	−1030.90	−1029.38	−1028.45	−1028.55	−1028.10
NADH4.	1378	−3815.49	−3809.11	−3805.06	−3801.43	−3798.72
NADH4l	297	−779.77	−777.73	−777.32	−776.86	−776.74
NADH5.	1812	−5458.31	−5447.56	−5443.47	−5440.97	−5438.70
NADH6.	525	−1336.29	−1333.70	−1330.15	−1333.34	−1330.65

NOTE.—Equal - equal distance between all amino acid pairs.

In summary, amino acids have different propensities to occur in different secondary structures, and they have different neighbor preferences. Amino acids with similar neighbor preferences tend to replace each other more frequently than amino acids with different neighbor preferences. The incorporation of the neighbor preference into the index of amino acid dissimilarities can substantially improve codon-based and amino acid-based substitution models.

Acknowledgments

This study is supported by a CRCG grant from the University of Hong Kong (10203043/27662/25400/302/01) and RGC grants from Hong Kong Research Grant Council (HKU7265/00M, HKU7212/01M) to X.X. Comments by F. X. Fu and two anonymous reviewers have significantly improved the manuscript.

LITERATURE CITED

BERMAN, H. M., J. WESTBROOK, Z. FENG, G. GILLILAND, T. N. BHAT, H. WEISSIG, I. N. SHINDYALOV, and P. E. BOURNE. 2000. The protein data bank. Nucleic Acids Res. **28**:235–242.

BRANDEN, C., and J. TOOZE. 1998. Introduction to protein structure. Garland Publishing, Inc., New York.

CAVALLI-SFORZA, L. L., and A. W. F. EDWARDS. 1967. Phylogenetic analysis: models and estimation procedures. Evolution **32**:550–570.

CHOU, P. Y., and G. D. FASMAN. 1974*a*. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry **13**:211–222.

———. 1974*b*. Prediction of protein conformation. Biochemistry **13**:222–245.

———. 1978*a*. Empirical predictions of protein conformation. Annu. Rev. Biochem. **47**:251–276.

———. 1978*b*. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. Relat. Areas Mol. Biol. **47**:45–148.

CLARKE, B. 1970. Selective constraints on amino-acid substitutions during the evolution of proteins. Nature **228**:159–160.

CREIGHTON, T. E. 1993. Proteins: structure and molecular properties. Freeman, New York.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in protein. Pp. 345–352 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. Natl. Biomed. Res. Found., Silver Spring, Md.

DAYHOFF, M. O., and W. C. BARKER. 1972. Mechanisms and molecular evolution: examples. Pp. 41–45 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. Natl. Biomed. Res. Found., Washington, D.C.

DAYHOFF, M. O., W. C. BARKER, and L. T. HUNT. 1983. Establishing homologies in protein sequences. Methods Enzymol. **91**:524–545.

EPSTEIN, C. J. 1967. Non-randomness of amino-acid changes in the evolution of homologous proteins. Nature **215**:355–359.

FASMAN, G. D., and P. Y. CHOU. 1974. Prediction of protein conformation: consequences and aspirations. Pp. 114–125 *in* E. R. BLOUT, F. A. BOVEY, M. GOODMAN, and N. LATAN, eds. Peptides, polypeptides and proteins. Wiley, New York.

GILLESPIE, J. H. 1991. The causes of molecular evolution. Oxford University Press, Oxford.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

GOUY, M., C. GAUTIER, M. ATTIMONELLI, C. LARAVE, and G. DIPAOLA. 1985. ACNUC–a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput. Appl. Biosci. **1**:167–172.

GRANTHAM, R. 1974. Amino acid difference formula to help explain protein evolution. Science **185**:862–864.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, United Kingdom.

MIYATA, T., S. MIYAZAWA, and T. YASUNAGA. 1979. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. **12**:219–236.

MORRIS, A. L., M. W. MACARTHUR, E. G. HUTCHINSON, and J. M. THORNTON. 1992. Stereochemical quality of protein structure coordinates. Proteins **12**:345–364.

NEI, M. 1972. Genetic distance between populations. Am. Nat. **106**:283–292.

NISHIZAWA, M., and K. NISHIZAWA. 1999. Local-scale repetitiveness in amino acid use in eukaryote protein sequences: a genomic factor in protein evolution. Proteins **37**:284–292.

NISHIZAWA, K., M. NISHIZAWA, and K. S. KIM. 1999. Tendency for local repetitiveness in amino acid usages in modern proteins. J. Mol. Biol. **294**:937–953.

RAMACHANDRAN, G. N., and V. SASISEKHARAN. 1968. Conformation of polypeptides and proteins. Adv. Protein Chem. **23**:284–438.

REYNOLDS, J. B., B. S. WEIR, and C. C. COCKERHAM. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics **105**:767–779.

SCHULZ, G. E., and R. H. SCHIRMER. 1979. Principles of protein structure. Springer, New York.

SINGH, J., and J. M. THORNTON. 1992. Atlas of protein side-chain interactions. IRL Press, Oxford.

SNEATH, P. H. A. 1966. Relations between chemical structure & biological activity in peptides. J. Theor. Biol. **12**:157–195.

THORNTON, J. M. 1992. Protein structures: the end point of the folding pathway. Pp. 59–82 *in* T. E. CREIGHTON, ed. Protein folding. Freeman, New York.

XIA, X. 1998. The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. Mol. Biol. Evol. **15**:336–344.

XIA, X. 2000. DAMBE (software package for data analysis in molecular biology and evolution). Version 4.0 Department of Ecology and Biodiversity, University of Hong Kong, Hong Kong.

XIA, X., and W.-H. LI. 1998. What amino acid properties affect protein evolution? J. Mol. Evol. **47**:557–564.

YANG, Z. 2000. PAML (phylogenetic analysis by maximum likelihood). University College, London.

YANG, Z., S. KUMAR, and M. NEI. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

YANG, Z., R. NIELSEN, and M. HASEGAWA. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. **15**:1600–1611.

ZUCKERKANDL, E., and L. PAULING. 1965. Evolutionary divergence and convergence in proteins. Pp. 97–166 *in* V. BRYSON and H. J. VOGEL, eds. Evolving genes and proteins. Academic Press, New York.

YUN-XIN FU, reviewing editor

Accepted August 27, 2001