J Mol Evol (1998) 47:557–564

JOURNAL OF MOLECULAR EVOLUTION

What Amino Acid Properties Affect Protein Evolution?

Xuhua Xia,[1] Wen-Hsiung Li[2]

[1] Department of Ecology and Biodiversity, University of Hong Kong, Pokfulam Road, Hong Kong
[2] Human Genetics Center, SPH, University of Texas, P.O. Box 20334, Houston TX 77225, USA

Received: 2 January 1998 / Accepted: 25 April 1998

Abstract. We studied 10 protein-coding mitochondrial genes from 19 mammalian species to evaluate the effects of 10 amino acid properties on the evolution of the genetic code, the amino acid composition of proteins, and the pattern of nonsynonymous substitutions. The 10 amino acid properties studied are the chemical composition of the side chain, two polarity measures, hydropathy, isoelectric point, volume, aromaticity, aliphaticity, hydrogenation, and hydroxythiolation. The genetic code appears to have evolved toward minimizing polarity and hydropathy but not the other seven properties. This can be explained by our finding that the presumably primitive amino acids differed much only in polarity and hydropathy, but little in the other properties. Only the chemical composition (C) and isoelectric point (IE) appear to have affected the amino acid composition of the proteins studied, that is, these proteins tend to have more amino acids with typical C and IE values, so that nonsynonymous mutations tend to result in small differences in C and IE. All properties, except for hydroxythiolation, affect the rate of nonsynonymous substitution, with the observed amino acid changes having only small differences in these properties, relative to the spectrum of all possible nonsynonymous mutations.

Key words: Polarity — Hydropathy — Isoelectric point — Aromaticity — Chemical composition — Volume — Genetic code — Protein evolution — Substitution rate

Correspondence to: X. Xia; *e-mail:* xxia@hkusua.hku.hk

Introduction

In the evolution of proteins, amino acid substitutions occur more frequently between similar amino acids than between dissimilar ones (Zuckerkandl and Pauling 1965; Epstein 1966; Sneath 1966; Clarke 1970; Grantham 1974; Miyata et al. 1979; Kimura 1983, p. 152). Different amino acids can differ in many physicochemical properties; indeed, 134 properties were listed by Sneath (1966). It is therefore desirable to know which properties of amino acids are more important than others in determining the rate and pattern of protein evolution. In spite of the rapid progress in molecular biology, we still know little about which properties of amino acids are important and which are not. Zuckerkandl and Pauling's (1965) statement that ''apparently chemists and protein molecules do not share the same opinions regarding the definition of the most prominent properties of a residue'' applies almost equally well today as it did more than 30 years ago.

We propose to evaluate the relative importance of amino acid properties with respect to (1) the evolution of the genetic code, (2) the amino acid composition of proteins, and (3) the pattern of nonsynonymous substitutions. The rationale of these approaches is outlined below.

First, it has long been proposed that the genetic code might have been arranged in such a way as to reduce the effect of nonsynonymous mutations involving single nucleotide changes (henceforth termed ''single-step nonsynonymous codon mutations''; SSNCMs). In particular, Zuckerkandl and Pauling (1965) stated, ''According to Eck's proposal for a complete genetic code, nearly all transitions between functionally closely related amino

acids can be brought about by one single mutational step.'' Evidence that the genetic code has evolved to minimize the effect of SSNCMs was indeed provided by the pioneering studies of Sonneborn (1965), Epstein (1966), Goldberg and Wittes (1966), and Alff-Steinberger (1969). Take polarity, for example. Most SSNCMs result in replacements between amino acids of similar polarity (Epstein 1966; Woese et al. 1966). The differences in polarity between amino acids that can mutate to each other through a SSNCM would be much larger had the genetic code been randomly generated (Alff-Steinberger 1969; Haig and Hurst 1991). These findings indicate that the genetic code has evolved toward minimizing the change in polarity when a SSNCM occurs.

However, other amino acid properties studied along the same line did not yield unequivocal results. Vogel and Zuckerkandl (1972) found no good correlation between polarity and evolutionary variability. Haig and Hurst (1991) concluded that the genetic code had evolved to minimize the differences in polar requirement and hydropathy (the two are positively correlated) but not the differences in size (measured by molecular volume) and isoelectric point. In contrast, Alff-Steinberger (1969) found the genetic code to have minimized the difference in all four properties mentioned above. We present here a more straightforward method for evaluating the relative importance of amino acid properties in shaping the evolution of the genetic code.

Second, the relative importance of amino acid properties can be evaluated through their effects on the amino acid composition of proteins. For example, suppose that certain sites in a protein require a polar amino acid to maintain the normal function. Consider Glu and Asp, both of which are polar. When a SSNCM occurs, Asp has a higher probability (0.56), but Glu a lower probability (0.44), of being replaced by another polar amino acid (Epstein 1966). If the protein gene uses more Asp codons and fewer Glu codons for such sites, the function of the protein would less likely be disrupted by mutation. It is therefore beneficial for the fictitious protein gene to use more Asp codons but fewer Glu codons.

Third, the relative importance of amino acid properties in protein evolution can be evaluated directly from the pattern of observed nonsynonymous substitutions. For illustration, suppose that a protein gene is made entirely of UUU codons (coding for Phe), which can mutate to GUU (Val), UUG (Leu), UUA (Leu), UCU (Ser), AUU (Ile), UGU (Cys), UAU (Tyr), and CUU (Leu) through a SSNCM. Using the polar requirement values of Woese et al. (1966), one can show that the differences in polar requirement between these eight alternative amino acids and Phe are 0.6, 0.1, 0.1, 2.5, 0.1, 0.2, 0.4, and 0.1, respectively. If all these SSNCMs are equally likely to occur, then the average difference in polar requirement for all eight possible nonsynonymous substitutions is 0.51. However, if polarity is important and purifying selection eliminates all mutations involving a change in polar requirement larger than, say, 0.1, then only those SSNCMs with a difference of 0.1 in polar requirement will have a chance to be fixed and observed. The mean difference in polar requirement for the observed nonsynonymous substitutions will then be 0.1, which is much smaller than 0.51, the expected value under the assumption that all SSNCMs are equally likely. If the observed mean is not different from the expected mean, then polarity is not important in determining the rate of nonsynonymous substitution.

In this paper we evaluate the relative importance of amino acid properties using the three approaches outlined above. Ten amino acid properties are considered: chemical composition of the side chain, polarity, volume [denoted C, P, and V, respectively (Grantham 1974)], polar requirement [PR (Woese et al. 1966)], hydropathy [HY (Kyte and Doolittle 1982)], isoelectric point [IE (Alff-Steinberger 1969)], and PC I, PC II, PC III, and PC IV (Sneath 1966). The values for polar requirement, hydropathy, and isoelectric point are taken from Table 1 of Haig and Hurst (1991). The last four properties are from Sneath (1966) and each of them is a principal component, which is a linear combination of a number of amino acid properties. Sneath (1966) found only PC III to be clearly interpretable, and he termed it aromaticity because all amino acids with an aromatic ring structure on the side chain (i.e., Trp, Phe, Tyr, and His) have high PC III values. We refer to PC III and aromaticity synonymously. The other three properties (PC I, PC II, and PC IV), which were termed by Sneath as aliphaticity, hydrogenation, and hydroxythiolation, respectively, are difficult to interpret and are included here for completeness.

Materials and Methods

The data consist of complete mitochondrial sequences from 19 mammalian species: hedgehog (GenBank accession number X88898), mouse (J01420), rat (X14848), cat (U20753), gray seal (X72004), harbor seal (X63726), horse (X79547), donkey (X97337), rhinoceros (X97336), cow (V00645), fin whale (X61145), blue whale (X72204), gibbon (X99256), Sumatran orangutan (X97707), Bornean orangutan (D38115), gorilla (X93347), pygmy chimpanzee (D38116), chimpanzee (D38113), and human (X93334). Of the 13 protein-coding mitochondrial genes, only 10 were used in this study, with the 3 shortest genes (ATPase 8, ND3, and ND4L) excluded.

There are now at least 23 completely sequenced mammalian mitochondrial genomes, several of which (e.g., opossum, wallaroo, and duckbill platypus) were not used in this study because of difficulties in sequence alignment. Although other authors seem to have succeeded in aligning sequences from diverse groups including mammalian, avian, and amphibian species (Cummings et al. 1995; Otto et al. 1996; Janke et al. 1997), we are not certain of the accuracy of such alignments.

The unrooted phylogenetic tree (Fig. 1) for the 19 mammalian species receives the strongest support from both traditional and molecular phylogenetics and represents our existing knowledge of mammalian evolution (Novacek 1988; Cao et al. 1994; Cummings et al. 1995; Janke et al. 1997). The only difference between the tree in Fig.

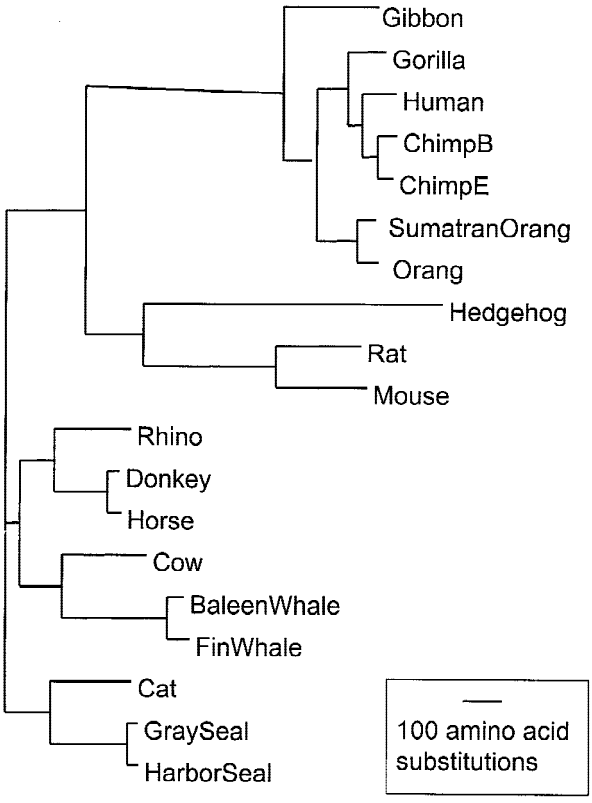

Fig. 1. Unrooted phylogenetic tree for the 19 mammalian species. The branch length is proportional to the number of nonsynonymous codon substitutions, counted from pairwise comparisons between neighboring nodes and summed over the 10 genes studied.

1 and that of Janke et al. (1997) is that Janke et al. grouped perissodactyls and carnivores together as sister taxa, whereas Fig. 1 groups perissodactyls and artiodactyls as sister taxa. This minor difference in topology has only a negligible effect on the result of our analysis. We used the topology for the reconstruction of ancestral states by using the BASEML program in the PAML package (Yang 1996), which implements the likelihood-based method detailed by Yang et al. (1995). The reconstruction is generally satisfactory, with the overall accuracy of the ancestral sequences being about 90%.

Nonsynonymous codon substitutions were counted by comparing the DNA sequences between two neighboring nodes. For example, for the unrooted tree in Fig. 1, there is a total of 35 pairwise comparisons, with 16 between neighboring internal nodes and 19 between the terminal nodes and their neighboring internal nodes. This counting procedure differs from some other studies that counted substitutions from all possible pairwise comparisons, many of which are nonindependent and would introduce biases (Felsenstein 1992; Nee et al. 1996; Xia et al. 1996).

The 35 pairwise, codon by codon, comparisons result in four categories of codon pairs: identical codon pairs, different but synonymous codon pairs; and nonsynonymous codon pairs differing at one, and at more than one, codon positions. These basic data are summarized in Table 1, together with the relevant sequence information.

Results and Discussion

Amino Acid Properties and Evolution of the Genetic Code

For the 20 amino acids, there are 190 possible pairwise distances ($D_{aa.ij}$, $i, j = 1, 2, \ldots, 20$; $i < j$) between amino acids. The mean ($\overline{D}_{aa}$) of these 190 $D_{aa.ij}$ values was calculated for each of the 10 amino acid properties (Table 2). These means represent the expected values if the genetic code does not constrain the interchange between amino acids, e.g., when we randomly assign amino acids to different codons so that each amino acid will have an equal chance of being a neighbor of any other amino acid in the genetic code.

Table 1. Basic information on codon differences from 35 pairwise comparisons between neighboring nodes on the phylogenetic tree (Fig. 1) for each of the 10 mitochondrial genes studied[a]

	N_{codon}	N_{same}	N_{syn}	N_1	$N_{>1}$	R
COI	513	15,376	2,364	161	54	0.42
COIII	261	7,814	1,105	151	65	0.83
COII	227	6,773	983	121	68	0.83
Cyt-b	379	11,298	1,485	352	130	1.27
ND1	315	9,408	1,204	311	101	1.31
ATPase 6	226	6,615	937	251	107	1.58
ND4	459	13,407	1,882	532	241	1.68
ND5	600	17,332	2,311	955	393	2.25
ND6	172	5,009	600	312	96	2.37
ND2	343	9,842	1,255	656	252	2.65

[a] N_{same}, number of identical codon pairs; N_{syn}, number of synonymous codon pairs; N_1 and $N_{>1}$, number of nonsynonymous codon pairs differing at one and more than one position, respectively. $R = (N_1 + N_{>1})/N_{codon}$ is an overall measure of amino acid divergence.

Each codon can, through a single nucleotide substitution, mutate to one of nine alternative codons, some of which are nonsynonymous. Designate the number of nonsynonymous codons which codon k can mutate to through a single nucleotide change N_k ($k = 1, 2, \ldots, 60$, e.g., N_{UUU} equals 8). For the genetic code of mammalian mitochondria, there are 190 ($=\sum_{k=1}^{60} N_k/2$) possible nonsynonymous codon pairs differing at one codon position (e.g., UUU–UUA). Each of these 190 nonsynonymous codon pairs is associated with one $D_{codon.kl}$ value ($k = 1, 2, \ldots, 60$; $l = 1, 2, \ldots, N_k$; codons k and l are nonsynonymous and differ at only one codon position). For example, $D_{UUU–UUA}$ (i.e., $D_{Phe–Leu}$) is 21 for molecular volume (Grantham 1974). The mean of these 190 $D_{codon.kl}$ values for each of the 10 amino acid properties is

$$\overline{D}_{codon} = \frac{\sum_{k=1}^{60}\sum_{l=1}^{N_k} D_{codon.kl}}{\sum_{k=1}^{60} N_k} \quad (1)$$

and is expected to be smaller than that for the mean $D_{aa.ij}$ values ($\overline{D}_{aa}$) if the genetic code constrains the interchange between amino acids in such a way as to reduce the effect of SSNCMs. Indeed, Table 2 shows that $\overline{D}_{codon} < \overline{D}_{aa}$ for each of the 10 properties. The probability for this to happen is $0.5^{10} = 0.001$, if the 10 amino acid

Table 2. Means of 190 pairwise $D_{aa.ij}$ values between amino acids for the 10 amino acid properties ($\overline{D}_{aa}$) and means of the 190 pairwise $D_{codon.kl}$ values ($\overline{D}_{codon}$)[a]

Property	$\overline{D}_{aa}$	$\overline{D}_{codon}$	$\Delta\overline{D}$ (%)	Prob
C	0.74	0.67	9.35	0.2147
P	3.13	2.13	31.92	0.0001*
V	50.06	43.73	12.65	0.0407
PC I	0.04	0.04	2.48	0.8782
PC II	0.27	0.23	14.56	0.0126
PC III	0.25	0.21	17.53	0.0297
PC IV	0.27	0.26	3.88	0.5977
PR	2.82	1.90	32.56	0.0001*
HY	3.46	2.63	23.88	0.0007*
IE	1.72	1.48	14.23	0.0839

[a] $\Delta\overline{D}$ (%) = $100(\overline{D}_{aa} - \overline{D}_{codon})/\overline{D}_{aa}$. Prob stands for the probability of $\overline{D}_{aa} = \overline{D}_{codon}$ based on the Wilcoxon two-sample test, unadjusted for the inflation of Type I error due to multiple comparisons.
* Significant when Type I error rate is controlled for.

Table 3. Differences in the variability of the 10 properties among 10 presumably primitive amino acids (with subscript 10) and the modern set of 20 amino acids (with subscript 20)[a]

Property	$Mean_{10}$	STD_{10}	$Mean_{20}$	STD_{20}	*P*	ΔD (%)
C	0.66	0.19	0.74	0.43	0.0000*	−54.89
P	3.17	5.11	3.13	4.99	0.3979	2.50
V	43.91	883.80	50.06	1162.69	0.0064	−23.99
PC I	0.05	0.00	0.04	0.00	0.4804	4.37
PC II	0.33	0.04	0.27	0.03	0.0125	28.59
PC III	0.20	0.02	0.25	0.04	0.0000*	−44.07
PC IV	0.22	0.03	0.27	0.04	0.0045*	−24.04
PR	3.24	5.12	2.82	4.91	0.3451	4.19
HY	3.54	9.16	3.46	6.19	0.0002*	47.96
IE	2.43	4.25	1.72	3.38	0.0188	25.86

[a] The *P* values are the probabilities that $Variance_{10} = Variance_{20}$. ΔD (%) = $10(STD_{10} - STD_{20})/STD_{20}$.
* Significant when Type I error is controlled for.

properties are independent of each other. However, because P, PR and HY are correlated, the probability is perhaps closer to $0.5^8 = 0.004$.

For significance tests involving multiple comparisons, one needs to control the Type I experimentwise error rate, which is the probability of making at least one Type I error for a set of hypothesis tests. If we designate α_e the Type I experimentwise error rate and set it to 0.05, then the comparisonwise error rate (designated α_c), with 10 comparisons in Table 2, can be obtained by solving $1 - (1 - \alpha_c)^{10} = 0.05$ for α_c, which yields $\alpha_c = 0.0051$. Thus, only P, PR, and HY have their $\overline{D}_{codon}$ significantly smaller than their $\overline{D}_{aa}$.

One puzzling aspect of the result is that some properties known to be important did not have their $\overline{D}_{codon}$ significantly smaller than $\overline{D}_{aa}$ (Table 2). C and V have long been considered to be important amino acid properties, and are two of the three amino acid properties Grantham (1974) chose to construct his amino acid distance matrix. It has been documented that pairwise differences in C and V between amino acids are negatively correlated with substitution rates between amino acids. Why, then, did the genetic code not minimize the differences in C and V to the same extent as in polarity measures and hydropathy? To use a concrete example, the exchange between Gly and Trp (the smallest and the largest amino acids, respectively) almost never occurs in protein genes, indicating strong purifying selection against such changes. So, why did the genetic code not constrain nonsynonymous substitutions so that codons for Gly and Trp cannot mutate to each other through a single nucleotide substitution? In other words, why were Gly and Trp not assigned to codon families that differ by two or three codon positions?

Another property, PC III, which Sneath (1966) termed aromaticity, also did not show a significant difference when the experimentwise error is controlled, i.e., the *P* value is not smaller than 0.0051 (Table 2). There are two lines of evidence that aromaticity is important. First, it has been experimentally shown that the replacement of tyrosine by phenylalanine (both having an aromatic ring on the side chain) at position 2 of oxytocin has little effect on the oxytocic activity of the peptide (Jaquenoud and Boissonnas 1959; Boissonnas and Guttmann 1960), but the replacement by serine (lacking the aromatic ring) reduces the oxytocic activity to an undetectable level (Guttmann and Boissonnas 1960). Second, substitution models based on Grantham's distance or Miyata's distance, neither of which takes aromaticity into consideration, invariably overestimate nonsynonymous substitutions involving codons coding for aromatic amino acids (our unpublished results). Therefore, it is surprising that the genetic code has not minimized aromaticity to the same extent as polarity measures and hydropathy.

A hypothesis for these unexpected results is that the genetic code was partially fixed before there was a complete set of 20 amino acids and that, for the subset of early amino acids, there was little variation in C, V, and aromaticity compared to the modern set of 20 amino acids. Crick (1968) suggested that the early amino acids were Gly, Ala, Ser, and Asp (which are the amino acids resulting from sparking the mixture of CH_4, NH_3, H_2O, and H_2). For these amino acids, there is indeed little difference in C, V, and aromaticity, consistent with the hypothesis.

One difficulty with this hypothesis is the assumption that the genetic code had already been fixed, even partially, when there were only four amino acids used in protein synthesis. However, subsequent sparking experiments produced more amino acids that still support the hypothesis. Miller (1986) argued that a mixture of CH_4, N_2, H_2O, and traces of NH_3 is a more realistic atmosphere for the primitive earth, and sparking this mixture produced 10 amino acids: Ala, Asp, Glu, Gly, Ile, Leu, Pro, Ser, Thr, and Val. We computed the standard deviation for each of the 10 properties among these 10 amino acids and compared it with that for the modern set

Table 4. The mean values of 190 pairwise $D_{\text{codon}.ij}$'s between nonsynonymous codons differing at one codon position ($\overline{D}_{\text{codon}}$) and the mean values of D_{ij}'s for all possible codon substitutions differing at one codon position ($\overline{D}_{\text{ACF}}$), adjusted for codon frequency [Eq. (3)]

		$\overline{D}_{\text{ACF}}$									
Property	$\overline{D}_{\text{codon}}$	ATPase6	COI	COII	COIII	Cyt-b	ND1	ND2	ND4	ND5	ND6
C	0.67	0.53	0.58	0.59	0.61	0.60	0.58	0.56	0.57	0.59	0.58
P	2.13	2.00	2.05	2.18	2.04	2.10	2.08	2.04	2.09	2.04	2.08
V	43.73	35.90	40.69	38.61	41.62	40.30	38.51	37.40	37.81	37.93	44.75
PC I	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
PC II	0.23	0.21	0.22	0.23	0.21	0.22	0.22	0.21	0.21	0.21	0.23
PC III	0.21	0.18	0.21	0.19	0.23	0.21	0.19	0.18	0.19	0.20	0.21
PC IV	0.26	0.28	0.26	0.26	0.26	0.27	0.27	0.29	0.28	0.28	0.26
PR	1.90	1.64	1.75	1.92	1.73	1.76	1.73	1.65	1.71	1.70	1.83
HY	2.63	2.73	2.62	2.67	2.56	2.69	2.74	2.70	2.75	2.64	2.63
IE	1.48	1.10	1.12	1.29	1.18	1.16	1.13	1.10	1.15	1.17	1.03
		Probability ($\overline{D}_{\text{codon}} \leq \overline{D}_{\text{ACF}}$)									
C		0.0001	0.0210	0.0215	0.0758	0.0477	0.0144	0.0033	0.0093	0.0255	0.0245
IE		0.0004	0.0006	0.0564	0.0042	0.0028	0.0012	0.0005	0.0021	0.0032	0.0001

of 20 amino acids (Table 3). For the 10 ''primitive'' amino acids, the standard deviations for C, V, and PC III are 0.19, 883.80, and 0.02, respectively. For the modern set of 20 amino acids, the corresponding values are 0.43, 1162.69, and 0.04, respectively (Table 3). Thus, the 10 ''primitive'' amino acids are much more similar in C, V, and PC III (aromaticity) than the modern set of 20 amino acids. In contrast, the 10 ''primitive'' amino acids are more different in polarity (P and PR in Table 2) and hydropathy (HY) than the modern set of 20 amino acids, with the standard deviations for P, PR, and HY being 5.12, 5.12, and 9.16, respectively, for the 10 amino acids, but 4.99, 4.91, and 6.19, respectively, for the 20 amino acids (Table 3). We conclude that early amino acids differed to some extent in polarity and hydropathy, and consequently the genetic code has evolved towards minimizing the difference in polarity and hydropathy when new amino acids were recruited into protein synthesis. In contrast, C, V, and aromaticity did not play a significant role in shaping the evolution of the genetic code because there was little variation in these properties among the ''primitive'' amino acids.

The early amino acids also differed more in PC II than the modern set of 20 amino acids. The genetic code did seem to have evolved to minimize the pairwise difference in PC II although the P value ($=0.013$) is not smaller than 0.0051 (Table 2).

Relative Importance of Amino Acid Properties in Affecting Amino Acid Usage

Proteins do not use amino acids with equal frequency. Some amino acids, such as Leu and Thr, are typical in that they have a number of similar alternative amino acids which they can mutate to through a SSNCM, whereas other amino acids, such as Trp and Cys, are idiosyncratic and have few similar alternative amino acids. If a protein-coding gene contains a large number of typical amino acids and few idiosyncratic amino acids, then the effect of SSNCM would be further reduced.

Designate the number of codon k in the DNA sequence under study as F_k ($k = 1, 2, \ldots, 60$), and recall that the number of nonsynonymous codons which codon k can mutate to through a SSNCM is N_k (e.g., N_{UUU} equals 8). The total number of possible SSNCMs for all codons is then

$$N_{\text{T}} = \sum_{k=1}^{60} F_k N_k \tag{2}$$

and the expected mean difference in a certain property for these N_{T} possible SSNCMs is

$$\overline{D}_{\text{ACF}} = \frac{\sum_{k=1}^{60} F_k \sum_{l=1}^{N_k} D_{\text{codon}.kl}}{N_{\text{T}}} \tag{3}$$

where the subscript ACF stands for ''adjusted for codon frequency.'' If the protein contains a large number of typical amino acids and few idiosyncratic ones, then $\overline{D}_{\text{ACF}}$ should be smaller than $\overline{D}_{\text{codon}}$. For amino acid properties C and IE, $\overline{D}_{\text{codon}}$ is consistently larger than $\overline{D}_{\text{ACF}}$ for all 10 genes, and the probability from t tests that $\overline{D}_{\text{codon}} \leq \overline{D}_{\text{ACF}}$ is shown at the bottom of Table 4 (nonparametric tests yield similar results). For the other eight amino acid properties, either the differences between $\overline{D}_{\text{codon}}$ and $\overline{D}_{\text{ACF}}$ are inconsistent for different genes or the probability value for the null hypothesis that $\overline{D}_{\text{codon}} \leq \overline{D}_{\text{ACF}}$ is larger than 0.05. In summary, among the 10 amino acid properties studied, only the chemical composition of the side chain (C) and isoelectric point

Table 5. The values of $\overline{D}_{obs}$ for the 10 amino acid properties for comparison with the corresponding $\overline{D}_{ACF}$ values in Table 4[a]

	Gene										
Property	ATPase6	COI	COII	COIII	Cyt-b	ND1	ND2	ND4	ND5	ND6	Mean (%)
C	0.43	0.38	0.40	0.44	0.38	0.46	0.37	0.43	0.43	0.38	29.09
	(0.0004)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	
P	1.33	1.17	1.42	1.33	1.26	1.55	1.42	1.49	1.65	1.37	32.40
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	
V	26.16	26.09	28.41	26.51	25.23	27.67	24.88	29.18	28.87	32.42	29.90
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	
PC I	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	22.38
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	
PC II	0.18	0.19	0.17	0.17	0.17	0.17	0.19	0.18	0.18	0.20	17.36
	(0.0001)	(0.0006)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.003)	
PC III	0.14	0.12	0.14	0.13	0.14	0.14	0.13	0.15	0.15	0.18	28.19
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0019)	
PC IV	0.32	0.26	0.26	0.29	0.28	0.31	0.30	0.30	0.31	0.25	−6.51
	(0.0001)	(0.8143)	(0.9142)	(0.0174)	(0.1474)	(0.0001)	(0.0228)	(0.0126)	(0.0001)	(0.4241)	
PR	1.01	1.01	1.15	0.98	0.95	1.08	1.05	1.14	1.25	1.26	37.49
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	
HY	2.14	1.87	1.83	2.10	1.95	2.28	2.10	2.09	2.22	1.86	23.58
	(0.0001)	(0.0001)	(0.0001)	(0.002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	
IE	0.57	0.67	0.71	0.58	0.55	0.54	0.64	0.74	0.74	0.71	43.48
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	

[a] The percentage change is calculated as $(\overline{D}_{ACF} - \overline{D}_{obs})/\overline{D}_{ACF}$, and the last column (mean %) is the average of percentage changes over the 10 genes for each of the 10 amino acid properties. Numbers in parentheses are the probability of $\overline{D}_{ACF} = \overline{D}_{obs}$. For PC IV, ($\overline{D}_{ACF} > \overline{D}_{obs}$) holds only for the ND6 gene.

(IE) show significant effects on the amino acid composition of proteins.

Effects of Amino Acid Properties on the Rate of Amino Acid Substitution

Let N_{obs} be the observed number of nonsynonymous codon substitutions in a protein-coding gene. Let $D_{obs.m}$ ($m = 1, 2, \ldots, N_{obs}$) be the difference in a property between the two amino acids coded by the two interchanged nonsynonymous codons. The mean difference is then

$$\overline{D}_{obs} = \frac{\sum_{m=1}^{N_{obs}} D_{obs.m}}{N_{obs}} \quad (4)$$

If a property is important, then purifying selection should eliminate those nonsynonymous mutations involving a large difference between the interchanged amino acids and tolerate only those involving a small difference, so $\overline{D}_{obs}$ should be smaller than $\overline{D}_{ACF}$ in Eq. (3). An unimportant property should have $\overline{D}_{obs}$ close to $\overline{D}_{ACF}$.

The number of nonsynonymous codon substitutions is tabulated in Table 1 for each of the 10 genes, and the mean D_{obs} values are displayed in Table 5 for comparison with the mean D_{ACF} values in Table 4. Nine of the 10 amino acid properties (with PC IV being the only exception) have their $\overline{D}_{Obs}$ significantly smaller than $\overline{D}_{ACF}$ for all 10 genes ($P \leq 0.003$; Table 5). The average percentage reduction, calculated as the mean of $(\overline{D}_{ACF} - \overline{D}_{Obs})/\overline{D}_{ACF}$, varies from 17.36 to 43.48% for the nine properties (Table 5). Parametric *t* tests and nonparametric tests yield similar results. For PC IV, $\overline{D}_{obs}$ is smaller than $\overline{D}_{ACF}$ only for the ND6 gene and is larger than $\overline{D}_{obs}$ for the other nine genes (Tables 4 and 5).

Our finding suggests that the existing measures of amino acid dissimilarity, such as Grantham's and Miyata's distances, have missed some important differences between amino acids. In particular, in spite of their strong effects on the rate of nonsynonymous substitution, both aromaticity (PC III) and IE have not been incorporated into these two popular amino acid dissimilarity measures. One can justify the exclusion of some of the 10 amino acid properties. For example, P from Grantham (1974) and PR from Woese et al. (1966) are highly correlated ($r = 0.963$, $P = 0.0001$), so there is no point in including PR when P has already been included in the construction of amino acid distances. However, IE is not significantly correlated with any of the three properties (C, P, and V) used in constructing amino acid distances ($r = -0.272$, 0.018, and 0.295, respectively, with the corresponding $P = 0.245$, 0.939, and 0.207, respectively). It is difficult to justify the exclusion of IE in the construction of genetic distances.

Aromaticity is positively correlated with volume ($r = 0.7163$, $P = 0.0004$). One might therefore think that differences between amino acids in aromaticity are very

much the same as differences in volume. This is not true. We have evaluated (data not shown) the sufficiency of Grantham's and Miyata's distances in predicting the rate of nonsynonymous substitution. Both distance measures overestimated the rate of nonsynonymous substitution for codons coding for amino acids with a ring structure (Pro, Trp, Phe, Tyr, and His). This suggests that aromaticity should be included in the construction of genetic distances, especially in light of the fact that these distance measures may be incorporated in phylogenetic programs [e.g., the CODEML program in the PAML package (Yang 1996)] for implementing the codon-based substitution model.

All three amino acid properties used in constructing Grantham's distance (C, P, and V) are important according to Table 5. Miyata et al. (1979) dropped C in their reconstruction of amino acid distances. Our results suggest that leaving out C (the chemical composition of the side chain) is undesirable.

There are a few factors that may confound our conclusions. First, we have interpreted our finding of $\overline{D}_{\text{Obs}} > \overline{D}_{\text{ACF}}$ to be due to purifying selection against nonsynonymous substitutions with deleterious effects. However, the finding can also be explained by transitional mutation bias. For example, transitions may occur more frequently than transversions, and nonsynonymous transitions may have smaller effects than nonsynonymous transversions. Of the 60 mammalian mitochondrial codons, there are 190 possible SSNCMs, of which 54 are transitions and 136 are transversions (Xia 1998). The average Grantham's distance is 78.81 for the transitions and 85.34 for transversions, and the overall average is 83.48. Obviously, if all mutations (and consequently substitutions) are transitions, then $\overline{D}_{\text{Obs}}$ would be 78.81 if there is neither purifying selection nor codon usage bias. In other words, our finding of $\overline{D}_{\text{Obs}} > \overline{D}_{\text{ACF}}$ may simply be a consequence of transitional mutation bias coupled with the fact that nonsynonymous transitions result in amino acid replacements with a smaller D_{aa} than nonsynonymous transversions. The relative contribution of mutation bias and purifying selection to the pattern of nonsynonymous substitutions awaits further investigation.

Another factor that may confound our conclusion is the selection favoring the use of ribonucleotide A in mitochondrial protein-coding genes because of the relative abundance of ATP in the mitochondria relative to the other three ribonucleotides (Xia 1996). We have argued in this paper that the protein genes should use more typical amino acids such as Thr and Leu and fewer idiosyncratic amino acids such as Trp and Cys as a consequence of minimizing the effect of SSNCMs. However, the idiosyncratic amino acids tend to be A-poor codons (i.e., codons with the first two codon positions having no A, such as UGY for Cys and UGR for Trp). If codons in mitochondrial genes tend to be A-rich, then the frequency of idiosyncratic amino acids will be reduced. More research is needed to evaluate this alternative explanation quantitatively.

In summary, all 10 amino acid properties except for PC IV are important in one way or another. The genetic code appears to have evolved toward minimizing polarity and hydropathy, but not the other important properties such as C, V, and aromaticity. This can be explained by our finding that the presumably primitive amino acids differed greatly only in polarity and hydropathy, and little in C, V, and aromaticity. Only C and IE appear to have affected the amino acid composition of a protein, i.e., proteins encoded by mtDNA tend to have more amino acids with typical C and IE values, so that a nonsynonymous mutation tends to result in a small difference in C and IE. All 10 properties except for PC IV affect the rate of nonsynonymous substitution, with the observed nonsynonymous codon substitutions involving only small differences in these properties. These amino acid properties deserve further investigation.

Acknowledgments. We thank Y.X. Fu, H.W. Deng, T. Spradling, and J. Demastes for discussion and comments. This project is supported by a CRCG grant from the University of Hong Kong (335/023/0022) and an RGC grant from the Hong Kong government (HKU 7259/97M) to X. Xia and NIH grants to W.-H. Li.

References

Alff-Steinberger C (1969) The genetic code and error transmission. Proc Natl Acad Sci USA 64:584–591

Boissonnas RA, Guttmann S (1960) Synthèse d'analogues de l'oxytocine et de la lysine-vasopressine contenant de la phénylalanine ou de la tyrosine en positions 2 et 3. Helv Chim Acta 43:190–200

Cao Y, Adachi J, Janke A, Pääbo S, Hasegawa M (1994) Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J Mol Evol 39:519–527

Clarke B (1970) Selective constraints on amino-acid substitutions during the evolution of proteins. Nature 228:159–160

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

Cummings M, Otto S, Wakeley J (1995) Sampling properties of DNA sequence data in phylogenetic analysis. Mol Biol Evol 12:814–822

Epstein CJ (1967) Non-randomness of amino-acid changes in the evolution of homologous proteins. Nature 215:355–359

Felsenstein J (1992) Estimating effective population size from samples of sequences: Inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet Res 59:139–147

Goldberg AL, Wittes RE (1966) Genetic code: Aspects of organization. Science 153:420–424

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Guttmann S, Boissonnas RA (1960) Synthèse de dix analogues de l'oxytocine et de la lysine-vasopressine contenant de la sérine, de l'histidine ou du tryptophane en position 2 ou 3. Helv Chim Acta 43:200–216

Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. J Mol Evol 33:412–417

Janke A, Xu X, Arnason U (1997) The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relation-

ship among Monotremata, Marsupialia, and Eutheria. Proc Natl Acad Sci USA 94:1276–1281

Jaquenoud PA, Boissonnas RA (1959) Synthèse de la Phé^2-oxytocine. Helv Chim Acta 42:788–793

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

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Miller SL (1986) Current status of the prebiotic synthesis of small molecules. In: Baltscheffsky H, Jörnvall H, Rigler R (eds) Molecular evolution of life. Cambridge University Press, Cambridge, pp 5–11

Miyata T, Miyazawa S, Yasunaga T (1979) Two types of amino acid substitution in protein evolution. J Mol Evol 12:219–236

Nee S, Holmes EC, Rambaut A, Harvey PH (1996) Inferring population history from molecular phylogenies. In: Harvey PH, Brown AJL, Maynard Smith J, Nee S (eds) New uses for new phylogenies. Oxford University Press, Oxford, pp 66–80

Novacek MJ, Wyss AR, McKenna M (1988) The major groups of eutherian mammals. In: Benton MJ (ed) The phylogeny and classification of the tetrapods, Vol 2. Clarendon Press, Oxford, pp 31–71

Otto SP, Cummings MP, Wakeley J (1996) Inferring phylogenies from DNA sequence data: The effects of sampling. In: Harvey PH, Brown AJL, Maynard Smith J, Nee S (eds) New uses for new phylogenies. Oxford University Press, Oxford, pp 103–115

Sneath PHA (1966) Relations between chemical structure and biological activity. J Theor Biol 12:157–195

Sonneborn TM (1965) Degeneracy of the genetic code: Extent, nature and genetic implications. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 377–397

Vogel H, Zuckerkandl E (1972) The evolution of polarity relationships in globins. In: Neyman J (ed) Darwinian, neo-Darwinian, and Non-Darwinian evolution. Proc 6th Berkeley Symp Math Stat Prob. University of California Press, Berkeley, pp 155–176

Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC (1966) On the fundamental nature and evolution of the genetic code. Cold Spring Labor Symp Quant Biol 31:723–736

Xia X (1996) Maximizing transcription efficiency causes codon usage bias. Genetics 144:1309–1320

Xia X (1998) The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. Mol Biol Evol 15:336–344

Xia X, Hafner MS, Sudman PD (1996) On transition bias in mitochondrial genes of pocket gophers. J Mol Evol 43:32–40

Yang Z (1996) Phylogenetic analysis by maximum likelihood (PAML). Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650

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