J Mol Evol (1985) 22:53–62

Journal of
Molecular Evolution

Amino Acid Composition and the Evolutionary Rates of Protein-Coding Genes

Dan Graur*

Center for Demographic and Population Genetics, University of Texas Health Science Center, PO Box 20334, Houston, Texas 77225, USA

Summary. Based on the rates of amino acid substitution for 60 mammalian genes of 50 codons or more, it is shown that the rate of amino acid substitution of a protein is correlated with its amino acid composition. In particular, the content of glycine residues is negatively correlated with the rate of amino acid substitution, and this content alone explains about 38% of the total variation in amino acid substitution rates among different protein families. The propensity of a polypeptide to evolve fast or slowly may be predicted from an index or indices of protein mutability directly derivable from the amino acid composition. The propensity of an amino acid to remain conserved during evolutionary times depends not so much on its being featured prominently in active sites, but on its stability index, defined as the mean chemical distance [R. Grantham (1974) Science 185:862–864] between the amino acid and its mutational derivatives produced by single-nucleotide substitutions. Functional constraints related to active and binding sites of proteins play only a minor role in determining the overall rate of amino acid substitution. The importance of amino acid composition in determining rates of substitution is illustrated with examples involving cytochrome c, cytochrome b_5, *ras*-related genes, the calmodulin protein family, and fibrinopeptides.

Key words: Rate of amino acid substitution — Amino acid composition — Glycine — Functional constraints

*Present address: Lehrstuhl für Populationsgenetik, Institut für Biologie II der Universität Tübingen, Auf der Morgenstelle 28, 7400 Tübingen 1, West Germany.

Introduction

One of the most valuable principles in molecular evolution is that functionally less important proteins or parts of proteins evolve, in terms of substitution rates, faster than the more important ones (Dickerson 1971; Kimura and Ohta 1974; Dayhoff 1972, 1978; for reviews, see Doolittle 1979; Kimura 1983; Nei and Koehn 1983). Noteworthily, this postulate is presented by Kimura (1983, pp. 98–113) as one of the four empirical principles that distinguish molecular evolution from phenotypic evolution. In practice, however, it is usually very difficult, if not impossible, to quantify the importance of a protein or a gene objectively except by illustrating the principle with such extreme examples as the evolution of fibrinopeptides vs the evolution of histones (Dickerson 1971), functional genes vs pseudogenes (Li et al. 1981), third vs first and second positions of codons (Miyata et al. 1980; Li et al. 1981), and silent changes vs amino acid replacement changes. Because of this difficulty, the "importance" of a protein or a site is frequently inferred from its rate of evolution, and the argument thus becomes a circular one. In other words, genes or parts of genes that are conserved during the evolutionary process are assumed a priori to be functionally more important than those that evolve faster.

But is it really so? Can we really say that cytochrome c, for instance, is twice as important as cytochrome b_5 just because the former evolves 2 times slower than the latter? Can we even be sure that conserved sites are functionally more important than variable sites? Taniguchi et al. (1980), for instance, concluded that the conserved amino acids in human leukocyte and fibroblast interferons are subjected to strong selective pressure because these sites are like-

Table 1. Rates of nucleotide substitution ($\times 10^{-9}$ per site per year) and indices of mutability in various genes

Protein	Nonsynonymous rate of substitution	Synonymous rate of substitution	I_1	I_2	I_c	I_m
Histone H4 (2)	0.004	6.13	−0.505	−0.658	−0.664	0.052
α-Actin (4)	0.01	3.67	0.754	0.461	0.592	0.149
Histone H3 (2)	0.04	6.38	1.000	0.279	0.539	0.031
Histone H2B (2)	0.07	3.59	0.931	0.396	0.081	0.087
Histone H2A (2)	0.08	2.06	0.342	0.635	0.560	−0.002
Hypoxanthine phosphoribosyl transferase (3)	0.13	2.13	0.765	0.763	0.154	0.011
Gastrin (2)	0.15	3.52	0.194	0.163	0.118	0.238
Insulin (5)	0.16	5.41	0.654	0.335	0.309	0.189
Neurophysin II (2)	0.43	2.23	−0.263	−0.172	0.569	0.558
Metallothionein II (3)	0.43	2.83	0.662	0.757	0.451	0.317
Parathyroid hormone (2)	0.44	1.73	1.007	1.630	0.808	0.506
Fibrinogen γ (2)	0.55	5.82	0.520	0.693	0.616	0.240
Hemoglobin α (9)	0.56	3.94	0.901	0.738	0.474	0.433
Metallothionein I (3)	0.61	3.38	0.518	0.802	0.576	0.796
α-Glycoprotein hormone (4)	0.67	6.23	1.058	0.526	0.434	0.636
Hemoglobin δ (8)	0.87	2.96	0.624	1.079	0.930	1.027
Serum albumin (4)	0.92	6.72	1.337	0.931	0.677	0.854
Growth hormone (4)	0.95	4.37	1.096	0.775	1.242	1.078
α-Lactalbumin (3)	1.10	4.16	1.183	1.083	0.970	1.394
α-Fetoprotein (3)	1.21	4.90	1.093	1.107	1.514	1.052
Prolactin (3)	1.29	5.59	1.084	1.255	1.145	1.141
Interferon α_1 (2)	1.41	3.53	1.400	1.412	1.374	1.336
Interferon α_2 (2)	1.47	3.15	1.401	1.246	1.300	1.460
Relaxin C peptide (3)	1.78	4.32	1.286	1.639	2.003	2.024
Interferon β_1 (2)	2.21	5.88	1.396	1.739	2.025	2.333
Relaxin (3)	2.51	7.49	1.075	1.080	1.377	1.907
Interferon γ (2)	2.80	8.59	1.306	2.182	2.688	2.758

Values in parentheses represent the numbers of sequences from different organisms used for computing amino acid frequencies. For definitions of indices of mutability, see text

ly to be essential for function. Recently, however, Valenzuela et al. (1985) checked this assumption, and found out that proteins that were mutated in several strictly invariable regions retained biological activities indistinguishable from those of the wild-type interferon.

Furthermore, the assertion that a certain protein is functionally important is often made even before the very function of the gene in question is known. Recently, for instance, *ras*-related genes were found to be extremely conserved evolutionarily, the degree of conservation being similar to that of histones (Shilo and Weinberg 1981; DeFeo-Jones et al. 1983; Gallwitz et al. 1983). Although the normal function of the cellular *ras*-related gene product or products is unknown, it has already become common lore (e.g., see Newmark 1983) that they are important.

In a preliminary survey of proteins, I found that the amino acid composition of a protein to a large extent determines, at least qualitatively, the rate of amino acid substitution. The larger the content of glycine, cysteine, and tyrosine, for instance, the slower the rate. In this study, I approach the problem of what determines rates of molecular evolution from a more quantitative point of view. I shall consider some measurable properties of proteins, namely different parameters of amino acid composition, and see whether or not they can explain the differences in rates of molecular evolution between protein families.

Data

This study is based on two independently derived sets of data. Li et al. (1985) compiled data on rates of synonymous (silent) and nonsynonymous (amino acid altering) nucleotide substitutions. The rates of substitution for 27 mammalian genes for which the number of codons compared is larger than 50 are listed in Table 1. The data for genes with less than 50 codons were not used in the present study to avoid major effects of sampling error. The mean nonsynonymous rate for these proteins is $(0.84 \pm 0.15) \times 10^{-9}$ per site per year, and the mean synonymous rate is $(4.45 \pm 0.34) \times 10^{-9}$. For unknown reasons, the rate of synonymous substitution is positively and significantly correlated with the rate of

nonsynonymous substitution ($r = 0.5126$, $P = 0.003$). We see from Table 1 that the rate of amino acid substitution varies greatly with the gene, with the highest rate being about 700 times higher than the lowest one. In comparison, the highest rate of synonymous substitution is only about 5 times higher than the lowest one. Similarly, the variation in both synonymous and nonsynonymous rates among different organisms for a given protein is very small (Wu and Li 1985).

The second set of data on amino acid substitution is taken from Dayhoff (1978) and Marshall and Brown (1975). We used the same criteria as with the previous set in choosing the proteins for this study, and in order to maintain the independence of this set of data, we excluded those proteins for which estimates on substitution rates were available in Li et al. (1985). The estimates of amino acid substitution for the 33 proteins included in this set are in PAM units (accepted point mutations per amino acid per 10^8 years). We point to the fact that PAM estimates are not straightforwardly comparable with Li et al.'s estimates of nonsynonymous substitutions. In general, the estimates by Li et al. are more reliable, since they are derived mostly from nucleotide sequences of closely related organisms (i.e., mammals) whose evolutionary histories are fairly well known. Li et al.'s method also takes into account the relative likelihoods of nucleotide and codon changes. Dayhoff's estimates, in comparison, are based on amino acid sequences from very divergent species, and problems concerning alignment uncertainties and multiple mutations at the same site are unavoidable. We shall thus use this second set of data only for qualitative purposes of checking the results and conclusions derived from the first set. In this set of data the highest rate of amino acid substitution is for κ-casein (33.0 PAM) and the lowest is for ubiquitin (<0.1 PAM). In both sets of data we excluded immunoglobulins, because their rate and pattern of nucleotide substitution was previously found to differ considerably from the pattern found in other proteins (Gojobori and Nei 1984; D. Graur, unpublished results).

Data Analysis

Amino Acid Composition and Substitution Rates

The high intergenic variation in the rates of amino acid substitution must be due at least partially to the amino acid compositions of proteins, since some amino acids are known to be highly mutable and some are highly conservative (Dayhoff 1978). In other words, amino acids differ from one another in their evolutionarily effective mutation rates (i.e., substitution rates), and they are fixed with different probabilities, due in part to differences in their intrinsic mutation rates, but mainly to differences in the stringency of purifying selection (Gojobori et al. 1982).

Table 2. Expected mean chemical distances (stability indices, S) for amino acids

Amino acid	S based on Grantham's (1974) index	S based on Miyata et al.'s (1979) index
Cys	168.14 (1)	2.42 (5)
Trp	150.57 (2)	3.34 (1)
Tyr	124.33 (3)	2.64 (2)
Gly	105.39 (4)	2.51 (3)
Ser	100.03 (5)	1.83 (8)
Arg	97.59 (6)	2.50 (4)
Asp	90.75 (7)	2.03 (6)
Asn	85.25 (8)	1.89 (7)
Glu	71.57 (9)	1.71 (10)
Lys	71.36 (10)	1.59 (11)
Ala	70.75 (11)	1.11 (19)
Pro	69.42 (12)	1.52 (12)
Val	68.25 (13)	1.78 (9)
Phe	64.88 (14)	1.14 (18)
Thr	62.83 (15)	1.32 (17)
His	60.63 (16)	1.44 (14)
Leu	59.06 (17)	1.52 (13)
Ile	58.43 (18)	1.39 (15)
Gln	51.71 (19)	1.37 (16)
Met	38.67 (20)	1.01 (20)

Values in parentheses represent rank

To see the effect of amino acid composition on the rates of substitution, we developed a measure of amino acid immutability or stability, denoted S, considering the changes in chemical properties resulting from a point mutation in each amino acid, and examined the relationship between the theoretical measure and the rate of amino acid substitution. Since chemically similar amino acids are known to be more interchangeable than dissimilar ones, due to the structure of the genetic code and the pattern of purifying selection (Clarke 1970; Jukes and King 1971, 1979; Gojobori et al. 1982; Graur 1985), we measure the stability of each amino acid by the average chemical distance (Grantham 1974; Miyata et al. 1979) between an amino acid and its mutational derivatives that can be produced by a single-nucleotide substitution. For example, methionine (Met) changes to arginine (Arg), isoleucine (Ile), leucine (Leu), Lysine (Lys), threonine (Thr), and valine (Val) with relative probabilities of 1/9, 3/9, 2/9, 1/9, 1/9, and 1/9, respectively, when a single-nucleotide substitution occurs at random (Nei 1975, p. 23). Grantham's chemical distances between Met, on the one hand, and Arg, Ile, Leu, etc., on the other, are 91, 10, 15, etc., respectively. Therefore, the average distance or the stability of Met is $S_{Met} = 1/9 \times 91 + 3/9 \times 10 + 2/9 \times 15 + \ldots = 38.7$. The stability

indices for the other 19 amino acids were computed in the same way, and the resulting values are presented in Table 2. This table also includes the stability indices obtained by using Miyata et al.'s (1979) distances. The mean indices of stability based on Grantham's and Miyata et al.'s distances are 83.48 ± 7.39 and 1.80 ± 0.14, respectively. The rankings of the stability indices for the 20 amino acids obtained by using the two chemical distances are to a certain extent different, but are highly correlated ($r = 0.870$). Both stability indices indicate that four amino acids (Gly, Try, Cys, and Trp) are highly stable (more than 2 standard errors above the mean for the entire group of amino acids), whereas six amino acids (Phe, Thr, His, Ile, Gln, and Met) are highly mutable (more than 2 standard errors below the mean).

For Li et al.'s (1985) data we calculated the correlation coefficient between the sum of the frequencies of the four most highly stable amino acids and the rate of amino acid substitution. The correlation coefficient (−0.509) was statistically significant and negative. The correlation coefficient (+0.328) between the sum of frequencies of the six most highly mutable amino acids and the rate of amino acid substitution was statistically significant and positive. The correlation coefficient between the sum of the frequencies of the ten amino acids with intermediate stability indices and the rate of amino acid substitution was 0.1961 ($P = 0.163$). The results indicate that proteins containing many highly immutable amino acids tend to evolve slower than those containing many highly mutable amino acids. Surprisingly, no significant correlation was found between any of these three variables and synonymous substitution rates, even though the two rates of substitution are correlated.

However, the correlation coefficients are not very high, indicating either that the stability indices we used are not appropriate or that factors other than amino acid composition play a role in determining substitution rates (e.g., the elusive "importance"). There may be two reasons why the direct usage of the indices of stability may not be appropriate. First, it should be noted that the chemical distances of Grantham and Miyata et al. are constructed by considering only three and two physicochemical properties of amino acids, respectively, which properties are chosen for their intuitive appeal. Sneath (1966), for instance, uses more than 50 characteristics for determining his distances, and his list may not be exhaustive either. It is thus probable that no measure of chemical distance measures immutability precisely. Second, different amino acids may contribute with different weights to the determination of the overall rate of amino acid substitution. It is reasonable to believe that the frequency of amino acids, for example, may have something to do with that. Of the 10 highly mutable or highly immutable amino acids we considered previously, only Gly and Thr are present in proteins in appreciable frequencies (> 5%). The observed frequencies of rare amino acids, e.g., Met and Trp, may be subject to large sampling errors, and thus be uninformative in predicting the rate of substitution.

To determine the contribution of amino acid composition to the rates of amino acid substitution, we first examined the correlation between substitution rate and the frequency for each of all 20 amino acids. We found that the frequencies of Leu, Ala, Gly, Asn, Gln, Glu, and Phe correlate significantly with amino acid substitution rate; the significance levels for Gly, Asn, Gln, and Phe were <1%. The highest absolute correlation (−0.6187) was obtained for Gly, and this amino acid alone explains about 38% of the total variation in amino acid substitution rate. The highest positive correlation (+0.4920) was obtained for Gln. (Note that Gly and Gln are a highly stable and a highly mutable amino acid, respectively.) The relationships between the frequencies of Gly and Gln and the rate of amino acid substitution are shown in Fig. 1a and b, respectively. For Dayhoff's set of data we obtained smaller correlation coefficients, −0.3774 and 0.3133 for Gly and Gln, respectively, but both of these are statistically significant.

We then fitted a multiple regression equation of m variables (m = number of amino acids) that maximized the correlation between the observed rate and the rate predicted from amino acid composition. We used a forward inclusion process with increasing numbers of amino acids, starting with the variable that explained the largest amount of the correlation, namely Gly (Nie et al. 1975, pp. 321–342). We call these multiple linear regression equations the empirical indices of mutability, and denote them as I_m, where m is the number of amino acids used. For example,

$$I_7 = 0.841 - 5.096f_{Gly} + 24.145f_{Asn} - 26.807f_{Tyr} - 7.398f_{Val} + 18.219f_{Phe} - 8.263f_{Asp} + 7.960f_{Ile}$$

where f_i is the frequency of amino acid i. The I_m values for m = 1, 3, 7, and 20 are presented in Table 1. The relationship between I_7 and the rate of amino acid substitution is shown in Fig. 2.

The change in the explainable fraction of variation in the amino acid substitution rates (r^2) as the number of amino acids considered in I_m is increased is shown in Fig. 3. (The regression functions for m = 1 to m = 10 are given in the Appendix.) As expected, by increasing the number of amino acids used to predict the rate of amino acid substitution, we can increase the fraction of the total variation that is explainable. As we mentioned previously, Gly alone explains 38% of the variation. Using the frequencies of two amino acids (Gly and Asn) we can explain

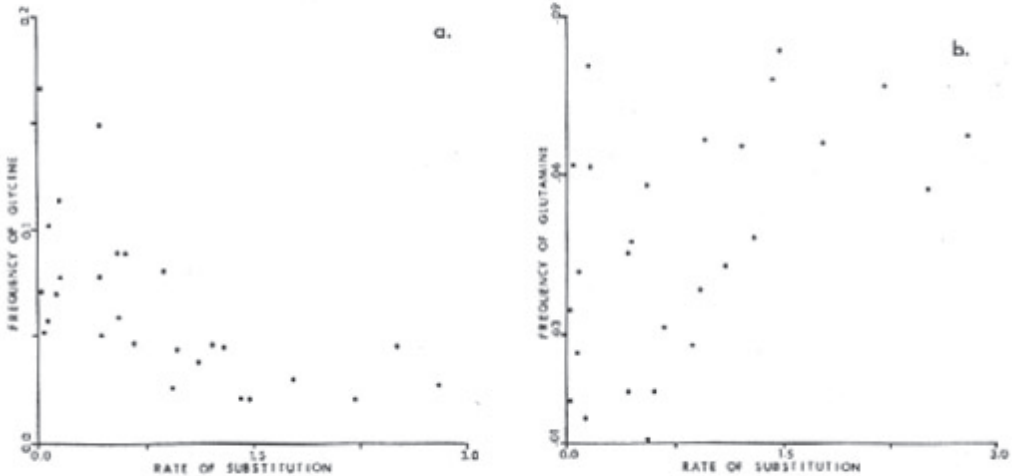

Fig. 1. Relationship between frequencies of glycine (a) and glutamine (b) and rates of amino acid substitution ($\times 10^{-9}$/year). Each point represents data for a different gene

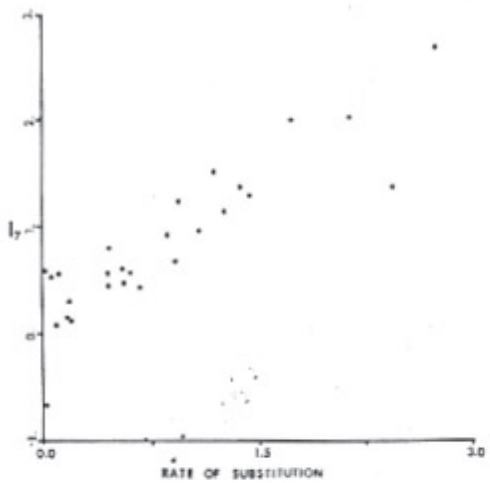

Fig. 2. Relationship between an empirical mutability index (I_r) and rates of amino acid substitution ($\times 10^{-9}$/year). Each point represents data for a different gene

about 50% of the variation. Using five amino acids we can explain about 73%, and using all 20 amino acids we can explain about 97% of the variation. However, with the increase in our ability to explain the variation in rates, it is expected that the intragenic variation should become larger. In other words, to be able to predict a difference in the rate of evolution between two proteins from information on their amino acid composition, we will need data from an increasingly large number of organisms. The question arises as to how many and which amino acids we should use in computing I_m to be able to say that protein A evolves faster, say 2 times faster, than protein B, based on a reasonably small number of sequences.

To study this problem, we used amino acid sequences of hemoglobins α and β taken from 63 different organisms (Dayhoff et al. 1983). We calculated the coefficients of variation for the 20 different indices of mutability. The results are presented in Table 3. As one can see, the coefficient of variation increases rapidly with the number of amino acids used, and at $m = 8$ it already exceeds 100%. We then calculated for each I_m the minimum difference in the rate of amino acid substitution that is detectable with 95% confidence between two proteins at the 5% level of significance, given one, two, or three sequences from each protein. We followed the procedure of Sokal and Rohlf (1969, p. 247). The results are given in Table 3. As one can see, the mutability indices based on seven or more amino acids have very poor resolving power even when three sequences from each protein are used. In general, with these indices it is difficult to detect small differences in the rate of amino acid substitution between two proteins based on amino acid composition only. Yet, since we have completely ignored functional constraints, it is remarkable that a prediction, rough as it may be, concerning rates of evolution of different proteins is feasible at all.

Using the I_m values derived from Li et al.'s data we checked our ability to predict the rates of amino acid substitution for the independent set of data (Dayhoff's). Only the I_m values for $m < 5$ gave statistically significant results. The correlation coeffi-

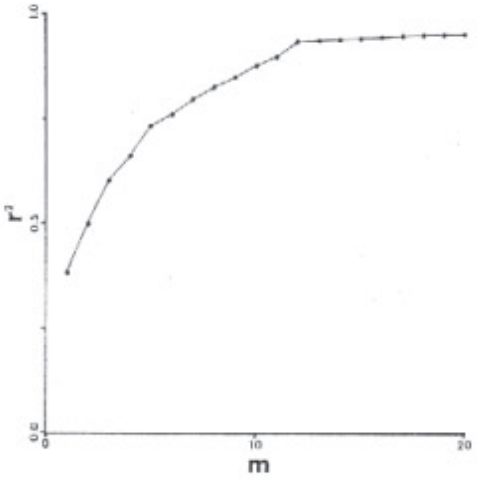

Fig. 3. The fraction of variation (r^2) in amino acid substitution rates explained by different indices of mutability (I_m)

cients for m = 1, 2, 3, and 4 were 0.3774, 0.3972, 0.3082, and 0.3167, respectively. Therefore, I_m does not perform as well for the second set of data. We feel, however, that these results reflect more the crudeness of Dayhoff's estimates than an intrinsic flaw in our method. The ratio of the rate of substitution for Dayhoff's data to that for Li et al.'s data varied from 3.5 to 65.3 for the 16 proteins for which both estimates were available. The mean ratio (±SE) was 18.56 ± 3.72. In comparison, the theoretical expectation for the ratio of the two estimates is 30 (rate per amino acid per 10^{-8} years/rate per nucleotide per 10^{-9} years). This discrepancy may be the reason why the results obtained with the second set of data are not as good as those obtained with Li et al.'s estimates. We admit, however, that our method is still very crude, and improvements should be made as more reliable data on substitution rates are accumulated.

Depending on the number of sequences available and the expected difference in rate of amino acid substitution between the proteins in question, we suggest using only those I_ms for which $m < 7$. Our experience is that for proteins of unknown function, a reasonably good qualitative prediction of the evolutionary rates (fast, moderate, or slow) can be made even with values of m as small as I_3.

Active Sites

It is fairly well established that active sites evolve slower than does the rest of the protein. The correlation with amino acid composition reported above

Table 3. Coefficient of variation (CV) and percentage minimum difference (100% = 1.0) in rates of amino acid substitution between two different proteins that can be detected with different I_m

m	CV (%)	Percentage difference		
		One sequence	Two sequences	Three sequences
1	16.3	4.4	1.2	0.7
2	30.2	8.1	2.2	1.4
3	41.0	11.0	3.0	1.9
4	85.6	23.0	6.2	3.9
5	44.2	11.9	3.2	2.0
6	56.7	15.2	4.1	2.6
7	71.3	19.2	5.2	3.2
8	115.2	31.0	8.3	5.2
9	117.6	31.6	8.5	5.3
10	139.6	37.6	10.1	6.3
11	101.4	27.3	7.3	4.6
12	93.8	25.2	6.8	4.2
13	107.4	28.9	7.8	4.9
14	119.8	32.2	8.7	5.4
15	121.5	32.7	8.8	5.5
16	124.8	33.6	9.0	5.6
17	156.2	42.0	11.3	7.1
18	192.2	51.7	13.9	8.7
19	191.7	51.6	13.9	8.7
20	177.8	47.8	12.8	8.0

may thus reflect the composition of active sites rather than that of the entire protein. In the following, we shall further establish the relationship between amino acid composition and substitution rate, and shall distinguish between composition effects and effects on rate due to active sites. For this purpose, we compared the frequencies of the 20 amino acids in variable, conservative, invariable, and active sites (Table 4). The protein sequences were taken from Dayhoff's (1976, 1978) and Dayhoff et al.'s (1983) compilations. Conservative sites were defined as sites containing the same amino acid in >80% of at least five homologous sequences derived from two or more different orders of organisms. Invariable sites are those that contain the same amino acid in all homologous sequences. Active sites were defined as regions involved in binding substrate, cofactor, coenzyme, or prosthetic moieties, as identified independently of either the rates of evolution or possible homology with other proteins.

Seven amino acids (Ser, Lys, Glu, Asp, Arg, Cys, and His) make up 81% of all active sites. In comparison, the overall frequency of these amino acids in proteins is only 35%. Of these amino acids only Cys is more common in invariable sites than its overall frequency in proteins. Cys is also the only amino acid in this group that, judging from its index of stability (Table 2), is a highly immutable amino acid. All the other six amino acids featuring prom-

inently in active sites of many different kinds of proteins either maintain the same frequency (Ser, Asp, Arg, and His), or are in fact significantly less abundant (Lys and Glu) in conserved and invariable regions than their overall frequencies in proteins. This indicates that active sites, while evolving very conservatively, constitute only a minor fraction of total proteins, such that they are not very important in determining the overall rate of molecular evolution of a protein. Gly, on the other hand, rarely occurs in active sites, yet its frequency in invariable sites is almost 2 times higher than its overall frequency in proteins.

We conclude at this point that the conservation of an amino acid in the evolutionary process depends not so much on its frequent occurrence in active sites, but on its propensity to mutate acceptably across the entire length of a protein.

Discussion

Our results indicate that to a great extent, the differences in the rates of amino acid substitution are attributable to differences in the primary structure of proteins, i.e., amino acid composition. We do not have to invoke differences in functional importance every time we find that two proteins evolve at different rates. Proteins of equivalent importance may evolve at different rates depending on their compositions. We acknowledge, however, that by studying the primary structure of proteins it may not always be possible to predict correctly the absolute or relative rates of molecular evolution. This is especially true when dealing with short polypeptides and proteins composed mostly of active sites. Nevertheless, it is clear that different proteins have different propensities to tolerate amino acid substitution depending on their amino acid constitutions. We have also shqwed that functional constraints related to active and binding sites of proteins play only a minor role in determining the overall rate of amino acid substitution.

I am obviously drawing too sharp a dichotomy between amino acid composition and protein function. The reasons are that while data on amino acid composition are available readily and in enormous quantities, data pertaining to functions of specific sites of proteins derived independently of degree of evolutionary conservation are extremely scarce and mostly inaccessible to the uninitiated. Optimally, one should remove the effects of site-specific functional constraints, and only then assess the importance of amino acid composition and its effects on evolutionary rates. Unfortunately, although excellent methods exist (e.g., see Zuckerkandl 1976), data are lacking.

Table 4. Frequencies of amino acids in different regions of proteins

Amino acid	Region			
	All	Conserved	Invariable	Active sites
Ala	8.6	6.9	6.8	1.1
Gly	8.4	9.7	14.1	2.0
Leu	7.4	6.5	6.9	2.0
Ser	7.0	6.9	5.1	8.7
Val	6.6	7.4	5.0	0.8
Lys	6.6	4.7	5.0	10.4
Thr	6.1	5.7	4.3	1.1
Glu	6.0	4.8	3.2	7.3
Asp	5.3	6.0	4.9	12.9
Pro	5.2	7.4	7.3	0.6
Arg	4.9	3.7	3.6	10.9
Ile	4.5	4.8	2.7	0.8
Asn	4.3	3.6	3.4	0.6
Gln	3.9	4.3	1.8	0.3
Phe	3.6	4.1	2.3	2.0
Tyr	3.4	4.9	3.2	4.8
Cys	2.9	4.2	15.5	15.4
His	2.0	1.1	2.2	15.4
Met	1.7	1.6	1.4	2.2
Trp	1.3	1.7	1.4	0.8

The sample sizes were 49,931 amino acids for all regions, 936 for conserved regions, 740 for invariable regions, and 357 for active sites

One thing in particular has become clear from this study, namely that the content of glycine residues is extremely important in determining the rate of amino acid substitution. It seems that Gly, which is the smallest amino acid and a "borderline" member of the group of amino acids with uncharged polar groups (Lehninger 1975, p. 72), is almost uninterchangeable with any other amino acid, possibly because radical changes in a protein's tertiary structure would be induced by such a mutation. The molecular volume of Gly is 3 $Å^3$, which is 10 times less than that of the next smallest amino acid (Ala). Moreover, taking into account the genetic code and assuming equal frequency of mutations, a Gly residue will, on the average, be replaced by an amino acid about 26 times larger. Such a radical change will tend to distort the tertiary structure and subsequently alter the function of many proteins, regardless of the position of Gly and its proximity to the active sites. No mutation in another amino acid can produce such a drastic result. Consequently, mutations involving Gly will be strongly selected against. According to Grantham's and Miyata et al.'s chemical distances Gly is not predicted to be the most immutable amino acid, but in their distance measures the effects of molecular volumes are grossly underestimated (more so in Grantham's). French and Robson (1983) found that "bulk" (volume) is one of the three most conserved properties in the evolution of proteins, and this may account

for the evolutionary features of glycine. Moreover, glycine rarely appears in active sites of proteins (Table 4), but it is one of the most conserved amino acids. The highly conserved cytochrome c and the rapidly evolving fibrinopeptides are good examples with which to make this point.

Cytochrome c has, on the average, 13 Gly residues per molecule (12.5%). Of the 34 invariant amino acids in this protein among 32 sequences taken from organisms such as yeast, *Drosophila*, human, etc., 8 (23.5%) are glycines. Noteworthily, only one of these eight invariant Gly residues occurs in any close proximity to the heme molecule (the main active site). Five glycines are located on the exterior of the molecule and two are in buried side chains with no known function (Baba et al. 1981). Interestingly, it has been established that not only do amino acids located on the surface of a globular protein not have any specific function, but they also contribute nothing to the overall stability of proteins, thus probably lacking even a general stabilizing function (Grütter and Hawkes 1983). M. Goodman (personal communication) suggested that the oxidase-reductase interaction domain is an important functional area. Interestingly, only 3 of the 16 sites involved in this domain are occupied by invariable amino acids in all 32 species, in all 3 cases by lysine. The degree of conservatism (18.8%) is thus much lower than the overall degree of conservatism across the entire cytochrome c molecule (33.6%). Hence, we refute the notion that this active site contributes to the low rate of molecular evolution of cytochrome c. In fact, using Zuckerkandl's (1976) formula, the functional density (FD) of cytochrome c is found to range between 0.26 and 0.39, depending on whether the side chains packed against the heme molecule (Baba et al. 1981) are included. In either case, the FD estimate for cytochrome c is lower than that for hemoglobin β (0.52), and this means that if one considers function only, one expects cytochrome c to evolve faster than hemoglobin β, contrary to the facts.

The rate of evolutionary change of fibrinopeptides is among the most rapid ever observed. In fact, fibrinopeptide B (19 amino acid) shows one amino acid difference even between the closely related species dog and fox. Interestingly, the only two invariant positions in fibrinopeptides from species ranging from lizards to humans are occupied by glycine, only one of which serves a function (the cleavage site between fibrinopeptides and fibrins).

Recently, Baba et al. (1984) studied the rates of molecular evolution of the various members of the calmodulin family of proteins. They concluded that the "spectrum of evolutionary tempos displayed by the members of the calmodulin family may mirror a variable spectrum of selective restraints related to the unique physiological role of each protein." The relative conservatism of calmodulin, for instance, is assumed to reflect the functional versatility of this protein in regulating diverse activities in cells, in comparison with the very few functions attributed to other members of the calmodulin family, such as the catalytic and regulatory light chains of myosin and parvalbumin. It is entirely possible that Baba et al. (1984) are right, and the differences in the rates of evolution are fully explainable by differences in function only. However, the FDs of these proteins are not known even approximately, and Baba et al., on discovering differences in evolutionary rate among the members of the calmodulin family of proteins, assigned differences in functional constraints post factum. Actually, the differences in the rates of evolution of these proteins are explainable by their amino acid compositions and our indices of mutability. If we consider only those proteins for which data from mammalian species exist, we see that calmodulin evolves the slowest, followed by troponin C, whose evolution is also slow but not nearly as slow as that of calmodulin. The catalytic light chain of myosin is next, followed by the regulatory light chain of myosin. (We excluded all proteins in the calmodulin family for which the rates of amino acid substitution are based on dates of divergence prior to the mammalian radiation, i.e., before about 80 million years ago, since we have little confidence in the accuracy of these dates, and inaccuracies in long times of divergence cause major biases in the estimates of the rates of molecular evolution.) Our first observation is that the content of glycine is approximately constant (6.3–8.2%) in these proteins, such that I_1 reveals no difference in expected rate of nucleotide substitution. On the other hand, all other indices of mutability reveal a potential difference in the rates of evolution of these proteins. The values of I_7, for instance, are 0.749, 0.874, 1.457, and 2.093 for calmodulin, troponin C, and catalytic and the regulatory light chains of myosin, respectively. The correlation between rate and amino acid composition is perfect.

As to the questions we posed at the beginning of this article regarding the rates of evolution of *ras*-related genes and those of cytochromes c and b_5, we see that by using amino acid composition we can explain the differences in rates of substitution satisfactorily. In the case of the two cytochromes, it is enough to consider their respective Gly contents to be able to predict their relative rates of molecular evolution. Cytochrome c contains 12.6% Gly residues, whereas cytochrome b_5 contains only 4.8%. Judging from the amino acid composition only, we expect cytochrome c to be much more conserved than cytochrome b_5, in complete agreement with the facts.

The story of the *ras*-related genes is similar. Although they do not contain extraordinary amounts of highly stable amino acids, they do contain conspicuously small amounts of highly mutable amino acids (e.g., Phe and Ile). All indices of mutability except I_1 reveal this fact. The I_7 values for the p30 and p21 proteins of Harvey and Kirsten murine sarcoma viruses, for instance, are −0.475 and −0.300, respectively. These are among the lowest values we have found (cf. Table 1). Consequently, these proteins are expected to undergo a slow rate of molecular evolution just because of their amino acid compositions, and regardless of their putative functions.

Acknowledgments. I thank Dr. Masatoshi Nei for informative discussions and helpful comments, Drs. W.-H. Li, C.-C. Luo, and C.-I. Wu for generously providing me with unpublished data, and Mr. Robert Schwartz for his help with computational facilities. Drs. E. Zuckerkandl and M. Goodman critically reviewed this paper, and provided many important insights, greatly improving this analysis. This work was supported by grants NIH GM-20293 and NSF BSF-8315115 to Dr. M. Nei.

Appendix

The indices of mutability (I_m) were calculated by a forward (stepwise) inclusion approach. The independent variables (the frequencies of the 20 amino acids) were entered one by one into the multiple regression function, the order of inclusion being determined by their respective added contributions to the explained variance in rates of amino acid substitution (for procedural details, see Nie et al. 1975, pp. 321–367). The first variable to be used was the frequency of the amino acid glycine. The I_m functions for m values from 1 to 10 are as follows:

$$I_1 = 1.674 - 13.008f_{Gly}$$

$$I_2 = 0.755 - 9.420f_{Gly} + 18.399f_{Asn}$$

$$I_3 = 0.863 - 8.629f_{Gly} + 25.420f_{Asn} - 14.623f_{Tyr}$$

$$I_4 = 1.063 - 7.719f_{Gly} + 28.306f_{Asn} - 13.096f_{Tyr} - 7.567f_{Val}$$

$$I_5 = 0.697 - 5.290f_{Gly} + 25.123f_{Asn} - 18.259f_{Tyr} - 8.738f_{Val} + 14.909f_{Pro}$$

$$I_6 = 1.001 - 5.720f_{Gly} + 26.727f_{Asn} - 20.211f_{Tyr} - 8.567f_{Val} + 15.311f_{Pro} - 6.390f_{Asp}$$

$$I_7 = 0.841 - 5.096f_{Gly} + 24.145f_{Asn} - 26.807f_{Tyr} - 7.398f_{Val} + 18.219f_{Pro} - 8.263f_{Asp} + 7.960f_{Ile}$$

$$I_8 = 1.908 - 6.546f_{Gly} + 16.495f_{Asn} - 31.439f_{Tyr} - 4.862f_{Val} + 17.965f_{Pro} - 9.899f_{Asp} + 7.614f_{Ile} - 7.487f_{Ala}$$

$$I_9 = 1.586 - 5.404f_{Gly} + 14.028f_{Asn} - 32.585f_{Tyr} - 6.041f_{Val} + 24.126f_{Pro} - 9.670f_{Asp} + 6.969f_{Ile} - 10.040f_{Ala} + 5.435f_{Lys}$$

$$I_{10} = 1.585 - 5.624f_{Gly} + 9.490f_{Asn} - 34.545f_{Tyr} - 3.920f_{Val} + 26.267f_{Pro} - 13.519f_{Asp} + 8.363f_{Ile} - 11.469f_{Ala} + 7.338f_{Lys} + 16.163f_{Trp}$$

References

Baba ML, Darga LL, Goodman M, Czelusniak J (1981) Evolution of cytochrome c investigated by the maximum parsimony method. J. Mol Evol 17:197–213

Baba ML, Goodman M, Berger-Cohn J, Demaille JG, Matsuda G (1984) The early adaptive evolution of calmodulin. Mol Biol Evol 1:442–455

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

Dayhoff MO (ed) (1972) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Silver Spring, Maryland

Dayhoff MO (ed) (1976) Atlas of protein sequence and structure, vol 5, suppl 2. National Biomedical Research Foundation, Washington, DC

Dayhoff MO (ed) (1978) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington, DC

Dayhoff MO, Hunt LT, Barker WC, Orcutt BC, Yeh LS, Chen HR, George DG, Blomquist MC, Johnson GC (1983) Protein sequence database (June release). National Biomedical Research Foundation, Washington, DC

DeFeo-Jones D, Scolnick EM, Koller R, Dhar R (1983) *ras*-Related gene sequences identified and isolated from *Saccharomyces cerevisiae*. Nature 306:707–709

Dickerson RE (1971) The structure of cytochrome c and the rates of molecular evolution. J Mol Evol 1:26–45

Doolittle RF (1979) Protein evolution. In: Neurath HD (ed) The proteins. Academic Press, New York, pp 1–118

French S, Robson B (1983) What is a conservative substitution? J Mol Evol 19:171–175

Gallwitz D, Donath C, Sander C (1983) A yeast gene encoding a protein homologous to the human c-*has*/*bas* proto-oncogene product. Nature 306:704–707

Gojobori T, Nei M (1984) Concerted evolution of the immunoglobulin V_H gene family. Mol Biol Evol 1:195–212

Gojobori T, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

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

Graur D (1985) Pattern of nucleotide substitution and the extent of purifying selection in retroviruses. J Mol Evol 21:221–231

Grütter MG, Hawkes RB (1983) Mutation and the conformational stability of globular proteins. Naturwissenschaften 70:434–438

Jukes TH, King JL (1971) Deleterious mutations and neutral substitutions. Nature 231:114–115

Jukes TH, King JL (1979) Evolutionary nucleotide replacements in DNA. Nature 281:605–606

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

Kimura M, Ohta T (1974) On some principles governing molecular evolution. Proc Natl Acad Sci USA 71:2848–2852

Lehninger AL (1975) Biochemistry. Worth Publishers, New York

Li W-H, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239

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

Marshall DR, Brown AHD (1975) The charge-state model of protein polymorphism in natural populations. J Mol Evol 6: 149–163

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

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328–7332

Nei M (1975) Molecular population genetics and evolution. North-Holland, Amsterdam

Nei M, Koehn RK (1983) Evolution of genes and proteins. Sinauer, Sunderland, Massachusetts

Newmark P (1983) More *rasmatazz*. Nature 306:642

Nie NH, Hull CH, Jenkins JG, Steinbrenner K, Bent DH (1975) SPSS. McGraw-Hill, New York

Shilo B-Z, Weinberg RA (1981) DNA sequences homologous to vertebrate oncogenes are conserved in *Drosophila melanogaster*. Proc Natl Acad Sci USA 78:6789–6791

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

Sokal RR, Rohlf FJ (1969) Biometry. WH Freeman, San Francisco

Taniguchi T, Mantei N, Schwarzstein M, Nagata S, Muramatsu M, Weissmann C (1980) Human leukocyte and fibroblast interferons are structurally related. Nature 285:547–549

Valenzuela D, Weber H, Weissmann C (1985) Is sequence conservation in interferons due to selection for functional proteins? Nature 313:698–700

Wu C-I, Li W-H (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA 82:1741–1745

Zuckerkandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. J Mol Evol 7:167–183

Received December 13, 1984/Revised March 20, 1985