J Mol Evol (1985) 21:233–257

Journal of
Molecular Evolution
© Springer-Verlag 1985

An Estimate on the Effect of Point Mutation and Natural Selection on the Rate of Amino Acid Replacement in Proteins

C. Frömmel and H.-G. Holzhütter

Institut für Physiologische und Biologische Chemie der Humboldt-Universität zu Berlin, Hessische Str. 3-4, DDR 1040 Berlin

Summary. We outline a method for estimating quantitatively the influence of point mutations and selection on the frequencies of codons and amino acids. We show how the mutation rate, i.e., the rate of amino acid replacement due to point mutation, can be affected by the codon usage as well as by the rates of the involved base exchanges. A comparison of the mutation rates calculated from reliable values of codon usage and base exchange probabilities with those that would be expected on the basis of chance reveals a notable suppression of replacements leading to tryptophan, glutamate, lysine, and methionine, and particularly of those leading to the termination codons.

If selection constraints are neglected and only mutations are taken into account, the best agreement between expected and observed frequencies of both codons and amino acids is obtained for $\alpha = 1.13$–1.15, where

$$\alpha = \frac{\text{rate of occurrence of base replacements affecting G–C pairs}}{\text{rate of occurrence of base replacements affecting A–U pairs}}.$$

The "selection values" of codons and amino acids derived by our method show a pattern that partially deviates from others in the literature. For example, the selection pressure on methionine and cysteine turns out to be much more pronounced than expected if only the discrepancies between their observed and expected occurrences in proteins are considered. To estimate to what extent randomly occurring amino acid replacements are accepted by selection, we constructed an "acceptability matrix" from the well-established matrix of accepted point mutations. On the basis of this matrix "acceptability values" of the amino acids can be defined that correlate with their selection values.

We also examine the significance of mutations and selection of amino acids with respect to their physicochemical properties and functions in proteins. The conservatism of amino acid replacements with respect to certain properties such as polarity can be brought about by the mutational process alone, whereas the conservatism with respect to other relevant properties—among them all measures of bulkiness—obviously is the result of additional selectional constraints on the evolution of protein structures.

Key words: Genetic code — Point mutations — Selection — Amino acid composition of proteins — Molecular evolution

Introduction

A fundamental problem in elucidating the structures and evolution of proteins concerns the functions of the 20 amino acids commonly occurring in proteins. Several attempts (Sneath 1966; Epstein 1967; Dunill 1968; Alff-Steinberger 1969; Papentin 1973; Grantham 1974; Batchinsky 1976; Salemme et al. 1977; Chou and Fasman 1978; Doolittle 1979; Sander and Schulz 1979; Wolfenden et al. 1979, 1981; Argyle 1980; Richard et al. 1980; Charton 1981; Hendry et al. 1981 a,b) have been made to set up principles that would permit the amino acids to be ordered in similarity groups. One possible approach is to compare the amino acids on the basis of their physicochemical properties (Sneath 1966; Dunill 1968;

Offprint requests to: H.-G. Holzhütter

Grantham 1974; Batchinsky 1976; Doolittle 1979; Wolfenden et al. 1979, 1981; Hendry et al. 1981a,b) but this approach requires knowledge of which properties actually determine the roles and specific functions of the amino acids in proteins. Another method is to analyze the transition probability matrix of accepted amino acid substitutions within a family of "modern" proteins constructed from a phylogenetic tree (Gamov 1954; Alff-Steinberger 1969; McLachlan 1971; Dayhoff et al. 1972; Grantham 1974; Yano and Hasegawa 1974; Salemme et al. 1977; Argyle 1980). This method is based on the assumption that amino acids having similar functions in proteins replace each other more frequently than others replace them. Finally, the genetic code may provide information about similarities between amino acids if it is postulated that similarity can be measured as the minimum number of point mutations necessary to replace a given amino acid by another (Alff-Steinberger 1969; McLachlan 1971; Dayhoff et al. 1972; Batchinsky 1976; Salemme et al. 1977; Sander and Schulz 1979; Argyle 1980). Only a rather crude classification can be obtained in this way.

As an extension and refinement of the last approach, we have examined the rate of occurrence of amino acid substitutions caused by point mutations. In the first part of this paper we demonstrate the extent to which the rate of amino acid substitution can be influenced by different kinds of point mutations and by codon usage. In the second part we calculate the steady-state amino acid composition of proteins expected on the basis of the mutational process only. From the discrepancies between expected and observed composition, we derive a quantitative estimate of the selectional constraints on the usage of codons and amino acids.

Methods

Our first approach to separating the effects of mutation and selection is based on the theory of natural self-organization developed by Eigen and Schuster (1979). Let us consider a system of mRNA sequences $\{C_{i_1}, C_{i_2}, \ldots, C_{i_N}\}$, where C_{i_α} denotes the codon in the α-th position and N is the length of the sequence. Here C_i refers to the base triplet $[B_1, B_2, B_3]$ with $i = 16[\mathrm{val}(B_1) - 1] + 4[\mathrm{val}(B_2) - 1] + \mathrm{val}(B_3)$, where the values val(B) of the four nucleotides B = A, G, C, and U are 1, 2, 3, and 4, respectively. For example, C_1 corresponds to [A,A,A] and C_{64} corresponds to [U,U,U] (cf. Table 1). The probability of finding at time t a sequence $\{C_{i_1}, C_{i_2}, \ldots, C_{i_N}\}$ in the considered system of mRNA sequences we denote by $P(i_1, i_2, \ldots, i_N)$. Taking the sum of this joint probability function over all 64 codons at certain positions we obtain so-called conditional probability functions of different orders:

$$1 = \sum_{\alpha=1}^{N} \sum_{i_\alpha=1}^{64} P(i_1, i_2, \ldots, i_N) \tag{1}$$

$$p(i_\alpha) = \sum_{\substack{\beta=1 \\ \beta \neq \alpha}}^{N} \sum_{i_\beta=1}^{64} P(i_1, i_2, \ldots, i_N) \tag{2}$$

$$p(i_\alpha, i_\beta) = \sum_{\substack{\gamma=1 \\ \gamma \neq \alpha,\beta}}^{N} \sum_{i_\gamma=1}^{64} P(i_1, i_2, \ldots, i_N) \tag{3}$$

Equation (1) simply represents the normalization condition for the probability function $P(i_1, i_2, \ldots, i_N)$. Equation (2) gives the probability $P(i_\alpha)$ of finding a sequence with codon C_{i_α} at the α-th position at time t. Similarly, $p(i_\alpha, i_\beta)$ is the probability of finding a sequence with the codon C_{i_α} at the α-th position and the codon C_{i_β} at the β-th position at time t.

For the temporal evolution of the system we propose the following kinetic equation:

$$\frac{d}{dt} P(i_1, \ldots, i_N) = \sum_{\alpha=1}^{N} \sum_{j_\alpha=1}^{64} [W_{i_\alpha j_\alpha} P(i_1, \ldots, i_{\alpha-1}, j_\alpha, i_{\alpha+1}, \ldots, i_N) - W_{j_\alpha i_\alpha} P(i_1, \ldots, i_{\alpha-1}, i_\alpha, i_{\alpha+1}, \ldots, i_N)] + [E(i_1, \ldots, i_N) - \bar{E}] P(i_1, \ldots, i_N) \tag{4}$$

In Eq. (4) the influence of point mutations is reflected by the first sum on the right-hand side; $W_{i_\alpha j_\alpha}$ denotes the probability of a codon replacement $C_{j_\alpha} \rightarrow C_{i_\alpha}$ occurring at position α per time unit. The presence of additional selective constraints is taken into account by the last term of the right-hand side of Eq. (4), where $E(i_1, \ldots, i_N)$ represents the selectional value of the sequence $\{C_{i_1}, \ldots, C_{i_N}\}$. The so-called excess productivity $\bar{E}$ is given by

$$\bar{E} = \sum_{\alpha=1}^{N} \sum_{i_\alpha=1}^{64} E(i_1, \ldots, i_N) P(i_1, \ldots, i_N), \tag{5}$$

which immediately follows from the normalization condition (1). If there is no mutation all sequences having a selectional value greater than $\bar{E}$ at time t will increase in frequency, whereas the others will die out.

Instead of $P(i_1, \ldots, i_N)$ containing all the information about the dynamics of the system of mRNA sequences, we will consider in the following the codon frequency F_i, which is defined as the probability of finding at time t codon C_i averaged over all sequences, i.e.,

$$F_i = \sum_{\alpha=1}^{N} p(i_\alpha = i). \tag{6}$$

A kinetic equation for F_i can be derived by summing up Eq. (4) using Eqs. (2) and (6):

$$\frac{d}{dt} F_i = \sum_{\alpha=1}^{N} \sum_{j_\alpha=1}^{64} [W_{i_\alpha=i, j_\alpha} p(j_\alpha) - W_{j_\alpha, i_\alpha=i} p(i_\alpha = i)] + \sum_{\beta=1}^{N} \sum_{\substack{\alpha=1 \\ \alpha \neq \beta}}^{N} \sum_{i_\alpha=1}^{64} \{[E(i_1, \ldots, i_{\beta-1}, [i_\beta = i], i_{\beta+1}, \ldots, i_N) - \bar{E}] \times P(i_1, \ldots, i_{\beta-1}, [i_\beta = i], i_{\beta+1}, \ldots, i_N)\}. \tag{7}$$

Considering the summation in Eq. (7) as an averaging procedure, the statistical mean of a product can be replaced in the lowest approximation by the product of the single means. Thus we obtain

$$\frac{d}{dt} F_i = \sum_{j=1}^{64} (Q_{ij} F_j - Q_{ji} F_i) + E_i F_i, \tag{8}$$

where

$$Q_{ij} = \sum_{\alpha=1}^{N} W_{i_\alpha=i, j_\alpha=j} \tag{9}$$

is the average probability of a codon replacement $C_j \rightarrow C_i$ occurring. The quantity

$$E_i = \sum_{\beta=1}^{N} \sum_{\substack{\alpha=1 \\ \alpha \neq \beta}}^{N} \sum_{i_\alpha=1}^{64} [E(i_1, \ldots, [i_\beta = i], \ldots, i_N) - \bar{E}] \tag{10}$$

Table 1. Numbering of codons and amino acids

Codon C_i	Codon number i	Amino acid A_k	Amino acid number k
GCA	25	alanine	1
GCG	26		
GCC	27		
GCU	28		
CGA	37	arginine	2
CGG	38		
CGC	39		
CGU	40		
AGA	5		
AGG	6		
AAC	3	asparagine	3
AAU	4		
GAC	19	aspartic acid	4
GAU	20		
UGC	55	cysteine	5
UGU	56		
GAA	17	glutamic acid	6
GAG	18		
CAA	33	glutamine	7
CAG	34		
GGA	21	glycine	8
GGG	22		
GGC	23		
GGU	24		
CAC	35	histidine	9
CAU	36		
AUA	13	isoleucine	10
AUC	15		
AUU	16		
CUA	45	leucine	11
CUG	46		
CUC	47		
CUU	48		
UUA	61		
UUG	62		
AAA	1	lysine	12
AAG	2		
AUG	14	methionine	13
UUC	63	phenylalanine	14
UUU	64		
CCA	41	proline	15
CCG	42		
CCC	43		
CCU	44		
AGC	7	serine	16
AGU	8		
UCA	57		
UCG	58		
UCC	59		
UCU	60		
ACA	9	threonine	17
ACG	10		
ACC	11		
ACU	12		
UGG	54	tryptophan	18
UAC	51	tyrosine	19
UAU	52		
GUA	29	valine	20
GUG	30		
GUC	31		
GUU	32		
UAA	49	terminator	21
UAG	50		
UGA	53		

can be regarded as the average selectional value of codon C_i. From the normalization condition (1) and definition (6) it follows that

$$\sum_{i=1}^{64} F_i = 1 \tag{11}$$

and

$$\sum_{i=1}^{64} E_i F_i = 0 . \tag{12}$$

The 20 amino acids commonly occurring in proteins we designate A_k (cf. Table 1). The relationship between amino acids and codons, i.e., the genetic code, can be expressed as a mapping L relating each codon C_i uniquely to one amino acid A_k, $k = L(i)$, whereas commonly n_k different synonymous codons C_i are assigned to one amino acid A_k, $i \in L^{-1}(k)$.

The frequency f_k of amino acid A_k is given by

$$f_k = \sum_{i \in L^{-1}(k)} F_i . \tag{13}$$

Summing up Eq. (8) according to Eq. (13) gives the following kinetic equation governing the time-dependent variation of f_k:

$$\frac{d}{dt} f_k = \sum_{l=1}^{21} \{M_{kl} f_l - M_{lk} f_k\} + e_k f_k . \tag{14}$$

Here the element

$$M_{kl} = \sum_{i \in L^{-1}(k)} \sum_{j \in L^{-1}(l)} Q_{ij} h_j \tag{15}$$

of the "mutation matrix" represents the probability of an amino acid replacement $A_l \to A_k$ caused by a point mutation occurring per time unit. In Eq. (15) the "codon usage"

$$h_i = \frac{F_i}{f_k} [k = L(i)] \tag{16}$$

denotes the relative probability that codon C_i is used to code for amino acid A_k.

The quantity

$$e_k = \sum_{i \in L^{-1}(k)} h_i E_i \tag{17}$$

can be interpreted as the average selectional value of amino acid A_k. To characterize the mean selective pressure affecting the amino acid frequencies in proteins, we will use the quantity

$$\langle e \rangle = \sqrt{\sum_{k=1}^{20} e_k^2 f_k} . \tag{18}$$

Table 2. Elements of the mutation matrix[a]

	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu
Ala		0	0	1	0	1	0	α	0	0	0
Arg	0		0	0	β	0	β	[21 + 22 + 23 + 24]α + [21 + 22]$\alpha\beta$	β	[13]	[45 + 46 + 47 + 48]
Asn	0	0		$\alpha\beta$	0	0	0	0	α	[15 + 16]	0
Asp	[27 + 28]α	0	β		0	2[17] + 2α[18]	0	[23 + 24]$\alpha\beta$	α	0	0
Cys	0	[39 + 40]$\alpha\beta$	0	0		0	0	[23 + 24]α	0	0	0
Glu	[25 + 26]α	0	0	2α[19] + 2[20]	0		α	[21 + 22]$\alpha\beta$	0	0	0
Gln	0	[37 + 38]$\alpha\beta$	0	0	0	α		0	2α[35] + 2[36]	0	[45 + 46]
Gly	α	[37 + 38 + 39 + 40]α + [15 + 16]β	0	β	1	β	0		0	0	0
His	0	[39 + 40]$\alpha\beta$	1	α	0	0	2[33] + 2α[34]	0		0	[47 + 48]
Ile	0	[5]α	1	0	0	0	0	0	0		[45 + 47 + 48]α + [61]
Leu	0	[37 + 38 + 39 + 40]	0	0	0	0	1	0	1	2[13 + 15 + 16]	
Lys	0	$\alpha\beta$[5 + 6]	2α[3] + 2[4]	0	0	$\alpha\beta$	α	0	0	[13]	0
Met	0	[6]α	0	0	0	0	0	0	0	β[13] + α[15] + [16]	α[46] + [62]
Phe	0	0	0	0	α	0	0	0	0	[15 + 16]	[47 + 48]$\alpha\beta$ + 2[61] + 2α[62]
Pro	α	[37 + 38 + 39 + 40]α	0	0	0	0	1	0	1	0	[45 + 46 + 47 + 48]β
Ser	α	2[5] + 2α[6] + [39 + 40]α	β	0	[55 + 56] + α[55 + 56]	0	0	[23 + 24]$\alpha\beta$	0	[15 + 16]	[61 + 62]β
Thr	$\alpha\beta$	[5 + 6]α	1	0	0	0	0	0	0	β	0
Trp	0	$\alpha\beta$[38] + [6]	0	0	α[55] + [56]	0	0	[22]α	0	0	[62]
Tyr	0	0	1	α	$\alpha\beta$	0	0	0	$\alpha\beta$	0	0
Val	$\alpha\beta$	0	0	1	0	1	0	α	0	β	[45 + 46 + 47 + 48]α + [61 + 62]
Ter	0	$\alpha\beta$[37] + [5]	0	0	α[55] + [56]	α	$\alpha\beta$	α[21]	0	0	2[61] + [62]

[a] The numbers in brackets refer to the codon usage h_i; for example, [23 + 24] = $h_{23} + h_{24} = h_{[G,U,C]} + h_{[G,G,U]}$. The diagonal elements M_{kk} of the mutation matrix (probabilities of self-replacement, i.e., silent mutations) are given in the last column. The factor $2m/[(1 + \alpha)(2 + \beta)]$ that should appear in each element according to (25) has been omitted in this table. For example, the probability of the replacement threonine → methionine occurring is given by

$$M_{13,14} = \frac{2m}{(1 + \alpha)(2 + \beta)}[10]\alpha\beta$$

Table 2. Extended

Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Ter	
0	0	0	α	[57 + 58 + 59 + 60]	β	0	0	β	0	$(\beta + 2)[25 + 28] + \alpha(\beta + 2)[26 + 27]$
β	1	0	α	$(1 + 2\alpha)[7] + 3[8]$	$\alpha[9 + 10]$	$1 + \beta$	0	0	$\beta[53] + [53]$	$(\beta + 1)[5] + (\beta + 2 + \alpha)[37] + (\beta + 3)[38] (\beta + 2)(\alpha[39] + [40])$
$2[1] + 2\alpha[2]$	0	0	0	$\alpha\beta[7 + 8]$	$\alpha[11 + 12]$	0	1	0	0	$\beta[4] + \alpha\beta[3]$
0	0	0	0	0	0	0	1	[31 + 32]	0	$\beta[20] + \alpha\beta[19]$
0	0	1	0	[7 + 8] + α[59 + 60]	0	2α	β	0	$2\beta[53]$	$\beta[56] + \alpha\beta[55]$
β	0	0	0	0	0	0	0	[29 + 30]	β[49 + 50]	$\alpha\beta[18] + \beta[17]$
1	0	0	α[41 + 42]	0	0	0	0	0	2[49]	$\alpha\beta[34] + \beta[33]$
0	0	0	0	$[7 + 8]\beta$	0	1	0	1	$\beta[53]$	$(\beta + 2)[21 + 24] + \alpha(\beta + 2)[22 + 23]$
0	0	0	α[43 + 44]	0	0	0	β	0	0	$\beta[36] + \alpha\beta[35]$
[1]	$\alpha(2 + \beta)$	1	0	$[7 + 8]\alpha$	$\alpha\beta$[9 + 11 + 12]	0	0	$\alpha\beta$[29 +31 + 32]	0	$\alpha(\beta + 1)[15] + (\beta + 1)[16] + 2[13]$
0	2	$(\beta + 2\alpha)[63] + (\beta + 2)[64]$	$\alpha\beta$	$\alpha\beta[57 + 58]$	0	α	0	$2\alpha[29 + 30] + \alpha[31 + 32]$	β[49] + β[50] + $\alpha\beta[53]$	$2(\beta + 1)([45] + \alpha[46]) + (\beta + 2)([48] + \alpha[47]) + 2\beta[61] + \beta(\alpha + 1)[62]$
	1	0	0	0	$\alpha[9 + 10]$	0	0	0	β[49 + 50]	$\alpha\beta[2] + \beta[1]$
[2]		0	0	0	$[10]\alpha\beta$	0	0	$[30]\alpha\beta$	0	0
0	0		0	$\alpha\beta$[59 + 60]	0	0	1	$\alpha\beta$[31 + 32]	0	$\beta[64] + \alpha\beta[63]$
0	0	0		β[57 + 58 + 59 + 60]	1	0	0	0	0	$(\beta + 2)[41 + 44] + \alpha(\beta + 2)[42 + 43]$
0	0	β	$\alpha\beta$		$(1 + \alpha)[11 + 12] + [9 + 10]$	α	1	0	β[49 + 50] + $\alpha\beta[53]$	$\alpha\beta[7] + \beta[8] + (\beta + 2)[57 + 60] + \alpha(\beta + 2)[58 + 59]$
1	β	0	α	α[7 + 8] + [57 + 58 + 59 + 60]		0	0	0	0	$(\beta + 2)[9 + 12] + \alpha(\beta + 2)[10 + 11]$
0	0	0	0	$[58]\alpha$	0		0	0	[50 + 53]	0
0	0	1	0	α[59 + 60]	0	0		0	2([49] + [50])β	$\beta[52] + \alpha\beta[51]$
0	β	1	0	0	0	0	0		0	$(\beta + 2)[29 + 32] + \alpha(\beta + 2)[30 + 31]$
1	0	0	0	$2\alpha[57] + \alpha[58]$	0	$2\alpha\beta$	$2\alpha[51] + 2[52]$	0		1

The codon replacement probability matrix Q_{ij} has the form

$$Q_{ij} = \begin{cases} T_{BB'} \begin{cases} \text{if } C_i = [B, B_2, B_3] \text{ and } C_j = [B', B_2, B_3] \\ \text{if } C_i = [B_1, B, B_3] \text{ and } C_j = [B_1, B', B_3] \\ \text{if } C_i = [B_1, B_2, B] \text{ and } C_j = [B_1, B_2, B'] \end{cases} \\ 0 \quad \text{otherwise,} \end{cases} \tag{19}$$

where $T_{BB'}$ denotes the probability of occurrence of the base replacement $B' \rightarrow B$, which obviously has to fulfill the normalization condition

$$\sum_{\substack{B \\ B \neq B'}} T_{BB'} = \alpha_{B'}, \tag{20}$$

α_B being the probability of base B being affected by a mutational event.

In principle, each of the four nucleotides B may have its own mutation probability α_B, but owing to the double-helical structure of DNA any base change is accompanied by a corresponding change of the complementary base. Hence we shall discriminate between the two probabilities α_{GC} and α_{AU}, which correspond to replacements of GC and AU base pairs, respectively. Furthermore, it seems reasonable to distinguish between transitions ($A \leftrightarrow G$, $C \leftrightarrow U$) and transversions ($A \leftrightarrow C$, $A \leftrightarrow U$, $G \leftrightarrow C$, $G \leftrightarrow U$), and so we assign them probabilities β_{TI} and β_{TV}, respectively, where

$$\beta_{TI} + 2\beta_{TV} = 1. \tag{21}$$

With these limitations, it holds that

$$T_{BB'} = \begin{array}{c} \begin{matrix} \quad A & \quad G & \quad C & \quad U \end{matrix} \\ \begin{bmatrix} 0 & \alpha_{GC}\beta_{TI} & \alpha_{GC}\beta_{TV} & \alpha_{AU}\beta_{TV} \\ \alpha_{AU}\beta_{TI} & 0 & \alpha_{GC}\beta_{TV} & \alpha_{AU}\beta_{TV} \\ \alpha_{AU}\beta_{TV} & \alpha_{GC}\beta_{TV} & 0 & \alpha_{AU}\beta_{TI} \\ \alpha_{AU}\beta_{TV} & \alpha_{GC}\beta_{TV} & \alpha_{GC}\beta_{TI} & 0 \end{bmatrix} \begin{matrix} A \\ G \\ C \\ U \end{matrix} \end{array}.$$

For example, the probability of the replacement $G \rightarrow A$ occurring is given by $T_{AG} = \alpha_{GC}\beta_{TI}$.

Defining the average base exchange probability m as

$$m = \frac{\alpha_{GC} + \alpha_{AU}}{2} \tag{22}$$

and the ratios

$$\alpha = \alpha_{GC}/\alpha_{AU} \tag{23}$$

and

$$\beta = \beta_{TI}/\beta_{TV}, \tag{24}$$

we can put the matrix $T_{BB'}$ into the following form:

$$T_{BB'} = \frac{2m}{(1+\alpha)(2+\beta)} \begin{bmatrix} 0 & \beta & \alpha & 1 \\ \beta & 0 & \alpha & 1 \\ 1 & \alpha & 0 & \beta \\ 1 & \alpha & \beta & 0 \end{bmatrix}. \tag{25}$$

From Eqs. (19) and (25) the mutation matrix M_{kl} has the general structure shown in Table 2.

Throughout this paper we will assume steady-state conditions, i.e.,

$$\sum_{j=1}^{64} \{Q_{ij}\hat{F}_j - Q_{ji}\hat{F}_i\} + E_i\hat{F}_i = 0, \tag{26.1}$$

$$\sum_{l=1}^{21} \{M_{kl}\hat{f}_l - M_{lk}\hat{f}_k\} + e_k\hat{f}_k = 0. \tag{26.2}$$

These equations can be solved numerically by an iterative procedure,

$$\hat{F}_i^{(\mu+1)} = \sum_{j=1}^{64} Q_{ij}\hat{F}_j^{(\mu)} \Big/ \left(\sum_{j=1}^{64} Q_{ji} - E_i\right), \tag{27}$$

$$\lim_{\mu \to \infty} \hat{F}_i^{(\mu)} = \hat{F}_i, \tag{28}$$

which is stopped if the results of two successive iterations (μ) and ($\mu + 1$) are nearly identical, i.e., if $|\hat{F}_i^{(\mu+1)} - \hat{F}_i^{(\mu)}| < \epsilon$ for all $i = 1, 2, \ldots, 64$ and a sufficiently small ϵ.

On the other hand, the selectional values E_i can be calculated from Eq. (26.1) if the codon replacement matrix Q_{ij} and the steady-state codon frequencies $\hat{F}_i$ are given:

$$E_i = \frac{1}{\hat{F}_i} \sum_{j=1}^{64} \{Q_{ji}\hat{F}_i - Q_{ij}\hat{F}_j\}. \tag{29}$$

For the particular case that no selectional constraints are present, so that the steady-state frequencies are the result of point mutations only ($E_i = 0$, $i = 1, 2, \ldots, 64$), Eq. (26.1) can be solved analytically employing the principle of detailed balance (Cercignany 1975):

$$Q_{ij}F_j = Q_{ji}F_i \; (i, j = 1, 2, \ldots, 64). \tag{30}$$

Specifying Q_{ij} according to Eqs. (19) and (25), relation (30) implies

$$\hat{A}/\hat{G} = \hat{U}/\hat{C} = \alpha, \tag{31.1}$$

$$\hat{A} = \hat{U}, \tag{31.2}$$

$$\hat{G} = C. \tag{31.3}$$

Together with the normalization condition $\hat{A} + \hat{G} + \hat{C} + \hat{U} = 1$, relations (31.1)–(31.3) immediately yield

$$\hat{A} = \hat{U} = \frac{\alpha}{2(1+\alpha)}, \tag{32.1}$$

$$\hat{G} = \hat{C} = \frac{1}{2(1+\alpha)}. \tag{32.2}$$

In addition to the selectional values e_k of amino acids A_k defined by Eqs. (17) and (29) we propose an "acceptability matrix" A_{kl} as a further measure of selective constraints influencing the amino acid composition of proteins. This acceptability matrix A_{kl} gives the probability that a randomly occurring replacement $A_l \rightarrow A_k$ will be accepted by selection. This can be expressed by writing the so-called transition probability matrix of accepted point mutations N_{kl}, which was constructed from phylogenetic trees (Yano and Hasegawa 1974), as a product

$$N_{kl} = M_{kl} \cdot A_{kl}. \tag{33}$$

According to Eq. (33), the probability of a replacement being accepted (observed) is determined by the probability this replacement occurring by chance as well as by the probability that this replacement is accepted by selection; these prerequisites are regarded as independent events.

To relate the mutation matrix M_{kl} to the matrix N_{kl} of accepted amino acid replacements, both matrices have to be referred to the same time scale; i.e., the mutation probability m determining the order of magnitude of the mutation matrix [cf. Eq. (22)] has to be adjusted to the evolutionary time interval of 2 PAMs (PAM = accepted point mutations per amino acid) usually chosen for computation of the elements of N_{kl} (Ohta and Kimura 1971; Dayhoff et al. 1972). Therefore we set

$$A_{kl} = \gamma \frac{N_{kl}}{M_{kl}}, \tag{34}$$

where γ serves as an adjustable scaling factor. From the acceptability matrix A_{kl} one can derive an "acceptability value" a_k of amino acid A_k by averaging the acceptability matrix over all replacements $A_l \rightarrow A_k$ that can result from single point mutations (i.e., $M_{kl} = 0$):

$$a_k = \sum_l A_{kl}\delta_{kl} \Big/ \sum_l \delta_{kl},$$

$$\delta_{kl} = \begin{cases} 1 \text{ if } M_{kl} \neq 0, \\ 0 \text{ if } M_{kl} = 0. \end{cases} \quad (35)$$

Results

Properties of the Mutation Matrix M_{kl}

The mutation matrix M_{kl} (Table 2) contains 188 nonzero elements if mutations via the terminator are included; 108 of these elements depend on the codon usage h_i. Among the 108 h_i-dependent elements are 10 that refer to terminator → amino acid replacements. Since such replacements are very unlikely to be observed, these ten elements will not be considered in the following. As an illustration of such dependency on codon usage the possible replacements of isoleucine by other amino acids are depicted in Fig. 1. For instance, the replacement of isoleucine by arginine or lysine due to one point mutation is only possible if the codon [AUA] is used. From the analysis of mRNA sequences it is well known that codons are not used randomly. It has been suggested that this nonrandomness might reflect metabolic discrimination between bases, codon–anticodon interaction energies (Grosjean et al. 1978; Pieczenik 1980), regulation of replication or transcription through degenerate base usage (Grantham et al. 1980, 1981; Conrad et al. 1983), harmonization of codon and anticodon populations in each cell type (Kafatos et al. 1977; Berger 1978; Grantham et al. 1980, 1981; Holland and Holland 1980; Ikemura 1981), mRNA secondary structure optimization and regulation (Fitch 1980; Wain-Hobson et al. 1981), and reduction in number of mutations with drastic effects (Fitch 1980; Holmquist and Pearl 1980; Modiani et al. 1981; Golding and Strobeck 1982). With reference to this last aspect it is interesting to point out the consequences of nonrandom condon usage with respect to the h_i-dependent elements of the mutation matrix. For that purpose we calculated the elements of the mutation matrix, taking the codon usage h_i from 162 different mRNA sequences published by Grantham et al. (1980, 1981).

Unfortunately, there is no convincing information about the actual values of the parameters α and β of the mutation matrix. Vogel and Kopun (1977) analyzed more than a thousand amino acid replacements in various phylogenetic trees and found a higher rate of transitions than of transversions, i.e., that $\beta > 1$. This observation is in agreement with other findings (Vogel and Röhrborn 1966; Fitch 1967; Jukes 1975; Wolkenstein 1979; Holmquist and Pearl 1980; Modiani et al. 1981; Brown 1982).

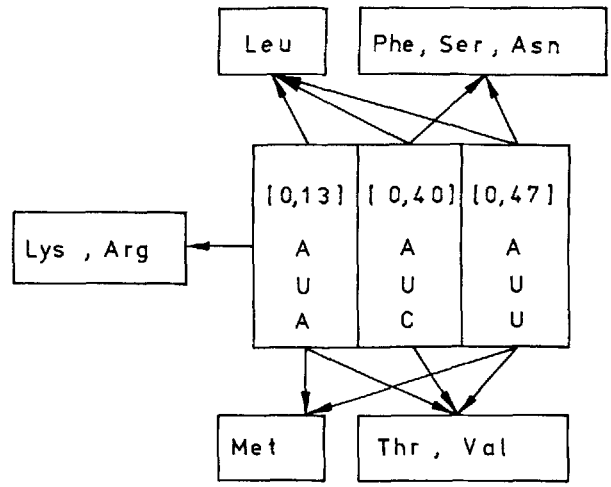


Fig. 1. Possible replacements of isoleucine by other amino acids due to single point mutations. The values in parentheses refer to the observed average codon usage (cf. Table 4)

It should be emphasized that the method used in these papers does not allow a determination to be made as to whether the observed ratio between transitions and transversions is the result of mutation or of selection, since the observed replacements represent the outcome of both processes.

Direct studies of mutation (Coulondre and Miller 1977; Fersht 1979; Fersht and Knill-Jones 1981) reveal that transitions predominate (β = 2–5) but that in about half of all cases this predominance is due to the G–C pairs of hotspots.

Fersht and Knill-Jones (1981) investigated the mutation rates of transitions and transversions starting from the bases A or T. They found that transversions arise from purine–purine mismatch, but not from pyrimidine–pyrimidine exchanges. Their study yielded for the base replacements T → A, G, C and G → A almost the same absolute value of about 4×10^{-7}; only the rate for the transversion A → C was considerably smaller (5×10^{-9}). Similar results have been obtained from model-building studies (Topal and Fresco 1976). The mutation rates can also be influenced by mutator genes (Cox 1976; Lehmann and Karran 1981; Weinberg et al. 1981) and hence by the concentrations of the nucleoside triphosphates (Fersht and Knill-Jones 1981). Furthermore, in most translation systems misincorporations of amino acids show a pattern that would also occur as a result of replication errors with a rate of transitions exceeding 3 to 5 times the rate of transversions (Davies et al. 1966; Yarus 1979; Ozoline et al. 1980). Therefore we will assume $\beta = 3$, bearing in mind all the uncertainties concerning this value. As shown in the next section, a reasonable α value can be estimated from the known average base composition of DNA, yielding $\alpha = 1.15$.

To differentiate between the influence of codon

usage and the two mutation parameters α and β on the elements of the mutation matrix, we will distinguish among the following three cases:

(1) Case I (reference case):
$\alpha = \beta = 1$; $h_i = 1/n_k$ [$k = L(i)$].
(2) Case II (semirealistic case):
$\alpha = \beta = 1$; h_i = observed codon usage.
(3) Case III (realistic case):
$\alpha = 1.15$, $\beta = 3$; h_i = observed codon usage.

To compare the elements of the mutation matrices evaluated for case I (M_{kl}^0) and for case III with respect to the observed codon usage in the p-th mRNA ($p = 1, 2, \ldots, 162$) (matrix elements designated by $M_{kl}^{(p)}$), we introduce the relative replacement probability

$$m_{kl}^{(p)} = 100 \frac{M_{kl}^{(p)}}{M_{kl}^0}. \tag{36}$$

The 98 × 162 numerical values of this relative replacement probability (corresponding to the 98 h_i-dependent elements of the mutation matrix—excluding terminator → amino acid replacements—evaluated with the codon usages observed in 162 different mRNAs) vary between 0 (if all codons that have to be occupied to make possible the replacement $A_l \to A_k$ are avoided) and the upper bound of 450. Dividing the total range [0,450] into 30 subintervals $\Delta_i = [15(i-1),15i]$ ($i = 1, 2, \ldots, 30$) of length 15, we define the distribution function $\phi_{kl}(i)$ as the portion of all 162 $m_{kl}^{(p)}$ values situated in the i-th interval:

$$\phi_{kl}(i) = \sum_{p=1}^{162} f_l^{(p)} \varphi_i^{(p)} \Big/ \sum_{p=1}^{162} f_l^{(p)},$$

$$\varphi_i^{(p)} = \begin{cases} 1 \text{ if } m_{kl}^{(p)} \in \Delta_i, \\ 0 \text{ otherwise.} \end{cases} \tag{37}$$

In Eq. (37) $f_l^{(p)}$ denotes the frequency of the initial amino acid A_l (which is replaced) in the protein assigned to the p-th mRNA.

The 98 distribution functions can be arranged empirically into the seven groups A–G shown in Fig. 2A–G. The peculiarities of the distribution functions lumped in each group are outlined in the legend of the figure. Table 3 lists the group labels (A–G) of the 98 h_i-dependent replacements and the mean values of the relative replacement probability

$$m_{kl} = \frac{1}{30} \sum_{i=1}^{30} \phi_{kl}(i) \tag{38}$$

calculated for cases II and III. Comparing the mean values obtained for cases II and III it can be seen that for almost all of the 98 replacements, nonrandom codon usage and biased mutation ($\alpha, \beta \neq 1$) shift the distribution function in the same direction. For example, with respect to extremely suppressed replacements leading to tryptophan, arginine, glutamate, methionine, and the termination codons, the leftward shift of the distribution caused by codon usage alone is additionally enhanced by biased mutations ($\alpha, \beta \neq 1$).

It is noteworthy that most of the suppressed amino acid replacements (group A, Fig. 2A) are known to give rise quite often to drastic changes within the protein structure. From this finding it can be concluded that point mutations play a "preselectional" role by lowering the probability of undesired amino acid replacements. Of course, only a rather crude preselection can be achieved in this way, since the variation of α and β simultaneously affects many elements of the mutation matrix and the avoidance of certain codons leads to the preference of others according to the normalization condition.

$$\sum_{i \in L^{-1}(k)} h_i = 1 \qquad (k = 1, 2, \ldots, 21), \tag{39}$$

which immediately follows from the definition (16). For example, a lowering of the probability of mutation from isoleucine to arginine or lysine by avoidance of the codon AUA inevitably increases the probability of mutation to asparagine, phenylalanine, threonine, serine, or valine (cf. Fig. 1). Arginine and lysine differ considerably from isoleucine in many physicochemical properties, including polarity (Woese 1973; Grantham 1974; Jones 1975), hydrophobicity (Aboderin 1971; Fendler et al. 1975; Sternberg and Thornten 1977; Manavalan and Ponnuswamy 1978; Olsen 1980), bulkiness (Jones 1975), and charge. Serine, threonine, and asparagine, on the other hand, are also slightly polar amino acids. Nevertheless, the replacement of isoleucine by arginine and lysine seems to be more dangerous to the protein structure than do the other possible replacements.

For a more systematic study it is convenient to derive from the (asymmetric) mutation matrix M_{kl} a "genetic distance matrix" D_{kl}. It seems reasonable to assume that the genetic distance D_{kl} between two amino acids A_k and A_l is small (large) if the probability of mutual replacement (expressed as the product of the unidirectional replacement probabilities) is large (small); i.e.,

$$D_{kl} = \begin{cases} \dfrac{D_0}{M_{kl} M_{lk}} & \text{if } M_{kl} \cdot M_{lk} \neq 0 \\ \infty & \text{if } M_{kl} \cdot M_{lk} = 0 \end{cases},$$

$$D_0 = \underset{k,l=1,\ldots,20}{\text{MAX}} (M_{kl} \cdot M_{lk}). \tag{40}$$

From the symmetric distance matrix D_{kl} one can derive a grouping of the amino acids with increasing distance. A cluster analysis performed by means of the unweighted pair-group (UPG) method (Li 1981)

yields the dendrogram shown in Fig. 3. It can be seen that the amino acids group into two main clusters representing the hydrophobic residues (with the exception of serine) and the polar ones.

Influence of Point Mutation on the Frequencies of Codons and Amino Acids

If there are no selectional constraints [$E_i = 0$ in Eq. (26.1), $i = 1, 2, \ldots, 64$], the steady-state nucleotide frequencies are given by Eqs. (32.1) and (32.2) and the frequencies of codons and amino acids can be calculated in a straightforward manner. The dependency of the expected frequencies $\hat{h}_i$ and $\hat{f}_k$ on the parameter α is shown in Table 4. This table also lists the observed frequencies $\langle h_i \rangle$ and $\langle f_k \rangle$ obtained by averaging over the 162 mRNA sequences published by Grantham et al. (1980, 1981).

Using as a measure of the distance between the expected composition $\hat{\mathcal{F}}_i$ and the observed composition $\langle \mathcal{F}_i \rangle$ ($i = 1, 2, \ldots, n$) the quantity (Laird and Holmquist 1975)

$$d(\hat{\mathcal{F}}, \langle \mathcal{F} \rangle) = \sum_{i=1}^{n} \frac{1}{\hat{\mathfrak{F}}_i} (\hat{\mathfrak{F}}_i - \langle \mathfrak{F}_i \rangle)^2, \qquad (41)$$

a minimum distance between expected and observed amino acid composition is obtained with $\alpha = 1.15$ (see Table 5). The most significant discrepancies are those for alanine, arginine, lysine, and serine.

For $\alpha = 1.15$ the expected average base frequencies according to Eqs. (32.1) and (32.2) are $\hat{A} = \hat{U} = 0.269$ and $\hat{G} = \hat{C} = 0.231$, which are in a good agreement with the values $\langle A \rangle = 0.258$, $\langle U \rangle = 0.266$, $\langle G \rangle = 0.240$, and $\langle C \rangle = 0.235$ observed in the 162 coding sequences. One might object that this result is trivial because the average base frequency $\langle B \rangle$ is related to the average amino acid frequencies $\langle f_k \rangle$ according to the equation

$$\langle B \rangle = \sum_{k=1}^{21} \langle f_k \rangle \langle B \rangle_k. \qquad (42)$$

Here $\langle B \rangle_k$ denotes the average proportion of base B in the subgroup $L^{-1}(k)$ of those codons assigned to amino acid A_k, satisfying $\sum_{B=A,G,C,U} \langle B \rangle_k = 1$ ($k = 1, 2, \ldots, 21$). But $\langle B \rangle_k$ essentially depends on the codon usage; in other words, proteins having identical amino acid compositions can be coded for by DNAs having completely different base compositions. By minimizing the distance $d(\hat{h}, \langle h \rangle)$ between expected and observed codon usage, we get $\alpha = 1.13$, which is very close to the value $\alpha = 1.15$ needed to optimize the expected amino acid composition. From the finding that biased mutation with $\alpha = 1.13$–1.15 yields the best agreement between expected and observed compositions of both nucleic acids and proteins (if selectional constraints are neglected), it can be concluded that the probability of a point mutation affecting an AU pair is slightly greater than that for a GC pair.

An alternative method of calculating the steady-state amino acid composition in the absence of selectional constraints is based on Eq. (26.2) with $e_k = 0$ ($k = 1, 2, \ldots, 21$):

$$\sum_{l=1}^{20} \{M_{kl} f_l - M_{lk} f_k\} = 0. \qquad (43)$$

This equation was solved by means of the iterative procedure (27) using a mutation matrix calculated with $\alpha = 1.15$, $\beta = 3$, and the average observed codon usages $\langle h_i \rangle$ given in Table 4. The frequencies so obtained are listed in Table 5. The same calculation was carried out based on Eq. (43) using the transition probability matrix N_{kl} of accepted point mutations (Yano and Hasegawa 1974) instead of the mutation matrix M_{kl}. In the Yano–Hasegawa matrix N_{kl} the numbers for all but one exchange of cysteine with other amino acids are equal to zero. This is probably due to the small number of observations and the conservativeness of disulfide bridges (Dayhoff et al. 1972). Therefore the fifth column and row of N_{kl} (corresponding to cysteine = A_5) were excluded from the computation to prevent an unrealistically high proportion of cysteine from being calculated. As can be seen from Table 5, the amino acid frequencies calculated from Eq. (43) with the Yano–Hasegawa matrix agree better with the observed frequencies than do those obtained with the mutation matrix. This is a plausible result because the observed amino acid composition represents the outcome of both mutation and selection, which determine the matrix N_{kl}, whereas the mutation matrix reflects the mutational process alone. Nevertheless, in both cases considerable discrepancies remain between expected and observed amino acid frequencies, indicating the presence of additional selectional constraints that have hitherto been neglected.

The Selectional Values of Codons and Amino Acids

The evolution of genes and related proteins is the result of two processes, mutation and selection. As shown in the previous section, and as pointed out by several authors (Jukes et al. 1975; Holmquist 1978; Coutelle and Hofacker 1982), the observed average amino acid frequencies in proteins cannot be explained satisfactorily by the mutational process alone.

To estimate to what extent the observed frequencies of codons and amino acids are affected by additional selective forces besides mutations, we introduce the "selectional values" calculated from Eqs.

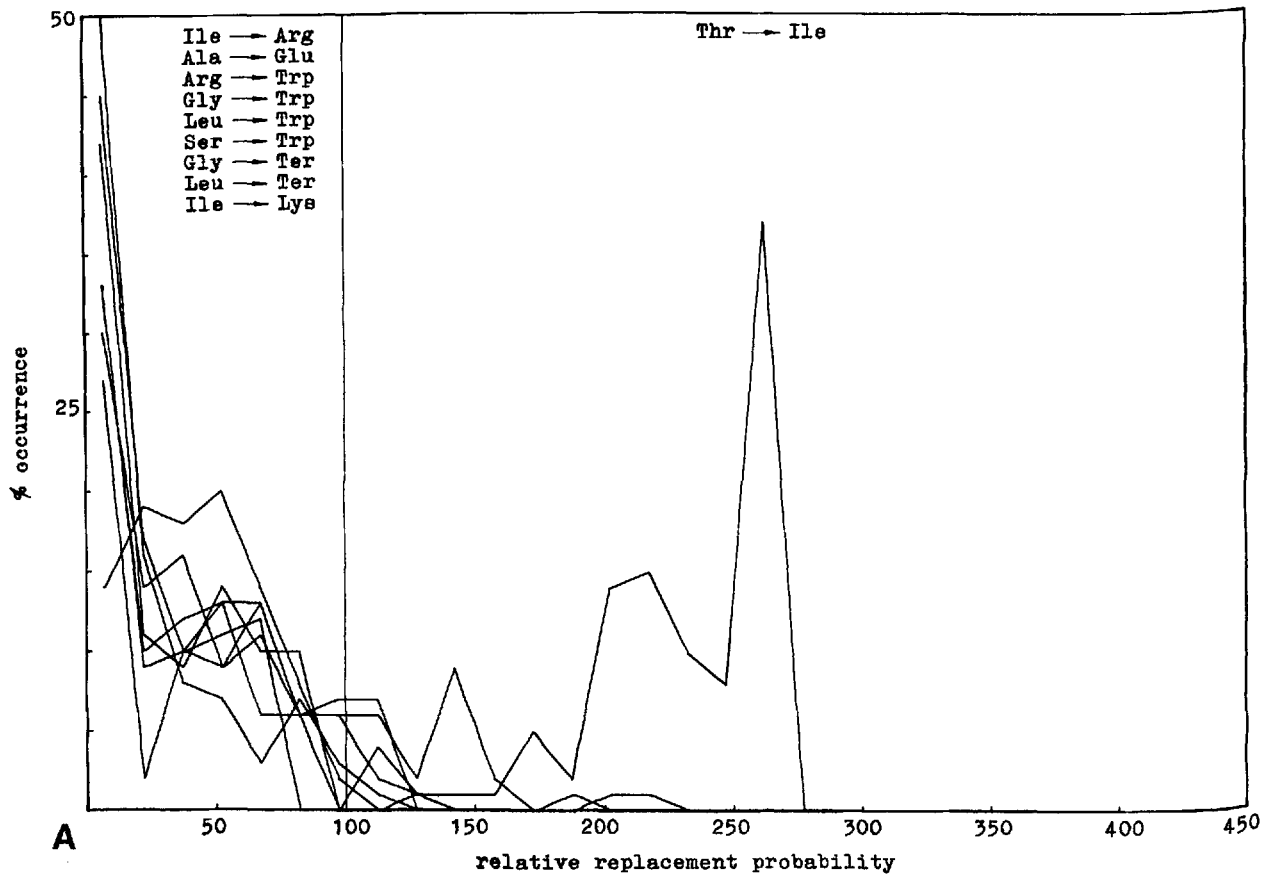
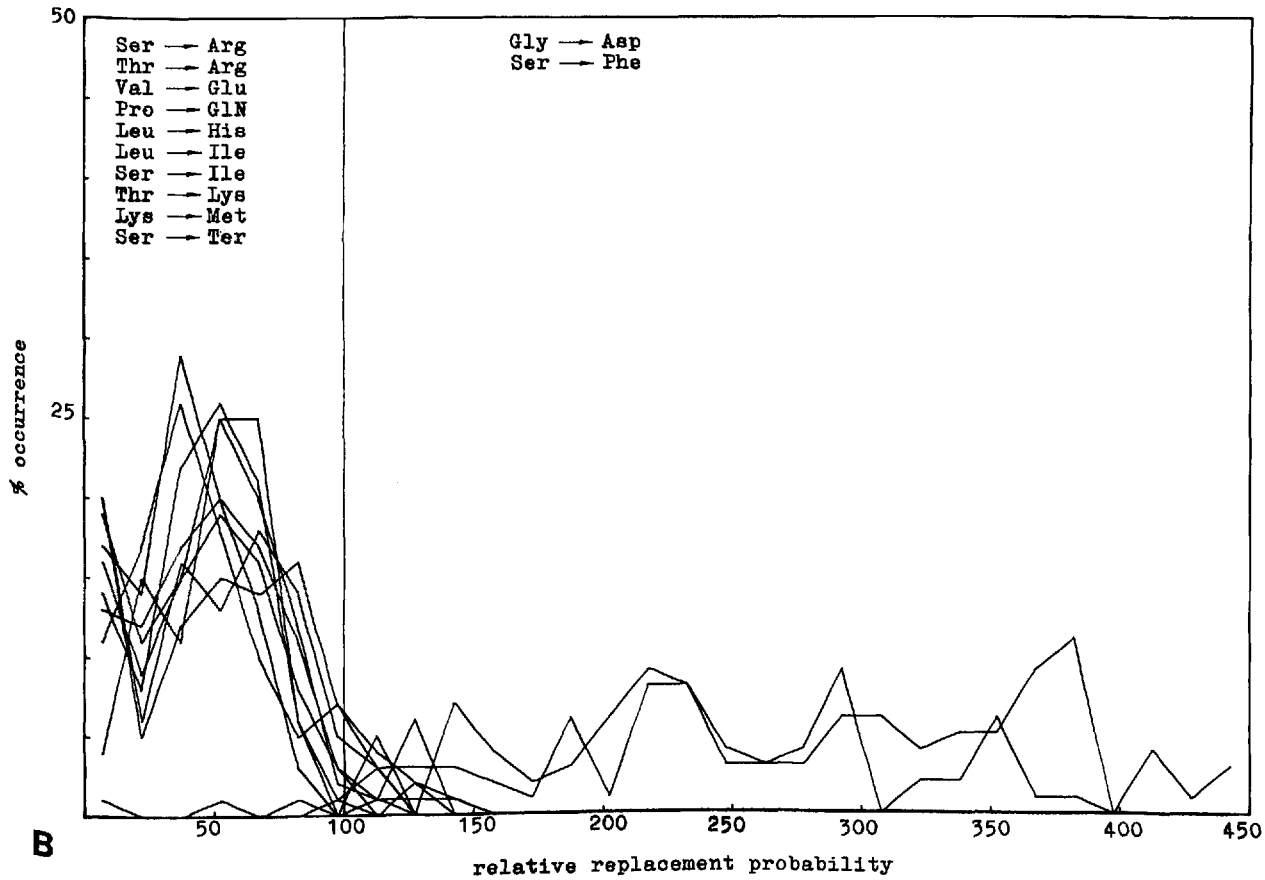


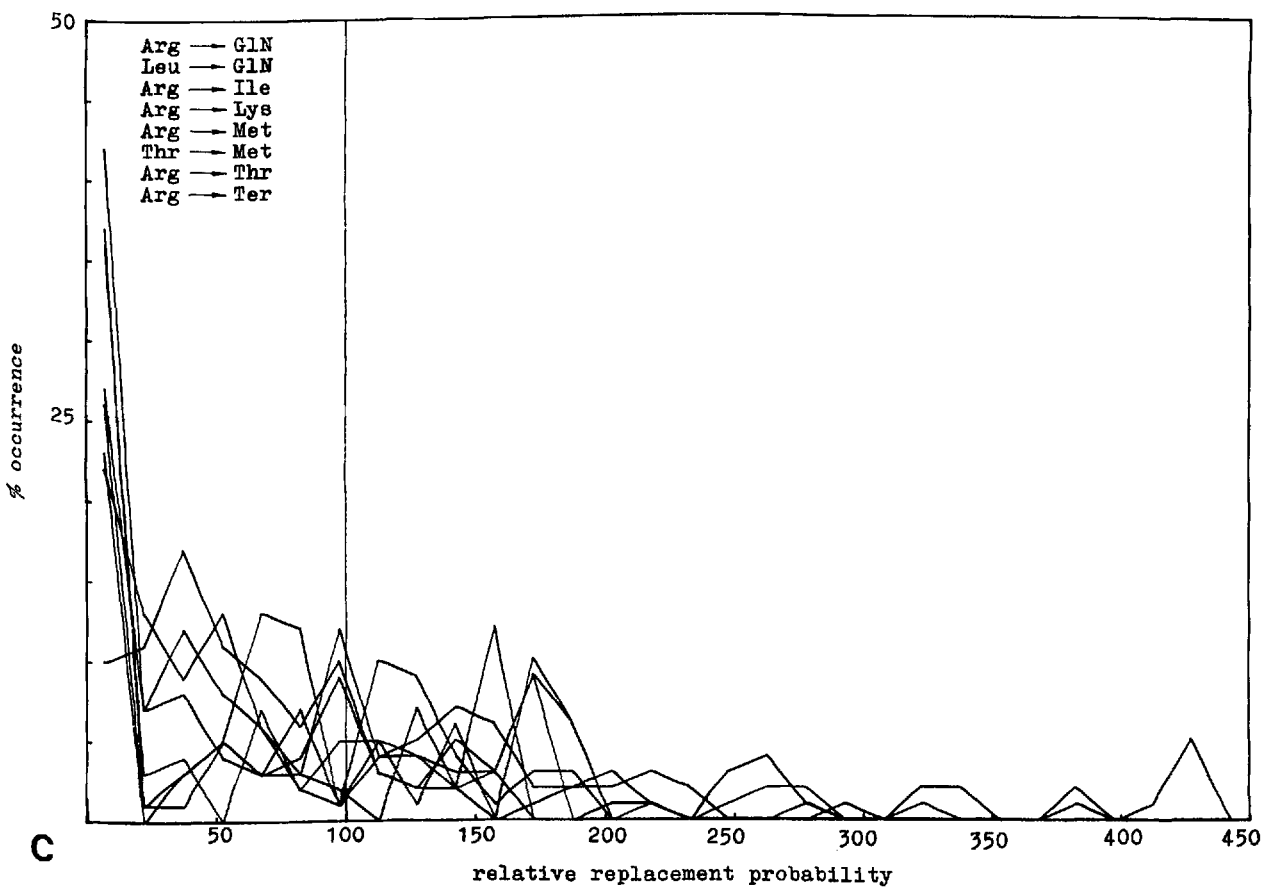
Fig. 2A-G. (B–G appear on subsequent pages.) Distribution of the relative replacement probabilities among 162 different mRNAs. The distributions were computed according to Eq. (37) for the 98 replacements that depend on codon usage (terminator → amino acid replacements excluded). The 98 distributions can be empirically arranged into seven groups according to their widths, the positions of the maxima (if there are any), and the mean values $\langle m_{kl} \rangle$ obtained for the semirealistic case II and the realistic case III. The capital letters A–G labeling the seven groups are specified for each replacement in Table 3. **A** (Group A) A considerable leftward shift of the distribution is caused by both codon usage and biased mutation parameters α and β. The maximum is located around zero; i.e., the replacement is avoided in many mRNAs. The single rightward shifted distribution corresponds to the replacement threonine → isoleucine. **B** (Group B) The distribution is shifted either leftward or rightward by both codon usage and biased mutation parameters. The leftward shifted distributions show a main maximum between 40 and 80 and a second optimum around zero; i.e., the replacement is avoided in several mRNA species. The rightward shifted distributions show no distinct maximum. **C** (Group C) The distribution is very broad and exhibits no distinct maximum. The replacement is avoided in several mRNAs. The mean values for cases II and III are both approximately equal to zero. **D** (Group D) The distribution is very broad and exhibits no distinct maximum. Codon usage gives rise to a leftward shift (mean value <100), whereas biased mutation parameters have the opposite effect (mean value >100). **E** (Group E) The distribution is shifted either leftward or rightward by biased mutation parameters only. **F** (Group F) The distribution is sharp (peaked) and the positions of the maxima are determined only by the mutation parameters. **G** (Group G) The distribution resembles a Gaussian one around 100. Codon usage often gives rise to a rightward shift (mean value >100), whereas biased mutation parameters cause a leftward shift (mean value <100)

(10) and (17). Under steady-state conditions these quantities can be calculated from Eqs. (29) and (17). Table 6 shows the selectional values E_i [$i = 1, 2, \ldots, 64$, $i \neq 49, 50, 53$ (termination codons)] and e_k ($k = 1, 2, \ldots, 20$) for the reference case I and the realistic case III considered in the previous section. We note that the "selective pressure" $\langle e \rangle$ defined by relation (18) is smaller for the realistic case III. This seems to be a further confirmation of the hypothesis of a "preselected" mutational process. The great variation among the selectional values E_i of codons assigned to a given amino acid (e.g., alanine, arginine, glutamate, leucine, proline, serine, threonine, tyrosine, and valine) can hardly be accounted for by selectional constraints present only at the phenotypic (protein) level. Hence, nonrandomness of codon usage must be due to another type of selectional constraint operative at the genotypic level (Conrad et al. 1983).

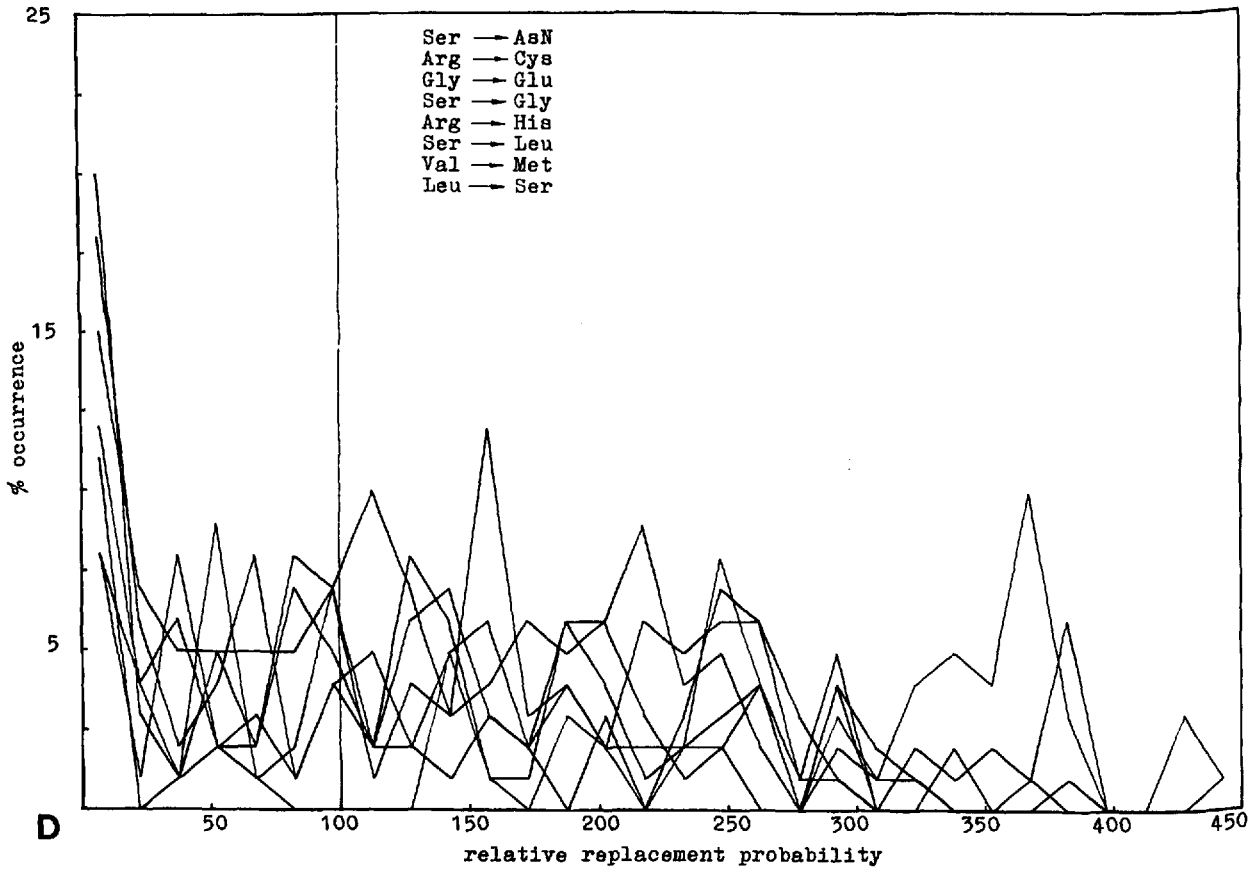
Roughly speaking, the selectional values E_i and e_k tell us whether the corresponding codons and amino acids are used on the average more or less frequently than would be expected on the basis of random point mutation. In general, however, the usefulness of a new amino acid in a protein essentially depends on the amino acid that has been re-


Ser → Arg
Thr → Arg
Val → Glu
Pro → GlN
Leu → His
Leu → Ile
Ser → Ile
Thr → Lys
Lys → Met
Ser → Ter
Gly → Asp
Ser → Phe
% occurrence
50
25
50
100
150
200
250
300
350
400
450
relative replacement probability

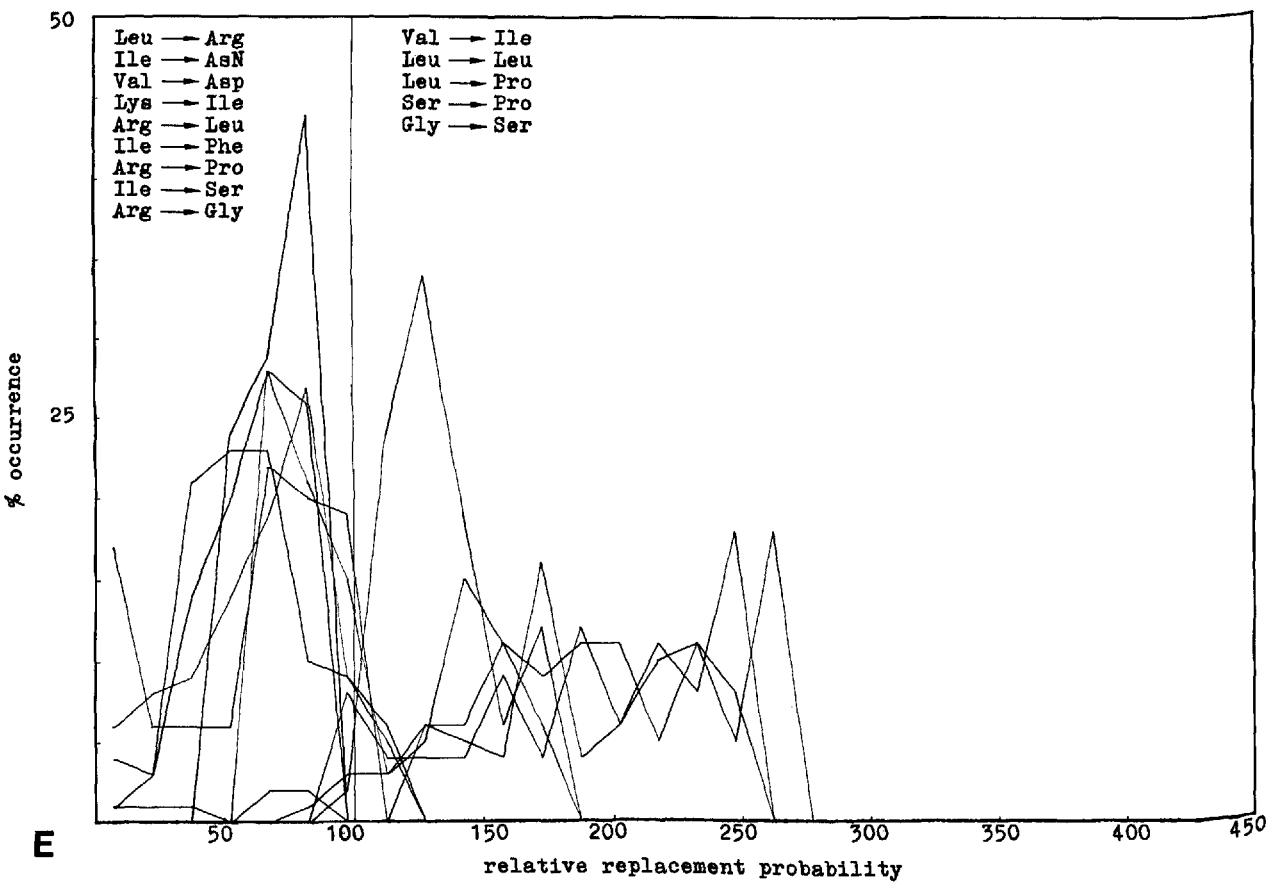
B


Arg → GlN
Leu → GlN
Arg → Ile
Arg → Lys
Arg → Met
Thr → Met
Arg → Thr
Arg → Ter
% occurrence
50
25
50
100
150
200
250
300
350
400
450
relative replacement probability

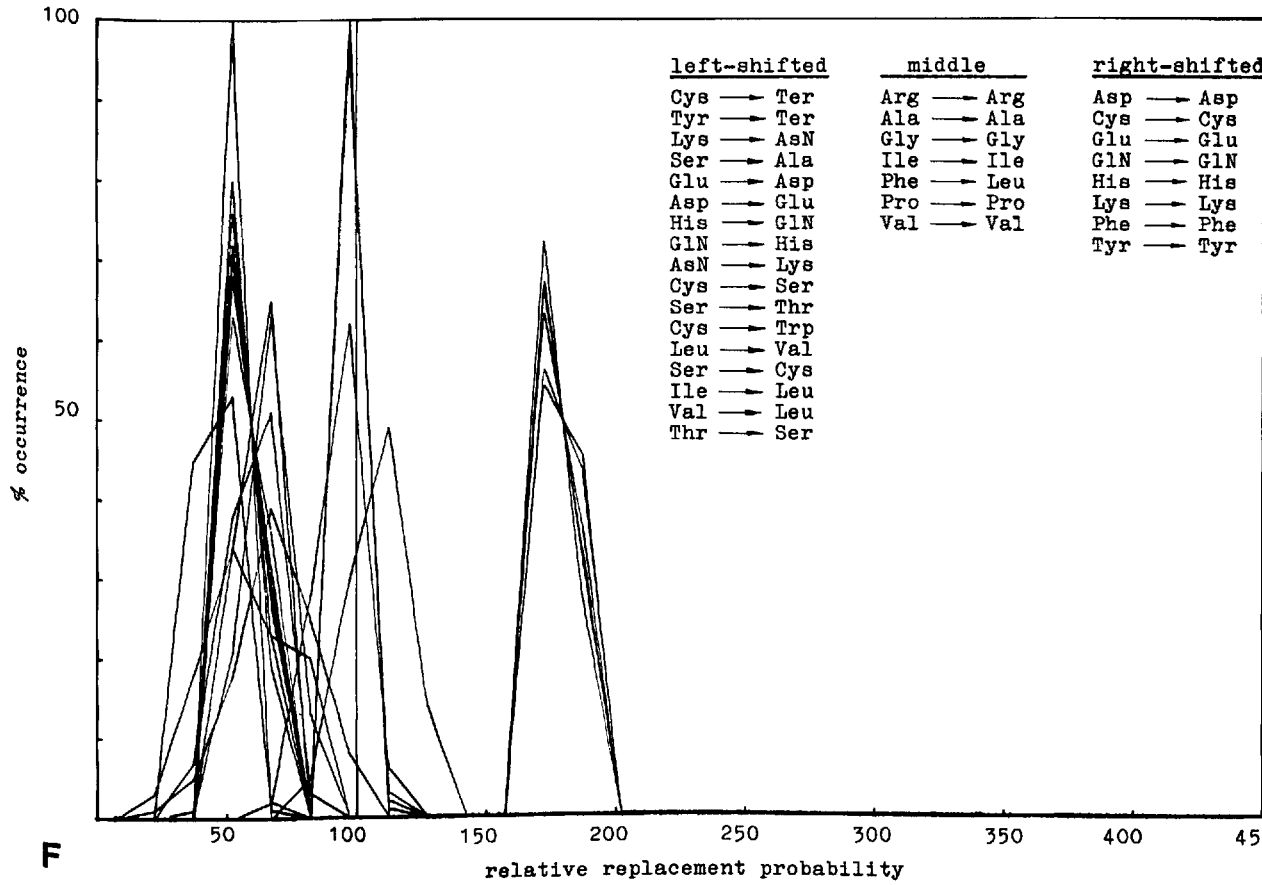
C


Ser → AsN
Arg → Cys
Gly → Glu
Ser → Gly
Arg → His
Ser → Leu
Val → Met
Leu → Ser
% occurrence
25
15
5
50
100
150
200
250
300
350
400
450
relative replacement probability

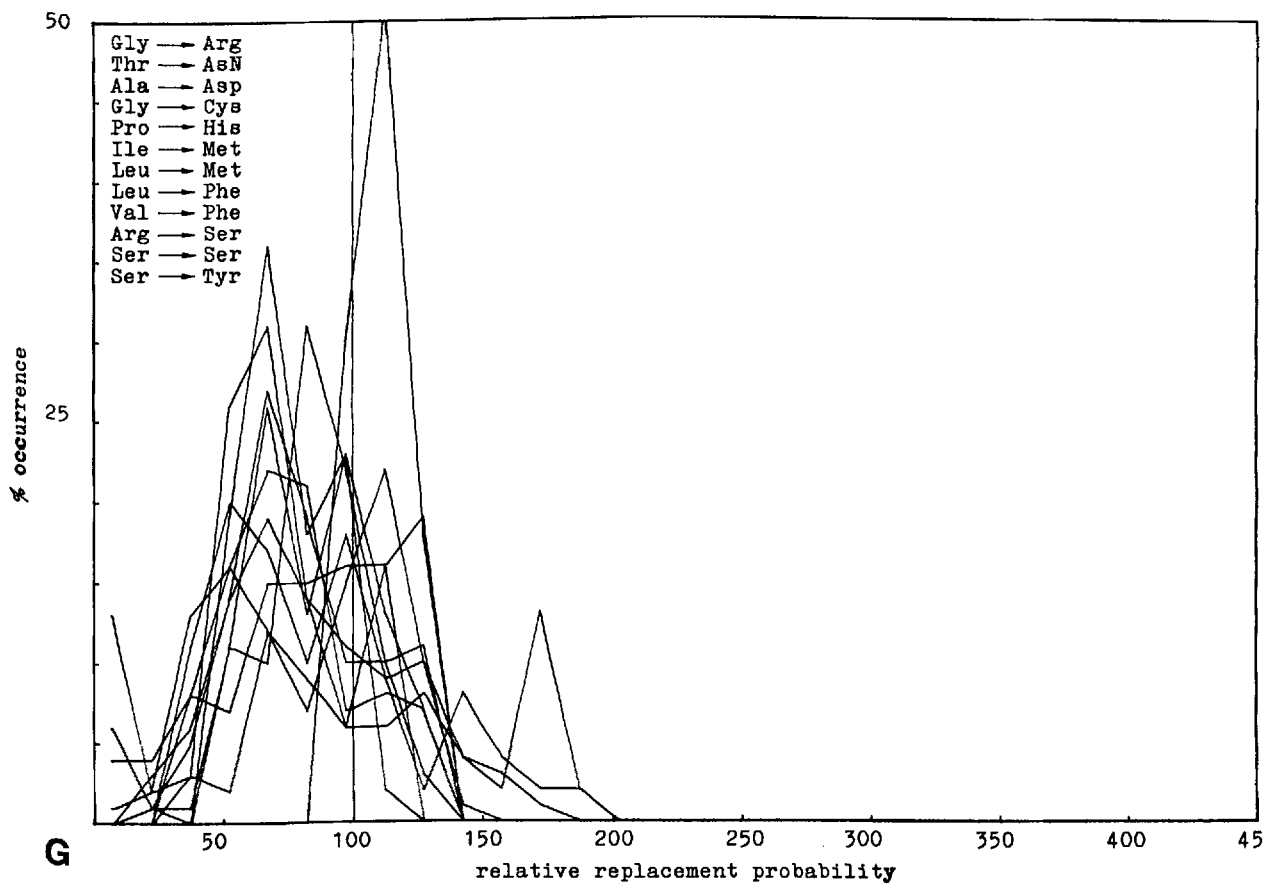
D


Leu → Arg
Ile → AsN
Val → Asp
Lys → Ile
Arg → Leu
Ile → Phe
Arg → Pro
Ile → Ser
Arg → Gly
Val → Ile
Leu → Leu
Leu → Pro
Ser → Pro
Gly → Ser
% occurrence
50
25
50
100
150
200
250
300
350
400
450
relative replacement probability

E


100
50
% occurrence
left-shifted
Cys → Ter
Tyr → Ter
Lys → AsN
Ser → Ala
Glu → Asp
Asp → Glu
His → GlN
GlN → His
AsN → Lys
Cys → Ser
Ser → Thr
Cys → Trp
Leu → Val
Ser → Cys
Ile → Leu
Val → Leu
Thr → Ser
middle
Arg → Arg
Ala → Ala
Gly → Gly
Ile → Ile
Phe → Leu
Pro → Pro
Val → Val
right-shifted
Asp → Asp
Cys → Cys
Glu → Glu
GlN → GlN
His → His
Lys → Lys
Phe → Phe
Tyr → Tyr
50
100
150
200
250
300
350
400
450
relative replacement probability

F


50
25
% occurrence
Gly → Arg
Thr → AsN
Ala → Asp
Gly → Cys
Pro → His
Ile → Met
Leu → Met
Leu → Phe
Val → Phe
Arg → Ser
Ser → Ser
Ser → Tyr
50
100
150
200
250
300
350
400
450
relative replacement probability

G

Table 3. Mean values of the relative replacement probabilities[a]

	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile
Ala	97/97 F			100		100		115		
Arg		90/93 F			300		300	85/79 G	300	46/26 A
Asn			97/177 F	345					115	124/70 E
Asp	132/84 G		300	97/177 F		97/55 F		142/273 B	115	
Cys		143/211 D			97/179 F			142/90 G		
Glu	66/42 A			97/55 F		97/177 F	115	56/109 D		
Gln		41/77 C				115	97/179 F		97/56 F	
Gly	115	97/102 C		300	100	300		97/97 F		
His		143/211 D	100	115			97/57 F		97/177 F	
Ile		111/82 C	100							97/108 F
Leu		94/60 E					100		100	83/45 F
Lys		106/129 C	97/56 F			345	115			46/26 A
Met		73/49 C								97/75 G
Phe					115					124/70 E
Pro	115	92/60 E					100		100	
Ser	115	122/74 G	300		97/52 F			142/273 E		124/70 E
Thr	345	106/69 C	100							300
Trp		60/61 A			97/57 F			49/32 A		
Tyr			100	115	345				345	
Val	345			100		100		115		300
Ter		84/72 C			97/57 F	115	345	62/41 A		

[a] The mean values of the relative replacement probabilities were calculated according to Eq. (36). They indicate whether the probability of an amino acid replacement is unaffected (≈ 100), increased (>100), or diminished (<100) when comparing the reference case I with the semirealistic case II (number to left of slash) and the realistic case III (number to right of slash). The letters A–G identify the group to which the distribution function of each relative replacement probability belongs (cf. Fig. 2A–G)

placed. It would be expected that the exchanges of amino acids exhibiting very similar properties are much more frequently accepted by selection than are exchanges of quite dissimilar residues.

To answer the question of the extent to which a randomly occurring replacement $A_l \rightarrow A_k$ is tolerated by selection, we set up an "acceptability matrix" A_{kl} defined by Eq. (34). Table 7 shows the elements of the acceptability matrix evaluated with respect to the reference case I and the realistic case III. The scaling factor γ was chosen such that for the realistic case III the highest value of A_{kl} ($A_{12,10}$, isoleucine $\rightarrow$ lysine) is equal to 1. The "acceptability values" a_k derived from A_{kl} according to Eq. (35)

Table 3. Extended

Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
				115	105/58 F	300			300
107/60 E	300	100		115	86/50 B	60/38 B	400		
	97/55 F				86/158 D	138/88 G		100	
								100	112/62 E
			100		113/69 F		230	300	
	300								83/45 B
121/68 C	100			87/57 B					
					86/144 D		100		100
80/46 B				111/72 G				300	
74/46 B	117/64 E	575	100		86/54 B	115/231 A			100/200 E
102/133 E		200	97/97 F	345	68/123 D		115		94/60 F
	97/176 F	100				60/38 B			
149/91 G	80/46 B					44/81 C			87/153 D
81/86 G			97/177 F		139/245 B			100	112/72 G
07/183 E				97/97 F	105/179 E				
77/123 D			300	345	103/110 G	111/65 F	300	100	
	100	300		115	97/52 F	97/97 F			
62/38 A					29/27 A				
			100		139/88 G			97/177 F	
97/62 F		300	100						97/97 F
78/45 A	100				75/50 B		690	97/56 F	

are listed in Table 6. For the realistic case III the acceptability values of most amino acids are higher than for case I, indicating again that preselection of mutations leads to a higher proportion of accepted mutations. This finding is in line with the lowering of the selective pressure pointed out above.

To compare the selectional values e_k and the acceptability values a_k of the 20 amino acids, we used a rank plot (Fig. 4). These measures give rise to similar rankings except for glutamate and aspartate. One might speculate that selection has favored and charged amino acids such as glutamate and aspartate in those (water-soluble) proteins that were used for the construction of the matrix N_{kl}.

On the basis of the ranking shown in Fig. 4, the amino acids can be crudely arranged into three groups. The first group (ranks 1–6) consists of those amino acids that are obviously suppressed by selection. Typical representatives of this group are cysteine, histidine, arginine, proline, and serine. A

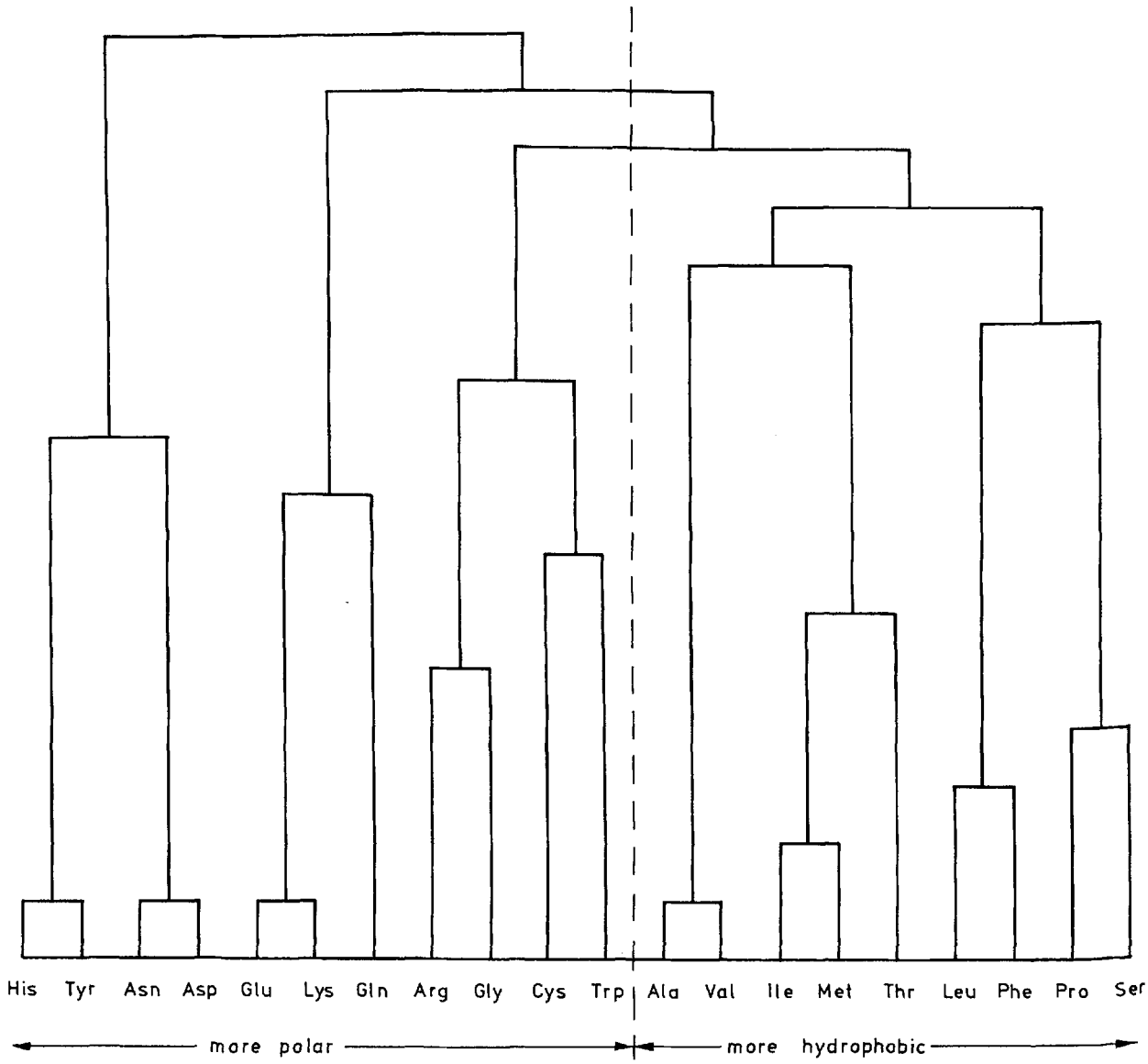


Fig. 3. Dendrogram obtained from the genetic distance matrix obtained using Eq. (40) with $\alpha = 1.15, \beta = 3$, and the average observed codon usage $\langle h_i \rangle$ given in Table 4

second group of "neutral" amino acids (ranks 7–12) apparently are not governed by special selective constraints. To this group belong such residues as valine, isoleucine, leucine, and tyrosine, all of which are nonpolar. Finally, there are some amino acids such as glutamine and lysine that are favored by selection, meaning that they occur more frequently than would be expected from a random distribution. Owing to the great discrepancies between the ranks corresponding to their e_k and a_k values, methionine, tryptophan, and asparagine cannot be uniquely placed in any one of these three groups.

It is important to point out that the selectional values e_k in some cases deviate considerably from those obtained with the measure

$$s_k = \frac{\hat{f}_k - \langle f_k \rangle}{\hat{f}_k} \tag{44}$$

(Holmquist 1978). A severe drawback of this measure is its neglect of interrelationships among the amino acid frequencies due to the mutational process. For instance, if the frequency of all precursors A_i convertible into A_k by a point mutation (i.e., for which $M_{kl} \neq 0$) is low due to any selective force, then A_k will occur with a low frequency regardless of any selective pressure on it. As shown in Fig. 4, notable discrepancies between the s_k and e_k rankings are found not only for the rare amino acids cysteine, tryptophan, and methionine, but also for valine, tyrosine, aspartate, and glutamate.

Table 4. Expected and observed amino acid frequencies $\hat{f}_k$, $\langle f_k \rangle$ and codon usages $\hat{h}_i$, $\langle h_i \rangle$[a]

Amino acid	$\langle f_k \rangle$	$\hat{f}_k$	Codon	$\langle h_i \rangle$	$\hat{h}_i$
		A. f_k decreases monotonously with α			
Ala	0.088	$\frac{1}{4(1+\alpha)^2}$	GCA	0.19	$\alpha/[2(1+\alpha)]$
			GCG	0.15	$1/[2(1+\alpha)]$
			GCC	0.25	$1/[2(1+\alpha)]$
			GCU	0.41	$\alpha/[2(1+\alpha)]$
Gly	0.077	$\frac{1}{4(1+\alpha)^2}$	GGA	0.18	$\alpha/[2(1+\alpha)]$
			GGG	0.12	$1/[2(1+\alpha)]$
			GGC	0.30	$1/[2(1+\alpha)]$
			GGU	0.40	$\alpha/[2(1+\alpha)]$
Pro	0.043	$\frac{1}{4(1+\alpha)^2}$	CCA	0.27	$\alpha/[2(1+\alpha)]$
			CCG	0.20	$1/[2(1+\alpha)]$
			CCC	0.20	$1/[2(1+\alpha)]$
			CCU	0.33	$\alpha/[2(1+\alpha)]$
Arg	0.055	$\frac{2+\alpha}{8(1+\alpha)^2}$	CGA	0.06	$\alpha/[(2+\alpha)(1+\alpha)]$
			CGG	0.07	$1/[(2+\alpha)(1+\alpha)]$
			CGC	0.22	$1/[(2+\alpha)(1+\alpha)]$
			CGU	0.30	$\alpha/[(2+\alpha)(1+\alpha)]$
			AGA	0.22	$\alpha^2/[(2+\alpha)(1+\alpha)]$
			AGG	0.13	$\alpha/[(2+\alpha)(1+\alpha)]$
		B. f_k has an optimum near $\alpha = 1$			
Asp	0.050	$\frac{\alpha}{8(1+\alpha)^2}$	GAC	0.49	$1/(1+\alpha)$
			GAU	0.51	$\alpha/(1+\alpha)$
Cys	0.030	$\frac{\alpha}{8(1+\alpha)^2}$	UGC	0.46	$1/(1+\alpha)$
			UGU	0.54	$\alpha/(1+\alpha)$
Glu	0.051	$\frac{\alpha}{8(1+\alpha)^2}$	GAA	0.55	$\alpha/(1+\alpha)$
			GAG	0.45	$1/(1+\alpha)$
Gln	0.041	$\frac{\alpha}{8(1+\alpha)^2}$	CAA	0.45	$\alpha/(1+\alpha)$
			CAG	0.55	$1/(1+\alpha)$
His	0.020	$\frac{\alpha}{8(1+\alpha)^2}$	CAC	0.48	$1/(1+\alpha)$
			CAU	0.52	$\alpha/(1+\alpha)$
Thr	0.058	$\frac{\alpha}{4(1+\alpha)^2}$	ACA	0.18	$\alpha/[2(1+\alpha)]$
			ACG	0.11	$1/[2(1+\alpha)]$
			ACC	0.33	$1/[2(1+\alpha)]$
			ACU	0.38	$\alpha/[2(1+\alpha)]$
Val	0.070	$\frac{\alpha}{4(1+\alpha)^2}$	GUA	0.19	$\alpha/[2(1+\alpha)]$
			GUG	0.25	$1/[2(1+\alpha)]$
			GUC	0.21	$1/[2(1+\alpha)]$
			GUU	0.35	$\alpha/[2(1+\alpha)]$
Met	0.021	$\frac{\alpha^2}{8(1+\alpha)^3}$	AUG	1.00	1
Ser	0.075	$\frac{3\alpha}{8(1+\alpha)^2}$	AGC	0.14	$1/[3\alpha(1+\alpha)]$
			AGU	0.15	$1/[3(1+\alpha)]$
			UCA	0.15	$1/[3(1+\alpha)]$
			UCG	0.07	$1/[3\alpha(1+\alpha)]$
			UCC	0.18	$1/[3\alpha(1+\alpha)]$
			UCU	0.31	$1/[3(1+\alpha)]$
Trp	0.012	$\frac{\alpha}{8(1+\alpha)^3}$	UGG	1.00	1
		C. f_k increases monotonously with α			
Asn	0.040	$\frac{\alpha^2}{8(1+\alpha)^2}$	AAC	0.57	$1/(1+\alpha)$
			AAU	0.43	$\alpha/(1+\alpha)$
Lys	0.063	$\frac{\alpha^2}{8(1+\alpha)^2}$	AAA	0.59	$\alpha/(1+\alpha)$
			AAG	0.41	$1/(1+\alpha)$
Phe	0.042	$\frac{\alpha^2}{8(1+\alpha)^2}$	UUC	0.48	$1/(1+\alpha)$
			UUU	0.52	$\alpha/(1+\alpha)$
Tyr	0.030	$\frac{\alpha^2}{8(1+\alpha)^2}$	UAC	0.53	$1/(1+\alpha)$
			UAU	0.47	$\alpha/(1+\alpha)$
Ile	0.052	$\frac{\alpha^2(2\alpha+1)}{8(1+\alpha)^3}$	AUA	0.13	$\alpha/(1+2\alpha)$
			AUC	0.40	$1/(1+2\alpha)$
			AUU	0.47	$\alpha/(1+2\alpha)$
Leu	0.097	$\frac{\alpha(2+\alpha)}{8(1+\alpha)^2}$	CUA	0.06	$\alpha/[(2+\alpha)(1+\alpha)]$
			CUG	0.29	$1/[(2+\alpha)(1+\alpha)]$
			CUC	0.14	$1/[(2+\alpha)(1+\alpha)]$
			CUU	0.19	$\alpha/[(2+\alpha)(1+\alpha)]$
			UUA	0.17	$\alpha^2/[(2+\alpha)(1+\alpha)]$
			UUG	0.15	$\alpha/[(2+\alpha)(1+\alpha)]$
Ter	—	$\frac{\alpha^2(2+\alpha)}{8(1+\alpha)^3}$	UAA	—	$\alpha/(2+\alpha)$
			UAG	—	$1/(2+\alpha)$
			UGA	—	$1/(2+\alpha)$

[a] For a graphical illustration of the α-dependency of the expected frequencies $\hat{f}_k$, see Cornish-Bowden and Marson (1977)

The Physicochemical Significance of Amino Acid Replacements

It is generally accepted that the observed interchangeabilities of amino acids can be related to their physicochemical properties (Gamov 1954; Epstein 1967; Alff-Steinberger 1969; McLachlan 1971; Dayhoff et al. 1972; Papentin 1973; Batchinsky 1976; Salemme et al. 1977; Doolittle 1979; Sander and Schulz 1979; Wolfenden et al. 1979, 1981; Argyle 1980; Richard et al. 1980; Charton 1981; Hendry et al. 1981a,b). Usually a more or less arbitrary set of properties such as size; shape; local concentration of electrical charge; and ability to form salt bonds, hydrophobic bonds, or hydrogen bonds is used a priori to classify the amino acids and the resulting classifications are used to predict which mutational interchanges are conservative. But these a priori classifications are not necessarily pertinent to the functions of amino acids in proteins. Measures of the interchangeability of proteinogeneous amino acids may provide information about which properties are of importance for their use in proteins. Some aspects of this approach are outlined in this paper; a much more detailed analysis based on more than 500 different amino acid properties will be presented in a subsequent paper (C. Frömmel and H.-G. Holzhütter, manuscript in preparation).

Let p_k designate any property of amino acid A_k. A very simple way to define a distance between the amino acids A_k and A_l with respect to that property is

$$P_{kl} = |p_k - p_l|. \qquad (45)$$

Table 5. Observed and expected amino acid frequencies

k	Amino acid A_k	Observed frequency $\langle f_k \rangle$ in those proteins related to the 162 mRNAs	Expected frequency $\hat{f}_k$ according to Eqs. (32.1) and (32.2) with $\alpha = 1.15$[a]	Expected frequency $\hat{f}_k$ according to Eq. (43)[b]	Expected frequency $\hat{f}_k$ according to Eq. (43) using the matrix N_{kl} of accepted point mutations instead of M_{kl}[c]
1	Ala	0.088	0.054	0.057	0.061
2	Arg	0.055	0.086	0.077	0.051
3	Asn	0.040	0.038	0.042	0.035
4	Asp	0.050	0.033	0.038	0.050
5	Cys	0.030	0.033	0.041	—
6	Glu	0.051	0.033	0.027	0.045
7	Gln	0.041	0.033	0.026	0.051
8	Gly	0.077	0.057	0.055	0.077
9	His	0.020	0.033	0.039	0.055
10	Ile	0.052	0.058	0.059	0.051
11	Leu	0.097	0.104	0.097	0.080
12	Lys	0.063	0.042	0.032	0.048
13	Met	0.021	0.018	0.015	0.035
14	Phe	0.042	0.038	0.045	0.055
15	Pro	0.043	0.057	0.060	0.027
16	Ser	0.075	0.098	0.105	0.077
17	Thr	0.058	0.066	0.064	0.091
18	Trp	0.012	0.015	0.011	0.015
19	Tyr	0.030	0.038	0.046	0.039
20	Val	0.070	0.066	0.064	0.070
Distance $d(\hat{f},\langle f \rangle)$ according to Eq. (41)			0.102	0.132	0.079

[a] Normalization condition: $\sum_{k=1}^{20} \hat{f}_k = 1$

[b] The mutation matrix in Eq. (43) was calculated with $\alpha = 1.15$, $\beta = 3$, and the average observed codon usages $\langle h_i \rangle$ given in Table 4

[c] Values for k = 1 = 5 were omitted to prevent an unrealistically high frequency of cysteine from being obtained

Table 6. Selectional and acceptability values

Codon C_i	Selective value E_i[a]	Amino acid A_k	Selective value e_k Case III	Selective value e_k Case I	Acceptability value a_k
GCA	0.26	Ala	0.53	0.48	0.20
GCG	−0.43				
GCC	0.28				
GCU	1.16				
CGA	−5.61	Arg	−1.02	−1.35	0.08
CGG	−6.73				
CGC	0.26				
CGU	0.87				
AGA	−0.79				
AGG	−2.68				
AAC	0.46	Asn	−0.02	0.06	0.11
AAU	−0.66				
GAC	0.39	Asp	0.26	0.59	0.10
GAU	0.13				
UGC	−1.85	Cys	−1.88	−1.54	0.01
UGU	−1.89				
GAA	0.42	Glu	0.34	0.22	0.19
GAG	0.26				
CAA	0.40	Gln	0.66	0.51	0.16
CAG	0.88				
GGA	−0.49	Gly	0.33	0.33	0.08
GGG	−1.68				
GGC	0.71				
GGU	1.06				
CAC	−1.62	His	−1.55	−1.85	0.07
CAU	−1.48				
AUA	−4.11	Ile	−0.19	−0.05	0.17
AUC	0.36				
AUU	0.45				
CUA	−4.90	Leu	0.16	0.04	0.11
CUG	1.64				
CUC	−0.32				
CUU	0.09				
UUA	0.56				
UUG	−0.54				
AAA	1.21	Lys	0.91	1.40	0.22
AAG	0.48				
AUG	0.89	Met	0.89	0.51	0.20
UUC	0.61	Phe	0.42	0.75	0.16
UUU	0.24				
CCA	0.40	Pro	−0.70	−0.77	0.06
CCG	−1.41				
CCC	−1.65				
CCU	−0.57				
AGC	−0.29	Ser	−0.62	−0.33	0.12
AGU	−3.22				
UCA	−0.36				
UCG	0.74				
UCC	−1.62				
UCU	−1.74				
ACA	−0.87	Thr	−0.43	−0.28	0.14
ACG	−3.55				
ACC	0.30				
ACU	0.07				
UGG	0.60	Trp	0.60	−0.11	0
UCA	0.24	Tyr	0.07	−0.13	0.10
UAU	−0.12				
GUA	−0.43	Val	−0.24	−0.25	0.11
GUG	0.09				
GUC	−1.24				
GUU	0.20				
Mean selective pressure			$\langle e \rangle = 0.61$	$\langle e \rangle = 0.64$	

[a] Replacements involving the terminator were not considered; i.e., $i \neq 49, 50, 53$ in Eq. (29)

Table 7. Elements of the acceptability matrix[a]

	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	–	0	0	0.10 0.18	0	0.11 0.20	0	0.17 0.27	0	0	0	0	0	0	0.17 0.26	0.31 0.51	0.17 0.10	0	0	0.15 0.09
Arg	0	–	0	0	0	0	0.09 0.06	0.03 0.03	0.09 0.05	0.07 0.33	0.04 0.06	0.11 0.07	0.09 0.16	0	0.02 0.03	0.04 0.07	0.06 0.15	0.03 0.03	0	0
Asn	0	0	–	0.18 0.09	0	0	0	0	0.15 0.23	0	0	0.04 0.06	0	0	0	0.47 0.28	0.14 0.15	0	0	0
Asp	0.06 0.07	0	0.20 0.12	–	0	0.25 0.42	0	0.15 0.06	0.03 0.05	0	0	0	0	0	0	0	0	0	0.05 0.08	0
Cys	0	0	0	0	–	0	0	0	0	0	0	0	0	0	0	0.07 0.10	0	0	0	0
Glu	0.20 0.45	0	0	0.19 0.32	0	–	0.17 0.26	0.12 0.10	0	0	0	0.09 0.05	0	0	0	0	0	0	0	0.08 0.16
Gln	0	0.27 0.35	0	0	0	0.22 0.34	–	0	0.07 0.12	0	0.03 0.05	0.09 0.16	0	0	0.07 0.12	0	0	0	0	0
Gly	0.19 0.29	0.04 0.04	0	0.06 0.04	0	0.09 0.06	0	–	0	0	0	0	0	0	0	0.32 0.22	0	0	0	0.05 0.09
His	0	0.13 0.04	0.08 0.14	0.04 0.07	0	0	0.06 0.10	0	–	0	0.06 0.10	0	0	0	0.04 0.05	0	0	0	0.07 0.04	0
Ile	0	0.40 0.48	0.04 0.07	0	0	0	0	0	0	–	0.24 0.47	0.02 0.03	0	0.09 0.17	0	0.14 0.26	0.09 0.04	0	0	0.44 0.23
Leu	0	0.03 0.05	0	0	0	0	0.05 0.08	0	0	0.26 0.55	–	0	0.14 0.24	0.04 0.04	0.13 0.07	0.03 0.03	0	0	0	0.09 0.14
Lys	0	0.60 0.30	0.09 0.15	0	0	0.03 0.02	0.05 0.07	0	0	0.22 1.00	0	–	0	0	0	0	0.08 0.22	0	0	0
Met	0	0.13 0.27	0	0	0	0	0	0	0	0.07 0.10	0.35 0.43	0.07 0.15	–	0	0	0	0.04 0.05	0	0	0.41 0.21
Phe	0	0	0	0	0	0	0	0	0	0.04 0.05	0.11 0.11	0	0	–	0	0.11 0.04	0	0	0.48 0.86	0.05 0.07
Pro	0.11 0.17	0.07 0.11	0	0	0	0	0.02 0.04	0	0.03 0.05	0	0.03 0.02	0	0	0	–	0.11 0.06	0.01 0.02	0	0	0
Ser	0.39 0.61	0.04 0.06	0.32 0.19	0	0.01 0.02	0	0	0.27 0.10	0	0	0.07 0.05	0	0	0.04 0.02	0.13 0.07	–	0.23 0.34	0	0.05 0.08	0
Thr	0.25 0.13	0	0.06 0.11	0	0	0	0	0	0	0.15 0.09	0	0.08 0.14	0.18 0.11	0	0.06 0.09	0.37 0.64	–	0	0	0
Trp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	–	0	0
Tyr	0	0	0.04 0.07	0.01 0.02	0	0	0	0	0.03 0.02	0	0	0	0	0.29 0.51	0	0.11 0.11	0	0	–	0
Val	0.10 0.05	0	0	0.03 0.05	0	0.10 0.17	0	0.02 0.04	0	0.39 0.23	0.17 0.27	0	0.09 0.05	0.06 0.10	0	0	0	0	0	–

[a] The normalization constant γ in Eq. (34) was chosen such that $\underset{k,l=1,\ldots,20}{\text{MAX}} A_{kl} = 1 = A_{12,10}$ (Ile → Lys) with respect to the realistic case III. The upper value refers to the reference case I; the lower value, to the realistic case III

This symmetric property-distance matrix P_{kl} is expected to correlate with the matrix N_{kl} of accepted point mutations. The results of a correlation analysis based on only some selected properties are shown in Table 8. Significant correlations are found for "classical" properties such as molecular weight, polarity, volume, accessible surface, and transfer energy, as well as for some other properties that cannot be interpreted so easily, such as nonbonded energies.

According to Eq. (33) we have split N_{kl} into two components M_{kl} and A_{kl} describing separately the influences of mutations and of selection on the probability of occurrence of accepted amino acid replacements. It is interesting to study the relative shares of these two components in the correlations between P_{kl} and N_{kl}. These correlation coefficients are also shown in Table 8. Note that for the reference case I and the realistic case III most of the correlations between P_{kl} and N_{kl} account for selection, i.e., that $r_A > r_M$, where r_A and r_M denote the cor-

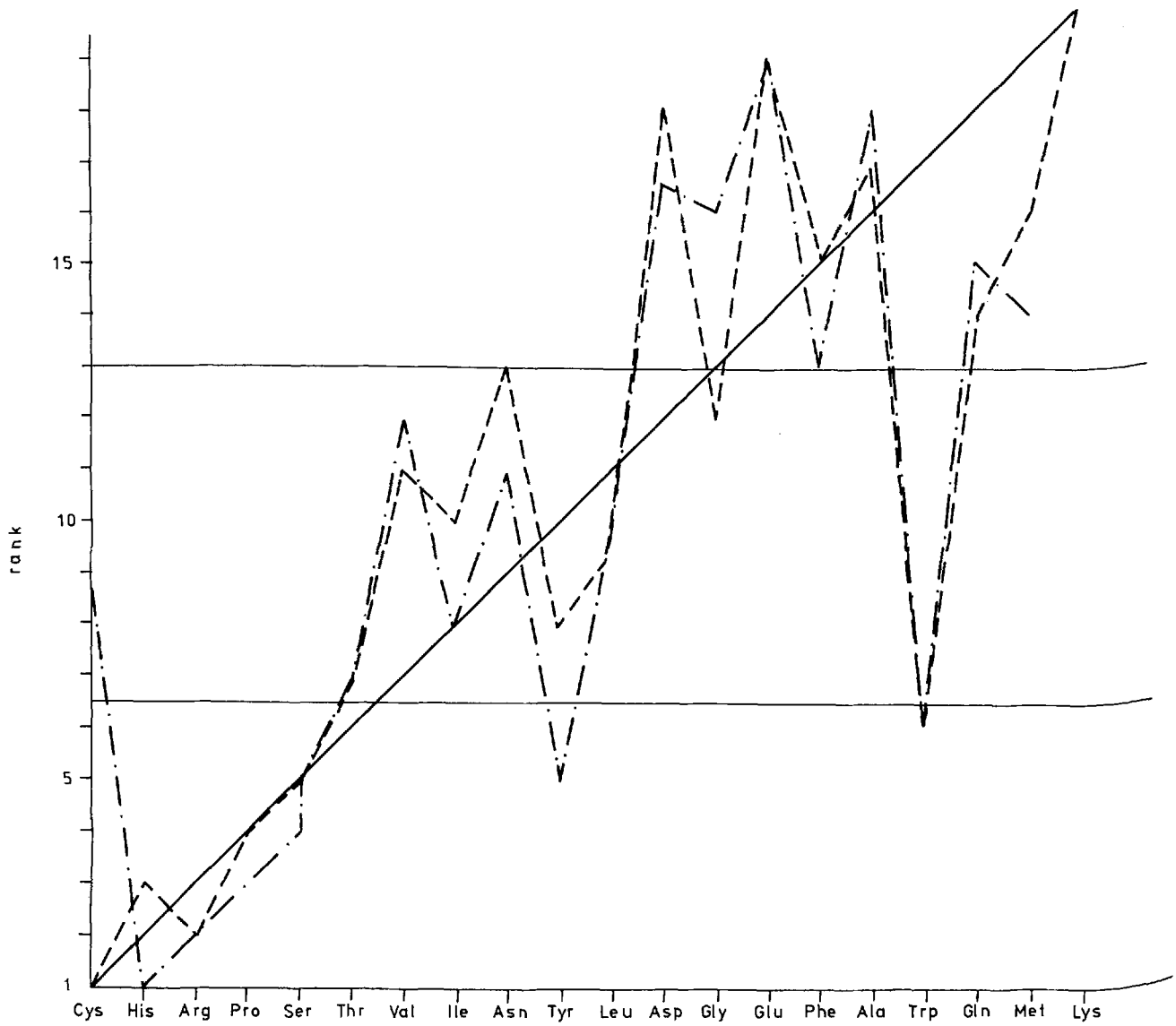


Fig. 4. Rank plot of the selectivity measures e_k, a_k, and s_k. Rank 1 is attributed to the lowest value and rank 20 to the highest value. ——, Ranking according to the selectional values e_k shown in Table 6; – – –, ranking according to the acceptability values a_k shown in Table 6; –·–·–, ranking according to the selectional values s_k calculated from Eq. (44) with $\hat{f}_k = 61/n_k$ and the observed frequencies $\langle f_k \rangle$ listed in Table 4

relation coefficients between the property-distance matrix P_{kl} and the matrices A_{kl} and M_{kl}, respectively. In particular, this holds for all measures of bulkiness, e.g., volume and molecular weight.

Discussion

As shown by Alff-Steinberger (1969) and Papentin (1973) the genetic code possesses a higher evolutionary efficiency (given by the mean number of generations required to reach identity in five of ten residues, if originally all ten were different) than do other possible codes. The genetic code fixes the minimum number of point mutations necessary to replace one amino acid by another, but the probability (or rate) of such a replacement depends on the probability of the involved base replacements occurring and to about 50% also on the codon used for coding the initial amino acid. In our approach the key quantities affecting the mutation probability are the two mutation parameters α and β and the so-called codon usage h_i. Variation of these quantities decreases or increases the mutation probabilities. Our analysis indicates that the capacity of the mutational process to favor or to suppress certain amino acid replacements represents an additional way for living systems to regulate their evolutionary efficiency. We have called this phenomenon "preselection" in order to stress that the stochastic component of evolution (mutation) is adapted to basic requirements imposed by selectional constraints. One of these ba-

Table 8. Correlation coefficients r_N, r_M, and r_A between the property-distance matrix P_{kl} and the (symmetrized) matrix of accepted point mutations N_{kl}, mutation matrix M_{kl}, and acceptability matrix A_{kl}, respectively[a]

Property	r_N	r_M		r_A	
		Case I	Case III	Case I	Case III
Nonbonded energy per atom (Oobatake and Ooi 1977)					
Total	−0.40†	−0.30***	−0.25*	−0.30***	−0.10
Short range	−0.27*	−0.10	−0.03	−0.15	−0.10
Long range	−0.41†	−0.28**	−0.29**	−0.31**	−0.16
Nonbonded energy per amino acid molecule (Oobatake and Ooi 1977)					
Total	−0.42†	−0.09	−0.09	−0.44†	−0.45†
Short range	−0.38†	−0.08	−0.07	−0.39†	−0.42†
Long range	−0.46†	−0.06	−0.17	−0.48†	−0.43†
Molecular weight (Zamyatnin 1972)	−0.41†	−0.20	−0.16	−0.36†	−0.36†
Specific van der Waals volume (Goldsack and Chalifoux 1973)	−0.41†	−0.09	−0.03	−0.46†	−0.49†
Transfer energy (Bull and Breese 1974; Olsen 1980)	−0.45†	−0.27*	−0.21	−0.39†	−0.28**
Polarity (Grantham 1974)	−0.44†	−0.42†	−0.35†	−0.36†	−0.18
Internal preference (Chothia 1975)	−0.21	−0.37†	−0.27*	−0.07	0.17
Accessible surface (Chothia 1976)	−0.41†	−0.17	−0.14	−0.39†	−0.38†
Interaction parameter (Krigbaum and Komoriya 1979)	−0.41†	−0.33***	−0.29**	−0.31**	−0.09
Probability of β-structure (Chou and Fasman 1978)	−0.31**	−0.18	−0.18	−0.28**	−0.13
Probability of α-helix (Chou and Fasman 1978)	0.12	0.03	0.06	0.04	−0.04
Information measure of β-structure (Robson and Suzuki 1976)	−0.31**	−0.19	−0.24*	−0.29**	−0.12
Information measure of α-helix (Robson and Suzuki 1976)	0.10	0.00	0.06	0.02	0.01

Significance levels: *5%; **1%; ***0.3%; †0.1%

[a]The correlation coefficients r_S (S = N, M, A) were calculated according to

$$r_S = \sum_{k=1}^{19} \sum_{l=k+1}^{20} (P_{kl} - \bar{P})([S]_{kl} - \bar{S}) \Big/ \sqrt{\sum_{k=1}^{19} \sum_{l=k+1}^{20} (P_{kl} - \bar{P}) \sum_{k=1}^{19} \sum_{l=k+1}^{20} ([S]_{kl} - [\bar{S}])^2},$$

where the summations run over the nonzero elements of the matrices P and S. Here [S] denotes the symmetric part of the matrix S, i.e., $[S]_{kl} = \frac{1}{2}(S_{kl} + S_{lk})$

sic requirements seems to be the conservation of the polar or hydrophobic character of the replaced amino acid as given in the grouping based on the genetic-distance matrix (see Fig. 3). Probably this adaptation was established during evolution of protein structures by selection of such copy-error-preventing mechanisms as would secure a higher proportion of accepted mutations and thus a higher evolutionary efficiency. This conclusion can be drawn from the general tendency of the acceptabilities of amino acid replacements to increase when switching from the reference case I to the realistic case III.

Several authors have pointed out a relationship between the nonrandomness of codon usage and the rate of protein synthesis (Grantham et al. 1980, 1981). On the assumption that the selectional value of a given protein population is determined by the selectional value (or functional fitness) of a single protein molecule (which is a function of the amino acid composition only) as well as by the number of protein molecules synthesized per time unit, the selection of codons might in fact be triggered by the rate at which an amino acid is incorporated into the nascent polypeptide chain. Obviously this rate may vary as a function of the codon and the related tRNA. But in our opinion it is more likely that the selectional value of a codon depends essentially on the availability of the associated tRNA, which means that the postulated selective constraints at the genotypic level are due to the pattern of tRNA concentration. In other words, during evolution or adaptation of organisms, a tRNA pattern has been

selected that requires a codon usage yielding a lower risk of producing dangerous amino acid replacements as a result of point mutation.

On the basis of the distribution functions ϕ_{kl} (i) obtained for the h_i-dependent elements of the mutation matrix M_{kl}, the amino acid replacements can be roughly classified as follows: Some replacements occur with notably lower or higher rates than would be expected from random codon usage and unbiased mutation (reference case I). Typical examples are the replacements leading to the terminator codon and the replacements isoleucine → arginine, leucine → tryptophan, serine → tryptophan, serine → arginine, proline → glutamine, serine → isoleucine, threonine → arginine, valine → glutamine, leucine → isoleucine(!), lysine → methionine, threonine → lysine (suppressed); and glycine → aspartate, serine → phenylalanine, valine → isoleucine, and threonine → isoleucine (increased) (see Fig. 2 and Table 3). If one believes in "preselected" mutations these pairs of amino acids should have very similar or dissimilar properties with respect to their functions in proteins. On the other hand, there are replacements the probabilities of which vary considerably around the mean value in both directions, e.g., glycine → glutamate, arginine → cysteine, serine → asparagine, serine → glycine, arginine → histidine, arginine → lysine, threonine → methionine, valine → methionine, arginine → threonine, and leucine → serine. One can speculate that the usefulness of these replacements essentially depends on the particular functions of the involved amino acids in the protein structure. Therefore a systematic investigation of codon usage as it relates to amino acid replacement probabilities should provide information about relevant amino acid properties and their significances for the strategy of protein evolution.

During the past few years much attention has been paid to the nonrandomness of amino acid composition. Jukes et al. (1975) described this phenomenon as "selection against the genetic code." In a recent paper Golding and Strobeck (1982) proposed a method for investigating how the forces of selection and mutation can shape the pattern of codon usage. As a counterpart to their approach, our procedure yields a quantitative estimate of the selective forces from the observed pattern of codon usage. The obtained selectional values e_k and acceptability values a_k confirm the assumption of Golding and Strobeck (1982) that there must be selection against cysteine and proline, both of which can radically alter the structure of proteins. It should be emphasized that the selectional values e_k reflect only quite general selective constraints on proteins irrespective of their special functions because the protein-specific selective pressure is canceled by our use of a steady-state composition, which represents the average composition of a large group of functionally different proteins.

Whereas the selectional value e_k of amino acid A_k represents a rather global measure of the selectional constraints that affect the usage of this amino acid in proteins, the acceptability A_{kl} of a replacement $A_l \rightarrow A_k$ yields a more detailed picture of favorable and unfavorable replacements. It is important to note that the acceptability matrix A_{kl} was derived from observed amino acid replacements within phylogenetic trees of functionally closely related proteins. For that reason the acceptability values a_k can be expected to reflect the selective pressures encountered during adaptation of proteins to special requirements of their environment. For example, most of the proteins considered in constructing the matrix N_{kl} are soluble in water. This might account for the pronounced selectivity of charged residues such as glutamate and aspartate.

Since the discovery of the genetic code many physicochemical relationships between amino acids and codons have been described. According to the coevolutionary hypothesis of Wong (1975), the evolutions of tRNA and of the genetic code are closely related. Such a coevolution seems plausible if we take into consideration that the amino-acyl tRNA synthetase is the only enzyme that interacts both with nucleic acids and with proteins. For a new amino amino acid to join the set of already existing proteinogeneous amino acids, the binding regions of the synthetase for both molecules—amino acid and tRNA—have to be altered. These necessary changes are small if a synthetase is employed that is related to an amino acid very similar to the new one. From our correlation analysis between amino acid properties and the processes of mutation and selection, it can be concluded that unbiased (nonpreselected) mutations (designated in this paper as the reference case I) would be insufficient to secure the conservation of certain properties, which after the establishment of the genetic code at an early stage of evolution would have been a prerequisite for efficient evolution of proteins. For the realistic case III about half of the correlations between P_{kl} and A_{kl} obtained for the reference case I disappear, indicating that preselected mutations reduce the importance of selection in bringing about proper amino acid replacements. Moreover, it can be suggested that after the establishment of the genetic code, further selective constraints for an efficient evolution of proteins have become important that influence the use of amino acids. Hence we have to look for different similarity relations between amino acids at different stages of protein evolution.

As shown in Table 4, the expected random amino acid frequencies strongly depend on the mutation parameter α. In particular, an increase in the GC

content of DNA, typical of thermophilic organisms (Amelunxen and Murdock 1978), should be accompanied by an increase in the amounts of arginine and glycine and a decrease in the amounts of serine and lysine. Such changes in the amino acid composition are characteristic of increases in thermostability among evolutionarily related proteins (Singleton et al. 1977; Argos et al. 1979). The question arises as to whether arginine and glycine are indeed responsible for the higher thermostability or whether they are only markers (i.e., a necessary consequence) of the higher GC content of DNA, with thermostability resulting from other factors. The latter hypothesis has been confirmed by a single-residue correlation analysis between melting points and amino acid compositions of proteins without consideration of their sources (Ponnuswamy et al. 1982), which showed that an abundance of glutamate and lysine is characteristic of thermostable proteins, whereas a high proportion of serine and valine is observed in low-melting proteins. With the aid of our method, it should be possible to examine how mutation and selection jointly participate in the change of amino acid composition between thermostable and thermolabile proteins.

Concluding Remarks

In this paper we have pointed out the existence of a close relationship between the probability (or rate) of amino acid replacement and the relative constancy of some fundamental properties of amino acids that are known to affect substantially the stability of the protein structure, e.g., polarity and hydrophobicity. The further elucidation of which properties of amino acids are invariant against amino acid replacements due to point mutations may yield deeper insight in the roles and functions of amino acids in proteins. Of course one must keep in mind that our averaging approach registers only such basic properties of amino acids as are important for all proteins irrespective of their specific functions.

The conservatism of the genetic code, which obviously can be reinforced by use of nonrandom codon usage in combination with the mutation probabilities of the bases in nucleic acids—a phenomenon we have called "preselection," should reduce variation in the properties of proteins resulting from single point mutations. For instance, cytochrome c has remained cytochrome c despite all the point mutations the original gene has undergone. The preselectional effect operative at the genetic level may be quite important in assuring the conservatism of mutational changes since it can diminish the proportion of dangerous amino acid replacements not only in the transmission of genetic information to the following generation (DNA replication) but also in the realization of the genetic information (transcription and translation).

Finally, it seems appropriate to point out that in our opinion a distinction should be made between molecular changes occurring during evolution that can be regarded as adaptations to environmental pressures and the "true" biological evolution that leads to new or more complex organisms having characteristic patterns of proteins and biological functions. Point mutations are probably relevant only in the first process and can therefore be regarded as the conservative or more neutral element of the biological evolution responsible for gradual changes of biological structures and functions, whereas true biological evolution is characterized by large-scale and coordinate changes.

Acknowledgment. The authors wish to thank Prof. Dr. S. M. Rapoport for many helpful discussions and for critical reading of the manuscript.

References

Aboderin AA (1971) An empirical hydrophobic scale for α-amino-acids and some of its applications. Int J Biochem 2:537–544

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

Amelunxen R, Murdock AL (1978) Mechanisms of thermophily. CRC Crit Rev Microbiol xx:343–393

Argos P, Rossman MG, Grau UM, Zuber H, Frank G, Tratschin JD (1979) Thermal stability and protein structure. Biochemistry 18:5698–5703

Argyle E (1980) A similarity ring for amino acids based on their evolutionary substitution rate. Orig Life 10:357–360

Batchinsky AG (1976) Structure and noise immunity of the genetic code. J Gen Biol 37:163–174

Berger EM (1978) Pattern and chance in the use of the genetic code. J. Mol Evol 10:319–323

Brown AL (1982) Evolution and molecular biology. Nature 298:793–794

Bull HB, Breese K (1974) Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. Arch Biochem Biophys 161:665–670

Cercignany C (1975) Theory and applications of the Boltzmann equation. Scottish Academic Press, Edinburgh London

Charton M (1981) Protein folding and the genetic code: an alternative quantitative model. J Theor Biol 91:115–123

Chothia C (1975) Structural invariants in protein folding. Nature 254:304–308

Chothia C (1976) The nature of the accessible and buried surfaces in proteins. J Mol Biol 105:1–14

Chou PY, Fasman GD (1978) Prediction of secondary-structure of proteins from their amino-acid sequence. Adv Enzymol 47:45–148

Conrad M, Friedlander C, Goodman M (1983) Evidence that natural selection acts on silent mutation. Biosystems 16:101–111

Cornish-Bowden A, Marson A (1977) Evaluation of the nonrandomness of protein composition. J Mol Evol 10:231–240

Coulondre C, Miller JH (1977) Genetic studies of the lac-repressor. Part IV: Mutagenic specifity in the LacI gene of *Escherichia coli*. J Mol Biol 117:577–606

Coutelle R, Hofacker GLC (1982) Influence of selective processes on the amino acid composition of proteins: collagen, cytochrome c, ferredoxin and α-crystallin. J Theor Biol 95: 615–639

Cox EC (1976) Bacterial mutator genes and the control of spontaneous mutation. Annu Rev Genet 10:135–156

Davies J, Jones DS, Khorano HG (1966) A further study of misreading of codons induced by streptomycin and neomycin using ribopolynucleotides containing two nucleotides in alternating sequence as templates. J Mol Biol 18:48–57

Dayhoff MO, Eck RV, Park CM (1972) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure. National Biomedical Research Foundation Washington, DC, pp. 89–99

Doolittle RF (1979) Protein evolution. In: Neurath H, Hill RL (eds) The proteins, 3rd edn, vol IV. Academic Press, New York, pp 1–118

Dunill P (1968) The use of helical net-diagrams to represent protein structures. Biophys J 8:865–875

Eigen M, Schuster P (1979) The hypercycle: a principle of natural self-organization. Springer-Verlag, Berlin Heidelberg New York

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

Fendler JH, Nome F, Nagyvary J (1975) Compartmentalization of amino acids in surfactant aggregates. Partitioning between water and aqueous micellar sodium dodecanoate and between hexane and dodecylammonium propeonate trapped in hexane. J Mol Evol 6:215–232

Fersht AR (1979) Fidelity of replication of phage ΦX174 DNA by DNA-polymerase III holoenzyme: spontaneous mutation by misincorporation. Proc Natl Acad Sci USA 76:4946–4950

Fersht AR, Knill-Jones JW (1981) DNA polymerase accuracy and spontaneous mutation rates: frequencies of purine–purine, purine–pyrimidine and pyrimidine–pyrimidine mismatches during DNA replication. Proc Natl Acad Sci USA 78:4251–4255

Fitch WM (1967) Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. J Mol Biol 26:499–507

Fitch WM (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: Comparison of several methods and three beta hemoglobin messenger RNAs. J Mol Evol 16:153–209

Gamov G (1954) Possible relation between deoxyribonucleic acid and protein structure. Nature 173:318–320

Golding GB, Strobeck C (1982) Expected frequencies of codon use as a function of mutation rates and codon fitness. J Mol Evol 18:379–386

Goldsack DE, Chalifoux RC (1973) Contributions of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins. J Theor Biol 39:645–651

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

Grantham R, Gautier C, Gouvy M (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according the genome type. Nucleic Acids Res 8: 1893–1912

Grantham R, Gautier C, Gouvy M, Jacobsone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r74

Grosjean H, Sankoff D, Jou WM, Fiers W, Cedergren RJ (1978) Bacteriophage MS2 RNA: a correlation between the stability of the codon anticodon interaction and the choice of code words. J Mol Evol 12:113–119

Hendry LB, Bransome ED, Petersheim M (1981a) Are there structural analogies between amino acids and nucleic acids? Orig Life 11:203–221

Hendry LB, Bransome ED, Hutson MS, Campbell LK (1981b) First approximation of a stereochemical rationale for the genetic code based on the topography and physicochemical properties of cavities constructed from models of DNA. Proc Natl Acad Sci USA 78:7440–7444 (see also refs 3–47 within)

Holland JP, Holland MJ (1980) Structural comparison of two non-tandemly repeated yeast glyceraldehyde-3-phosphate dehydrogenase genes. J Biol Chem 255:2596–2605

Holmquist R (1978) Evaluation of compositional nonrandomness in proteins. J Mol Evol 11:349–360

Holmquist R, Pearl D (1980) Theoretical foundations for quantitative paleogenetics. Part III: The molecular divergence of nucleic acids and proteins for the case of genetic events of unequal probability. J Mol Evol 16:211–267

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

Jones DD (1975) Amino acid properties and side chain orientations in proteins. J Theor Biol 50:167–183

Jukes TH (1975) Mutation in proteins and base changes in codons. Biochem Biophys Res Commun 66:1–18

Jukes TH, Holmquist R, Moise H (1975) Amino acid composition of proteins: selection against the genetic code. Science 189:50–51

Kafatos FC, Efstratiadis A, Forget BE, Weissmann SM (1977) Molecular evolution of human and rabbit β-globin mRNAs. Proc Natl Acad Sci USA 74:5618–5622

Krigbaum WR, Komoriya A (1979) Local interactions as a structure determinant for protein molecules. Biochim Biophys Acta 576:204–228

Laird M, Holmquist R (1975) Tables of critical values for examining compositional non-randomness in proteins and nucleic acids. J Mol Evol 4:261–276

Lehmann AR, Karran P (1981) DNA repair. Int Rev Cytol 72: 101–146

Li, W-H (1981) Simple method for constructing phylogenetic trees from distance matrices. Proc Natl Acad Sci USA 78: 1085–1089

Manavalan P, Ponnuswamy PK (1978) Hydrophobic character of amino acid residues in globular proteins. Nature 275:673–674

McLachlan AD (1971) Tests for comparing related amino acid sequences: cytochrome c and cytochrome c_{551}. J Mol Biol 61: 409–424

Modiani G, Battistuzzi G, Motulsky AG (1981) Nonrandom patterns of codon usage and of nucleotide substitutions in human and β-globin genes: An evolutionary strategy reducing the rate of mutations with drastic effects? Proc Natl Acad Sci USA 78:1110–1114

Ohta T, Kimura M (1971) Amino acid composition of proteins as a product of molecular evolution. Science 174:150–153

Olsen KW (1980) Internal residue criteria for predicting three dimensional protein structure. Biochim Biophys Acta 622: 259–267

Oobatake H, Ooi T (1977) An analysis of nonbonded energy of proteins. J Theor Biol 67:567–587

Ozoline ON, Oganesjan HG, Kamzolova SG (1980) On the fidelity of transcription by *Escherichia coli* RNA polymerase. FEBS Lett 110:123–125

Papentin F (1973) A Darwinian evolutionary system. Part II: Experiments on protein evolution and evolutionary aspects of the genetic code. J Theor Biol 39:417–430

Pieczenik G (1980) Predicting coding function from nucleotide sequence or survival of "fitness" of tRNA. Proc Natl Acad Sci USA 77:3539–3543

Ponnuswamy PK, Muthusany R, Manavalan P (1982) Amino acid composition and thermal stability of proteins. Int J Biol Macromol 4:186–190

Richard AB, Barry NJ, Divulet FE, Garner WH, Lehmann LD, Gurd FRN (1980) Evolution of the amino acid substitution in the mammalian myoglobin gene. J Mol Evol 15:197–218

Robson B, Suzuki E (1976) Conformational properties of amino acid residues in globular proteins. J Mol Biol 107:327–356

Salemme FR, Miller MD, Jordan SR (1977) Structural convergence during protein evolution. Proc Natl Acad Sci USA 74:2820–2824

Sander C, Schulz GE (1979) Degeneracy of the information contained in amino acid sequence: evidence from overlaid genes. J Mol Evol 13:245–252

Singleton R, Middaugh CR, McElroy RD (1977) Comparison of proteins from thermophilic and nonthermophilic sources in terms of structural parameters inferred from amino acid composition. Int J Pept Protein Res 10:39–50

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

Sternberg HJE, Thornten YM (1977) On the conformation of proteins: hydrophobic ordering of strands in β-pleated sheets. J Mol Biol 115:1–17

Topal MD, Fresco JR (1976) Complementary base pairing and the origin of substitution mutations. Nature 263:285–289

Vogel F, Kopun M (1977) Higher frequencies of transitions among point mutations. J Mol Evol 9:159–180

Vogel F, Röhrborn G (1966) Amino-acid substitution in hemoglobins and the mutation process. Nature 210:116–117

Wain-Hobson S, Nussinov R, Brown RJ, Sussman JL (1981) Preferential codon usage in genes. Gene 13:355–364

Weinberg G, Ullman B, Martin DW (1981) Mutator phenotypes in mammalian cell mutants with distinct biochemical defects and abnormal deoxyribonucleoside triphosphate pools. Proc Natl Acad Sci USA 78:2447–2451

Woese CR (1973) Evolution of the genetic code. Naturwissenschaften 60:447–459

Wolfenden RV, Cullis PM, Southgate CCF (1979) Water, protein folding and the genetic code. Science 206:575–577

Wolfenden RV, Andersson L, Cullis PM, Southgate CCF (1981) Affinities of amino acid side chains for solvent water. Biochemistry 20:849–855

Wolkenstein MV (1979) Mutation and the value of information. J Theor Biol 80:155–169

Wong JT (1975) A co-evolution theory of genetic code. Proc Natl Acad Sci USA 72:1909–1912

Yano T, Hasegawa M (1974) Entropy increase of amino acid sequence in proteins. J. Mol Evol 4:179–187

Yarus M (1979) The accuracy of translation. Prog Nucleic Acid Res Mol Biol 23:195–225

Zamyatnin AA (1972) Protein volume in solution. Prog Biophys Mol Biol 24:107–123

Received June 9, 1983/Revised August 13, 1984