J. Mol. Evol. 11, 349–360 (1978)

Journal of Molecular Evolution

Evaluation of Compositional Nonrandomness in Proteins

R. Holmquist

Space Sciences Laboratory, University of California at Berkeley, Berkeley, California 94720, USA

Summary. Cornish-Bowden and Marson have recently suggested that the finite sampling component of Q, a measure of nonrandomness in the amino acid composition of proteins , may have been underestimated because it was calculated on the basis of the genetic code table frequencies rather than on the basis of the average natural abundance with which the twenty amino acids actually occur in proteins. This underestimate would lead to an overestimate of Q_c a measure of selective effects above and beyond those imposed by the average natural abundance of the amino acids. In this paper the finite sampling component of Q is quantitatively estimated on the basis of these natural abundances and found to reduce Q_c from its previous average value of 24.3 to the lower value of 9.7, with the standard deviation of the population of Q_c values being 12.5. Individual Q_c values are given for 81 protein families of mean composition per 61 codons of $Ala_{5.3}$ $Arg_{2.4}$ $Asn_{3.0}$ $Asp_{3.6}$ $Cys_{1.5}$ $Gln_{2.6}$ $Glu_{3.5}$ $Gly_{4.7}$ $His_{1.3}$ $Ile_{3.4}$ $Leu_{4.5}$ $Lys_{4.2}$ $Met_{1.0}$ $Phe_{2.3}$ $Pro_{2.3}$ $Ser_{4.2}$ $Thr_{3.6}$ $Trp_{0.8}$ $Tyr_{2.6}$ $Val_{4.2}$. The mean Q_c value of 9.7 is notably small, and indicates that quantitatively minimal adjustments away from the average protein composition are necessary to maintain many different biological functions. This small value, however, is shown to differ significantly from the value of zero expected were the natural abundances of the amino acids the only selective constraint. These small deviations from the natural abundances are thus effectively selected for in the Darwinian sense.

Key words: Amino acid composition of proteins – Compositional selection for biological function

Introduction

The relationships between the chi-square distributed statistic $^{19}\chi^2$ (Laird and Holmquist, 1975) – designated Z by Cornish-Bowden and Marson (1977) – which describes the overall nonrandomness in amino acid composition for proteins, and

0022-2844/78/0011/0349/$ 02.40

the twenty binomially distributed statistics ${}_p\chi_i^2$ defined for each amino acid type are simply summarized:

$$Z = {}^{19}\chi^2 = T \sum_{i=1}^{20} {}_p\chi_i^2 = T \sum_{i=1}^{20} \frac{(p_i - e_i)^2}{e_i} \qquad (1)$$

where T is the number of amino acid residues in the protein and p_i and e_i are the observed and expected (under the hypothesis of interest) proportions of amino acid i.

Another statistic which has been used (Holmquist, 1975; Holmquist and Moise, 1975) to describe overall compositional nonrandomness is

$$Q = 100 \sum_{i=1}^{20} |p_i - e_i'| \quad , \qquad (2)$$

where e_i' was taken as $N_i/61$, N_i being the codon multiplicity of amino acid i in the genetic code table (N_i is 3 for isoleucine, for example). The expectation values, $\langle Z \rangle$ and $\langle Q \rangle$, of Z and Q are nonzero, the expectation of the former being 19 and independent of protein length, and that of the latter being closely proportional to $1/\sqrt{T}$ and hence dependent on protein length.

Biologically Q is the sum of the excess or deficit of the experimentally observed proportion of each amino acid relative to the proportion expected from the genetic code table. Because of this direct conceptually simple interpretation, for which there is no clear analogue in Z, our preference has been to deal directly with Q or its related measure Q_c:

$$Q_c = Q - \langle Q \rangle \quad , \qquad (3)$$

this form being chosen because Q_c has an expectation value of zero if the amino acids in the protein examined are multinomially distributed with their expected proportion being given by the genetic code table.

Distributions of ${}^{19}\chi^2$ and Q

Despite the conceptual simplicity of Q, the occurence of the absolute values in the summands in Eq. 2 sufficiently complicates the calculation of statistical tables of critical values so that none have yet been published for determining the significance level for a given observed value of Q. On the other hand ${}^{19}\chi^2$ has the classical chi-square distribution with 19 degrees of freedom for which tables of critical values are readily available. Moreover, the significance level of each of the twenty components, ${}_p\chi_i$, of ${}^{19}\chi^2$, can be determined as described in Laird and Homquist (1975). Because of this simplicity in the mathematics of $Z = {}^{19}\chi^2$, Z may be more amenable to further development. Cornish-Bowden and Marson (1977) have made an interesting start

along this line in determining <Z> for what they call a "partially random protein": that is a protein which contains a total of T residues for which M are essential in the sense that any mutation other than a synomymous one among these M residues is fatal, while the compositon of the remaining T−M residues are multinominally determined.

Estimating the Effects of Natural Selection

Because any pattern of amino acid compostion, nonrandom or random, could have resulted from the superposition of various selective factors, it is not in principle. possible to isolate the selective component from the stochastic (Holmquist, 1976) using protein sequence data alone.

It is meaningful however to consider one by one the individual constraints upon amino acid composition which result from the interplay between individual well-defined components of natural selection and particular well-defined stochastic processes, and to ask, and answer, the question "How much of the experimentally observed data can be explained by these posited components?"

As a concrete illustation of this approach let us posit for the stochastic component that amino acid composition is multinomially determined. For the selective component we consider two possibilities: the expected frequency of each amino acid is
Hypothesis 1) *proportional to its codon multiplicity in the genetic code table.*
Hypothesis 2) *proportional to its average natural abundance in proteins.*

It is worth noting that because the multinomial distribution is a function of the expected frequencies of the twenty amino acids, even at this elementary level, which is the most simple possible consistent with remaining at all faithful to the biology, the stochastic and selective components are not separable and it is accordingly meaningless to ask which is more important: both are.

Taking as a measure of the overall nonrandomness Q (Eq. 2), because we are familiar with it (other measures such as the Z values of Cornish-Bowden and Marson (1977) or the compositional entropy D_1 of Gatlin (1974) or the somewhat similar measure S of Vogel (1975) lead to analogous results), the expectation value of Q, <Q>, and the population standard deviation σ_Q can be calculated under the two hypotheses by Monte-Carlo simulation (Holmquist and Moise, 1975), or analytically (Cornish-Bowden and Marson, 1977). Expected values of Q under the two hypotheses are in Table 1. The average natural abundance of each amino acid was calculated for 81 protein families (Table 2) comprising 189 individual sequences, and is per 61 codons $Ala_{5.3}$ $Arg_{2.4}$ $Asn_{3.0}$ $Asp_{3.6}$ $Cys_{1.5}$ $Gln_{2.6}$ $Glu_{3.5}$ $Gly_{4.7}$ $His_{1.3}$ $Ile_{3.4}$ $Leu_{4.5}$ $Lys_{4.2}$ $Met_{1.0}$ $Phe_{2.3}$ $Pro_{2.3}$ $Ser_{4.2}$ $Thr_{3.6}$ $Trp_{0.8}$ $Tyr_{2.6}$ $Val_{4.2}$. These frequencies do *not* reflect the codon multiplicities for each amino acid as given by the genetic code table. We have described this phenomenon for a somewhat smaller collection (68 families) as selection against the genetic code (Jukes, Holmquist, and Moise, 1975). For the present collection this phenomenon is graphically depicted in Figure 1.

From its definition Q must lie between 0 and 200. The highest value achievable in practice is 197 for polytryptophan (Cornish-Bowden and Marson, 1977). Note in Table 1 that even under the first hypothesis expected values of Q are greater than zero, because the finite length of real sequences prevents the observed number of residues

Table 1. Expectation value <Q> and polulation standard deviation σ_Q (in parentheses) for different protein lengths T (in residues) under two hypotheses

T	Code-Random[a]	Abundance-Random[b]
1	187.3(5.35)	188.8(5.00)
2	175.7(7.66)	177.9(7.31)
3	165.0(9.67)	168.5(9.06)
4	154.9(10.66)	159.0(10.55)
5	146.1(11.84)	151.0(11.69)
6	136.7(12.66)	142.1(12.77)
7	128.7(13.24)	134.7(13.30)
8	121.6(14.03)	127.8(14.46)
9	114.5(14.09)	120.7(14.59)
10	107.7(14.91)	114.5(15.27)
11	103.9(13.93)	109.9(14.69)
15	88.2(13.30)	95.1(14.04)
20	77.1(11.94)	83.8(12.82)
40	53.7(9.98)	63.4(9.87)
60	42.2(7.88)	53.9(8.98)
80	37.8(6.81)	50.2(7.92)
100	33.7(5.97)	47.4(7.25)
140	28.4(4.82)	43.8(6.23)
220	22.6(3.92)	40.3(5.01)
300	19.3(3.52)	38.7(4.37)
360	17.6(3.32)	37.8(3.92)
400	16.7(3.19)	37.4(3.92)
500	15.0(2.60)	36.7(3.68)
∞	0(0)	33.86(0)[c]

[a] Amino acids were generated with a frequency proportional to their codon multiplicity in the genetic code table. The results are plotted in Figure 2 (lower solid curve)

[b] Amino acids were generated with a frequency proportional to their average natural abundance in proteins (see text). These results are plotted in Figure 2 (upper solid curve)
For each protein of T residues, 1000 sequences of length T were generated and the average Q and standard deviation for these 1000 sequences were calculated. Convergence was satisfactory, the values at 500 and 1000 iterations being within about 1% of each other

[c] This value is exact as it was calculated directly from the differences between the natural abundance frequencies and the genetic code table frequencies and not by simulation

of a particular amino acid from always assuming a value which is an integral multiple of its codon multiplicity. Even for proteine whose lengths are integral multiples of 61, statistical scatter will rarely permit all twenty amino acids to be present at exact multiples of their codon multiplicities as given in the genetic code table. The value for <Q> at T = ∞ (last entry, 33.86, of column 3 in Table 1) for the second hypothesis (abundance-random) does not tend to zero because in the definition of Q (Eq. 2) the e_i' are the code-random not abundance-random expected frequencies. This convention was chosen because the e_i' are stable, i.e. true constants, whereas the natural abundances

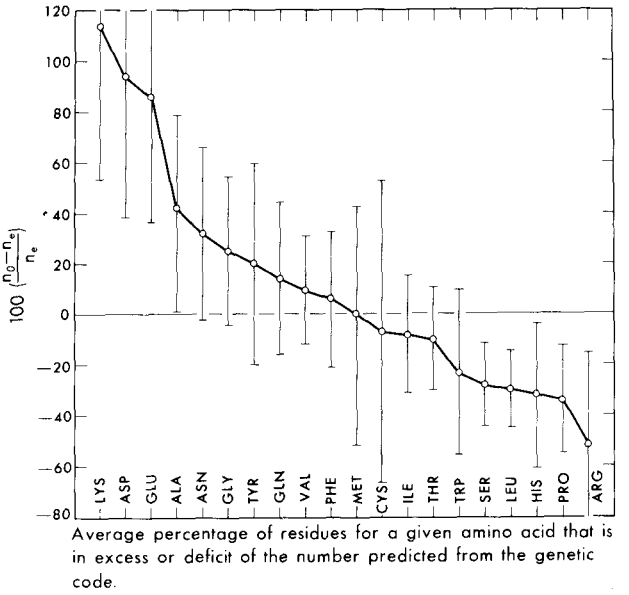

Fig. 1. Selection against the genetic code. The ordinate expresses the excess or deficit of the observed number n_o of residues of each of the twenty amino acids relative to the expected number n_e of the collection of proteins in Table 2. The vertical bars span one standard deviation

shift somewhat depending on the data collection from which they are culled. The approximate inverse dependence of Q on the square-root of protein length is illustrated by the solid curves in Figure 2.

Experimental values of Q (the open circles in Figure 2) were calculated for this collection of 81 protein families, no family being represented more than once in this collection which represents a cross-section of biological function and specificity

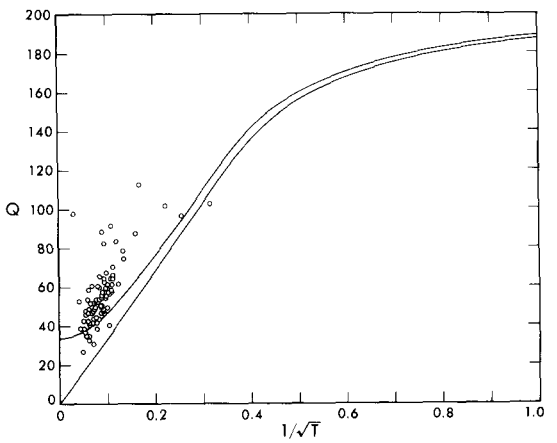

Fig. 2. Q as a function of protein length; open circles are the experimental values for 81 protein families, the lower solid curve represents the values expected for a protein of the given length under Hypotheses 1 in text, and the upper solid curve the values expected under Hypothesis 2

Table 2. Q_c' values for 81 protein families comprising 189 individual sequences

Protein[a]	Length in Residues	Q_c'
Actin	357	0
Acyl carrier protein	77	19
Adrenodoxin	114	2
Alcohol dehydrogenase	374	4
α-S1-casein	199	6
Aspartate aminotransferase	412	-11
Amyloid protein	76	16
Avidin	128	10
Azurin	129	5
β–casein	209	19
Carbonic anydrase B	260	-8
Carboxypeptidase A	307	-4
Coat protein	117	1
Chymotrypsinogen A & B	245	0
Cytochrome b_5	90	7
Cytochrome c	106	12
Cytochrome c_2	112	18
Cytochrome c_3	102	20
Cytochrome c_{551}	82	12
Cytochrome $c_{551.5}$	68	31
Cytochrome c_{553}	82	41
Deoxyribonuclease A	257	-5
DNA binding protein	88	-9
Elastase	240	1
Endolysin	154	-5
Ferredoxin plants	97	2
Ferrodoxin bacterial	55	22
Fibrinopeptide A	15	1
Fibrinopeptide B	20	17
"Fibrous" proteins	983	62
Flavodoxin	137	21
Gastric juice peptide	10	-13
γ-crystallin	165	2
Glyceraldehyde-3-PO_4-dehydrogenase	133	9
α-l-acid glycoprotein	181	7
Glutamate dehydrogenase	499	1
Haptoglobin	114	18
Hemerythrin	113	16
Hemoglobin	143	9
High potential iron protein	86	14
Histones	123	4
Immunoglobulin	160	-2
Keratin	98	11
Lactalbumin	123	12
Lactogen	191	2
Lactoglobulin	162	10
Lac-repressor	347	-2
Leghemoglobin	142	15
Lipoprotein	79	14
Luteinizing hormone α-chain	96	9
Lysozyme	141	1
α-lytic protease	198	10

Table 2 (continued)

Protein[a]	Length in Residues	Q_c'
Muscle calcium binding protein	108	35
Myelin membrane protein	170	6
Myoglobin	152	17
Nerve growth factor	118	2
Neurotoxins	64	8
Neurophysin II	97	14
Protease inhibitor proteins	84	8
Papain	212	5
Penicillinase *S.aureus*	257	18
Penicillinase *B.licheniformis*	265	7
Pepsin	327	6
Phospholipase A	129	2
Proinsulin	81	7
Prolactin	198	-11
Protamines	34	45
Nuclease	149	9
Ribonuclease	125	4
Ribonuclease (Barnase)	110	0
Ribonuclease T	104	9
Ribosomal protein 50S A1	120	42
Rubredoxin	53	18
Serum albumins	580	16
Steroid-Δ-isomerase	125	10
Subtilisin	275	14
Thermolysin	316	9
Thioredoxin	108	8
Tryptophan synthetase	267	3
Triose phosphate isomerase	248	8
Trypsinogen	229	11

[a]The specific sequences used are the 78 in Table 1 of Holmquist and Moise (1975) plus DNA binding protein (Nakashima et al., 1974), "fibrous proteins" (myosin, collagen, elastin, and silk fibroin; see Table 5–3 in Lehninger, 1970), and the serum albumins (bovine fraction V, Brown, 1975; and human serum albumin, Behrens et al., 1975)

(see Table 2). It is clear that neither hypothesis adequately explains the experimental data, though as might be expected, the second does better than the first.

In practical terms Figure 2 tells us that the stochastic component determining protein composition is more complex than a simple multinomial distribution would lead us to believe and/or there are additional selective constraints determining protein composition beyond the limits imposed by the natural-abundance of each amino acid. The distinction between "essential" and "non-essential" amino acid residues mentioned above in connection with the recent work of Cornish-Bowden and Marson (1977) may help in understanding these additional constraints.

The excess of the experimentally determined Q value and that expected from the natural abundance (Hypothesis 2) of each amino acid (the difference between

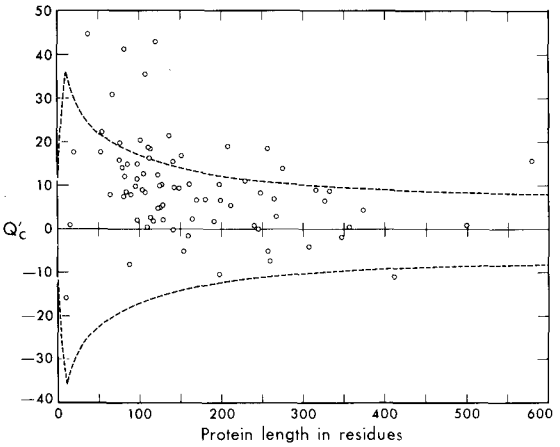

Fig. 3. The excess, Q'_c, of the experimentally determined Q value and that expected from the natural abundance of each of the twenty amino acids and finite protein length, as a function of protein length; open circles are the experimental values. Dashed lines represent the 2.326 σ margins and are given to give some idea of the expected scatter at each length. Because Q'_c is not normally distributed, these limits are not meant to be used to determine statistical significance levels

the ordinates of the experimental points and the upper curve in Fig. 2), designated Q'_c is plotted in Figure 3; the values for particular proteins are in Table 2. The mean experimental Q'_c value is 9.7 and the population standard deviation is 12.5.

The mean excess, $<Q_c>$, of the experimentally determined Q values over that expected on the basis of Hypothesis 1) was previously shown (Holmquist and Moise, 1975) to be 24.3, with a population standard deviation of 10.3. Cornish-Bowden and Marson (1977) are thus correct in their observation that Holmquist and Moise (1975) probably underestimated the magnitude of the finite sampling component of Q by using code-random frequencies rather than frequencies proportional to the natural abundance of each amino acid, the quantitative effect on the finite sampling component being given in Table 1, and leading to the lower mean of 9.7 vs. 24.3 and lower individual Q'_c values (Table 2) relative to the Q_c values calculated by Holmquist and Moise (1975).

Is the Observed Q'_c Mean Significantly Different from That Expected from the Natural Abundance of the Amino acids?

In the preceding section the basic experimental results of measurements on amino acid composition were set forth, and it was suggested that there are additional selective constraints determining protein composition above and beyond those imposed by the natural abundance of each amino aicd.

In this section the justification for the existence of these additional selective constraints is given.

As the population of Q'_c values has a finite variance and mean, and here the sample size is large (n = 81), the central-limit theorem applies (Mood, 1950a) and the

distribution of the mean approaches the normal distribution with variance σ^2/n. Estimating σ by the (unbiased) population standard deviation, the standard deviation of the mean is $\sqrt{[(12.5)^2/81]}$ or 1.39. The 99% confidence limits on this mean are given by 2.58 σ_{mean}. Thus $\langle Q_c' \rangle = 9.7 \pm 3.6$. To check that the sample size was indeed large enough for these calculations to be valid, one may note that in the limit of large n, the near normalcy of the distribution of the mean implies that the sample means be distributed approximately symmetrically about the true population mean. In this limit the expectation value of the median of the distribution must approach the mean itself. And provided the distribution can be reasonably approximated by a continuous distribution function which is not grossly asymmetrical, the sample median will be near the sample mean. To see if this is true for the sample considered here, the (unrounded) Q_c' values, from which the rounded values in Table 2 were obtained, were ordered from the largest to smallest; the median, ν, was found to be $\nu = 8.2$. It is not necessary to know the distribution of Q_c' to find confidence limits for the median as these are readily found by "distribution-free" or "non-parametric" methods which in the present case place 99% confidence limits on the median of $\nu = 8.2 \pm 2.9$ (Mood, 1950b), in reasonable agreement with the values found from the central-limit theorem above.

From this anaysis, it is clear that the mean of the experimental distribution of Q_c' is not zero as would be expected under Hypothesis 2. However in order to show conclusively that the observed mean differs *significantly* from zero, i.e., could not with reasonable probability have arisen from sampling error if Hypothesis 2 were true, the mean and standard deviation of a sample set of 81 proteins, with the specified lengths in Table 2, was calculated by Monte-Carlo methods under Hypothesis 2. Four such replications gave mean Q_c' values of -0.50, 0.73, 1.12, and 1.04. The combined mean and population standard deviation were 0.60 and 13.17 respectively; the standard deviation of the mean was 0.75 in satisfactory agreement with the central-limit theorem value of the population standard deviation divided by the square root of the sample size: $13.17 \sqrt{(81 \times 4)} = 0.73$. (This serves as an additional check that the sample size is large enough for the central-limit theorem to hold.) The probability that the observed difference in means of 9.1 (i.e. 9.7−0.6) could have arisen under Hypothesis 2 is less than 10^{-8} ($t = 5.76$).

The natural abundances of the amino acids thus does not suffice to explain the experimentally observed range of amino acid compositions. Additional selective constraints exist. Only in a few fortunate situations are there sufficient data to point directly at the nature of these additional constraints: one such case is in the hemoglobins, where a consideration of the contact regions between the α- and β- chains, clearly shows some of the additional constraints that can be involved (Vogel and Zuckerkandl, 1972; Goodman et al., 1975).

Discussion

In interpreting the above values, it is necessary to distinguish clearly between the magnitude of selective effects, and the effectiveness of selective effects. A small structural stress in an airplane wing can have a quite effective (and fatal) "selective" effect. Whereas in population genetics it makes sense to talk about the strength of selection in terms of the magnitude of the selection coefficient s, because the magnitude of s implies

a definite rate of *dynamic* changes in the population leading to some final equilibrium values, in discussing compositional changes, the measures Z, Q, or S do not play a role analogous to that of s. These measures do not imply anything about the dynamic changes in amino acid composition as evolution proceeds; rather they summarize the resultant of these changes at the present time as observed experimentally in protein sequences isolated from contemporary organisms. It is thus necessary to clearly maintain both the real and semantic distinctions between the terms *magnitude, strength*, and *effectiveness*. They are not synonymous, and particular restraint should be exercized in drawing analogies and using terms developed in other disciplines in situations in molecular studies to which they do not correctly apply. Just as the observation of randomness or near-randomness in some measured character does not imply selective neutrality or near-neutrality, the observation of a character of small magnitude does not imply its impotence.

A Q_c value of 24, the excess of the observed Q value beyond that expected from the genetic code table, is reasonably called moderate, and a Q_c' value of 9.7, the excess of the observed Q value beyond that expected from the natural abundance of the amino acids is reasonably called small.

It is not reasonable to "wave these figures away" by stating that these low values "....may simply reflect the fact that grossly abnormal compositions are required to produce high values of the index." (Cornish-Bowden and Marson, 1977). The compositional experimental data are quite clear: glossly abnormal compositions simply *do not occur*, with the infrequent exceptions of highly specialized proteins such as the histones. The compositional deviations *are* small, on the average, and this is fully compatible with selection pressure to maintain these small deviations. Bigger is clearly not better in this case.

Why Large Deviations from Compositional Norms Are So Rare

Earlier we have stated that the small magnitude of compositional deviations may be the result of selection for molecules with high potentials for adapting to environmental changes and thus is no evidence for selective neutrality or otherwise weak selective effects. (Holmquist, 1975, Holmquist and Moise, 1975). In this view selection has been for those structures having compositions not far displaced from the compositional norm given by the natural abundances of the amino acids, and adequate selective reasons exist to explain these deviations (Jukes, Holmquist, and Moise, 1975).

Another reason also exists. The function of a protein is determined importantly by its three-dimensional structure. There exists a subset of sequences, consisting of a great many primary amino acid sequences, and hence compositions, which are compatible with this three dimensional structure within a given range of compositional limits not far from the norm. It is thus unneceassary to select for molecular forms outside this subset expect for the most specialized functions. Indeed since such grossly compositionally abnormal forms are rare, they have clearly been selected against in most cases.

Within a given protein family, selection has been still stronger, narrowing the subset of near abundance-random sequences still more as evidenced by the covarion estimates of Fitch and Markowitz, (1970), our own estimates of the number of varions T_2 (Moore et al., 1976; Holmquist et al., 1976) and the calculation of the

number of essential residues M for a partially random protein by Cornish-Bowden and Marson (1977).

Conclusions

Darwinian selection is fully compatible with the observed moderate deviations from the genetic code table frequencies and small deviations from the natural abundances of each amino aicd. These small to moderate deviations are not an artifact of the mathematical structure of the various measures of compositional nonrandomness now in use, but reflect positive Darwinian selection for molecular structures not far from the norm. All of these quantitative measures D_1 (Gatlin, 1974); Q (Holmquist, 1975; Holmquist and Moise, 1975), compositional entropy S (Vogel, 1975), and Z, i.e., chi-square, (Cornish-Bowden, 1977) are consistent with respect to the weak to moderate magnitude of the compositional deviation, and in attributing them to positive Darwinian selection for functional structures. It appears that quantitatively minimal adjustments away from the average composition are necessary to maintain very different protein functions. Though small in magnitude, these deviations are effectively selected for; otherwise the occurence of proteins with compositions more deviate from the norm would be commonplace rather than rare.

Acknowledgments. This work was supported by NSF grant PCM 76–18627 and NASA grant NGR 05–003–460.

References

Behrens, P., Spiekerman, A., Brown, J. (1975). Fed. Proc. **34**, 591. Abstract 2106
Brown, J. (1975). Fed. Proc. **34**, 591. Abstract 2105
Cornish-Bowden. A., Marson, A. (1977). J. Mol. Evol. **10**, 231–240
Fitch, W., Markowitz, E. (1970). Biochemical Genetics, **4**, 579–593
Gatlin, L. (1974). J. Mol. Evol. **3**, 189–208
Goodman, M., Moore, G., Matsuda, G. (1975). Nature **253**, 603
Holmquist, R. (1975). J. Mol. Evol. **4**, 277–306
Holmquist, R., Moise, H. (1975). J. Mol. Evol. **6**, 1–14
Holmquist, R. (1976). Random and Nonrandom Processes in the Molecular Evolution of Higher Organisms. In: Molecular Anthropology, M. Goodman, R. Tashian, J. Tashian eds., pp. 89–116. New York: Plenum
Holmquist, R., Jukes, T., Moise, H., Goodman, M., Moore, G. (1976). J. Mol. Biol. **105**, 39–74
Jukes, T., Homquist, R., Moise, H. (1975). Science **189**, 50–51
Laird, M., Holmquist, R. (1975). J. Mol. Evol. **4**, 261–276
Lehninger, A.L. (1970). Biochemistry, p. 94, New York: Worth
Mood, A.M. (1950a). Introduction to The Theory of Statistics, p. 136, New York: McGraw-Hill
Mood, A.M. (1950b). Introduction to The Theory of Statistics, p. 385, New York: McGraw-Hill

Moore, G., Goodman, M., Callahan, C., Holmquist, R., Moise, H. (1976). J. Mol. Biol. **105**, 15–37

Nakashima, Y., Dunker, A., Marvin, D., and Konigsberg, W. (1974). FEBS letters **40**, 290–292

Vogel. H. (1975). J. Mol. Evol. **6**, 271–283

Vogel, H., Zuckerkandl, E. (1972). The Evolution of Polarity Relations in Globins. In: Proceedings of the Sixth Berkely Symposium on Mathematical Statistics and Probability, Volume 5: Darwinian, Neo-Darwinian, and Non-Darwinian Evolution, L.M. LeCam, J. Neyman, E. Scott eds., pp. 155–176. Berkeley: University of California Press

Received February 13, 1978; Revised April 4, 1978