Non-randomness of nucleotide bases in mRNA codons

By R. F. NASSAR
Department of Statistics, Kansas State University

and R. D. COOK
School of Statistics, University of Minnesota, St Paul

(*Received* 14 *January* 1975)

SUMMARY

Maximum likelihood estimates of codon and base frequencies from observed amino acid composition of proteins were obtained based on models capable of revealing dependency between base arrangements in the three positions of a codon. Results showed that many of the proteins analysed revealed dependency between base arrangements in the first and second codon positions (first-order interaction). Also, in a number of proteins the interactions between base arrangements seemed to involve simultaneously more than one first order interaction and/or a second-order interaction (among base arrangements in the three codon positions). It was of interest to observe that the model of random base arrangements did not fit the observed amino acid data in almost all of the proteins that were analysed. More than ten amino acids contributed to this deviation from randomness.

1. INTRODUCTION

Since the discovery in 1966 of a large amount of genetic variability in natural populations for genes at the molecular level (Lewontin & Hubbey, 1966; Harris, 1966), population geneticists have been attempting to explain how such variability is maintained and to determine its evolutionary importance. According to Kimura & Ohta (1971*a*, *b*) and King & Jukes (1969), who postulated what has been termed the 'neutral' or 'non-Darwinian' hypothesis, genetic variability revealed by electrophoresis for enzymes and other proteins is selectively neutral and the differences in amino acid composition of proteins arise largely without natural selection.

In the study of evolution at the molecular level, much attention has been directed at estimating base frequencies in RNA codons from the amino acid composition of proteins. Conclusions in support of the neutral hypothesis have been drawn from the finding that the amino acid composition of proteins can be predicted fairly well from knowledge of base frequencies and by assuming a random arrangement of the four kinds of nucleotide bases (King & Jukes, 1969; Kimura, 1968; Ohta & Kimura, 1970). But even if the base arrangement were random, some

* Contribution no. 229, Department of Statistics and the Statistical Laboratory of the Kansas Agricultural Experiment Station.

https://doi.org/10.1017/S0016672300016578 Published online by Cambridge University Press

researchers doubt the validity of the implication that amino acid substitutions in evolution could have resulted from random fixation of selectively neutral mutants (Stebbins & Lewontin, 1972; Clarke, 1969). Unfortunately, however, because of the *ad hoc* nature of the methods used in estimating base frequencies, the conclusion drawn in support of randomness is equivocal. The need for better methods of estimation is of major importance and has been recognized by others (Kimura & Ohta, 1972).

In a previous communication (Cook & Nassar, 1975) we have discussed four loglinear models that described the distribution of bases (by position) within a codon; and presented a maximum-likelihood procedure to estimate codon and base frequencies in the light of these models. In this study, we apply the procedure to 26 different proteins for estimation of codon and base frequencies and for revealing relationships among base arrangements in the three positions within a codon.

2. METHODS

The procedure we consider will be directed at estimating the individual cell probabilities in Table 1 from observed amino acid frequencies. Note that Table 1 is a condensation of the original $4 \times 4 \times 4$ standard RNA code table by combining U and C in the third codon position. This was done because no estimation procedure can distinguish between U and C in this position. Because multiple codons can code for the same amino acid, the code table contains considerable indeterminacy in the sense that frequencies of codons that code for the same amino acid will be mixed up. The problem is further confounded by the three chain-terminating codons representing *a priori* zeros in the table. Also, in some table cells, additional zeros might have to be invoked for proteins where an amino acid is not present. Such zeros are referred to as structural zeros to distinguish them from cells containing observed zeros as a result of sampling variation (Fienberg, 1972). A contingency table containing structural zeros is said to be incomplete.

The problem may now be considered as one of estimating cell probabilities in a $4 \times 4 \times 3$ incomplete contingency table with mixed up frequencies. This problem has been considered by Cook & Nassar (1975) and the interested reader may wish to consult their article for details on the theory and estimation procedure.

In our procedure of estimation we rely on four loglinear models for the underlying cell probabilities. In these models we make assumptions only about the joint probability of occurrence of bases in the three codon positions. From Table 1, it is clear that data on the amino acid composition of proteins can be used to obtain information on the distribution of the bases in the three positions within a codon.

The interest in model 1 is to determine whether the observed frequencies of the amino acids in a protein can be predicted by assuming independence between base arrangements in positions 1, 2 and 3 within a codon. Because of the structural zeros in the code table the term independence between bases should refer to that portion of the table not containing the structural zeros. This phenomenon is referred to as quasi independence (Cook & Nassar, 1975). In fitting model 2, we

https://doi.org/10.1017/S0016672300016578 Published online by Cambridge University Press

ask whether the observed frequencies of the amino acids in a protein could be predicted by assuming independence between base arrangement in the third position in a codon and arrangements in the first and second positions taken together. In other words the model assumes a dependency between base arrangements in the first and second positions only. Model 3 assumes a dependency between base arrangements in positions 1 and 3 only, and in model 4 we assume that a dependency exists only between base arrangements in positions 2 and 3.

Table 1. *RNA code table with uracil and cytosine combined in the third codon position*

1 \ 2	U	C	A	G	2 / 3
U	Phe	Ser	Tyr	Cys	U+C
	Leu	Ser	Term.	Term.	A
	Leu	Ser	Term.	Try	G
C	Leu	Pro	His	Arg	U+C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn.	Ser	U+C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U+C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

In these four loglinear models the word dependency in a model can be interpreted also as an interaction. We can be talking, for example, about a dependency between base arrangements in position 1 and 2 and that could be taken to mean a first order interaction between bases in these two positions.

For each model we estimate the cell or codon frequencies and from that we obtain estimates of amino acid frequencies by adding the appropriate cell probabilities. The goodness of fit to each model is tested using the Pearson chi-square statistic on the amino acid classes.

The degrees of freedom for each model are obtained by substracting the number of free parameters for a model from the number of amino acids in a protein (Table 2 lists the number of amino acids in each protein). The number of free parameters are 9 for model 1, 18 for model 2 and 15 for models 3 and 4. For more details the reader is referred to Cook & Nassar (1975) In Table 2 it is seen that, for haptoglobin $\alpha 1$, Proinsulin, and Trypsin inhibitor, model 2 cannot be used because of the lack of degrees of freedom.

In this analysis, we are interested in determining if model 1 (random model)

https://doi.org/10.1017/S0016672300016578 Published online by Cambridge University Press

can predict adequately the amino acid composition of most proteins as has been claimed in the literature. We are also interested in isolating a model that can describe best the amino acid data of proteins. From such a model we can make inferences about the relationships among base arrangements in the three codon positions. It is of course possible that none of the models may be sufficient to completely describe a set of amino acid data. This is so, since for a complete description of the relationships between base arrangements one needs to consider the whole class of models obtainable from the general model (2) in Cook & Nassar (1975). We are confined to the four models because of the restricted degrees of freedom available. It is nevertheless informative to be able to say that none of the four models describe adequately an amino acid composition of a protein. This would imply that interactions between base arrangements do exist, but that they are not exactly what is specified in the models. In such a case, an idea of the type of interactions involved can be obtained from the general model by elimination.

Table 2. *Chi-square goodness of fit (calculated from the amino acid classes in a protein) to models 1, 2, 3 and 4*

(Degrees of freedom for each model are obtained by substracting the number of free parameters for that model from the number of amino acids in a protein. The number of free parameters are: model 1 = 9, model 2 = 18, model 3 = 15, model 4 = 15.)

		Model			
Protein	Amino acids	1	2	3	4
Cytochrome *c* (human)	20	26·72**	3·19	21·83**	25·66**
Cytochrome C551 (*Pseud. fluor.*)	20	17·31	2·19	13·94*	14·53*
Haemoglobin gamma (human)	20	21·78*	4·69	18·16**	20·76**
Haemoglobin-α (mouse)	20	38·54**	14·11**	18·63**	38·62**
Haemoglobin delta (human)	19	16·55	0·804	14·16**	13·66**
Myoglobin (horse)	19	42·65**	10·92**	29·93**	38·02**
Haemoglobin β (rabbit)	20	31·27**	5·08	25·86**	30·62**
Haptoglobin α1 (human)	18	19·45*		14·89*	17·91**
Nuclease (*staph. aureus*)	19	12·49	1·11	10·19*	11·76*
Ribonuclease (bovine)	19	20·53*	0·99	15·87**	12·16**
Bence Jones Kappa (human cum)	20	14·20	6·47*	6·78	13·71*
Bence Jones λ (human SH)	19	9·55	1·89	6·97	7·77
Subtilisin (*Bacillis subt.*)	19	37·56**	2·50	29·17**	26·52**
Typtophane synthetase α (*E. coli*)	19	30·05**	2·81	25·51**	22·91**
Proinsulin (pig)	17	11·76		8·29*	9·55**
Tobacco mosaic virus (Dahl.)	19	19·82*	7·75**	8·42	15·00**
Basic trypsin inhibitor (bovine)	18	19·36*		11·15*	19·51**
Glyceraldehyde-3-phosphate DH (pig)	20	23·92*	10·24**	13·62*	20·27**
Lysozyme (bacteriophage T_4)	20	14·42	3·34	11·67*	11·06*
Chymotrypsinogen A (bovine)	20	36·65**	7·30*	24·52**	28·88**
Trypsinogen (bovine)	20	50·94**	10·13**	35·85**	36·70**
Growth hormone (human)	20	20·32*	0·88	18·80**	17·63**
Elastase (pig)	20	47·06**	15·95**	23·47**	43·65**
Azurin (*Bordetella bronchiseptica*)	20	26·15**	10·02**	16·70**	23·52**
Hemerythrin (sipunculid worm)	20	24·73**	9·73**	15·76**	12·83**
Thioredoxin (*E. coli*)	20	27·41**	7·16*	18·14**	24·25**

* Significant at the 5 % level. ** Significant at the 1 % level.

3. RESULTS

We analysed 26 different proteins (Table 2) for mRNA codon frequencies (cell frequencies in Table 1) and for the marginal base frequencies in each codon position for each of models 1, 2, 3 and 4. Proteins were taken from Dayhoff (1972). For a large number of species, some of the proteins, like cytochrome *c* and haemoglobins, were very similar in their amino acid compositions. To avoid repetition, we chose in such instance one representative protein. Table 2 presents the chi-square goodness of fit for each model. Judging from the non-significant chi-squares, it is seen that the data best fit model 2. Seven proteins showed good fit to model 2; one protein to all 4 models; four to models 1 and 2; one to model 3; one to model 1; and one to models 1 and 3. Eleven proteins did not fit any of the models, as indicated by the significant chi-squares.

Model 1 can be considered a special case of models 2, 3 or 4. In such a case a method is known whereby the expected values (for each of the amino acid classes) under model 1 and each of 2, 3 and 4 can be compared to find out which model best fits the data (Fienberg, 1970). If E_{1i} and E_{2i} are the expected values in each amino acid class i ($i = 1, 2, \ldots, k$) for models 1 and 2, then the test statistic

$$X^2 = 2 \sum_{i=1}^{k} E_{2i} \ln \frac{E_{2i}}{E_{1i}} \tag{1}$$

has an asymptotic chi-square distribution with degrees of freedom equal to the difference between the degrees of freedom for models 1 and 2. The same equation can be used to compare models 1, 3 and 1, 4. Table 3 presents the comparisons for all proteins in Table 2 that showed good fit to more than one model or that did not fit any of the models. For the protein that fits all models (Bence Jones λ), it is seen that model 1 is as good as 2, 3 or 4 in describing this amino acid composition. For the delta haemoglobin and myoglobin, model 2 is a better fit than model 1; and 2 and 3 are better than 1 for haemoglobin α. For chymotrypsinogen, trypsinogen, elastase, azurin, hemerythrin, cyt. *c* and thioredoxin it is seen that models 2, 3 and 4 are better fits than is model 1. For the six remaining proteins in Table 3, equal fit by all four models seems possible, although model 2 is favoured over 1 for nuclease, G3 pdh and lysozyme, and 3 over 1 for Bence Jones kappa (as judged by the P values for these chi-squares).

The estimates of base frequencies (Table 4, model 2) showed on the average an excess of A to U and of G to C in codon position 1, and an excess of A to U and C to G in codon position 2. In position 3, however, it seems likely that all 4 bases are equal in their frequency of occurrence. Table 5 presents the average cell frequencies (codons) over the 12 proteins adequately fitted by model 2. It is seen that there is a good deal of heterogeneity in the probability of occurrence of the codons that code for one amino acid.

DISCUSSION

It is clear from this analysis that the assumption of independence, or rather quasi-independence (because of the three terminal codons and other zero entries attributed to proteins with fewer than 20 amino acids), among the bases in codon

https://doi.org/10.1017/S0016672300016578 Published online by Cambridge University Press

Table 3. *Comparisons between model 1 and each of 2, 3 and 4 to determine which of the two best fit the observed amino acid data in a protein*

(Entries are chi-square values obtained from expression 1. Degrees of freedom for a chi-square are obtained as the difference between the degrees of freedom for the two models that are compared.)

Model ...	1,2	1,3	1,4
Haemoglobin delta (human)	15·19*	1·65	1·70
Haemoglobin-α (mouse)	19·96**	14·04*	0·464
Myoglobin (horse)	30·63***	9·38	2·35
Bence Jones λ (human SH)	7·71	2·32	2·97
Bence Jones Kappa (human cum)	7·50 (p = 0·58)	7·35 (p = 0·3)	0·56
Nuclease (*Staph. aureus*)	11·30 (p = 0·25)	2·18	0·79
Glyceraldehyde-3-phosphate dehydrogenase (pig)	12·07 (p = 0·22)	8·61 (p = 0·21)	1·26
Lysozyme (bacteriophage T_4)	11·70 (p = 0·23)	3·95	3·04
Chymotrypsinogen A (bovine)	20·88**	12·47*	5·32
Trypsinogen (bovine)	47·35***	17·28***	11·77*
Elastase (pig)	28·06***	16·93**	6·12
Azurin (*Bordetella bronchiseptica*)	16·38*	9·69	2·39
Hemerythrin (sipunculid worm)	12·73	5·87	7·73*
Thioredoxin (*E. coli*)	21·76***	8·36	1·41
Cytochrome C551 (*Pseud. fluor.*)	15·58*	16·07**	18·23**
Haptoglobin α1 (human)		4·64	4·48
Basic trypsin inhibitor (bovine)		0·69	0·09

* Significant at the 10 % level. ** Significant at the 5 % level.
*** Significant at the 1 % level.

positions 1, 2 and 3 is inadequate in explaining the majority of the observed amino acid data. Thus, randomness, as implied by model 1, does not seem to be as widely spread as suggested by the proponents of 'non-Darwinian' evolution. From considerations of the data in Tables 2 and 3, it seems clear that among the 26 proteins analysed perhaps four were adequately fitted by model 1, but only proinsulin was fitted by model 1 to the exclusion of the other models.

Model 2, postulating a dependency between bases in codon positions 1 and 2, seems to fit the largest portion of the proteins analysed. For some proteins (those having significant chi-squares for all our models, Table 2), the interactions between bases in the three positions are seen to be more complicated than the assumptions of the models. This implies that one is involved with the simultaneous presence of more than one first-order interaction (an interaction between bases in any of two positions) and/or a second-order interaction (among bases in the three positions). The emerging picture is that the base arrangement in the three positions of a codon is not random. Non-randomness is also reflected in the estimated base frequencies of Table 4. In base position 1, A was more frequent than U and G more frequent than C. In base position 2, A was still more frequent than U, but G less frequent than C. In base position 3, if one assumes that U and C are equally frequent as

https://doi.org/10.1017/S0016672300016578 Published online by Cambridge University Press

Table 4. *Estimates of marginal frequencies (from model 2) for each of the four bases in codon positions 1 and 2 and for A, G, U+C in codon position 3*

	Codon position										
	1				2				3		
	U	C	A	G	U	C	A	G	U+C	A	G
Cytochrome *c* (human)	0·179	0·097	0·406	0·317	0·221	0·179	0·423	0·176	0·337	0·420	0·241
Cytochrome C551 (*Pseud. fluor.*)	0·165	0·152	0·230	0·451	0·219	0·288	0·341	0·150	0·351	0·261	0·386
Haemoglobin γ (human)	0·230	0·157	0·248	0·363	0·301	0·236	0·315	0·146	0·450	0·126	0·423
Haemoglobin α (mouse)	0·196	0·222	0·198	0·382	0·262	0·309	0·297	0·129	0·593	0·684	0·338
Haemoglobin δ (human)	0·191	0·205	0·199	0·404	0·308	0·208	0·328	0·154	0·491	0·0830	0·425
Myoglobin (horse)	0·183	0·171	0·273	0·372	0·274	0·198	0·398	0·128	0·359	0·450	0·190
Haemoglobin β (rabbit)	0·225	0·177	0·199	0·397	0·315	0·216	0·342	0·125	0·449	0·067	0·483
Haptoglobin $\alpha 1$ (human)	0·136	0·208	0·292	0·361	0·168	0·196	0·469	0·164	0·539	0·303	0·156
Nuclease (*Staph. aur.*)	0·149	0·151	0·350	0·348	0·221	0·228	0·436	0·113	0·324	0·248	0·427
Ribonuclease (bovine)	0·256	0·133	0·334	0·274	0·169	0·313	0·379	0·138	0·531	0·0016	0·467
Bence Jones Kappa (human)	0·292	0·179	0·239	0·289	0·235	0·302	0·307	0·154	0·397	0·421	0·181
Bence Jones λ (human)	0·249	0·189	0·227	0·333	0·183	0·367	0·286	0·162	0·404	0·384	0·210
Subtilisin (*Bacillis subt.*)	0·175	0·130	0·252	0·441	0·240	0·352	0·244	0·161	0·626	0·055	0·318
Tryptophane synth. (*E. coli*)	0·185	0·191	0·232	0·389	0·299	0·285	0·280	0·133	0·405	0·393	0·200
Proinsulin (pig)	0·232	0·262	0·147	0·357	0·250	0·179	0·309	0·261	0·442	0·284	0·273
Tob. mosaic virus (Dahl.)	0·208	0·217	0·282	0·291	0·278	0·298	0·272	0·150	0·592	0·100	0·306
Trypsin inhibitor (bovine)	0·260	0·163	0·296	0·280	0·210	0·211	0·263	0·315	0·534	0·0	0·465
G3pdh (pig)	0·172	0·118	0·300	0·409	0·289	0·242	0·307	0·161	0·539	0·157	0·302
Lysozyme (bact. T_4)	0·151	0·174	0·350	0·323	0·274	0·195	0·335	0·195	0·483	0·161	0·355
Chymotrypsinogen A (bovine)	0·248	0·136	0·280	0·334	0·244	0·324	0·236	0·193	0·470	0·268	0·260
Trypsinogen (bovine)	0·280	0·148	0·273	0·296	0·227	0·280	0·296	0·195	0·537	0·321	0·140
Growth hormone (human)	0·318	0·149	0·234	0·297	0·297	0·211	0·351	0·139	0·436	0·350	0·213
Elastase (pig)	0·198	0·213	0·259	0·329	0·250	0·246	0·262	0·240	0·586	0·139	0·273
Azurin (*Bord. bronch.*)	0·181	0·128	0·279	0·410	0·271	0·232	0·341	0·155	0·493	0·039	0·467
Hemerythrin (sipunculid worm)	0·264	0·149	0·293	0·292	0·274	0·140	0·442	0·142	0·542	0·198	0·259
Thioredoxin (*E. coli*)	0·160	0·163	0·287	0·388	0·296	0·236	0·333	0·133	0·476	0·349	0·174
Mean	0·197	0·169	0·267	0·351	0·253	0·248	0·330	0·166	0·477	0·240	0·305

they seem to be in positions 1 and 2, the arrangement of the four bases would be close to random. That perhaps was to be expected, because in the RNA code table, the redundancy in the code results only from changes in bases of the third codon position. Averaged over the first 2 base positions, A is found to be more frequent than U, but G and C are about equal in relative frequency. A reflexion of the non-randomness in the nucleotide base arrangement is also seen in the average codon frequencies of Table 5. Aside from all codons with (U + C) in the third position, there are deviations from equality among the several codons coding for one amino acid (consider, for instance, the four leucine codons UUA, UUG, CUA and CUG, or the two valine codons GUA and GUG). It is of interest to determine the extent of non-randomness and if the deviations from randomness in model 1 can be attributed primarily to arginine as has been claimed in the literature. Table 6 presents the percentage contribution (over 5%) of each amino acid to the chi-square value for each of the proteins in Table 2 that deviated significantly from model 1. Results show that, based on the average percentage contribution for all proteins, more than ten amino acids seem to be equally important in causing deviations from randomness. Thus arginine does not seem to be the primary cause of non-randomness in the base arrangements.

Table 5. *Cell frequencies (codons) averaged over the 12 proteins adequately fitted by model 2*

(Entries correspond to the cell positions in Table 1.)

0·0381	0·0243	0·0332	0·0256
0·0210	0·0109	0·0	0·0
0·0299	0·0174	0·0	0·0088
0·0145	0·0070	0·0289	0·0102
0·0056	0·0104	0·0131	0·0042
0·0118	0·0136	0·0236	0·0074
0·0261	0·0229	0·0484	0·0119
0·0167	0·0124	0·0373	0·0039
0·0268	0·0191	0·0489	0·0021
0·0325	0·0417	0·0458	0·0373
0·0134	0·0209	0·0254	0·0176
0·0281	0·0357	0·0364	0·0267

The non-randomness in the codon base arrangements could arise as a result of (1) non-random mutation and no selection or (2) non-random survival of mutants due to natural selection. At this stage there is no basis for accepting one and not the other as a possible cause. It seems clear, however, that the extent of non-randomness of base arrangements is hard to justify solely on the basis of the neutral hypothesis. It is likely that both non-random mutations and natural selection played a role in the evolution of non-randomness of the genetic code.

There are indications in the literature that some non-randomness exists in the amino acid substitutions in evolution (Clarke, 1970; Epstein, 1967). Also, the analysis of Josse, Kaiser & Kornberg (1961) on nearest-neighbour base frequencies showed non-randomness existing in the base sequence of a DNA chain. Their analysis, however, does not distinguish between within- and between-codon base

Table 6. *Percentage contribution of each amino acid to the chi-square value for each protein that deviated significantly from model 1*

Protein	Phe	Leu	Ile	Met	Val	Ser	Pro	Thr	Ala	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Try	Arg	Gly
Cytochrome *c* (human)	—	—	—	—	7·5	12·2	8·5	—	—	—	—	—	—	7·11	—	—	—	—	14·7	30·5
Haemoglobin γ (human)	9·9	—	—	5·0	—	—	—	—	—	11·7	12·2	—	—	14·0	—	—	—	—	10·8	14·2
Haemoglobin α (mouse)	—	5·9	—	—	—	—	—	—	—	—	14·0	17·1	—	34·6	—	—	—	—	—	6·3
Haemoglobin β (rabbit)	5·7	—	9·6	6·2	—	—	—	—	—	5·6	12·8	—	7·9	17·5	9·72	—	—	—	—	3·6
Myoglobin (horse)	10·7	—	—	—	8·9	—	—	—	5·4	—	17·7	—	—	11·8	—	—	—	—	11·9	12·7
Haptoglobin α_1 (human)	—	—	—	—	—	10·1	25·9	—	—	—	6·6	—	—	9·1	—	9·7	9·8	—	7·5	6·1
Ribonuclease (bovine)	—	12·3	—	—	12·9	—	—	—	—	—	—	16·6	—	—	—	—	31·2	—	—	5·4
Subtilisin (*B. subt.*)	7·0	—	—	—	—	—	—	—	—	—	—	14·6	11·0	—	14·7	8·7	—	—	15·7	9·4
Tryptophane synth. α (*E. coli*)	16·1	—	6·5	—	18·0	9·6	—	11·9	15·6	—	6·3	—	—	—	—	—	—	—	—	—
Basic trypsin inhibitor (bovine)	8·1	5·4	—	9·3	7·8	28·8	8·1	—	11·1	—	—	—	—	—	—	—	8·2	—	—	
G-3 PDH (pig)	12·7	—	—	—	—	—	—	—	—	—	5·1	—	7·1	33·1	—	9·4	—	—	8·5	13·3
Chymotrypsinogen A (bovine)	—	—	—	6·2	—	—	—	—	—	7·7	—	19·6	8·6	—	—	10·7	—	—	22·6	—
Trypsinogen (bovine)	6·9	—	—	—	—	5·1	—	—	—	—	—	24·7	—	—	—	12·5	—	—	24·2	7·3
Growth hormone (human)	13·8	—	—	—	19·5	—	—	—	—	8·8	—	—	—	—	41·1	11·4	—	—	—	—
Elastase (pig)	6·2	—	—	—	6·9	—	—	—	—	—	—	33·4	10·4	6·2	7·7	6·4	—	—	5·8	—
Azurin (*Bord. bronch.*)	9·1	—	—	—	—	—	—	—	—	5·3	—	—	—	20·6	—	13·4	—	—	9·9	9·7
Hemerythrin (sip. worm)	18·2	—	5·3	—	9·3	—	—	—	—	5·3	—	—	—	7·1	—	—	—	21·0	—	—
Thioredoxin (*E. coli*)	—	9·1	—	—	14·3	9·0	—	—	—	—	—	—	—	8·6	10·1	—	—	9·3	10·9	6·0
Tobacco mosaic virus (Dahl.)	5.4	—	—	6·2	—	—	6·4	—	—	—	—	23·2	—	16·4	—	—	6·5	7·8	—	—
Mean	9·98	8·2	7·1	6·6	11·7	12·5	12·2	11·9	10·7	7·4	10·7	21·3	9·0	15·5	16·6	10·3	13·9	12·7	12·9	10·6

https://doi.org/10.1017/S0016672300016578 Published online by Cambridge University Press

arrangements. Ohta & Kimur (1970), using a random model for estimation, found inequality in frequency among the bases in position 1. A widely held belief, however, is that the base arrangement in a codon is largely random and any non-randomness is perhaps due primarily to arginine. Our analysis reveals, for the first time, that what has been believed to be minor in occurrence (non-randomness) is so widespread in occurrence as to be, seemingly, the rule rather than the exception. Besides, this is the first time that codon frequencies, as well as base frequencies (in all three codon positions), have been estimated based on underlying models capable of revealing the kinds of interactions among bases in the three codon positions.

The authors wish to thank a referee for his comments.

REFERENCES

CLARKE, B. (1969). Darwinian evolution of proteins. *Science* **168**, 1009–1011.

CLARKE, B. (1970). Selective constraint on amino acid substitutions during the evolution of proteins. *Nature* **228**, 159–160.

COOK, R. D. & NASSAR, R. F. (1975). The amino acid composition of proteins: A method of analysis. *Theoretical Population Biology* **67**, 64–83.

DAYHOFF, M. O. (1972). *Atlas of Protein Sequence and Structure*, vol. v. Silver Spring, Maryland: National Biometrical Research Foundation.

EPSTEIN, C. J. (1967). Nonrandomness of amino acid changes in the evolution of homologous proteins. *Nature* **215**, 355–359.

FIENBERG, S. E. (1970). The analysis of multidimensional contingency tables. *Ecology* **51**, 419–433.

FIENBERG, S. E. (1972). The analysis of incomplete multi-way contingency tables. *Biometrics* **28**, 177–202.

HARRIS, H. (1966). Enzyme polymorphisms in man. *Proceedings of the Royal Society, London* B **164**, 298–310.

JOSSE, J., KAISER, A. D. & KORNBERG, A. (1961). Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequencies in deoxyribonucleic acid. *Journal of Biological Chemistry* **236**, 864–875.

KIMURA, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* **11**, 247–269.

KIMURA, M. & OHTA, T. (1971*a*). Protein polymorphism as a phase of molecular evolution. *Nature* **229**, 467–469.

KIMURA, M. & OHTA, T. (1961*b*). *Theoretical Aspects of Population Genetics*. Princeton, New Jersey: Princeton University Press.

KIMURA, M. & OHTA, Y. (1972). Population genetics, molecular biometry, and evolution. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, v, 43–65.

KING, J. L. & JUKES, T. H. (1969). Nondarwinian evolution: random fixation of selectively neutral mutations. *Science* **164**, 788–798.

LEWONTIN, R. C. & HUBBY, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amounts of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54**, 595–609.

OHTA, T. & KIMURA, M. (1970). Statistical analysis of the base composition of genes using data on the amino acid composition of proteins. *Genetics* **64**, 387–395.

STEBBINS, G. L. & LEWONTIN, R. C. (1972). Comparative evolution at the levels of molecules, organisms and populations. *Proceedings of the Sixth Berkeley Symposium on Mathematica Statistics and Probability*, v, 23–42.

https://doi.org/10.1017/S0016672300016578 Published online by Cambridge University Press