J. Mol. Evol. 10, 231–240 (1977)

Journal of Molecular Evolution

Evaluation of the Non-Randomness of Protein Compositions

Athel Cornish-Bowden and Alan Marson

Department of Biochemistry, University of Birmingham, P.O. Box 363, Birmingham B 15 2TT, England

Summary. A method is described for assessing the non-randomness of protein compositions, based on the chi-squared statistic for the differences between the observed numbers of residues of each type and the numbers expected for a random distribution of codons. The analysis indicates that changes in at least 30% of the residues in natural proteins are selected against.

Key words. Evolutionary Invariants – Non-random Amino Acid Compositions of Proteins – Proportion of Covarions

Introduction

For any protein or peptide, the non-randomness of the amino acid composition can be assessed by comparing the proportion of residues of each type with the corresponding proportion of codons in the genetic code. Combining the results for the 20 types of amino acid specified in the code, Holmquist (1975) has defined a net non-randomness index Q as follows:

$$Q = 100 \sum_{i=1}^{20} \left| \frac{n_i}{N} - \Phi_i \right| \qquad (1)$$

where n_i is the number of residues of the ith type of amino acid, $N = \Sigma n_i$ is the total number of residues in the protein, and Φ_i is the proportion of codons in the genetic code that specify the ith type of amino acid (out of 61, as the three termination codons are ignored). Q has a range of possible values from zero to about 197, in which zero corresponds to a perfectly "code-random" protein, i.e. one with $n_i = N\Phi_i$ for all i, and 197 is the value for polymethionine or polytryptophan.

Holmquist and Moise (1975) determined Q for 169 proteins and peptides, and reported values ranging from 26 for pig aspartate aminotransferase to 113 for tuna protamine. Before this range could be interpreted, it was necessary to correct for the sampling component of Q that results from the finite lengths of natural proteins: because every term in

the summation is non-negative the sum has an expected value greater than zero even for a perfectly random protein. Holmquist and Moise (1975) estimated the sampling component of Q by Monte Carlo computation (though direct calculation is more precise and requires less computer time; see below) and found it to be approximately proportional to $N^{-\frac{1}{2}}$. By subtracting the appropriate correction from each observed value of Q they obtained the "biologically significant" index of non-randomness, Q_c, for each protein. The resulting values of Q_c were clustered about a mean of 24 with a standard deviation of 10, and Holmquist and Moise (1975) argued that the constancy of Q_c for widely diverse proteins indicated that it was an "evolutionary invariant", i.e. that natural selection had ensured values of Q_c close to 24 for all proteins.

It seemed to us that this analysis could be criticized on several grounds. First, it is difficult to find a clear biological meaning that can be attached to Q_c. Accordingly one cannot judge whether the mean of 24 indicates a smaller or a larger amount of selection away from randomness than had been supposed previously. Although 24 is much closer to zero than to the potential maximum of about 180, this may simply reflect the fact that grossly abnormal compositions are required to produce high values of the index.

Second, there is no obvious reason why the finite sampling component of Q and the biologically significant component should be additive as scalars. Rather, one would expect them to be largely uncorrelated, so that the observed value might better be regarded as the resultant of two orthogonal vectors. If so, the scalar subtraction used by Holmquist

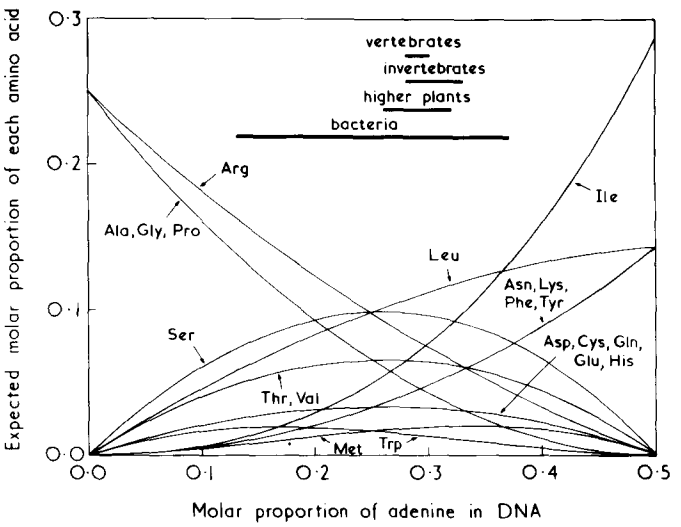

Fig. 1. Expected frequencies of the amino acids in the proteins of different organisms. For each molar proportion of adenine, the expected frequencies of the 61 codons that code for amino acids were calculated by the method of Kimura [1968; his "Expected (1)"], assuming that the proportions of adenine and thymine were equal and that the proportions of cytosine and guanine were equal. Nearest neighbour frequencies (Josse et al., 1961) were not used, largely because of Kimura's (1968) finding that they gave a worse rather than a better fit to the observed frequencies than that given by the simpler calculation. The codon frequencies were used to calculate the expected amino acid frequencies. The bars showing the ranges of DNA compositions observed for different types of organism were drawn from data of Sueoka (1961)

and Moise (1975) was probably an excessive correction. On the other hand, their Monte Carlo estimation of the magnitude of the finite sampling component probably led to underestimation, for a different reason. In assuming that a code-random protein was one in which the amino acid frequencies reflected the proportions of codons in the genetic code, they tacitly assumed that all codons (apart from termination codons) would be equally frequent in a random DNA sequence. This is not correct, however, because the four DNA bases are not equally abundant: for example, in vertebrate DNA, adenine and thymine are about 40% more abundant than cytosine and guanine (Sueoka, 1961). So in a truly random protein the types of amino acid should occur with frequencies f_i that are not necessarily equal to the codon proportions Φ_i, but instead vary with the species of organism, as illustrated in Figure 1. The consequence of this is that the expected limit of Q for a random protein as N approaches infinity is not zero but a finite positive value, and at finite values of N the sampling components are larger than the values calculated by Holmquist and Moise (1975).

In an effort to overcome these problems we have examined the properties of a different index of non-randomness Z, defined as follows:

$$Z = \sum_{i=1}^{20} \frac{(n_i - Nf_i)^2}{Nf_i} \tag{2}$$

where n_i , N and f_i have the same meanings as before. If amino acids were truly distributed at random with frequencies f_i this index would be distributed as χ^2, and in general it is much easier to analyse than Q because the use of squares avoids the intractable algebra characteristic of absolute values. Although Laird and Holmquist (1975) defined an index of this type they did not apply it widely, and objected that for small samples (i.e. short sequences in the present context) χ^2 does not possess its simple asymptotic properties; in particular, its distribution is skew and discrete. However, as Q is open to similar (and more serious) objections for small values of N, Z seems to be a more convenient index.

Theory

Expected Values of Q and Z for a Random Protein

We define a random protein as one in which the n_i values are distributed multinomially with probability f_i of finding the ith type of amino acid at any position. The expected value of Q, $E(Q)$, is then given by Eq. 3:

$$E(Q) = 100 \sum_{i=1}^{20} \sum_{n_i=0}^{N} \left[\frac{N! f_i^{n_i} (1-f_i)^{N-n_i}}{n_i!\,(N-n_i)!} \left| \frac{n_i}{N} - \Phi_i \right| \right] \tag{3}$$

This expression can readily be evaluated by computation for any set of f_i values, but it never leads to a simple dependence of $E(Q)$ on N. The line shown in Figure 2 is calculated from Eq. 3 with $f_i = \Phi_i$ for all i, i.e. the special case assumed by Holmquist and Moise

(1975), and agrees to within sampling variation with the $N^{-\frac{1}{2}}$ dependence found by them in Monte Carlo trials. However, it is clear from the abrupt changes in slope at certain values of N, especially multiples of 30.5, that $E(Q)$ is not exactly proportional to $N^{-\frac{1}{2}}$. The fluctuations occur because Q has a discrete (quantized) distribution, with eigenvalues that vary in a complex way with N. This is most strikingly evident in the baseline of Figure 2, which touches the axis only when N is a multiple of 61.

$E(Z)$, the expected value of Z, can also be expressed by an equation similar to Eq.3, but it is more instructive to expand it as follows:

$$E(Z) = \sum_i [E(n_i^2)/Nf_i - 2E(n_i) + Nf_i] \tag{4}$$

because $E(n_i^2)$ and $E(n_i)$ are readily obtainable from the following general formula for the factorial moments of the binomial distribution (Johnson and Kotz, 1969, pp. 51–53):

$$E[n_i!/(n_i - r)!] = N!f_i^r/(N - r)! \tag{5}$$

From this equation, $E(n_i) = Nf_i$ and $E(n_i^2) = N(N-1)f_i^2 + Nf_i$, so Eq. 4 simplifies to

$$E(Z) = 20 - \Sigma f_i = 19 \tag{6}$$

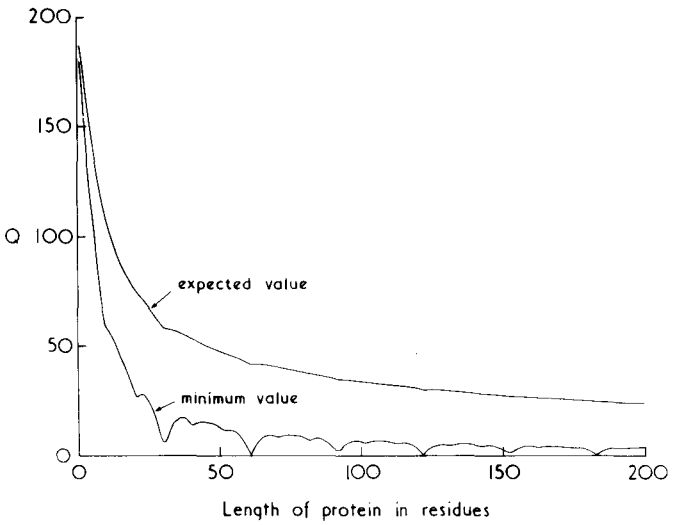

Fig. 2. Expected value of Q for a random protein. The upper line shows the expected value of Q, the index of non-randomness proposed by Holmquist (1975), for random proteins in which the probability of finding each type of amino acid at any locus is proportional to the number of codons for that type of amino acid. The lower line is a baseline and shows the smallest value that Q can have for each length of protein

in agreement with the standard result that the expected value of χ^2 is equal to the number of degrees of freedom. It is independent of the length of the sequence, though for very short sequences ($N < 10$) the distribution is so highly skewed that knowledge of the expected value is not very helpful.

Variance of Z for a Random Protein

The variance of Z can be expressed as $\sigma^2\ (Z) = E\ (Z^2) - [E(Z)]^2$. Now $[E(Z)]^2 = 361$ follows from Eq. 6, and $E(Z^2)$ can be found by expanding the expression for Z^2 as follows:

$$Z^2 = [\sum_i (n_i - Nf_i)^2 / Nf_i]^2 = N^{-2} \sum_i \sum_j (n_i^2 n_j^2 / f_i f_j - 2N n_i^2 n_j / f_i$$

$$- 2N n_i n_j^2 / f_j + \ldots + N^4 f_i f_j) \qquad (7)$$

$E(Z^2)$ is then obtained by substituting $E(n_i^2 n_j^2)$ for $n_i^2 n_j^2$, etc., using the following general formula for the mixed factorial moments of the multinomial distribution (Johnson and Kotz, 1969, pp. 284–285):

$$E[\, n_i! n_j! / (n_i - r_i)!(n_j - r_j)!\,] = N! f_i^{r_i} f_j^{r_j} / (N - r_i - r_j)! \qquad (8)$$

For terms in the summation for which $i = j$, Eq. 5 must be used rather than Eq. 8, because Eq. 8 applies only when $i \neq j$. Evaluation of Eq. 7 is tedious, but it involves no new principles and leads to the following expression for the variance of Z:

$$\sigma^2\ (Z) = 38 - (438 - \Sigma 1/f_i)/N \qquad (9)$$

For reasonable sets of f_i values, this expression does not differ appreciably from 38 (i.e. twice the number of degrees of freedom) except for small values of N. For example, if one puts $f_i = 0.05$ for all i Eq. 9 reduces to $\sigma^2\ (Z) = 38\ (1 - 1/N)$, so $34.2 < \sigma^2\ (Z) < 38$ for $N > 10$; more realistically, for a "code-random" sequence $\Sigma 1/f_i = 523.6$, so Eq. 9 reduces to $\sigma^2\ (Z) = 38 + 85.6/N$, which gives $38 < \sigma^2\ (Z) < 46.6$ for $N > 10$.

Expected Values of Q and Z for a Partially Random Protein

Consider a protein containing N residues, of which M are "essential", in the sense that any mutation (other than a synonymous one) is fatal, and $(N–M)$ are randomly distributed according to a multinomial distribution as described above for a random protein. Further, suppose that $M = \Sigma m_i$, where m_i is the number of essential residues of the ith type of

amino acid. In reality there is likely to be a continuum of selectivity, from sites with an absolute requirement for one type of amino acid, through sites at which any of several types of amino acid will serve, to sites with no selectivity at all. However, a two-class model is probably adequate as a first approximation.

The expected value of Q is given for this model by

$$E(Q) = 100 \sum_{i=1}^{20} \sum_{n_i = m_i}^{N-M+m_i} \left[\frac{(N-M)! f_i^{n_i - m_i} (1-f_i)^{N-M-n_i+m_i}}{(n_i - m_i)!\,(N-M-n_i+m_i)!} \left| \frac{n_i}{N} - \Phi_i \right| \right] \quad (10)$$

Unfortunately this equation is too unmanageable for any information about the magnitude of M to be simply extracted from measurements of Q. It is more practical to consider Z, for which the expected value may be derived by a procedure similar to that used for Eq. 6, and is as follows:

$$E(Z) = 19 + [\Sigma(m_i^2/f_i) - 19M - M^2]/N \quad (11)$$

Comparison of this with Eq. 6 shows that the constant term is identical with $E(Z)$ for a fully random protein. Thus the deviation of Z from 19 provides information about the extent of selection. Unfortunately one cannot estimate M from Z without introducing an assumption about how M is partitioned among the m_i. However, as we discuss below, it is possible to make a reasonable assumption as a first approximation that permits some conclusions to be drawn about the minimum values of M for real proteins.

Results and Discussion

Holmquist and Moise (1975) determined the values of Q for proteins from a wide variety of organisms. We have preferred to consider a single species, to avoid complications due to the species dependence of f_i. Accordingly, values of Z are shown in Table 1 for bovine peptides and proteins with values of N ranging from 9 to 500. Two examples from other ungulate species are included in order to fill some gaps between $N = 307$ and $N = 500$.

It is immediately evident from Table 1 that nearly all of the Z values are greater than 19, and nearly all deviate by several standard deviations. This establishes that natural proteins do not have random compositions, but this is hardly a novel conclusion, and was anyway evident from the studies of Holmquist and Moise (1975). More important is the information that the results provide about the extent of non-randomness in natural proteins, and two minimum estimates of M are shown in the Table, derived from two different assumptions about the partitioning of M among the m_i.

The first estimate is obtained by solving Eq. 11 for M after replacing $E(Z)$ with Z, putting $f_i = 0.05$ for all i, $m_1 = M$, and $m_i = 0$ for $i \neq 1$. This effectively assumes that all selection is concentrated in one "average" type of amino acid. For about half the examples this estimate of M exceeds the largest value of n_i for that protein, and so selec-

Table 1. Non-randomness of natural proteins

Protein[a]	N	Z [b]	M estimates [c]	[d]	Largest n_i	Reference[e]
Oxytocin	9	20.45	1.5	8.4	2 (Cys)	D–192
Met-Lys-bradykinin	11	28.91	2.9	7.2	3 (Pro)	D–214
Glucagon	29	18.79	f	f	4 (Ser)	D–208
Calcitonin	32	11.15	f	f	4 (Ser)	D–206
Corticotropin	39	25.97	4.3	28.5	4 (Pro)	D–196
Posterior pituitary peptide	48	40.01	7.8	22.8	8 (Gly)	D–194
Basic trypsin inhibitor	58	33.48	7.2	32.9	6 (Gly)	D–168
Proinsulin	81	55.11	12.9	27.9	12 (Gly)	D–209
Parathyroid hormone	84	35.59	9.1	44.9	9 (Lys)	D–205
Cytochrome b_5	93	51.12	13.0	34.6	12 (Glu)	D–31
Histone IV	102	66.21	16.4	29.3	17 (Gly)	D–280
Cytochrome c	104	84.26	19.4	23.5	18 (Lys)	D–14
Ribonuclease	124	44.85	13.5	52.6	15(Ser)	D–130
Histone IIB2	125	62.89	17.5	37.8	20 (Lys)	D–279
Histone III	135	57.42	17.0	44.7	18 (Arg)	S–69
Haemoglobin α-chain	141	71.33	20.2	37.5	20 (Leu)	D–61
Haemoglobin β-chain	145	67.73	19.8	40.7	18 (Val)	D–72
Myoglobin	153	107.02	27.1	27.2	19 (Lys)	D–82
Lactoglobulin	162	67.70	20.9	45.5	22 (Leu)	S–83
γ-Crystallin	165	53.56	17.8	58.6	19 (Arg)	S–74
Myelin membrane enceph-alitogenic protein	170	88.51	25.4	36.6	25 (Gly)	D–324
Growth hormone	189	43.69	16.2	82.2	24 (Leu)	D–203
α_{s1}-Casein	199	92.74	28.3	40.8	25 (Glu)	D–320
β-Casein	209	149.83	38.4	26.5	35 (Pro)	S–82
Trypsinogen	229	74.30	26.3	58.6	34 (Ser)	D–105
Chymotrypsinogen A	245	69.55	26.0	66.9	28 (Ser)	D–107
Carboxypeptidase A	307	38.44	18.2	151.7	32 (Ser)	D–126
Glyceraldehyde 3-phosphate dehydrogenase (pig)	332	105.92	39.5	59.4	34 (Val)	D–147
Alcohol dehydrogenase E-chain (horse)	374	98.98	40.2	71.8	39 (Val)	D–145
Glutamate dehydrogenase[g]	500	92.70	44.5	102.5	47 (Gly)	S–35
Combined data	4594	33.79	60.3	137.8	378 (Gly)	

[a] Bovine, except where stated otherwise

[b] In all cases the expected value of Z for a random composition is 19.00, with a standard deviation of 6.2 for proinsulin and below, the remainder decreasing from 6.8 for oxytocin to 6.3 for basic trypsin inhibitor

[c] Calculated by solving Eq. 11 for M with the assumption that all selection is concentrated in one type of amino acid

[d] Calculated from $M = 19N^2/[\Sigma(n_i^2/f_i) - N^2]$. This provides the smallest M value that gives $E(Z) > 19$, assuming that the proportion of "essential" residues is the same for each type of amino acid

[e] Page numbers in Dayhoff (1972) if prefixed D-, in Dayhoff (1973) if prefixed S-

[f] Calculation of M is not valid if $Z < 19$

[g] A single Glx residue is scored as glutamine; none of the other sequences contain incompletely characterized residues

tion cannot be confined to one type of amino acid, even if one were willing to entertain such an extreme hypothesis.

The second minimum estimate of M is obtained by supposing, as a first approximation, that the proportion of "essential" residues is the same for each type of amino acid, i.e. m_i is proportional to n_i. One can readily calculate $M = 19N^2 / [\Sigma(n_i^2/f_i) - N^2]$ as the smallest M value for each protein that gives $E(Z) > 19$. We note that, with only two exceptions – glucagon and calcitonin, both with such small N values that they are barely proteins – the observed Z value exceeds 19. We assume that this is not a coincidence and that $E(Z)$ in fact exceeds 19 for nearly all proteins.

The second method of calculating M does not lead to any contradictions (except in the cases of glucagon and calcitonin, for which the low values of Z suggest that the assumption that $E(Z) > 19$ may be invalid). Thus it seems to be the more reasonable of the two methods, and suggests that on average at least 30% of the residues in bovine proteins are "essential", This might be interpreted as a crude way of confirming the estimates (Fitch, 1976) of the numbers of "covarions" (Fitch and Markowitz, 1970) in various proteins: these indicate that for several proteins the proportion of covarions is much less than 70% (i.e. the proportion of "essential" residues is much more than 30%). However, our method is based on different assumptions from that of Fitch (1976), and so any agreement can be regarded as complementary. Moreover, although Fitch's method is more accurate it requires much more information, and can thus be applied to only a few proteins.

Before undertaking this work we suspected that an explanation of the consistent values of Q_c reported by Holmquist and Moise (1975) might be found in their ignoring of variations in base frequencies in the DNA of different species. In this study, however, we found that the general character of the results was not appreciably affected by the set of f_i values that were used. This is fortunate, because it suggests that more complex effects, such as the non-randomness of nearest-neighbour frequencies (Josse et al., 1961) and of utilization of synonymous codons (Sanger et al., 1977) can be neglected without serious error. Our analysis, far from eliminating the small amount of selection implied by Holmquist and Moise (1975), indicates a substantial amount of selection against changes in particular types of amino acid and is consistent with the high proportions of "essential" residues estimated by Fitch (1976).

In reality selection is more complicated than the simple model we have adopted. Selection against change is recognizable in the unvaried residues for which M is the parameter of consequence, but it is unclear how to introduce into the equations the consequential effects when the selection is not against change. One should note that although M is significant of selection, it does not bear on the neutralist controversy because it does not allow for these more complicated effects. It would be useful to find a way of allowing for these, to permit a more precise assessment of the amount of randomness in protein structure, and, indeed, in protein evolution. With our present model, however, we do not believe this to be possible.

One might expect that if one combined the composition data for many proteins the resulting value of Z would be relatively small and that the two estimates of M would be much smaller than the sums of the estimates for the individual proteins. This is indeed observed (bottom line of Table 1), and Z for the combined set of data is one of the smallest values in the table. Nonetheless it would be wrong to suppose that as one averages the results for a large number of proteins they tend towards a code-random composition.

Table 2. Expected and observed frequencies of the amino acids

Amino acid	Frequency per 100 residues		Difference (%)
	Expected[a]	Observed[b]	
Alanine	4.69	7.71	64.4
Arginine	7.93	4.72	−40.4
Asparagine	4.47	3.83	−14.3
Aspartate	3.24	4.79	47.8
Cysteine	3.24	2.11	−34.9
Glutamate	3.24	5.75	77.5
Glutamine	3.24	4.01	23.8
Glycine	4.69	8.23	75.5
Histidine	3.24	2.90	−10.5
Isoleucine	7.07	4.83	−31.7
Leucine	10.95	7.86	−28.2
Lysine	4.47	7.51	68.0
Methionine	1.88	1.96	4.3
Phenylalanine	4.47	4.16	− 6.9
Proline	4.69	4.96	5.8
Serine	9.72	7.62	−21.6
Threonine	6.48	5.62	−13.3
Tryptophan	1.36	1.09	−19.9
Tyrosine	4.47	3.40	−23.9
Valine	6.48	6.99	7.9

[a] i.e. $100f_i$ for each type of amino acid, calculated as described by Kimura (1968), assuming that adenine and thymine each account for 29% [rather than 29.5%, as assumed by Kimura (1968)], and guanine and cytosine 21% each, of the base content of vertebrate DNA

[b] Averages for the 30 proteins listed in Table 1

Table 2 shows that there are considerable differences between the expected and observed frequencies of the amino acids, and they are distributed in a way that is by no means haphazard. Two of the largest excesses occur with the two simplest types of amino acid, glycine and alanine, which may be preferred in some positions for reasons of metabolic economy; three of the four types of amino acid with sidechains that are charged at pH 7.4, i.e. glutamate, lysine and aspartate, also show large excesses; the fourth, arginine, shows the largest deficit in the table. The significance of these observations is not, however, obvious.

Acknowledgements. We thank Dr. J.S. Gale for suggesting some improvements to this paper and for help with the statistical calculations. We are also greatly indebted to the Reviewers for suggesting further improvements.

References

Dayhoff, M.O. (1972). Atlas of protein sequence and structure, Vol. V, Silver Spring: National Biomedical Research Foundation

Dayhoff, M.O. (1973). Atlas of protein sequence and structure, Vol. V, suppl. 1, Silver Spring: National Biomedical Research Foundation

Fitch, W.M. (1976). J. Mol. Evol. **8**, 13
Fitch, W.M., Markowitz, E. (1970). Biochem. Genet. **4**, 579
Holmquist, R. (1975). J. Mol. Evol. **4**, 277
Holmquist, R., Moise, H. (1975). J. Mol. Evol. **6**, 1
Johnson, N.L., Kotz, S. (1969). Discrete distributions. Boston: Houghton Mifflin
Josse, J., Kaiser, A.D., Kornberg, A. (1961). J. Biol. Chem. **236**, 864
Kimura, M. (1968). Genet. Res. **11**, 247
Laird, M., Holmquist, R. (1975). J. Mol. Evol. **4**, 261
Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddles, J.C., Hutchison, C.A., III, Slocombe, P.M., Smith, M. (1977). Nature **265**, 687
Sueoka, N. (1961). J. Mol. Biol. **3**, 31

Received July 25, 1977