J. Genet. Vol. 72, Nos. 2 & 3, December 1993, pp. 47–58. © Printed in India.

Logic of the genetic code: Conservation of long-range interactions among amino acids as a prime factor

N. V. JOSHI*, VIVEK V. KORDE† and V. SITARAMAM†

*Centre for Ecological Sciences, Indian Institute of Science, Bangalore 560012, India

† Biotechnology Centre, Department of Zoology, University of Poona, Pune 411 007, India

MS received 27 October 1993

Abstract. Any statement on the optimality of the existing code ought to imply that this code is ideal for conserving a certain hierarchy of properties while implying that other codes may have been better suited for conservation of other hierarchies of properties. We have evaluated the capability of mutations in the genetic code to convert one amino acid into another in relation to the consequent changes in physical properties of those amino acids. A rather surprising result emerging from this analysis is that the genetic code conserves long-range interactions among amino acids and not their short-range stereochemical attributes. This observation, based directly on the genetic code itself and the physical properties of the 20 amino acids, lends credibility to the idea that the genetic code has not originated by a frozen accident (the null hypothesis rejected by these studies) nor are stereochemical attributes particularly useful in our understanding of what makes the genetic code 'tick'. While the argument that replacement of, say, an aspartate by a glutamate is less damaging than replacement by arginine makes sense, in order to subject such statements to rigorous statistical tests it is essential to define what constitutes a random sample for the genetic code. The present investigation describes one possible specification. In addition to obvious statistical considerations of testing hypotheses, this procedure points to the more exciting notion that alternative codes may have existed.

Keywords. Genetic code; mutation rates; evolution; frozen accident; stereochemical properties.

1. Introduction

The origin, organization and evolution of the genetic code have continued to be sources of fascination since the discovery of the code in the sixties. That it had to be at least a triplet-based code was dictated by the fact that four bases code for 20 amino acids. However, explanations put forward for specific assignment of codons to amino acids were based on diverse speculations. These have invoked the role of structural and stereochemical features (Woese 1965), the role of chance (Jukes 1966) following an initial adaptation phase—the 'frozen accident'—(Crick 1968), kinetic mechanisms involving hypercycles and bootstrapping (Eigen and Winkler-Oswatitsch 1981; Crothers 1982), and others. Some of the more recent studies have explored how deviations from the 'universal' code (e.g. as seen in mitochondria) are brought about (Osawa and Jukes 1989), while others have analysed the intriguing patterns of complementary hydropathy (Konecny *et al.* 1993), which has a bearing on the similarity between the two polypeptides encoded in the sense and antisense strands of DNA.

Another kind of approach to understanding the logic of the genetic code is to

*For correspondence

examine the sensitivity of the different properties of amino acids to point mutations in the DNA (e.g. Salemme *et al.* 1977). The code is expected to have evolved in such a way as to make the more important properties of amino acids less sensitive to mutations. Sitaramam (1989) has suggested an approach whereby one can examine to what extent the genetic code protects or conserves a particular property of amino acids. For a given rate of point mutation, the probability of conversion of any amino acid into any other can be computed from the genetic code (an explicit definition is given in the next section). For the 20 amino acids, one thus obtains a set of 400 conversion probabilities. These may be termed 'genetic distances' between the amino acids. The *effect* of one amino acid changing into another as a result of one or more mutations can be quantified, for a given property, by the absolute value of the difference between the numerical values of the property for the two amino acids. For every property, this leads to a set of 400 values; these may be referred to as the 'physical distances' between the amino acids for that property. Sitaramam (1989) has pointed out that the correlation coefficient between genetic and physical distances for any property of amino acids may be used to measure the robustness of the property to mutations. A large negative value of the coefficient implies that conversions that are more probable lead to relatively small changes in the value of the property, and conversely, conversions that lead to large changes have a low probability of occurrence (see also Di Giulio 1989). By examining the values of the correlation coefficients between the genetic and physical distances for various properties of amino acids, one can thus identify the important properties, i.e. *the properties that the genetic code is designed to protect.* Alternatives to correlation coefficient, such as mean square error (Haig and Hurst 1991), have also been subsequently proposed, though correlation coefficient seems to have an advantage in being scale independent.

From such an analysis, Sitaramam (1989) demonstrated that the genetic code preferentially conserves the long-range nonbonded interaction energy. His approach, however, had two limitations. Firstly, the probability of a codon (triplet) undergoing one, two or three base changes was taken to be fixed at 1/9, 1/81 and 1/729 respectively. While this is a little better than considering rather *ad hoc* parameters like number of base changes needed (Di Giulio 1989) or exclusively single base changes (Haig and Hurst 1991), a more rigorous measure that takes into account probabilities of 0, 1, 2,... mutations would be more appropriate.

Secondly, the significance of the correlation coefficient was assessed using Fisher's z transformation (see e.g. Sokal and Rohlf 1981). This test implicitly assumes the existence of an infinite population of bivariate measurements from which a random sample is drawn. In the present context, what constitutes such a population? Indeed, the 400 values of physical and genetic distances are not independent. These considerations suggest that the formal application of the test of significance of the correlation coefficient may lead to a biased estimate of the p value.

These two limitations can be taken care of by allowing for various rates of base substitution and also by considering the population of all theoretically possible genetic codes (as done by Haig and Hurst 1991), from which a sample of codes can be drawn. The values of correlation coefficient (r) for each of the sample codes can be obtained, and the frequency distribution of these r values can be used to judge the statistical significance of the r value obtained for the existing genetic code.

We describe here an algorithm to compute the conversion probabilities for any intensity of mutation rate and also a Monte Carlo technique (Haig and Hurst 1991) for performing suitable tests of statistical significance.

2. Methods and models

2.1 *Computation of conversion probabilities*

Let y denote the average number of mutations (base substitutions) that a codon undergoes, say during DNA replication, or during the formation of germ-line cells which eventually give rise to the next generation. A given codon may undergo 0, $1, 2, \ldots, k$ such mutations with appropriate probabilities (defined later). As a result, the codon may code for the same amino acid (either because no mutations took place, or because multiple mutations restored the original triplet, or because the changed triplet coded for the same amino acid), or a different amino acid. The probability of conversion of, say, glycine to alanine is then defined as the probability that a triplet has mutated to one that codes for alanine, given that it originally coded for glycine. We compute the probabilities of conversion of amino acids from the conversion probabilities for codons. The details of the procedure are described below.

2.1a *Conversion probabilities for codons*: With reference to a given codon, say UGC, all the 64 codons can be classified into four groups:

(i) UGC : There is only one such codon
(ii) Differing from UGC in only one base : There are nine such codons
(iii) Differing from UGC in two bases : There are 27 such codons
(iv) Differing from UGC in three bases : There are again 27 such codons.

Let the groups be denoted by T0, T1, T2 and T3 respectively.

To begin with, the codon is in class T0. If it undergoes a single mutation, it moves into the class T1. If it now undergoes *one more* mutation, there will be three possible outcomes:

(i) The mutation may be at a different position than the first mutation. The codon then moves to class T2. The (conditional) probability of such an outcome is 2/3.
(ii) The mutation is at the same position as the first one, but the codon does not revert to UGC, i.e. it remains in class T1. The probability of such an outcome is $1/3 \times 2/3$, i.e. 2/9.
(iii) The mutation is at the same position as the first one, and the codon reverts to UGC, i.e. it moves into class T0. The probability of such an outcome is $1/3 \times 1/3$, i.e. 1/9.

In an analogous manner, one can compute the probability of a codon moving from any class to any other class as a result of a single mutation. The results can be expressed in the form of a square matrix M of order 4, such that the elements m_{ik} denote the probability of conversion from class Tk to class Ti. The matrix M is seen to be

$$M = \begin{bmatrix} 0 & 1/9 & 0 & 0 \\ 1 & 2/9 & 2/9 & 0 \\ 0 & 6/9 & 4/9 & 3/9 \\ 0 & 0 & 3/9 & 6/9 \end{bmatrix}$$

Let $P(N)$ be the column vector $(p_0, p_1, p_2, p_3)^T$ whose elements denote the probabilities

that the codon will be in the class T0, T1, T2 or T3 respectively after N mutations have taken place. It is seen that

$$P(N) = M^N \cdot U,$$

where U is the column vector $(1,0,0,0)^T$. Since the possible numbers of codons in classes T0, T1, T2, T3 are 1, 9, 27 and 27 respectively, as N tends to infinity $P(N)$ tends to $P(\infty) = (1/64, 9/64, 27/64, 27/64)^T$.

As defined earlier, the parameter y denotes the mean number of mutations experienced by a codon. Assuming mutations to be random events, the probability that a codon undergoes exactly k mutations ($k = 0, 1, 2, \ldots$) is given by the Poisson probability ($e^{-y} y^k/k!$).

Hence F, the column vector $(f_0, f_1, f_2, f_3)^T$, whose elements denote the probabilities that a codon will be in the class T0, T1, T2 or T3, is given by

$$F = e^{-y} \sum_{k=0}^{\infty} P(k) \cdot y^k/k!.$$

For any given value of y, the column vector F can be calculated as follows.

Let X denote the matrix of right eigenvectors of the matrix M, and B denote the diagonal matrix of eigenvalues. Then it can be shown that

$$M^N = X \cdot B^N \cdot X^{-1}.$$

The expression for F then becomes

$$\begin{aligned} F &= e^{-y} \cdot X \cdot \sum_k B^k \cdot y^k/k! \cdot X^{-1} \cdot U \\ &= X \cdot C \cdot X^{-1} \cdot U, \end{aligned}$$

where C is a diagonal matrix with

$$c_{ii} = \exp(-y) \cdot \exp(b_{ii} \cdot y) = \exp((b_{ii} - 1) \cdot y),$$

where b_{ii} is the ith eigenvalue of the matrix M.

The eigenvalues of M are seen to be equal to 1, 5/9, 1/9 and $-1/3$, while the matrix X of (unnormalized) corresponding eigenvectors is seen to be

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 9 & 5 & 1 & -3 \\ 27 & 3 & -5 & 3 \\ 27 & -9 & 3 & -1 \end{bmatrix}$$

It can also be shown that X^{-1} is equal to $X/64$. Thus, for any value of y, the column vector F can be computed.

The probability of conversion z_{ik} for two codons i and k in a specified time interval, when the mutation rate is y, is then seen to be, from the above expressions,

$z_{ik} = f_0$ if the two codons are identical,

$z_{ik} = f_1/9$ if the two codons differ in one base.

This is because f_1 is the probability of changing to a codon differing in one base. There are nine different codons which differ from i in exactly one base, so to obtain the probability of i going to k, which is one of the nine possibilities, f_1 has to be divided by 9. Arguing in a similar manner, it can be seen that

$$z_{ik} = f_2/27 \text{ if they differ in two bases, and}$$
$$= f_3/27 \text{ if they differ in all the three bases.}$$

Notice that $z_{ik} = z_{ki}$. It can also be seen that when $y = 0$ (no mutations), $z_{ii} = 1$ and $z_{ik} = 0$ for $i \neq k$. Also, as $y \to \infty$, $z_{ik} \to 1/64$ independently of i and k, as expected.

2.1b *Conversion probabilities for amino acids*: Let the degeneracies of two amino acids be n_1 and n_2, i.e. they are coded for by the codons $i_1, i_2, \ldots, i_{n_1}$ and $k_1, k_2, \ldots, k_{n_2}$ respectively. Assuming codon usage frequencies to be equal, q_{12}, the probability of conversion of amino acid 1 to amino acid 2 is

$$q_{12} = \frac{1}{n_1} \sum_{a=1}^{n_1} \sum_{b=1}^{n_2} z_{i_a k_b},$$

whereas

$$q_{21} = \frac{1}{n_2} \sum_{a=1}^{n_1} \sum_{b=1}^{n_2} z_{i_a k_b}.$$

Thus, q_{12} and q_{21} are not necessarily equal; they are equal only if $n_1 = n_2$.

2.2 *Physical distances*

A set of 20 representative properties of amino acids (Hutchens 1976; Prabhakaran and Ponnuswamy 1979) was used in the analysis. The physical distance between a pair of amino acids for a chosen property was defined to be the absolute value of the difference between the numerical values of the property for the two amino acids under consideration.

2.3 *Tests of significance for r*

As mentioned before, the test based on Fisher's z transform for determining the statistical significance of r is not very appropriate, owing to the non-independence and non-normality of the values of genetic and physical distances. The question we wish to ask is: how does the value of r obtained here compare with what may be obtained by *chance alone*?

To obtain the various values of r that might occur by chance, we need to consider alternative genetic codes. Unconstrained realizations of possible genetic codes (e.g. UUU, GGG, AAA and CCC coding for the same amino acid) are clearly inappropriate. We therefore keep the existing level of degeneracy and assignment of codons unchanged. Thus GCU, GCC, GCA and GCG are assumed to code for the same amino acid, but not necessarily alanine; AAA and AAG are assumed to code for the same amino acid, but not necessarily lysine; and so on. We thus keep the 20 sets of codons unchanged, and obtain realizations of alternative genetic codes by assigning the 20

amino acids randomly, one to each set of codons (Haig and Hurst 1991). For every such genetic code, one can compute the conversion probabilities, and correlate them with physical distances.

A FORTRAN program implementing the above approach has been developed. The necessary computational details are described here.

The genetic distances were stored in a 20×20 matrix G, and the physical distances in a matrix D. If these 400 elements are denoted by variables (in the program, arrays) X and Y respectively, the correlation coefficient r is given by

$$r = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{s_X \cdot s_Y},$$

where the bar denotes the mean and s the standard deviation.

An alternative genetic code is simulated by randomly permuting the serial numbers of the amino acids in the genetic code. This amounts to generating a new 20×20 matrix, say G'. The elements of G' are obtained by appropriate permutations of the rows and columns of G. The array (say X') obtained from G' thus contains the same elements as in X, but in a different order.

We have made use of this fact, and linearly transformed the elements of the original matrices G and D such that each set of 400 values has mean zero and standard deviation unity. The computation of r is thus much more efficient, since r is simply $\overline{XY}$.

Generating all the possible genetic codes even under the constraint imposed here would involve $20!$ ($\cong 2{\cdot}4 \times 10^{18}$) permutations, which is unrealistic. We have restricted our analysis to between 500 and 1000 permutations. For each permutation, the array X' was generated by looking up the elements of G in an appropriate order, and r was computed. We also computed the mean and standard deviation of the values of r so determined. For a given property and a given G matrix, the mean and standard deviation stabilized by about 400 simulations, which indicates that this is an adequate sample size. We fixed our sample size to be 1000 simulations. For any property, we can now compare the r value for the existing genetic code with the distribution of r values obtained from the simulations.

3. Results and discussion

3.1 *Transition probabilities*

The codon transition probabilities p_0, p_1, p_2 and p_3 are given in table 1 as a function of y, the mutation rate. As expected, p_0 monotonically decreases as y increases, while p_1 increases initially with y, and then decreases as y increases still further. It is interesting to note that when the value of p_1 (single base change) is highest (for $y \cong 1$), the value of p_0 (probability of no change) is still appreciably high ($\cong 0{\cdot}20$ to $0{\cdot}40$), emphasizing the importance of taking into account the probability of 'no change' in estimating the effects of mutation. On the other hand, for higher values of y, probabilities of two- and three-base changes (p_2 and p_3) also become appreciably high. In the light of these results, considering single-base changes alone as an indicator of the effect of mutations, as in the earlier investigations (e.g. Haig and Hurst 1991), does not seem to be fully justified.

Table 1. Codon transition probabilities for no change (p_0), one base change (p_1), two changes (p_2) and three changes (p_3) for different values of y, the average number of mutations.

y	p_0	p_1	p_2	p_3
0·1000	0·90534	0·09154	0·00309	0·00003
0·2000	0·82058	0·16773	0·01143	0·00026
0·3000	0·74461	0·23073	0·02383	0·00082
0·4000	0·67648	0·28240	0·03930	0·00182
0·5000	0·61532	0·32435	0·05699	0·00334
0·7500	0·48818	0·39544	0·10677	0·00961
1·0000	0·39043	0·43128	0·15880	0·01949
2·0000	0·17405	0·41305	0·32674	0·08616
3·0000	0·08973	0·33210	0·40969	0·16847
4·0000	0·05348	0·26540	0·43903	0·24208
5·0000	0·03636	0·22016	0·44443	0·29905
7·5000	0·02120	0·16619	0·43430	0·37832
10·0000	0·01734	0·14894	0·42654	0·40719
20·0000	0·01564	0·14072	0·42193	0·42170
30·0000	0·01563	0·14063	0·42188	0·42187

Table 2. Amino acid properties examined (from Hutchens 1976; Prabhakaran and Ponnuswamy 1976).

1. Thermodynamic transfer hydrophobicity
2. Protein environment or bulk hydrophobicity
3. Polarity
4. Isoelectric point pH_i
5. pK, Equilibrium constant with respect to the COOH group
6. Molecular weight
7. Bulkness
8. Chromatography index
9. Refractivity index
10. Short- and medium-range nonbonded energy
11. Long-range nonbonded energy
12. Protein environment total nonbonded energy
13. Power to be at the N-terminal of alpha helix
14. Power to be at the C-terminal of alpha helix
15. Power to be at the middle of alpha helix
16. Power to adopt beta sheet
17. Power to adopt beta bend
18. Heat capacity
19. Absolute entropy
20. Entropy change for formation from elements

3.2 *Correlations between genetic and physical distances*

The 20 physical properties of amino acids (Hutchens 1976; Prabhakaran and Ponnuswamy 1979; Sitaramam 1989) we examined are listed in table 2. For a property, the physical distance between a pair of amino acids is, as mentioned before, the absolute value of the difference between the numerical values for the amino acids.

Table 3. Correlation coefficients for the 20 properties as a function of mutation rate y.

property \ y -->	0.10	0.20	0.30	0.40	0.50	0.75	1.00	2.00	3.00	4.00	5.00	7.50	10.0	20.0	30.0
1	-0.31	-0.31	-0.31	-0.31	-0.31	-0.31	-0.31	-0.30	-0.26	-0.21	-0.17	-0.10	-0.08	-0.07	-0.07
2	-0.30	-0.31	-0.31	-0.32	-0.32	-0.33	-0.33	-0.33	-0.28	-0.21	-0.14	-0.03	0.01	0.03	0.03
3	-0.19	-0.19	-0.19	-0.19	-0.19	-0.19	-0.19	-0.18	-0.16	-0.13	-0.10	-0.06	-0.04	-0.03	-0.03
4	-0.21	-0.21	-0.21	-0.21	-0.20	-0.20	-0.20	-0.17	-0.12	-0.08	-0.04	0.02	0.03	0.04	0.04
5	-0.27	-0.27	-0.26	-0.26	-0.26	-0.26	-0.25	-0.22	-0.18	-0.13	-0.09	-0.03	-0.01	0.00	0.00
6	-0.30	-0.30	-0.30	-0.30	-0.30	-0.31	-0.31	-0.29	-0.24	-0.17	-0.11	-0.02	0.01	0.03	0.03
7	-0.28	-0.29	-0.29	-0.29	-0.29	-0.29	-0.29	-0.25	-0.18	-0.09	-0.01	0.09	0.13	0.14	0.14
8	-0.31	-0.32	-0.32	-0.33	-0.33	-0.34	-0.34	-0.34	-0.31	-0.25	-0.19	-0.10	-0.07	-0.05	-0.05
9	-0.29	-0.30	-0.30	-0.30	-0.31	-0.31	-0.32	-0.32	-0.29	-0.24	-0.19	-0.10	-0.07	-0.06	-0.06
10	-0.31	-0.31	-0.31	-0.30	-0.30	-0.30	-0.29	-0.25	-0.17	-0.09	-0.02	0.08	0.11	0.12	0.12
11	-0.32	-0.33	-0.34	-0.34	-0.35	-0.36	-0.37	-0.40	-0.40	-0.36	-0.32	-0.23	-0.19	-0.17	-0.17
12	-0.31	-0.32	-0.32	-0.33	-0.33	-0.34	-0.34	-0.35	-0.31	-0.26	-0.19	-0.09	-0.06	-0.04	-0.04
13	-0.31	-0.31	-0.31	-0.31	-0.31	-0.31	-0.31	-0.26	-0.18	-0.09	-0.01	0.10	0.13	0.15	0.15
14	-0.26	-0.26	-0.26	-0.27	-0.27	-0.27	-0.26	-0.24	-0.20	-0.15	-0.10	-0.03	-0.01	0.00	0.00
15	-0.31	-0.31	-0.31	-0.31	-0.31	-0.30	-0.30	-0.27	-0.23	-0.18	-0.14	-0.08	-0.06	-0.05	-0.05
16	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.27	-0.23	-0.18	-0.13	-0.07	-0.05	-0.04	-0.04
17	-0.31	-0.31	-0.31	-0.31	-0.30	-0.30	-0.29	-0.23	-0.16	-0.10	-0.05	0.00	0.02	0.03	0.03
18	-0.26	-0.26	-0.26	-0.26	-0.26	-0.26	-0.25	-0.19	-0.11	-0.02	0.06	0.16	0.19	0.21	0.21
19	-0.31	-0.32	-0.32	-0.32	-0.33	-0.33	-0.34	-0.35	-0.33	-0.29	-0.25	-0.17	-0.14	-0.12	-0.12
20	-0.14	-0.14	-0.14	-0.14	-0.14	-0.14	-0.13	-0.10	-0.04	0.01	0.06	0.12	0.13	0.14	0.14

The correlation coefficient between physical and genetic distances was computed for each property for 15 different values of the mutation intensity y, ranging from $y = 0{\cdot}10$ to $y = 30{\cdot}0$ (which represents $y \cong \infty$). The results are presented in table 3.

As mentioned before, a large negative value indicates that the property is less sensitive to mutational changes. For low rates of mutation, the value of the correlation coefficient (r) is negative for *all* the properties. Since the number of degrees of freedom is nearly 400, any value of r less (more negative) than $-0{\cdot}103$ is significant at 1%

level by a one-tailed test based on Fisher's z transform. By this criterion, even up to a mutation intensity $y = 3$, nineteen of the 20 properties (table 3, all except property 20, viz. entropy of formation) show significant negative correlation. In fact, 10 of the 20 properties have r less than $-0{\cdot}20$. By $y = 7{\cdot}5$, however, only properties 11 (long-range nonbonded interaction energy) and 19 (absolute entropy) retain such low values of r.

Variation of r with y indicates two broad patterns. For most of the properties, r increased (became less negative) with increasing y. For some of the properties, e.g. 2 (protein environment or bulk hydrophobicity), 8 (chromatography index), 11 (long-range nonbonded energy) and 19 (absolute entropy), r decreased with increasing y initially, and then increased with further increase in y. The minimum value of r corresponded to y between 1 and 2·5 for these properties. It seems that these properties are more resistant to change at an intermediate level of mutation than at low levels! The initial decrease in r for these properties was, however, small in magnitude (e.g. 0·08 or 25% for property 11, long-range nonbonded interaction energy).

Table 4 shows the proportions of r values in the simulations that were smaller than the value obtained for the existing genetic code. A number less than 0·05 indicates a 95% level of significance. It is seen from the table that even for the low mutation rate $y = 0{\cdot}50$, for two of the properties (5, pK for COOH group; 17, power to adopt beta bend) a *majority* of the simulation r values were *less* than the actual value! When the mutation rate was 3·0, the following properties also gave r values with the same feature: 4, isoelectric point; 5, pK for COOH; 7, bulkness; 10, short- and medium-range nonbonded energy; 13, power to be at the N-terminal of alpha helix; 18, heat capacity; 20, entropy of formation from elements. In other words, though the z test suggested that these values are significant (table 3), the rigorous randomization tests revealed a more than even possibility of obtaining such low values by chance. This result underscores the necessity of exercising care in the choice of tests of significance.

One can infer from table 4 that at low levels of mutation rate, the genetic code conserves properties 2 (protein environment or bulk hydrophobicity), 8 (chromatography index), 9 (refractivity index), 11 (long-range nonbonded energy), 12 (protein environment total nonbonded energy) and 19 (absolute entropy). In fact, for even lower rates of mutation, ($y = 0{\cdot}001$, 0·005, 0·01, 0·05; detailed results not shown), these are the only properties that show a negative correlation coefficient significant at 1% level. For high rates of mutation ($y > 4$), only property 11, long-range nonbonded interaction energy, is seen to be significant at 1% level. The result obtained by Sitaramam (1989) thus remains valid under this rigorous scrutiny.

In a similar study but with a different (and more approximate) formalism, Di Giulio (1989) examined 18 physicochemical properties of amino acids, and showed that polarity was a highly conserved property, followed by molecular volume. In a subsequent study of four properties, Haig and Hurst (1991), with minimization of the mean square error as the design criterion and a Monte Carlo technique for testing significance, obtained evidence against molecular volume as a conserved property, but confirmed the observation for polarity (polar environment) and demonstrated that hydropathy was also a highly conserved property. On the other hand, our results, based on what we believe to be a more rigorous measure of conversion probabilities, run counter to these in that polarity does not seem to be a conserved property. In fact, 25% or more of the randomly generated codes show a higher degree of conservation for polarity than the natural code. This point needs further, detailed scrutiny.

Table 4. Probability that the correlation coefficient for a property from a random genetic code is *less* than that obtained from the existing genetic code, for different values of mutation rate y.

property \ y -->	0.100	0.200	0.300	0.400	0.500	0.750	1.000	2.000	3.000	4.000	5.000	7.500	10.00	20.00	30.00
1	0.113	0.108	0.084	0.101	0.087	0.093	0.086	0.096	0.090	0.130	0.140	0.208	0.219	0.228	0.207
2	0.004	0.002	0.002	0.003	0.004	0.003	0.003	0.010	0.030	0.114	0.268	0.544	0.637	0.643	0.673
3	0.244	0.215	0.222	0.239	0.217	0.246	0.219	0.245	0.243	0.274	0.332	0.328	0.422	0.420	0.436
4	0.463	0.456	0.510	0.513	0.517	0.555	0.535	0.578	0.562	0.586	0.623	0.574	0.648	0.613	0.633
5	0.765	0.738	0.752	0.716	0.745	0.692	0.632	0.555	0.547	0.519	0.509	0.540	0.554	0.566	0.541
6	0.031	0.029	0.022	0.022	0.022	0.039	0.036	0.099	0.237	0.375	0.438	0.566	0.533	0.610	0.600
7	0.043	0.048	0.035	0.054	0.043	0.085	0.090	0.267	0.554	0.691	0.798	0.845	0.850	0.870	0.852
8	0.006	0.012	0.006	0.012	0.007	0.009	0.005	0.003	0.009	0.010	0.028	0.108	0.159	0.194	0.183
9	0.000	0.003	0.001	0.001	0.000	0.000	0.002	0.008	0.033	0.091	0.157	0.291	0.285	0.329	0.326
10	0.372	0.353	0.358	0.382	0.398	0.382	0.429	0.565	0.690	0.772	0.808	0.844	0.854	0.879	0.875
11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.006	0.013
12	0.001	0.002	0.006	0.004	0.004	0.004	0.003	0.001	0.011	0.024	0.088	0.205	0.293	0.313	0.336
13	0.094	0.113	0.127	0.139	0.133	0.177	0.217	0.401	0.624	0.746	0.840	0.916	0.929	0.933	0.926
14	0.114	0.133	0.150	0.142	0.150	0.180	0.179	0.273	0.331	0.381	0.438	0.490	0.507	0.524	0.505
15	0.320	0.295	0.295	0.308	0.296	0.292	0.308	0.331	0.334	0.317	0.328	0.327	0.322	0.315	0.312
16	0.191	0.176	0.202	0.199	0.192	0.226	0.205	0.268	0.296	0.324	0.317	0.341	0.384	0.368	0.371
17	0.704	0.728	0.735	0.756	0.748	0.767	0.769	0.814	0.823	0.793	0.736	0.689	0.659	0.650	0.654
18	0.354	0.395	0.381	0.436	0.445	0.511	0.543	0.703	0.824	0.876	0.905	0.909	0.928	0.937	0.941
19	0.010	0.002	0.003	0.008	0.001	0.004	0.004	0.005	0.003	0.005	0.019	0.066	0.079	0.095	0.082
20	0.293	0.326	0.353	0.418	0.415	0.528	0.559	0.740	0.761	0.770	0.797	0.816	0.802	0.816	0.811

We have examined three more aspects of these results.

(i) Since our main interest is in looking at change in a physical property when there is a change of amino acid, it may be worth looking at the correlation coefficient for the 380 pairs of values of genetic and physical distances, omitting the 20 cases where an amino acid does not change ('conversion to itself'). The results of such an analysis show that even for a low mutation rate of $y = 0{\cdot}1$, and even by Fisher's z transform-

based test, 12 of the 20 values of correlation coefficient are not significantly different from zero. Properties 11 (long-range nonbonded energy) and 19 (absolute entropy) continue to have large negative values of r, however.

(ii) In addition to the correlation between the genetic distance (G) and the physical distance (D), we have examined the correlations between $\log(G)$ and D, G and $\log(D)$, and $\log(G)$ and $\log(D)$. All four correlation coefficients have by and large similar values. Here too, a large negative correlation coefficient is obtained only for the two properties 11 and 19 over the entire range of mutation rate examined.

(iii) For each property, one can compute the correlation coefficient between the genetic and physical distances for each amino acid separately, from the 19 values of probability of conversion to other amino acids and the 19 values of the physical distances from the other amino acids. For each property, one obtains 20 values of r, one for each amino acid (Di Giulio 1989). A property that is conserved by the genetic code should have all or most of these 20 values negative. We see that long-range nonbonded energy shows a negative correlation for at least 19 of the 20 values over the entire range of mutation rate. The only exception is tyrosine—perhaps a hint that it may not have been well integrated into the genetic code.

The approach adopted here can be extended in several ways. Clearly, a larger number of properties of amino acids should be examined. Secondly, one can try to unravel the logic of the genetic code in the two-letter code itself. Thirdly, a finer analysis may assign different relative probabilities of mutation for transitions and transversions. These investigations are in progress.

Acknowledgements

We wish to acknowledge helpful discussions with Dr A. P. Gore and Dr. Vivek Borkar. The work was supported in part by a grant to V. S. from the Department of Science and Technology, Government of India.

References

Crick F. H. C. 1968 The origin of the genetic code. *J. Mol. Biol.* 38: 367–379

Crothers D. M. 1982 Nucleic acid aggregation geometry and the possible evolutionary origin of ribosomes and the genetic code. *J. Mol. Biol.* 162: 379–391

Di Giulio M. 1989 Some aspects of the organization and evolution of the genetic code. *J. Mol. Evol.* 29: 191–201

Eigen M. and Winkler-Oswatitsch R. 1981 Transfer-RNA, an early gene? *Naturwissenschaften* 68: 282–292

Haig D. and Hurst L. D. 1991 A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* 33: 412–417

Hutchens J. O. 1976 Heat capacities, absolute entropies and entropies of formation of amino acids and related compounds. In *Handbook of biochemistry and molecular biology. D. Physical and chemical data* (ed.) G. D. Fasman (Cleveland, Ohio: CRC Press) pp. 109, 110

Jukes T. H. 1966 *Molecules and evolution* (New York: Columbia University Press)

Konecny J., Eckert M., Schoniger M. and Ludwig Hofacker G. 1993 Neutral adaptation of the genetic code to double-strand coding. *J. Mol. Evol.* 36: 407–416

Osawa S. and Jukes T. H. 1989 Codon reassignment (codon capture) in evolution. *J. Mol. Evol.* 28: 271–278

Prabhakaran M. and Ponnuswamy P. K. 1979 The spatial distribution of physical, chemical, energetic and conformational properties of amino acid residues in globular proteins. *J. Theor. Biol.* 80: 485–504

Salemme F. R., Miller M. D. and Jordan S. R. 1977 Structural convergence in protein evolution. *Proc. Natl. Acad. Sci. USA* 74: 2820–2824

Sitaramam V. 1989 Genetic code preferentially conserves long-range interactions among the amino acids. *FEBS Lett.* 247: 46–50

Sokal R. R. and Rohlf F. J. 1981 *Biometry* (New York: W. H. Freeman)

Woese C. R. 1965 On the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA* 82: 1160–1164