Volume 247, number 1, 46–50 FEB 06994 April 1989

# Genetic code preferentially conserves long-range interactions among the amino acids

V. Sitaramam

*Biotechnology, Department of Zoology, University of Poona, Pune 411 007, India*

Received 27 January 1989

The physical properties of amino acids were investigated in order to evaluate their possible relationship to the assignment of codons for amino acids in the genetic code. A comparison of the interconversion probability between amino acids and the distances between the amino acids for individual physical properties revealed a striking hierarchy among the physical properties. Surprisingly, it is the long-range/solvent interactions and not the short-range/stereochemical properties which are preferentially conserved in the genetic code.

Physical properties; Frozen accident; Stereochemical hypothesis/fluctuation

## 1. INTRODUCTION

The assignment of specific codons to amino acids has remained an enigma since the discovery of the genetic code despite very logical arguments put forward to account for the nearly universal code. The first was the 'frozen accident' hypothesis according to which there could be no discernible pattern in the assignment of codons to amino acids [1]. This hypothesis was essentially a null hypothesis in that the discovery of any 'significant' pattern would falsify the (null) hypothesis. The second argument was represented by a number of 'stereochemical' hypotheses, e.g. a recognition between the anticodons of the tRNAs and the amino acids [2–4]. A convincing proof remains to be demonstrated, particularly since there exists no a priori consensus as to what level of significance or universality one can attach to any specific claim for stereochemical recognition. The third pertains to the possibility that the codes are so distributed for the amino acids as to conserve the physical properties such that the cost of mutations is reduced. This approach also derives from information-theoretic approaches such as the Gray code, with the purpose of minimizing sources of mutational error [5]. A major piece of evidence of this kind was based on calculations of the probabilities of occurrence of substitutions in proteins to the extent that sequences are available [6].

Since each amino acid is characterized by a number of physical properties, it would be inconceivable that all properties can be equally conserved. Are structural properties, p*K*/ionization, short-range interactions more important, these being the determinants of the presence of activity? Are thermodynamic properties more important and, if so, which are of greater significance, the short-range or long-range interactions? Obviously, such an answer must be sought from the genetic code itself rather than from its products. The significance of such a hierarchy, if found, is two-fold. Firstly, it confirms the intuition that some ordering should be present. Secondly, if any meaningful association among the hierarchically ordered physical properties of the amino acids becomes preponderant, this will have a two-fold effect of providing the much sought-after 'logical link' in the assignment of codons to the amino acids and also in providing stronger confirmation than tests of correlation alone.

*Correspondence address:* V. Sitaramam, Biotechnology, Department of Zoology, University of Poona, Pune 411007, India



*Published by Elsevier Science Publishers B.V. (Biomedical Division)*
00145793/89/$3.50 

## 2. METHODS

### 2.1. *Determination of amino acid interconversion probabilities*

The nucleotide substitution among codons represents a trivial case wherein the probabilities of 1, 2 or 3 substitutions are 1/9, 1/81 and 1/729, respectively. However, since different amino acids are coded by one or more codons, the interconversion probability among amino acids requires to be defined a priori. I adopted a simple strategy of determining the shortest distance (i.e. highest probability of occurrence) between the starting and ending amino acids for any given starting codon and merely determined the average path for all starting codons. That is to say, if AAA and AAG code for the starting amino acid and GAA and GGA for the ending amino acid, the interconversion probability would be 1/9 for AAA and 1/81 for AAB, such that the sum is 10/81 or 0.1234. Since the number of starting and ending amino acids need not be equal, the 20 × 20 interconversion probability matrix would not be fully symmetric. Fig.1 summarizes the frequency distribution of these interconversion probabilities. From this, it was clear that, among the 400 probabilities, low probabilities dominate nearly exponentially indicating the low odds in chance correlations despite the large number of probabilities involved. It is generally observed that an amino acid remaining unchanged despite a single nucleotide substitution, i.e. probability for self-mutations for each amino acid, tends to be relatively high.

### 2.2. *Calculation of physical distances*

Physical distances were calculated as the relative distance in units of the physical property taking both negative and positive values among the 400 physical distances. The absolute distance represents a modulus of this measure. The properties and the sources are listed in table 1.

## 3. RESULTS

Fig.1 shows that the genetic code imposes low interconversion probabilities among amino acids. In order to test the hypothesis as to whether the genetic code is so arranged as to conserve one or several physical properties, a number of tests can be devised. The simplest and most direct test is to determine whether the interconversion probabilities between the 20 amino acids and the physical distances among these amino acids are related. Since $n = 400$, one would expect any correlations, if present, to be readily apparent. The presence of a significant correlation suggests that, in view of low interconversion probabilities among amino acids, this significant correlation could be attained by distributing the available, few higher interconversion probabilities at shorter physical distances for important physical variables by an appropriate assignment of codons to amino acids. This can be tested readily by checking whether the variance associated with the probabilities varies

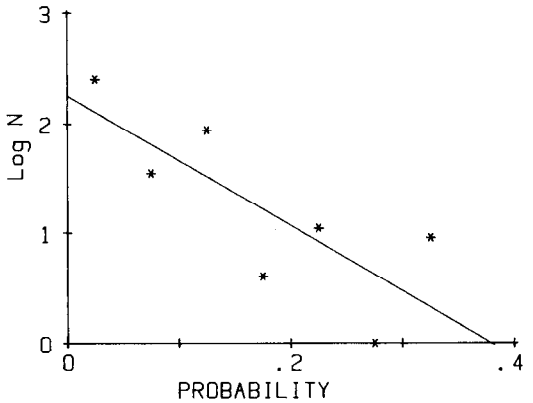


Fig.1. The interconversion probability matrix data for amino acids were obtained as explained in the text. The 20 × 20 matrix yielded 400 probabilities whose frequency distribution was plotted as a histogram with 7 class intervals. It should be noted that the matrix is significantly different from that of Dayhoff [6] because the interconversion probability matrix did not require calculations for a fixed number of mutations. The regression equation was log $N$, $Y = 2.251 - 5.941X$, coefficient of correlation, $r = 0.787$, $P < 0.05$.

with the physical distance. Important negative results include the absence of any relation to the nature of distribution of physical properties among the amino acids.

Table 1 shows that there exists a definitive correlation between the interconversion probabilities and the physical distances for several physical properties. Interestingly, there also exists a correlation for several of these physical properties that the variance associated with probabilities decreases with the physical distance, indicating the importance of the assignment of codons to impose higher probabilities of interconversion. Taken together for 18 physical properties (legend to table 1), the rank order of direct linear correlation ($r_1$) exhibits a reasonable ($P \leq 0.05$, by nonparametric tests of significance) association with the decrease in variance of interconversion probabilities with increase in physical distances. The frequency distributions of physical properties among the amino acids, also listed in table 1, offer no specific relationship to $r_1$ or $r_2$.

Fig.2C,D illustrates the correlations obtained for long-range non-bonded energy, in terms of the absolute distances vs probabilities as well as relative distances. For comparison, the physical property was substituted by random numbers and

Table 1

Physical properties of amino acids: their frequency distributions and the hierarchy of correlations between the physical distances and the interconversion probabilities for 20 amino acids

| Property | Correlation | | | | Frequency distribution | | | |
|---|---|---|---|---|---|---|---|---|
| | $r_1$ | $P_1$ | $r_2$ | $P_2$ | Mean | C.V.(%) | Skewness | Kurtosis |
| (1) Long-range non-bonded energy | −0.345 | $7.72 \times 10^{-13}$ | −0.797 | $3.0 \times 10^{-4}$ (14) | 0.550 | 27.59 | 0.108 | 1.939 |
| (2) Refractivity index | −0.309 | $1.74 \times 10^{-11}$ | −0.187 | $2.7 \times 10^{-1}$ (13) | 19.26 | 52.78 | 0.470 | 3.035 |
| (3) Protein environment or bulk hydrophobicity | −0.305 | $3.62 \times 10^{-10}$ | −0.785 | $7.0 \times 10^{-4}$ (13) | 12.88 | 12.24 | 0.495 | 1.850 |
| (4) Protein environment total non-bonded energy | −0.302 | $5.32 \times 10^{-10}$ | −0.765 | $7.0 \times 10^{-4}$ (14) | 1.755 | 11.33 | −0.352 | 1.881 |
| (5) $M_r$ | −0.289 | $3.04 \times 10^{-9}$ | −0.402 | $7.7 \times 10^{-2}$ (14) | 136.8 | 22.54 | 0.131 | 2.963 |
| (6) Chromatography index | −0.281 | $8.64 \times 10^{-9}$ | −0.607 | $1.1 \times 10^{-2}$ (14) | 10.03 | 55.96 | 0.142 | 1.519 |
| (7) Thermodynamic transfer hydrophobicity | −0.226 | $4.65 \times 10^{-6}$ | −0.728 | $1.6 \times 10^{-3}$ (14) | 1.414 | 83.07 | 0.424 | 1.984 |
| (8) Heat capacity[a] | −0.221 | $7.46 \times 10^{-6}$ | −0.592 | $1.7 \times 10^{-2}$ (13) | 42.98 | 27.43 | 0.417 | 2.619 |
| (9) Bulkness | −0.217 | $1.13 \times 10^{-5}$ | −0.118 | $4.3 \times 10^{-2}$ ( 5) | 15.36 | 30.13 | −0.556 | 3.414 |
| (10) Power to be at the N-terminal of $\alpha$-helix | −0.198 | $0.67 \times 10^{-5}$ | −0.476 | $5.0 \times 10^{-2}$ (13) | 0.947 | 50.36 | 0.060 | 2.418 |
| (11) Short- and medium-range non-bonded energy | −0.198 | $0.67 \times 10^{-5}$ | −0.585 | $1.4 \times 10^{-2}$ (14) | 1.201 | 10.51 | −0.382 | 2.555 |
| (12) $\beta$-Sheet adopting power | −0.184 | $2.02 \times 10^{-4}$ | −0.757 | $2.2 \times 10^{-3}$ (12) | 1.065 | 35.43 | −0.132 | 2.473 |
| (13) Power to be at the C-terminus of the $\alpha$-helix | −0.179 | $3.12 \times 10^{-4}$ | −0.590 | $1.3 \times 10^{-2}$ (14) | 1.032 | 53.07 | 1.033 | 2.849 |
| (14) Power to be at the middle of the $\alpha$-helix | −0.161 | $1.19 \times 10^{-4}$ | −0.612 | $1.3 \times 10^{-2}$ (13) | 1.048 | 52.59 | 0.104 | 2.490 |
| (15) Bend adopting power | −0.123 | $1.37 \times 10^{-2}$ | | | 1.065 | 42.43 | −0.076 | 1.712 |
| (16) Isoelectric point ($pH_i$) | −0.0836 | $9.5 \times 10^{-2}$ | 0.053 | $4.3 \times 10^{-1}$ (14) | 6.03 | 29.36 | 1.003 | 4.942 |
| (17) $pK_a$[b] | −0.1761 | $1.14 \times 10^{-1}$ | | | | | | |

Physical properties were obtained from [12]; units: (nos 3,7) kcal·mol$^{-1}$; (16,17) pH units; (15) Å$^2$; (1,4,11) kcal·mol$^{-1}$ atom; (8) cal·deg$^{-1}$·mol$^{-1}$; the remaining properties are dimensionless quantities. [a] Data from Handbook of Biochemistry and Selected Data for Molecular Biology, 2nd edn, CRC Press (1970). $r_1$, coefficient of correlation between absolute physical distances and interconversion probabilities (as in fig.2B,D). $P_1$, probability for the significance of the correlation using Fisher's Z transformation. $r_2$, coefficient of correlation between standard deviation in probabilities and absolute physical distances, obtained by grouping the physical distances into finite class intervals ($n$, in parentheses; analysis omitted when $n < 3$). $P_2$, corresponding probability for the significance of the correlation, using Fisher's Z transformation. [b] Only 7 charged amino acids considered. A number of other less important properties, such as entropy of formation, polarity and absolute entropy, are not included, being less significant. The ranking of properties based on $r_1$ was such that the first half of the properties tend to be general, long-range thermodynamic, environment (solvent)-based, as opposed to the second half which are increasingly related to the stereochemical attributes of the amino acids. Indeed, Fisher's Z transformation showed that an $r_1$ of 0.345 differs from any other $r_1$ (at a cut-off of 5%) of −0.239 or larger, i.e. from any property below the 6th property in the table

similar plots were obtained. In a series of 12 such trials with random numbers, one instance was observed at $P < 0.05$ (fig.2A,B) as would be expected. It may be noted that most of the correlations for true physical properties were way above this level of significance (table 1). The physical properties could be clearly ranked based on the level of significance as shown in table 1.

A systematic evaluation of the physical properties by a multiple regression analysis as well as principle component analysis (of both the primary data as well as the covariance matrix) did not reveal any intrinsic bias towards a hierarchy among these physical properties per se.

## 4. DISCUSSION

The starting premise for these analyses was the simplistic notion that, if physical properties were to be conserved at all, one would expect the short-

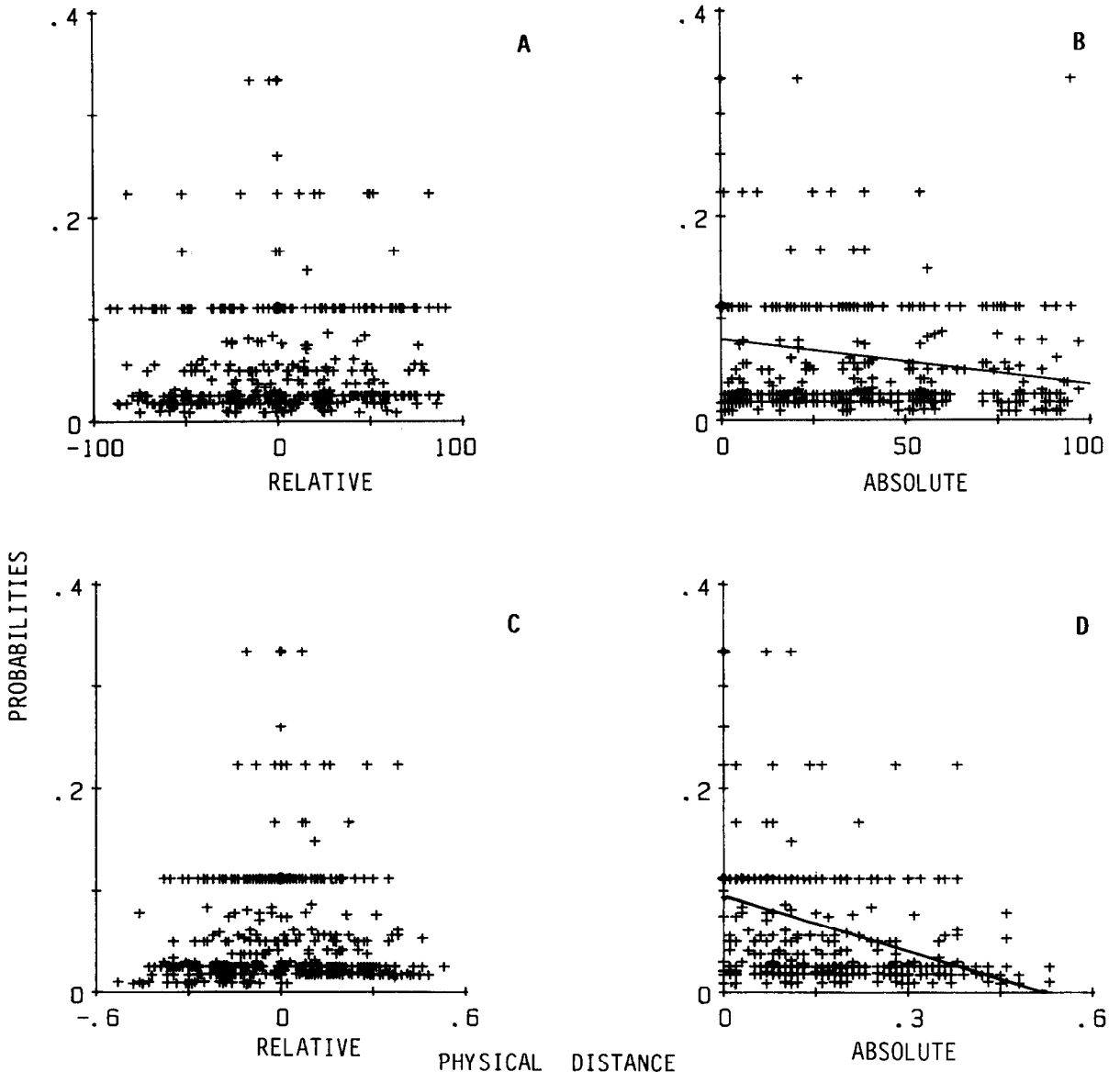


Fig.2. Plots of physical distances vs interconversion probabilities. (A,B) Random numbers substituted for physical properties, (C,D) long-range non-bonded energy (see table 1). B,D represent the absolute distances (modulus of distance) whereas A,C represent the actual distances. The lines in B,D represent linear regression by the method of least squares.

range interactions, i.e. the stereochemical attributes, such as pH and p$K$, to be the dominant modes. Such a premise was based on the logical expectation that protein folding and activity would be intimately related via short-range interactions which would be critical for the specificity of the protein function. This was not to be the case!

The pattern of hierarchy, in itself, is coherent enough to exclude the frozen accident hypothesis as the basis for the assignments of codons. The pattern of hierarchy clearly indicated that the stereochemical hypotheses do not appear to be of importance. While one associates mutations that affect such short-range interactions with lethal

mutations in the sense that the activity of such proteins would be seriously compromised, the long-range and solvent interactions would be associated with the mechanism of catalysis itself. A fluctuational view of catalysis and membrane phenomena associated with enzyme/transport/energy transduction processes based on enthalpy-entropy compensation mechanisms is gaining ground [7–10]. The assignment of codons in the genetic code also appears to specify a greater emphasis on the modulation of activity rather than on mere preservation of activity, coevolution of codons and amino acids by trial and error in the process of optimizing the conservation of physical properties rather than de novo (as with a frozen accident) and that global influences would dominate rather than local changes, which should find expression in thermodynamic properties of the enzymes, as amply demonstrated with regard to site-directed mutagenesis in lysozyme [11].

*Acknowledgements:* The author is grateful to Dr R. Gadagkar, Professor A.P. Gorey and Professor J. Barnabas for helpful discussions, V.V. Korde and N.M. Rao for technical assistance, and the Department of Science and Technology, India for a Grant-in-Aid.

## REFERENCES

[1] Crick, F.H.C. (1968) J. Mol. Biol. 38, 367–379.
[2] Woese, C.R. (1967) The Genetic Code, Harper & Row, New York.
[3] Dunnill, P. (1966) Nature 210, 1267.
[4] Root-Bernstein, R.S. (1982) J. Theor. Biol. 94, 895–904.
[5] Swanson, R. (1984) Bull. Math. Biol. 46, 187–203.
[6] Dayhoff, M.O. (1978) Atlas of Protein Sequence and Structure, vol.5, suppl.3, National Biomedical Research Foundation, Washington, DC.
[7] Somogy, B., Welch, G.R. and Damjanovich, S. (1985) Biochim. Biophys. Acta 768, 81–112.
[8] Sitaramam, V. and Sarma, M.K.J. (1981) Proc. Natl. Acad. Sci. USA 78, 3441–3445.
[9] Sitaramam, V. (1987) in: Interactions of Water in Ionic and Nonionic Hydrates (Kleeberg, H. ed.) pp.213–216, Springer, Berlin.
[10] Gavish, B. and Werber, M.M. (1979) Biochemistry 18, 1269–1275.
[11] Alber, T., Dao-Pin, S., Wilson, K., Wozniak, J.A., Cook, S.P. and Matthews, B.W. (1978) Nature 330, 41–46.
[12] Prabhakaran, M. and Ponnuswamy, P.K. (1979) J. Theor. Biol. 80, 485–504.