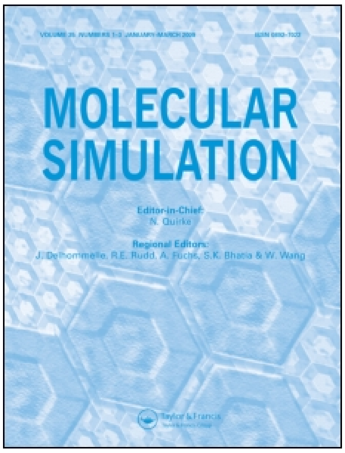


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Molecular Informatics: Quantifying Information Patterns in the Genetic Code

Dónall A. Mac Dónaill[a]; Matthew Manktelow[a]
[a] Department of Chemistry, Trinity College, Dublin, Ireland

To cite this Article Dónaill, Dónall A. Mac and Manktelow, Matthew(2004) 'Molecular Informatics: Quantifying Information Patterns in the Genetic Code', Molecular Simulation, 30: 5, 267 — 272

To link to this Article: DOI: 10.1080/08927020310001638749

URL: http://dx.doi.org/10.1080/08927020310001638749

Molecular Simulation, Vol. 30 (5), 30 April 2004 pp. 267–272

Molecular Informatics: Quantifying Information Patterns in the Genetic Code

DÓNALL A. MAC DÓNAILL* and MATTHEW MANKTELOW

Department of Chemistry, Trinity College, Dublin 2, Ireland

(Received September 2003; In final form October 2003)

The Genetic Code appears to be a non-random triplet code in which both the position of a nucleotide within a codon, as well as its physicochemical nature, contribute to the identity of the expressed amino acid. The non-randomness of the code is manifested in apparent patterns in the mapping from codon to amino acid; some of the patterns seem quite clear, while other more subtle patterns are less obvious or certain. Discussion in the literature has been largely qualitative in nature. In this study, we employ evolution similarity data, widely employed in the field of bioinformatics, to explore the patterns relating nucleotide features to amino acids. The results support a hierarchical order based on position and physicochemical features proposed by Jimenez-Montaño *et al.*, ["The Hypercube Structure of the Genetic Code Explains Conservative and Non-Conservative Amino Acid Substitutions *in vivo* and *in vitro*" *Biosystems* (1996) 39, pp. 117–125]. The method also provides a quantitative approach to testing the importance of other putative patterns.

Keywords: Genetic code; Nucleotide; Bioinformatics; Amino acid

INTRODUCTION

The relationship between nucleotide sequence and protein composition is captured in the Genetic Code. Amino acid identity is directed by nucleotide triplets or codons, most of which specify a particular amino-acid, and a few, which not coding for any amino acid, serve to signal termination. The Genetic Code is a non-random code, containing a variety of degeneracies and symmetries [1,2]; any model purporting to explain the structure of the code must also offer an explanation of the way information is structured. Perhaps the most immediately obvious and long recognised informational feature is that the significance of a nucleotide, N, is strongly dependent on its position within the codon, and as inspection will readily verify, generally follows the order $N_2 > N_1 > N_3$. In many instances changing the final nucleotide makes no difference to the expressed amino acid, and where it does, the new amino acid is often physicochemically similar. Thus, for example, changing the third nucleotide in codon GAU from U to A yields a closely related amino acid (Fig. 1a), whereas, by contrast, changing the second nucleotide from A to U changes the expressed amino acid from aspartic acid to valine, having quite different hydrophobicities (Fig. 1b).

The significance of a nucleotide appears to depend not only on its position but also on its chemical nature (C), that is, whether the nucleotide is a pyrimidine (Y), or a purine (R); this may be most easily appreciated by observing the effect of nucleotide substitution in the third position, where in so far as changing the nucleotide makes a difference, it is the size of the nucleotide which is important, with U and C in most instances proving equivalent, as do A and G. The hydrogen bonding strength (H) between a nucleotide and its complement also appears to have informational significance; *ab initio* calculations at the HF/6-31G* level of approximation yield *in vacuo* binding energies for A:U and C:G of 11.54 kCal mol^{-1} and 25.94 kCal mol^{-1} respectively [3], so that nucleotides A and U may be labelled weak (W), with C and G being labelled strong (S). Combining the roles of nucleotide position with physicochemical markers of size and strength of hydrogen-bonding, Swanson [4] proposed the following order of significance

$$C_2 > C_1 > H_2 > H_1 > C_3 > H_3 \qquad (1)$$

*Corresponding author. E-mail: dmcdonll@tcd.ie

ISSN 0892-7022 print/ISSN 1029-0435 online © 2004 Taylor & Francis Ltd
DOI: 10.1080/08927020310001638749

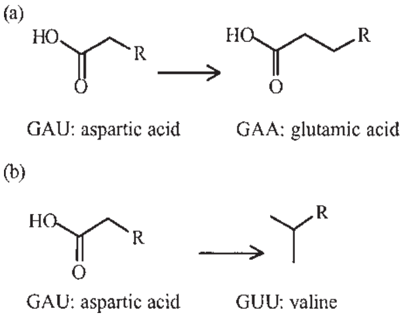


FIGURE 1 (a) Changing the third nucleotide in codon GAU changes the expressed amino acid from aspartic acid to the closely related glutamine; (b) changing the nucleotide in the second position yields valine.

whereas Jimenez-Montaño *et al.* [5] suggested a slightly different structure

$$C_2 > H_2 > C_1 > H_1 > C_3 > H_3 \qquad (2)$$

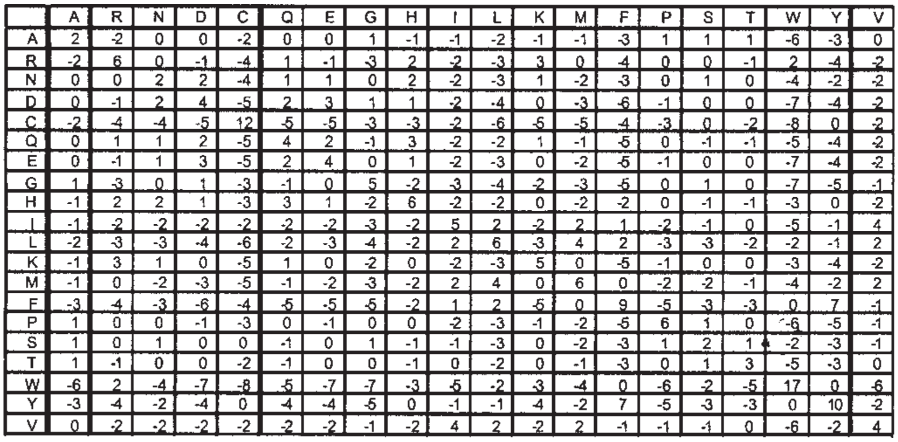
Both patterns are broadly similar, and have been used by their proposers to explore the Genetic Code as a "Gray code," more properly a one-bit change code, with Jimenez-Montaño *et al.* proceeding to develop a description of the code in terms of a six-dimensional binary hypercube [5]. However, the suggested hierarchical order in inequalities such as (1) and (2), and other structural insights such as the conjugate model [2], are essentially qualitative in nature, as a consequence of which the comparison of competing models, and particularly similar structures such as those just noted, is problematic. Accordingly, and with a view to further exploring the possible informational structures within the genetic code very much in mind, we outline below our approach for quantitatively determining the significance of nucleotide features on amino acid identity. The approach adopted is quite straightforward, and simply involves changing a selected nucleotide feature, and observing and quantifying the effect of the change, as judged by some measure of the difference between the amino acids expressed. This introduces the concept of amino-acid similarity, itself quite a broad topic. The next section, therefore, briefly rehearses the concept of amino-acid similarity, with particular reference to the PAM250 matrix, a measure widely employed in sequence similarity analysis. The methodology section outlines how such measures of similarity might be harnessed to assess the significance of a nucleotide feature in determining the identity of the expressed amino acid. The method outlined might be employed to assess the correlation between any nucleotide feature considered relevant. However, for the purposes of this study we explore the features considered in inequalities (1) and (2), although other features such as hydrogen donor–acceptor patterns might also be considered.

AMINO ACID SIMILARITY

There are a variety of ways in which a measure of the difference between amino acids might be constructed. A physicochemical perspective might favour a weighted mix based on chemical, functional, structural and other properties of amino acids, the approach adopted by Karlin and Ghandour [6]. The difficulty here of course is that while it is relatively easy to arrive at some measure of distance between amino acids based on a given property, the combination of the various property related measures of similarity, to yield some overall measure of distance is considerably more difficult. In the context of the Genetic Code the optimum weighting might be expected to reflect the contribution of various physicochemical properties in protein biochemical function, and at present our knowledge is limited. However, the more similar two amino acids are, the more frequently one substitutes for the other during evolution, so that amino acid evolutionary substitution rates afford an empirical measure of amino acid similarity. The concept of percent accepted mutations (PAM), due to Dayhoff [7], is the basis of one of the most commonly employed measures of similarity. The PAM250 matrix derives from evolutionary amino acid substitution data (Fig. 2). It is based on a data base of 1572 changes in 71 groups of closely related proteins. Each element, $s_{\alpha\beta}$ in the PAM250 matrix is a score the substitution probability of α by β over an evolution period, and ultimately reflects the similarity between a pair of amino acids, α and β; more positive scores indicate similarity, while less positive and negative scores reflect dissimilarity or difference. Alternative matrices such as the updated data sets of Jones [8] or Gonnet *et al.* [9], or indeed the widely used BLOSUM [10] data set, might equally well have been employed. However, pilot calculations indicate that the particular choice of matrix does not significantly affect the results.

METHODOLOGY

The purpose of this study is to explore the extent to which nucleotide features contribute, as a function of position, to determine amino acid expression. We report two experiments; in the first we explore the significance of nucleotide codonic position without reference to particular nucleotide features such as the purine/pyrimidine (C), weak/strong (H), or other properties. The second experiment incorporates these features, widely viewed as having particular informational significance. As the purine/pyrimidine and weak/strong features, each having just two states, are binary in nature, it proves convenient for both descriptive and coding



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

FIGURE 2 THE PAM250 matrix due to Dayhoff [7]. Ultimately derived from substitution rates the data can be interpreted as reflecting amino acid similarity. More positive numbers indicate strongly similar amino acids, with dissimilar amino acids have less positive or negative scores. Amino acids are indicated by the conventional one-letter code.

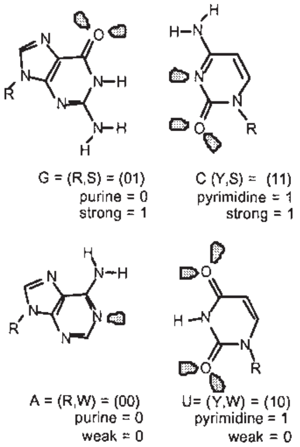


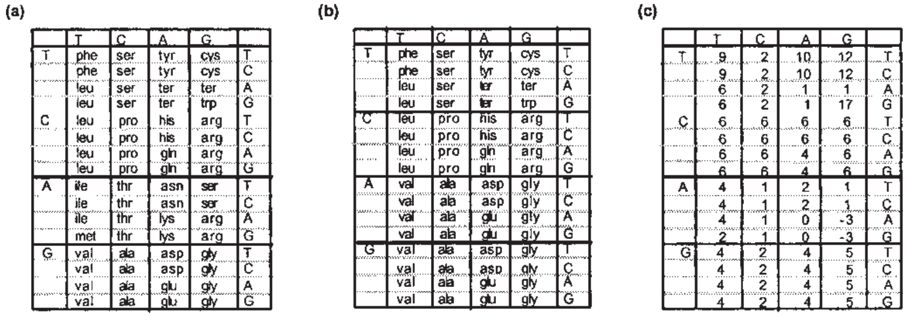
FIGURE 3 Binary interpretations of nucleotides as apparently manifest in the genetic code. Purines (R) = 0, and pyrimidines (Y) = 1, similarly, weak (W) = 0 and S (strong) = 1.

purposes to map the nucleotides to a two bit binary representation reflecting these features (Fig. 3) [11]. As codons have three nucleotides, each with two features of interest, there are a total of six features for investigation: C_1, C_2, C_3, H_1, H_2, and H_3.

Nucleotide Position

The approach adopted is as follows:

1. A nucleotide position, 1, 2, or 3, is selected. The effects of all twelve possible nucleotide mappings are considered, namely, (i) U → C, A and G, (ii) C → A, G and U, (iii) A → G, U and C, and (iv) G → U, C and A.
2. The mappings are applied to all 64 codons, in many instances changing the sense of the codon.
3. A modified genetic code table is constructed, in which the usually expressed amino acid or termination signal is replaced by the effect of the selected mapping. For example the mapping A → G in position one, maps the codon AAU (asparagine) to GAU (aspartic acid). The effect of the corresponding change across the genetic code is depicted in Fig. 4b. Figure 4a may be interpreted as a tensor **A**, with elements a_{ijk}, where codon coordinates, i, j and k span the values U, C, A and G. The effect of the selected mapping is captured in the modified tensor **B**, with elements b_{ijk} (Fig. 4b).



(a)

	T	C	A	G	
T	phe	ser	tyr	cys	T
	phe	ser	tyr	cys	C
	leu	ser	ter	ter	A
	leu	ser	ter	trp	G
C	leu	pro	his	arg	T
	leu	pro	his	arg	C
	leu	pro	gln	arg	A
	leu	pro	gln	arg	G
A	ile	thr	asn	ser	T
	ile	thr	asn	ser	C
	ile	thr	lys	arg	A
	met	thr	lys	arg	G
G	val	ala	asp	gly	T
	val	ala	asp	gly	C
	val	ala	glu	gly	A
	val	ala	glu	gly	G

(b)

	T	C	A	G	
T	phe	ser	tyr	cys	T
	phe	ser	tyr	cys	C
	leu	ser	ter	ter	A
	leu	ser	ter	trp	G
C	leu	pro	his	arg	T
	leu	pro	his	arg	C
	leu	pro	gln	arg	A
	leu	pro	gln	arg	G
A	val	ala	asp	gly	T
	val	ala	asp	gly	C
	val	ala	glu	gly	A
	val	ala	glu	gly	G
G	val	ala	asp	gly	T
	val	ala	asp	gly	C
	val	ala	glu	gly	A
	val	ala	glu	gly	G

(c)

	T	C	A	G	
T	9	2	10	12	T
	9	2	10	12	C
	6	2	1	1	A
	6	2	1	17	G
C	6	6	6	6	T
	6	6	6	6	C
	6	6	4	6	A
	6	6	4	6	G
A	4	1	2	1	T
	4	1	2	1	C
	4	1	0	-3	A
	2	1	0	-3	G
G	4	2	4	5	T
	4	2	4	5	C
	4	2	4	5	A
	4	2	4	5	G

FIGURE 4 (a) the genetic code—tensor **A**., (b) a modified genetic code where the sense of a codon is changed by substituting a specified nucleotide in a single position, in this case the mapping A → G is effected in position one; the table corresponds to tensor **B** in the text (c) the similarity of amino acids exchanged by effecting the A → G mapping, as estimated from PAM250—tensor **C**.

4. The elements of the PAM250 similarity matrix $S_{\alpha\beta}$ give a numerical measure of the similarity of amino acids α and β. The significance of the change effected by the selected mapping on each codon is expressed in tensor **C** (Fig. 4c) with elements c_{ijk}. Data for substitutions involving the termination codons were taken from ref. [12].

$$c_{ijk} = S_{a_{ijk}b_{ijk}} \tag{3}$$

In the example considered above, the mapping A → G in nucleotide 1 replaces, among others, isoleucine with valine; as these are similar amino acids the corresponding similarity score of 4 is high. Similarity scores for amino acids exchanging with termination codons, and for changes mapping one termination codon to another are taken from Ref. [11].

5. The overall significance S of a mapping may be estimated from the scores c_{ijk}. Summing across the full set of 64 codons yields a global measure of the effect of the selected mapping.

$$S = \sum_{\substack{i,j,k \\ (i,j,k=\mathrm{U,C,A,G})}} c_{ijk} = \sum_{\substack{i,j,k \\ (i,j,k=\mathrm{U,C,A,G})}} S_{a_{ijk}b_{ijk}} \tag{4}$$

It is convenient to scale S such that, where no change is effected, a score of 1 is obtained. A normalisation constant N may be defined as the sum of the self-similarity scores across the code, that is

$$N = \sum_{\substack{i,j,k \\ (i,j,k=\mathrm{U,C,A,G})}} S_{a_{ijk}a_{ijk}} \tag{5}$$

A modified Score S' is given by

$$S' = \frac{1}{N} \sum_{\substack{i,j,k \\ (i,j,k=\mathrm{U,C,A,G})}} c_{ijk} = \frac{1}{N} \sum_{\substack{i,j,k \\ (i,j,k=\mathrm{U,C,A,G})}} S_{a_{ijk}b_{ijk}} \tag{6}$$

6. The procedure is repeated so that all 12 possible mappings specified in step 1 above are considered. An average value for S' is obtained for each of the three nucleotide positions.

Nucleotide Features

The second experiment is similar in structure to the first, except that instead of considering all possible mappings at a given position, we examine instead the effect of modifying a particular nucleotide feature.

1. A nucleotide feature, C or H, on a single nucleotide position, 1, 2, or 3, is selected. Flipping the 'chemical' state of a nucleotide changes its pyrimidine/purine nature, and may be expressed as a mapping.

$$\begin{aligned} f_{\mathrm{C}} &: \{\mathrm{R} \rightarrow \mathrm{Y}, \mathrm{Y} \rightarrow \mathrm{R}\} \\ &= f_{\mathrm{C}} : \{\mathrm{A} \rightarrow \mathrm{U}, \mathrm{C} \rightarrow \mathrm{G}, \mathrm{G} \rightarrow \mathrm{C}, \mathrm{U} \rightarrow \mathrm{A}\} \end{aligned} \tag{7}$$

Numerically, this is equivalent to flipping the first bit in the 2-bit representation. Similarly, flipping the weak/strong feature while preserving the pyrimidine/purine nature may also be expressed as a mapping

$$\begin{aligned} f_{\mathrm{H}} &: \{\mathrm{W} \rightarrow \mathrm{S}, \mathrm{S} \rightarrow \mathrm{W}\} \\ &= f_{\mathrm{H}} : \{\mathrm{A} \rightarrow \mathrm{G}, \mathrm{C} \rightarrow \mathrm{U}, \mathrm{G} \rightarrow \mathrm{A}, \mathrm{U} \rightarrow \mathrm{C}\} \end{aligned} \tag{8}$$

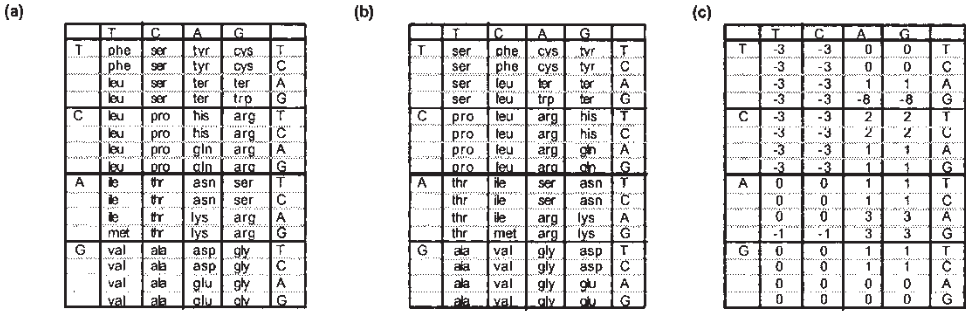
equivalent to flipping the second bit in the 2-bit representation.

2. The state of the selected feature is flipped in all codons; in most cases this results in a new amino acid, occasionally the expressed amino acid is left unchanged, while in a few cases nonsense codons signalling termination swap with codons expressing amino acids.
3. A modified genetic code table is constructed, in which the usually expressed amino acid or termination signal is replaced by the sense of mappings (7) or (8). The result of flipping hydrogen bonding or W/S nature of nucleotide 2 throughout the genetic code is depicted in Fig. 5b.
4. The significance of the change effected (Fig. 5c) is estimated using Eqs. (4) and (6).
5. The process is executed for all six nucleotide features; that is, the C and features in the three nucleotide positions.

RESULTS

We begin by testing the significance of the position of a nucleotide in a codon, as detailed in the 'nucleotide position' experiment described above. All 12 possible substitutions were effected at each codon position, and the similarity of the initial and modified codons assessed in terms of the similarity of the expressed amino acids (and termination signals). The data was averaged over all substitutions, that is, an average **C** tensor was obtained. Both the crude similarity score (unnormalised) from Eq. (4) and the normalised similarity scores as defined in Eq. (6) were determined (Table I).

The data indicates that, on average, perturbing nucleotides in the second position, N_2, results in a lesser overall similarity than perturbing in other positions. Modifying a nucleotide in the third position yields the greatest similarity, reflecting the fact that the third nucleotide position is the least significant. Averaged across the genetic code the informational significance of nucleotides follows



(a)

	T	C	A	G	
T	phe	ser	tyr	cys	T
	phe	ser	tyr	cys	C
	leu	ser	ter	ter	A
	leu	ser	ter	trp	G
C	leu	pro	his	arg	T
	leu	pro	his	arg	C
	leu	pro	gln	arg	A
	leu	pro	gln	arg	G
A	ile	thr	asn	ser	T
	ile	thr	asn	ser	C
	ile	thr	lys	arg	A
	met	thr	lys	arg	G
G	val	ala	asp	gly	T
	val	ala	asp	gly	C
	val	ala	glu	gly	A
	val	ala	glu	gly	G

(b)

	T	C	A	G	
T	ser	phe	cys	tyr	T
	ser	phe	cys	tyr	C
	ser	leu	ter	ter	A
	ser	leu	trp	ter	G
C	pro	leu	arg	his	T
	pro	leu	arg	his	C
	pro	leu	arg	gln	A
	pro	leu	arg	gln	G
A	thr	ile	ser	asn	T
	thr	ile	ser	asn	C
	thr	ile	arg	lys	A
	thr	met	arg	lys	G
G	ala	val	gly	asp	T
	ala	val	gly	asp	C
	ala	val	gly	glu	A
	ala	val	gly	glu	G

(c)

	T	C	A	G	
T	-3	-3	0	0	T
	-3	-3	0	0	C
	-3	-3	1	1	A
	-3	-3	-8	-8	G
C	-3	-3	2	2	T
	-3	-3	2	2	C
	-3	-3	1	1	A
	-3	-3	1	1	G
A	0	0	1	1	T
	0	0	1	1	C
	0	0	3	3	A
	-1	-1	3	3	G
G	0	0	1	1	T
	0	0	1	1	C
	0	0	0	0	A
	0	0	0	0	G

FIGURE 5 (a) the genetic code; (b) an apparently modified genetic code where the sense of a codon is changed by flipping the state of a specified nucleotide feature in a single position, in this case the state of the W/S nature in position 2; (c) the similarity of amino acids exchanged by effecting the W/S flip, as estimated from PAM250.

the order $N_2 > N_1 > N_3$. The result is well known but here is confirmed quantitatively, and serves as a simple test for the methodology adopted.

We turn now to the more interesting problem of the significance of nucleotide features, and not just of position. The effect of flipping the chemical (C) or purine/pyrimidine state, and the weak/strong or hydrogen-bonding state (H), was explored in each of the three nucleotide positions (Table II).

The normalised similarity score S' reflects the average similarity of codons, following a change in some feature, to the unperturbed codon. A value of $S' = 1$ would indicate perfect identity. The less positive, and by extension more negative, the value S', the less similar and more significant is the change in the meaning of the codon brought about by the changed feature. The data confirms that weak/strong nature of position 3, H_3, is the least significant. As inspection of the genetic code readily reveals, flipping the state of this feature yields a synonymous codon in almost all cases. Flipping the purine/pyrimidine nature of the nucleotide in position 2 yields codons which are less similar than those resulting from a change in any other feature, and we may conclude that C_2 is therefore the most significant information feature. The significance of both these features is well known; however, the data also indicates a clear hierarchy of significance; $C_2 > H_2 > C_1 > H_1 > C_3 > H_3$, clarifying the relative significance of H_2, C_1, and H_1 which is less obvious. The order of significance is in agreement with the analysis of Jimenez-Montaño *et al.* [5], favouring his structure over that proposed by Swanson [4].

TABLE I Unnormalised (S) and normalized similarity scores (S') for nucleotides in codon positions 1, 2 and 3 (N_1, N_2, N_3)

	N_1	N_2	N_3
S	235.9	220.1	281.1
S'	0.744	0.694	0.887

Normalisation constant $N = 317$ (see equation 5).

TABLE II Unnormalised (S) and normalized similarity scores (S') for nucleotides as a function of both codon position and nature of the feature changed

	C_1	H_1	C_2	H_2	C_3	H_3
S	−16	16	−60	−32	122	276
S'	−0.05	0.05	−0.19	−.10	0.38	0.87

Normalisation constant $N = 317$.

It is noteworthy that for all positions the chemical nature of a nucleotide proves more significant than its hydrogen-bonding nature, in line with experimental observation of the lower frequency of transversions (purine/pyrimidine exchanges) relative to transitions (exchange between pyrimidines or between purines). However, this does not apply to comparison between different positions, and notwithstanding the greater facility for transitions, a transition in position 2 will prove more damaging, and more significantly alter the expressed amino acid, than will a transversion in position 1.

Finally, we may obtain an alternative estimate of the relative significance of nucleotide positions by averaging the similarity data associated with the chemical and hydrogen-bonding features (Table III).

SUMMARY

The phenomenon of protein translation is one of the most fundamental informational possessing phenomena in molecular biology, and the elucidation of the underlying physicochemical or informatics factors, which might serve to shape the particular structure of the code, is an essential component in developing a more complete understanding of the emergence of living systems. The analysis confirms

TABLE III Overall similarity scores as a function of the feature changed

	C_1 and H_1	C_2 and H_2	C_3 and H_3
S	0	−46	199
S'	0	−0.15	0.63

Normalisation constant $N = 317$.

the long recognised significance of the purine/pyrimidine nature of the nucleotide at codon position 2, and almost total insignificance of the weak/strong hydrogen-bonding nature of the nucleotide in position 3. The relative significance of features H_2, C_1, and H_1, which is somewhat less obvious, is clarified. The hierarchy of significance is $C_2 > H_2 > C_1 > H_1 > C_3 > H_3$, in agreement with Jimenez-Montaño *et al.* [5]. We note however that the recent study of Ardell [13] suggests that a yet more subtle pattern may underlie this structure. We are currently pursuing this and related issues, such as the effect of amino acid frequencies, and differences between standard and primordial codes.

References

[1] Watson, J.D., Hopkins, N.H., Roberts, J.W., Steits, J.A. and Weiner, A.M. (1987) *Molecular Biology of the Gene: Chapter 15 The Genetic Code* (Benjamin/Cummings, California).

[2] Jayaram, B. (1997) "Beyond the Wobble: The Rule of Conjugates", *J. Mol. Evol.* **45**, 704–705.

[3] Mac Dónaill, D.A. and Brocklebank, D. (2003) "An *ab initio* quantum chemical investigation of the error-coding model of nucleotide alphabet composition", *Mol. Phys.* **101**, 2755–2763.

[4] Swanson, R. (1984) "A unifying concept for the amino acid code", *Bull. Math. Biol.* **46**, 187.

[5] Jimenez-Montaño, M.A., *et al.* (1996) "The hypercube structure of the genetic code explains conservative and non-conservative amino acid substitutions *in vivo* and *in vitro*", *Biosystems* **39**, 117.

[6] Karlin, S. and Ghandour, S. (1985) *Proc. Natl Acad. Sci.* **82**, 8597.

[7] Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) "A model for evolutionary change in proteins", In: Dayhoff, M.O., ed, *Atlas of Protein Sequence and Structure* (National Biochemical Research Foundation, Washington DC) Vol. **5**, pp 345–352.

[8] Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) "The rapid generation of mutation data matrices from protein sequences", *CABIOS* **8**, 275–282.

[9] Gaston, H., Gonnet, M.A., Cohen and Benner, S.A. (1992) "Exhaustive matching of the entire protein sequence database", *Science* **256**, 1443–1445.

[10] Henikoff, S. and Henikoff, J.G. (1992) "Amino acid substitution matrices from protein blocks", *Proc. Natl Acad. Sci.* **89**, 10915–10919.

[11] Mac Dónaill, D.A. and Buttimore, N.H. (1996) "The Exploitation of Assembly Language Instructions in Biological Text Manipulation: I Nucleotide Sequences", *Computers Maths. Applic.* **32**, 29–38.

[12] Pearson, W.A. (1990) "Rapid and sensitive sequence comparison with FASTP and FASTA", *Methods in Enzymology* (Academic Press, San Diego) **183**, pp 63–98.

[13] Ardell, D.H. (1998) "On Error Minimization in a Sequential Origin of the Standard Genetic Code", *J. Mol. Evol.* **47**, 1–13.