CODE DEPENDENT CONSERVATION OF THE PHYSICO-CHEMICAL PROPERTIES IN AMINO ACID SUBSTITUTIONS

MARIA PIEBER and JOSÉ TOHÁ C. **

Biofísica, Departamento de Física, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile,
Casilla 5487, Santiago de Chile

(Received 5 January; in revised form 24 May, 1983)

Abstract. The frequency of amino acid replacements in families of typical proteins has been elegantly analyzed by Argyle (1980) showing that the most frequent replacements involve a conservation of the amino acid chemical properties. The cyclic arrangement of the twenty amino acids resulting from the most frequent replacements has been described as an amino acid chemical ring.

In this work, a novel amino acid replacement frequency ring is proposed, for which a conservation of over 90 % of the most general physico-chemical properties can be deduced.

The amino acid chemical similarity ring is also analyzed in terms of the genetic code base probability changes, showing that the discrepancy that exists between the standard deviation value of the amino acid replacement frequency matrix and its respective ideal value is almost equal to that deduced from the corresponding base codon replacement probability matrices. These differences are finally evaluated and discussed in terms of the restrictions imposed by the structure of the genetic code and the physico-chemical dissimilarities between some codons of amino acids which are chemically similar.

1. Introduction

The amino acid sequences of typical proteins have changed slowly during evolution, as can be observed from the analysis of sequence data. Based on the amino acid replacement frequency in families of modern proteins, Argyle (1980) has constructed an amino acid transition probability matrix. The amino acid arrangement of this matrix which contains the replacement frequencies as matrix elements, was deduced by rearranging these elements in such a way as to provide the minimum variance about the main diagonal. In this amino acid arrangement the last amino acid is next to the first, generating the so-called amino acid ring. In this ring, amino acids which are chemically similar appeared grouped together. The amino acid sequence proposed by Argyle is Asp-Asn-Gly-Ser-Ala-Thr-Pro-Cys-Val-Ile-Leu-Met-Phen-Trp-Tyr-Arg-His-Gln-Lys-Glu, where Glu, is next to Asp generating the amino acid ring. As pointed out by Argyle, nearly 56 % of all replacements for an amino acid involved its nearest or second nearest neighbor in the ring.

Recently, it has been evaluated the amino acid chemical similarities for that ring in terms of some physico-chemical properties (Soto and Tohá, 1983). In this amino acid arrangement it was found that the standard deviation, σ_0, for the refractivity, bulkiness and hydrophobicity agree with the σ_i value of the corresponding ideal distributions in a 93 %, 92 % and 83 % respectively, $\sigma_i/\sigma_0 \times 100$. However, polarity and optical rotation appeared less conserved, as the σ_i/σ_0 values were found to be 77 % and 59 % respectively.

** This work was partially supported by OEA and Departamento de Desarrollo de la Investigación.

In this work, a new amino acid ring is proposed, for which a higher amino acid chemical similarity can be deduced.

Furthermore, the amino acid replacement frequency matrix is analyzed in terms of the corresponding codon replacement matrix. This would allow the assessment of the possible interrelationship between the observed most frequent amino acid replacements and the expected most probable base codon exchanges that can be infered from base transition transversion probability considerations. The expected most probable codon exchanges where considered as those which involve the fewest number of base exchanges and the fewest number of base transversions. The similarity of the standard deviation value of this codon matrix with respect to that of the amino acid replacement frequency matrix should give a measure of the amino acid codon dependent exchanges.

Finally, the most frequent amino acid codon dependent replacements are analyzed in terms of the restrictions imposed by the genetic code structure and the larger physicochemical discrepancy that exist between some codons corresponding to amino acids which are chemically similar, i.e. codons related by one transversion or more than one base change, rather than only one base transition.

2. Methods

A. In all cases, except that of Table I, the variance of the matrices were minimized about the main diagonal and was calculated according to the described equation (Argyle, 1980)

$$\sigma^2 = \frac{\sum_{k=-9}^{10} k^2 D_k}{\sum_{k=-9}^{10} D_k} \quad \text{for} \quad k \neq 0, \tag{1}$$

where

$\sigma^2 =$ variance and $\sigma =$ standard deviation.

$k =$ the distance of the lesser diagonal from the main diagonal (main diagonal $k = 0$).

$D_k =$ the sum of the matrix elements F_{ij} in the k-th diagonal.

$$D_k = \sum_{j=1}^{20} F_n + 1, j$$

where

$$n = (j + k - 1) \bmod 20.$$

B. In the case of σ values for matrices of Table I (where the matrix elements correspond to differences in chemical indexes), the variance appeared minimized about $k = 10$ diagonal and not about the main diagonal along which the smallest values are grouped.

In this case:

$$\sigma^2 = \frac{(10)^2 D_{10} + \sum_{k=-9}^{9} (10 - |k|)^2 D_k}{\sum_{k=-9}^{10} D_k} \quad \text{for} \quad k \neq 0. \tag{2}$$

In Table I each matrix element corresponds to the difference between corresponding chemical index values (refractivity, polarity, hydrophobicity and bulkiness) of two specific amino acids. The refractivity, bulkiness, hydrophobicity and polarity indexes were taken from Prabhakaran and Ponnuswamy (1979). The optical rotation indexes from Fasman (1976).

C. The respective ideal matrices were calculated by assuming the element of each column to be shifted vertically so as to minimized the variance about the diagonal element in that column and independently of all other columns. The σ_i-value of these matrices are calculated according to the above Equations (1) and (2), respectively.

3. Results

Figure 1 shows the replacement probability matrix with amino acids reordered as to provide the minimum variance about the main diagonal. The standard deviation, σ_0, of this matrix is 3.97, being 2.74 its respective ideal σ_i-value. It is interesting to point out that, although this new amino acid array resembles the one proposed previously (Argyle, 1980) with $\sigma_0 = 3.98$, the differences are important, as the new arrangement gives a conservation of over 90 % for all the most general physico-chemical properties, as shown Table I for the respective σ ratios. Moreover, in this table it can be observed that, from all σ values comparisons, the σ value for the refractivity distribution, is almost equal to that of an ideal distribution, (σ ratio = 1.0). This may suggest that the amino acid ring from Figure 1 is the best or at least extremely close to the best similarity ring attainable from amino acid refractivity similarities. On the other hand it appears quite improbable to find an amino acid arrangement for which all σ ratios of all physico-chemical properties could have been equally well improved.

These results are in agreement with the view that the amino acid substitutions during evolution were produced through a conservation of the amino acid chemical similarities, favouring in this way the maintenance of the structural and functional properties of the proteins:

If the most frequent amino acid replacements among proteins involve their more chemically alike amino acids, why is the σ value of the amino acid substitution probability matrix with respect to its ideal value equal to 0.69 (see Figure 1) and not above 0.9 as expected from the amino acid chemical properties similarities (see Table I)? As we shall see below this can be understood in terms of the codon base changes that might have been most probably involved during amino acid replacements. Figure 2 shows a codon substitution probability matrix with a codon array corresponding to the

	Glu	Asn	Gly	Ser	Ala	Thr	Pro	Val	Ile	Leu	Met	Phen	Tyr	Trp	Cys	Arg	His	Lys	Gln	Asp
Glu	X	32	32	26	53	0	30	21	0	0	0	0	0	0	0	12	0	47	89	206
Asn	8	X	23	84	21	37	10	0	0	0	0	10	0	0	0	0	79	41	0	95
Gly	50	64	X	58	100	22	20	27	0	10	0	0	25	0	0	24	0	24	64	32
Ser	0	170	73	X	211	186	69	27	0	13	49	21	25	0	16	24	16	36	25	32
Ala	59	64	92	110	X	89	89	82	27	10	0	11	25	0	0	12	0	24	13	55
Thr	42	32	23	200	137	X	30	34	80	5	97	21	0	0	0	0	48	41	38	16
Pro	0	11	0	39	58	6	X	0	27	10	0	0	0	0	0	24	16	6	13	0
Val	50	32	13	19	53	52	30	X	212	91	49	31	12	33	0	0	0	24	13	16
Ile	8	21	5	26	5	37	20	178	X	87	0	51	0	33	0	36	0	6	0	8
Leu	25	11	5	6	21	6	69	69	186	X	146	62	37	0	0	12	0	6	25	0
Met	0	0	0	0	0	6	0	55	40	63	X	10	0	33	0	12	0	18	25	0
Phen	0	0	0	19	11	6	10	13	13	58	49	X	258	100	0	0	0	6	0	0
Tyr	8	21	0	19	11	6	0	13	13	5	0	154	X	133	0	12	16	6	0	8
Trp	0	0	0	0	0	0	0	0	0	0	0	0	12	X	0	0	0	0	0	0
Cys	0	0	0	26	0	15	0	13	27	0	0	0	0	0	X	0	0	0	0	0
Arg	8	21	23	19	5	15	10	0	13	13	49	0	0	33	0	X	48	59	51	0
His	0	42	5	26	5	0	10	0	0	10	0	72	37	33	0	24	X	18	64	24
Lys	17	95	18	0	16	22	10	6	40	5	0	0	0	0	0	108	16	X	25	24
Gln	118	11	9	13	21	6	20	21	27	5	0	0	0	0	0	48	79	47	X	8
Asp	269	106	41	44	16	6	30	0	0	0	0	0	25	0	0	0	16	12	0	X

The σ value of this matrix, $\sigma_0 = 3.97$
The σ value for its respective ideal matrix, $\sigma_i = 2.74$
The σ ratio (σ_i/σ_0), $\sigma_r = 0.69$

Fig. 1. Amino acid frequency replacement matrix.

TABLE I

Matrices σ values for the proposed amino acid ring, evaluated in terms of different physico-chemical indexes

	σ observed σ_0	σ ideal σ_i	σ ratio $\sigma_r = \sigma_i/\sigma_0$
Refractivity	5.39	5.40	$\approx$ 1.00
Bulkiness	5.52	5.35	0.97
Optical rotation	5.86	5.63	0.96
Hydrophobicity	5.51	5.06	0,92
Polarity	5.39	4.90	0.91

The amino acid ring in all these matrices correspond to that given in Figure 1.

amino acid ring of Figure 1. The base codon changes within the matrix were evaluated as a substitution probability value, taking a base substitution probability value of 10/11 and 1/11 for one base transition and one base transversion, respectively. These probability values have been obtained from theoretical estimations based on isomeric keto-enol equilibria of nucleotides and from experimental evidences of base substitution reversion in the tryptophan synthetase A gene of E. coli (Topal and Fresco, 1976a, b), showing that base transitions occur ten times more frequently than base transversions. Consequently, the total substitution probability value of each element of the matrix will assume one of the following specific values, (in decreasing order): 10/11 for one transition (T), 100/121 for two transitions (T × T); 1/11 for one transversion

(V), 10/121 for one transition and one transversion (T × V), 100/1331 for two transitions and one transversion (T × T × V), 1/121 for two transversions (V × V), 10/1331 for two transversions and one transition (V × V × T) and 1/1331 for three transversions (V × V × V). Due to the degeneracy of the genetic code, a very large number of codon rings for the described amino acid ring could be deduced. However, in Figure 2, we have constructed the codon substitution probability matrix so that every particular codon in the codon arrangement may imply any of its degenerate codons that gives the highest probability exchange value with any other particular codon. For instance, alanine and glutamic codons are related by one transversion as the highest possible substitution probability value, when codons GCA and GCG of alanine are related to GAA and GAG of glumatic acid, respectively. There is no other comparison between these amino acid codons that gives a higher probability exchange value. However, codons GCC and GCU of alanine are the codons that give the highest possible substitution probability value when exchanging with AAU and AAC of asparagine, respectively. Therefore, the alanine codons to be used in the matrix in Figure 2 have to be read according to the most favorable exchange. This criterion has been used for evaluating all matrix elements in Figure 2 considering that all amino acid codons are represented in the proteins. As can be seen, the σ value of this matrix is 4.2 and its corresponding ideal σ value of 2.9. The ratio between these two values $2.9/4.2 = 0.68$ is very close to 0.69, the ratio displayed by the respective amino acid replacement probability matrices in Figure 1. Therefore, the deviation from ideality of the σ value of the amino acid substitution probability matrix can be interpreted as a consequence of the deviation from ideality of the respective codon substitution probability matrix. From Figure 2 it can be also infered that the degenerate codons of each particular amino acid are not equally represented. For instance, in the case of the two types of degenerate codons of leucine, serine, and arginine, the four fold degenerate type is more often represented than the two fold degenerate one. For leucine, the CUX codons have been used in nine exchanges, whereas the UUA or UUG codons were used only twice; (when exchanging with tryptophan and serine codons). For the remaining eight leucine codon exchanges, either codons CUX or UUA(G) are equally suited for obtaining the highest probability exchange value. In summary, for leucine, arginine and serine highest probability codon exchanges the four fold type is used 4.5, 1.13, and 1.14 times respectively, more often than the two fold type. From these last results one should then expect to find at the genome level of proteins, the codons of the four fold type more frequently represented; by analyzing the available nucleotide sequence data (Grantham, 1978; Grantham *et al.*, 1980; and Grantham, *et al.*, 1981) of the coding region of 23 different proteins (corresponding to 12 different families) it was found that, on a total average, the four fold type codons of leucine, arginine and serine are used 5.7, 1.1, and 1.5 times respectively, more often than the two fold type. On the other hand, we have constructed several other codon substitution probability matrices similar to that showed in Figure 2, maintaining the hereby described codon array, but using only one of the two types of degenerate codons of leucine, arginine and serine. Since each one of these amino acids has two types of codons, there will be $2^3 = 8$ different codon type combinations within the matrix, Table II shows the σ values for these eight different matrices, including their respectives ideal σ values and σ rations. As can be seen, all these σ ratios (except for the II-Serine IV-Leucine IV-Arginine, matrices) are below the σ ratio obtained from Figure 2, (σ ratio = 0.683). As the latter is more similar to the 0.69 ratio displayed by the amino acid replacement frequency matrices of Figure 1, it can be concluded that within the statistical dispersion, the most probable codon substitution matrix corresponding to the most frequent amino acid substitution ring, is the one shown in Figure 2.

TABLE II

σ values for codon substitution probability matrices, in various type of Leucine, Arginine and Serine codon combinations

Codon Type combination in the matrix	σ observed σ_0	σ ideal σ_i	σ ratio (σ_i/σ_0) $\sigma_r = \sigma_i/\sigma_0$
IV Ser II Leu II Arg	4.53	2.68	0.59
IV Ser II Leu IV Arg	4.24	2.67	0.63
IV Ser IV Leu II Arg	4.27	2.72	0.64
IV Ser IV Leu IV Arg	4.21	2.71	0.64
II Ser II Leu II Arg	4.25	2.73	0.64
II Ser IV Leu II Arg	4.22	2.74	0.65
II Ser II Leu IV Arg	3.93	2.68	0.68
II Ser IV Leu IV Arg	3.92	2.71	0.69

All these matrices correspond to the same amino acid ring proposed in Figure 1, II and IV means the two of four fold degenerate codon type of Serine, Leucine and Arginine.

At this stage it may be worthwhile to mention, that the high similarity of the σ ratio of the most probable codon and amino acid substitution matrices, cannot be attained when this comparison is made using Argyle's (1980) amino acid ring (Table III). It can be also seen that wereas the σ ratios for two slightly different amino acid replacements frequency rings are very similar, (A and A′), the σ ratios for the respectives codon substitution probability rings are more dissimilar (B and B′).

From all these results we can further suggest that the discrepancy between the σ value of the codon substitution probability matrix and its ideal value (Figure 2) can be explained mostly by:

(a)The larger structural dissimilarity between some codons which codify chemically alike amino acids.

(b)The inherent structure of the genetic code, its non-uniform degeneracy and the two types of nitrogen bases conforming the codons.

In an attempt to evaluate these factors, we have built up a codon substitution probability matrix where codons were ordered along each other by only one base exchange and regardless of the most frequent amino acid replacements (Table IV). Within this codon ring the minimum possible number of base exchanges and the

	Glu	Asn	Gly	Ser	Ala	Thr	Pro	Val	Ile	Leu	Met	Phen	Tyr	Trp	Cys	Arg	His	Lys	Gln	Asp
Glu	X	TV	T	TTV	V	TV	VV	V	TV	VV	TV	VVV	VV	TV	VVT	TT	VV	T	V	V
Asn	TV	X	TT	T	TV	V	VV	TV	V	VV	VV	VV	V	VVT	TV	TV	V	V	VVV	T
Gly	T	TT	X	T	V	TV	VV	V	TV	VV	TV	VV	TV	V	V	T	TV	TT	TV	T
Ser	TTV	T	T	X	V	V	T	TV	V	T	TV	T	V	V	V	V	TV	TV	TV	TT
Ala	V	TV	V	V	X	T	V	T	TT	TV	TT	TV	VV	VV	VV	TV	VV	TV	VV	V
Thr	TV	V	TV	V	T	X	V	TT	T	TV	T	TV	VV	VV	VV	V	VV	V	VV	VT
Pro	VV	VV	VV	T	V	V	X	TV	TV	T	TV	TT	TV	TV	TV	V	V	VV	V	VV
Val	V	TV	V	TV	T	TT	TV	X	T	V	T	V	VV	VV	VV	TV	VV	TV	VV	V
Ile	TV	V	TV	V	TT	T	TV	T	X	V	T	V	VV	VVT	VV	V	VV	V	VV	TV
Leu	VV	VV	VV	T	TV	TV	T	V	V	X	V	T	TV	V	TV	V	V	VV	V	VV
Met	TV	VV	TV	TV	TT	T	TV	T	T	V	X	VV	VVV	VV	VVV	V	VVV	V	VV	TVV
Phen	VVV	VV	VV	T	TV	TV	TT	V	V	T	VV	X	V	VV	V	TV	TV	VVV	VVT	VV
Tyr	VV	V	TV	V	VV	VV	TV	VV	VV	TV	VVV	V	X	TV	T	TT	T	VV	TV	V
Trp	TV	VVT	V	V	VV	VV	TV	VV	VVT	V	VV	VV	TV	X	V	T	TTV	TV	TT	VVT
Cys	VVT	TV	V	V	VV	VV	TV	VV	VV	TV	VVV	V	T	V	X	T	TT	VVT	TTV	TV
Arg	TT	TV	T	V	TV	V	V	TV	V	V	V	TV	TT	T	T	X	T	T	T	TV
His	VV	V	TV	TV	VV	VV	V	VV	VV	V	VVV	TV	T	TTV	TT	T	X	VV	V	V
Lys	T	V	TT	TV	TV	V	VV	TV	V	VV	V	VVV	VV	TV	VVT	T	VV	X	V	TV
Gln	V	VV	TV	TV	VV	VV	V	VV	VV	V	VV	VVT	TV	TT	TTV	T	V	V	X	VV
Asp	V	T	T	TT	V	TV	VV	V	TV	VV	VVT	VV	V	VVT	TV	TV	V	TV	VV	X

The σ value of this matrix, $\sigma_0 = 4.24$
The σ value for its ideal matrix, $\sigma_i = 2.90$
The σ ratio (σ_i/σ_0), $\sigma_r = 0.68$
Where T and V are the transition and transversion probabilities of base changes repectively.

T = 10/11 and V = 1/11

Fig. 2. Amino acid codon replacement probability matrix.

TABLE III

Comparison of the σ values of the amino acid and codon substitution probability matrices for two slightly different amino acid rings

		σ observed	σ ideal	σ ratio
σ_0	σ_i	$\sigma_r = \sigma_i/\sigma_0$		
Earlier ring (Argyle, 1980)	(A) Amino acid substitution frequency matrix	3.98	2.74	0.688
	(B) Codon substitution probability matrix	4.75	2.90	0.610
This ring	(A′) Amino acid substitution frequency matrix	3.97	2.74	0.690
	(B′) Codon substitution probability matrix	4.24	2.90	0.684

minimum of base transversions has been used. It is important to point out, that there will be many similar codon rings satisfying the above restrictions. However, for the purpose of this discussion, any of these codon substitution probability matrices will be suitable, as all of them should give relative similar σ values. Moreover to build up this matrix, the two types of degenerate codons of leucine, arginine and serine were considered. The codon ring and the σ value of this matrix including its ideal σ_i-value is shown in Table IV. As can be seen, the σ_0-value of this matrix with respect to its ideal one is 0.86. Therefore, for the expected most probable codon substitution matrix regardless of the amino acid substitutions the deviation of the σ value from ideality reflects the inherent restrictions of the genetic code itself. Hence, the discrepancy between this 0.86 ratio and 0.68 of Figure 2 can be explained almost by the fact that in some cases the codons of the more chemically alike amino acids are not related by the expected most probable codon exchanges; thus, this may suggest an independent generation of some primordial codons and amino acids.

TABLE IV

Codon replacement probability matrix where codons in the ring are related to each other by only one base exchange

σ_0	σ_i	$\sigma_r(\sigma_i/\sigma_0)$
3.77	3.23	0.86

The codon ring for this matrix written as amino acid sequence is:

```
Asp-Glu-Ala-Thr-Ile-Met-Val-Leu II-Phen-Leu IV-Pro-Ser IV
|                                                     |
Gly- Ser II- Asn-Lys-Arg II-Trp-Arg IV-Gln-His-Tyr-Cys.
```

II and IV means the two or four fold degenerate codon type, of leucine, arginine and serine codons.

In conclusion, there is a large conservation of the amino acid physico-chemical properties for the hereby described amino acid ring derived from the amino acid frequency replacement matrix. The relative irregular distribution of the amino acid frequency replacement values in the matrix is similar to that obtained from a corresponding codon probability replacement matrix. These distribution irregularities can be explained to a large extent by the restrictions imposed by the genetic code structure and the physico-chemical disimilarities of some codons that codify chemically alike amino acids. This should not mean that other restrictions as those involved in the transcription and translation process may also play some role in the amino acid codon replacements (Grosjean and Fiers, 1982).

References

Argyle, E.: 1980, *Origins of Life* **10**, 357.

Fasman, G. (ed.): 1976, *Handbook of Biochemistry and Molecular Biology*. 3rd Edition, Proteins Vol. I, CRC Press.

Grantham, R.: 1978 *FEBS Letters* **95**, 1.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R.: 1981, *Nucleic Acid Research* **9**, 243.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavé, A.: 1980, *Nucleic Acid Research* **8**, 249.

Grosjean, H., and Fiers, W.: 1982, *Gene* **18**, 199.

Prabhakaran, M., and Ponnuswamy, P. K.: 1979, *J. Theor. Biol.* **80**, 485.

Soto, M. A. and Tohá, J.: 1983, *Origins of Life*, this issue, 147.

Topal, D. M. and Fresco, R. J.: 1976a, *Nature* **263**, 285.

Topal, D. M. and Fresco, R. J.: 1976b, *Nature* **263**, 289.