ELSEVIER

Gene 261 (2000) 63–69

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code

Maria Luisa Chiusano[a,b], Fernando Alvarez-Valin[c], Massimo Di Giulio[d], Giuseppe D'Onofrio[a], Gaetano Ammirato[b], Giovanni Colonna[b], Giorgio Bernardi[a,*]

[a]*Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples, Italy*
[b]*Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Universitá, via Costantinopoli 16, 80138 Naples, Italy*
[c]*Sección Biomatemática, Facultad de Ciencias, Iguá 4225, Montevideo 11400, Uruguay*
[d]*International Institute of Genetics and Biophysics, CNR, via G. Marconi 10, 80125 Naples, Italy*

Accepted 30 October 2000
Received by T. Gojobori

Abstract

The nucleotide frequencies in the second codon positions of genes are remarkably different for the coding regions that correspond to different secondary structures in the encoded proteins, namely, helix, β-strand and aperiodic structures. Indeed, hydrophobic and hydrophilic amino acids are encoded by codons having U or A, respectively, in their second position. Moreover, the β-strand structure is strongly hydrophobic, while aperiodic structures contain more hydrophilic amino acids. The relationship between nucleotide frequencies and protein secondary structures is associated not only with the physico-chemical properties of these structures but also with the organisation of the genetic code. In fact, this organisation seems to have evolved so as to preserve the secondary structures of proteins by preventing deleterious amino acid substitutions that could modify the physico-chemical properties required for an optimal structure.

Keywords: Genetic code; Nucleotide frequencies; Protein secondary structures; Second codon positions

1. Introduction

Different secondary structures of proteins exhibit remarkable differences in amino acid frequencies (Szent-Gyorgyi and Cohen, 1957; Guzzo, 1965; Havsteen, 1966; Prothero, 1966; Cook, 1967; Goldsack, 1969; Chou and Fasman, 1974; Levitt, 1978). Some amino acids are, indeed, more prone to be present in specific secondary structures while others tend to disrupt them (Bahar et al., 1997). Propensities of amino acids for a secondary structure correlate with their physico-chemical properties. These properties have provided the basic information used in prediction methods. Protein secondary structures reflect the physico-chemical properties of the most frequent amino acids in those structures. For example, the β-strand structure is strongly hydrophobic, while aperiodic structures contain more hydrophilic amino acids. Therefore, constraints on the secondary and tertiary structures tend to limit accepted mutations to those in which an amino acid is replaced by another amino acid with similar properties (Epstein, 1967; Grantham, 1974; Chirpich, 1975; Zhang, 2000).

Several investigations have addressed the possible correlation between the nucleotides at each codon position and the properties of amino acids (Goodman and Moore, 1977; Wolfenden et al., 1979; Sjostrom and Wold, 1985; Bernardi and Bernardi, 1986; Taylor and Coates, 1989; Tolstrup et al., 1994). In particular, hydrophobic amino acids are encoded by codons having U in the second position, while hydrophilic amino acids are encoded by triplets with A in the second position. However, previous attempts to link protein secondary structures to the organisation of the genetic code (see Di Giulio, 1996) have been unsuccessful (Salemme et al., 1977; Goodman and Moore, 1977), except in the case of β-turns (Jurka and Smith, 1987) and β-strands (Di Giulio, 1996). Recently Gupta et al. (2000) have reported that the average frequencies of U and A at the second codon position are markedly different between α-helix and β-strand, but these authors did not perform any more detailed analysis.

The present work shows that the nucleotide distributions

* Corresponding author. Tel.: +33-081-5833215; fax: +39-081-5833402.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

0378-1119/00/$ - see front matter
PII: S0378-1119(00)00521-7

Table 1
Mean, minimum and maximum values of the nucleotide frequencies at second codon positions of coding regions corresponding to the three secondary structures

		β-strand				Helix				Aperiodic			
		U2	A2	C2	G2	U2	A2	C2	G2	U2	A2	C2	G2
Human	Mean	0.41	0.25	0.18	0.15	0.26	0.38	0.19	0.17	0.17	0.36	0.23	0.24
	Min	0.19	0.10	0.00	0.00	0.00	0.22	0.00	0.00	0.04	0.15	0.05	0.10
	Max	0.66	0.53	0.30	0.30	0.50	0.67	0.40	0.43	0.28	0.53	0.38	0.50
Prokaryotic	Mean	0.44	0.27	0.16	0.13	0.30	0.39	0.19	0.12	0.20	0.38	0.22	0.20
	Min	0.20	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.05	0.19	0.05	0.00
	Max	0.80	0.61	0.36	0.32	0.50	0.78	0.67	0.28	0.32	0.63	0.46	0.47

in second codon positions are strongly related to the average physico-chemical properties of protein secondary structure, and this relationship sheds light on the origin of the genetic code.

2. Materials and methods

Two data sets (available upon request) of experimentally determined structures comprising 77 human and 232 prokaryotic proteins were used in order to investigate the relationships between base composition of coding sequences and the secondary structures of the encoded proteins. Complete coding region information was available only for the human data set (Adzhubei et al., 1998). Thus, serine was excluded from the analysis in the case of prokaryotic proteins, because of its ambiguity in the second codon position (UCN or AGY). The secondary structures, assigned by the DSSP program (Kabsch and Sander, 1983), were described in terms of β-strand, helix (including 3_{10} helices and α-helices), and aperiodic structure (including the turn structure and the protein segments that are not defined and/or lack periodicity). The average hydrophobicity levels, calculated by the Gravy scale (Kyte and Doolittle, 1982), and molecular weights of the amino acids in each of the three structures were also calculated.

3. Results

In Table 1, we report the mean values of the nucleotide frequencies at second codon positions in human and prokaryotic proteins in coding regions corresponding to different secondary structures.

The three structures show marked differences in the frequency of U in second codon position (U_2), the aperiodic structure showing the lowest values, the β-strand structure the highest ones in both groups of organisms. A_2 is also different among the three structures, with higher values in helix and aperiodic structures. G_2 and C_2 have consistently lower values in all the three structures compared to A_2 and U_2, with higher figures in aperiodic structure in comparison to both helix and β-strand structures.

The differences in nucleotide frequency in the second codon positions can be accounted for by the different amino acid composition in the three structures (see Table 2). As expected, the amino acids have different propensities for each structure. Interestingly, all amino acids having U in the second positions exhibited a clear hierarchy, the highest frequencies being observed in the β-strand structure. In turn, helix exhibited higher frequencies for these amino acids compared to the aperiodic structure. The amino acids that contribute more to these differences are valine, isoleucine and phenylalanine, while leucine is strongly differentiated only between the aperiodic structure and the other two structures. With the only exception of tyrosine, amino acids that have A in the second position are more frequent in aperiodic

Table 2
Amino acid content in the secondary structures of human proteins[a]

	nt in 2nd position	Total	β-strand	Helix	Aperiodic
Leu	U	0.092	0.107	0.105	0.063
Val	U	0.077	0.122	0.062	0.047
Ile	U	0.054	0.084	0.052	0.027
Phe	U	0.047	0.067	0.044	0.031
Met	U	0.021	0.027	0.027	0.011
Glu	A	0.068	0.056	0.089	0.059
Lys	A	0.067	0.055	0.076	0.071
Asp	A	0.049	0.025	0.050	0.074
Asn	A	0.041	0.022	0.037	0.064
Gln	A	0.039	0.032	0.046	0.039
Tyr	A	0.034	0.044	0.032	0.027
His	A	0.023	0.022	0.024	0.023
Ala	C	0.068	0.054	0.098	0.052
Thr	C	0.06	0.07	0.047	0.062
Pro	C	0.041	0.023	0.026	0.073
Gly	G	0.066	0.047	0.040	0.112
Arg	G	0.046	0.037	0.054	0.047
Cys	G	0.027	0.032	0.023	0.026
Trp	G	0.015	0.019	0.016	0.011
Ser	G/C	0.064	0.056	0.055	0.080

[a] The amino acids are grouped according to the nucleotide present in the second codon position of their codons and then ordered by their average frequency in the proteins (Total).

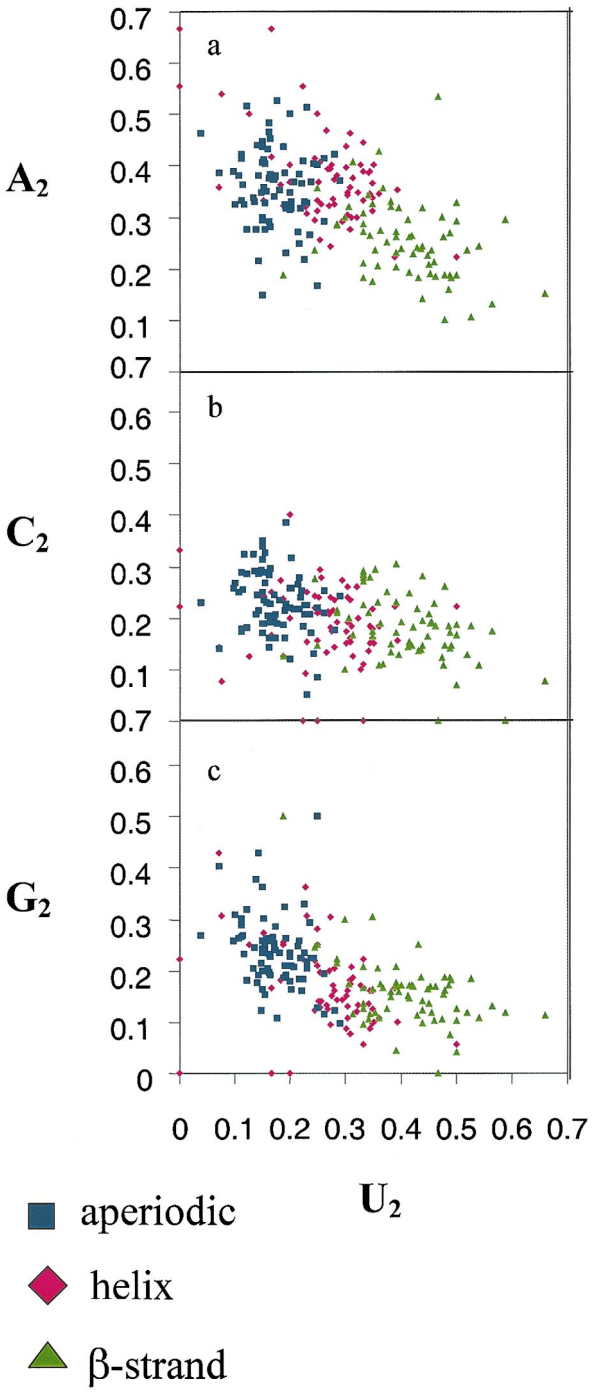

Fig. 1. Scatterplots of A_2, C_2 and G_2 versus U_2 in human data.

and helix structures. These results indicate that the differences among structures for U_2 and A_2, are not due to any particular amino acid that is very frequent in one structure and seldom present in the others. Instead, they demonstrate that there is a coherent behaviour. Amino acids having G_2 in their codons do not exhibit preferences for any particular structure, with the only exception of glycine, which is very frequent in the aperiodic structure. Amino acids with C_2 also exhibit no common preference, although alanine is most frequent in the α-helix and proline is very frequent in the aperiodic structure.

A further step in the analysis of nucleotide preferences in the secondary structures of proteins was to analyse their distributions by considering the intergenic variability. For this purpose all amino acids belonging to the same type of secondary structure were pooled together for each gene. The different distributions of the base frequencies in the three structures is clearly visible in both human (Fig. 1) and prokaryotic (Fig. 2) data sets: three separated 'clouds' can be distinguished, each one corresponding to a given structure. The frequency that best separates the 'clouds' is U_2.

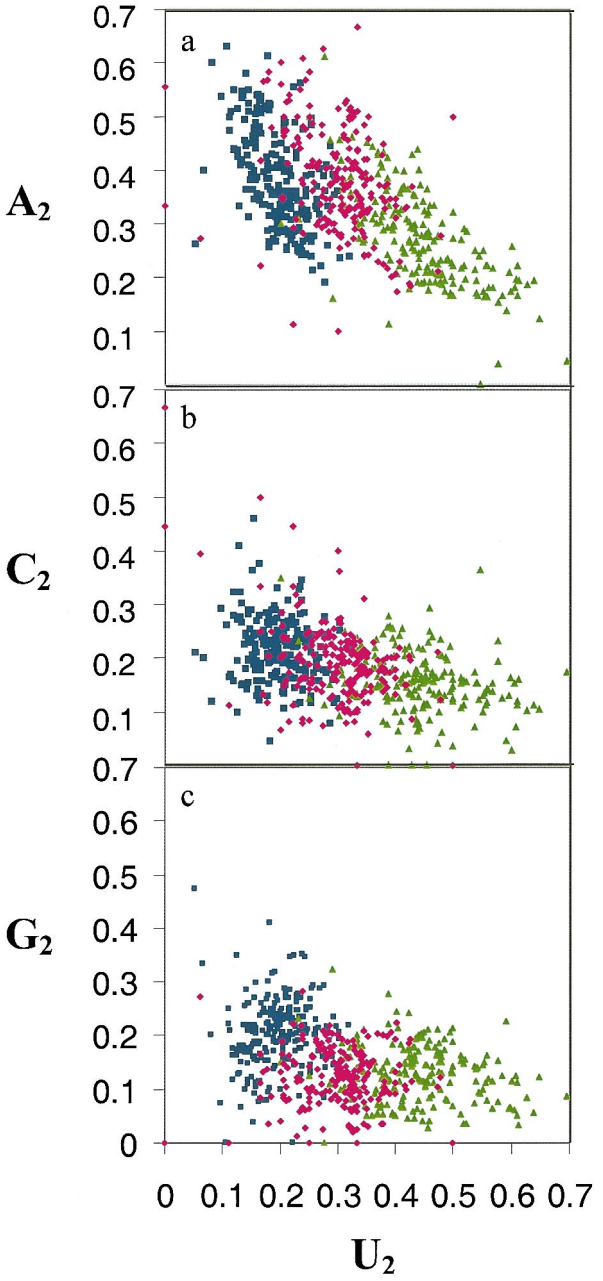

Fig. 2. Scatterplots of A_2, C_2 and G_2 versus U_2 in prokaryotic data. Symbols are as in Fig. 1.

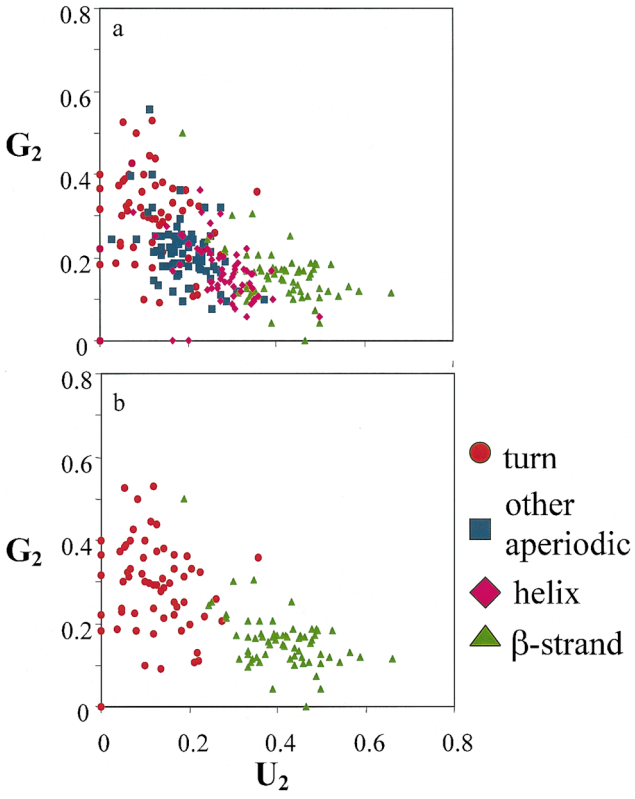

Fig. 3. Scatterplots of G_2 versus U_2 in the secondary structures (a) and in turn and β-strand structures (b). The contribution of the turn structure is separated from the remaining aperiodic structure.

The β-strand structure shows a U_2 distribution that is never lower than 0.19 (human data) or 0.20 (prokaryotic data), and that overlaps only marginally with the aperiodic structure distribution. Although the distribution of U_2 frequencies in the helix structure shows some overlap with both aperiodic and β-strand structures, it occupies a clearly differentiated, intermediate, position.

Figs. 1a and 2a both show a separation of the A_2 distributions between aperiodic and helix structures on one hand, with higher A_2 values, and β-strand on the other, with lower A_2 values.

Although G_2 and C_2 distributions do not distinguish the three structures as clearly as A_2 and U_2, their values tend to be higher for the aperiodic structure compared to helix and β-strand structures, especially in the case of G_2. In this latter case, the higher values of G_2 in the aperiodic structure are due to the contribution of turns. To show this, we plotted G_2 versus U_2 in human proteins where the turn structure is separated from the remaining types of aperiodic structure (Fig. 3). Fig. 3 shows that G_2 frequency is higher in the turn structure than in the other types of aperiodic structure (the mean difference of the two distributions being 0.083, and the two distributions being statistically different with a *P*-value which is less than 0.0001 on the basis of a paired comparison test), while U_2 is lower (the mean difference is −0.077 and the *P*-value is less than 0.0001), making the separation between turn and β-strands even stronger (Fig. 3b) than the one between aperiodic and β-strand struc-

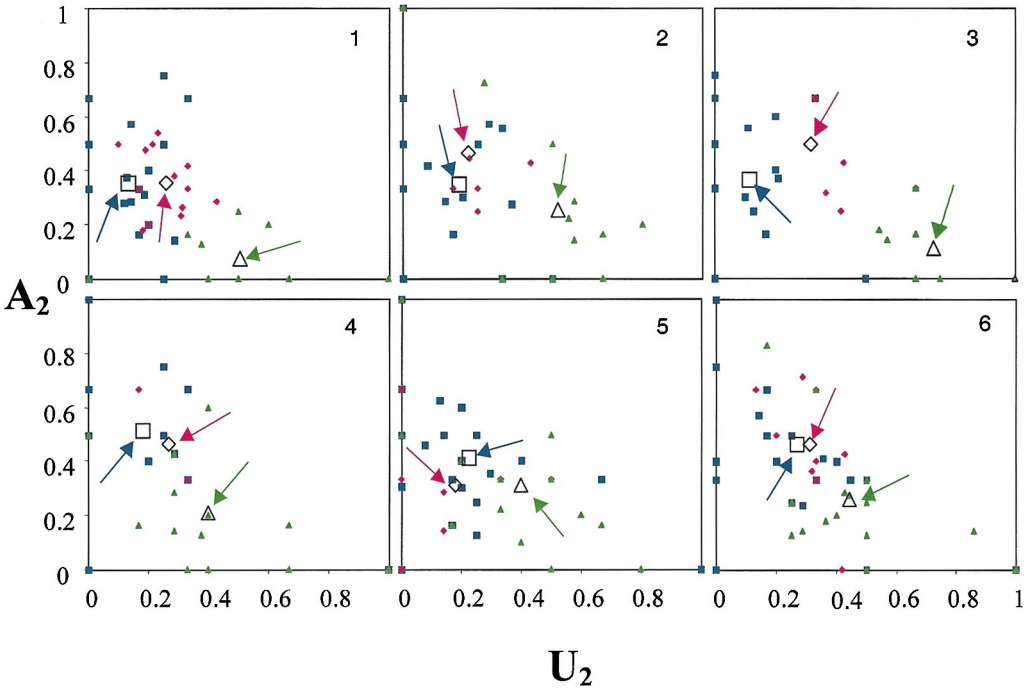

Fig. 4. A_2 versus U_2 for example proteins (PDB/GenBank accession numbers and names are listed below) included in the intragenic analysis. Coloured symbols represent average nucleotide frequencies in individual element of structure. Open symbols represent average nucleotide frequencies of all the elements of structure of the same type in the protein (see Fig. 1). Symbols are as in Fig. 1. (1) 1ALD/178350, Aldolase A; (2) 1DRF/30774, Dihydrofolate reductase; (3) 1HMP_A /32449, Hypoxanthine guanine phosphoribosyltransferase; (4) 1ILR_1/186291 Interleukin 1 receptor antagonist protein; (5) 1PPB_H/37128, Alpha-Thrombin; (6) 7API_A/28965, Alpha-1-Antitrypsin.

ture (Fig. 1c). The same comparison for prokaryotic proteins yields similar results (not shown).

An analysis at the 'intragenic level', made in order to assess the contribution of each single structure element to the average nucleotide frequencies for that type of structure (in a given gene), suggests that each individual element follows the general behaviour described for the intergenic analysis. However, the intragenic analysis shows a larger variability that is probably due to the smaller size of the sampling points. Fig. 4 shows some representative genes.

4. Physico-chemical properties of structures

To understand better the reasons for the nucleotide preferences in the three structures in the context of the relationship between the physico-chemical properties of amino acids and the genetic code, we investigated the average hydrophobicity values and molecular weights of the amino acids in the secondary structures under analysis.

By plotting the molecular weights versus hydrophobicity, in both human and prokaryotic proteins (Fig. 5, panels a and b, respectively), we observed that the distributions of the values is different for the three secondary structures. The β-strand structure has higher hydrophobicity values when compared to aperiodic and helix structures. In particular, the intervals of hydrophobicity variation for aperiodic and β-strand structures show only a small overlap. Molecular weights show different, yet widely overlapping distributions. β-strands and helix structures exhibit molecular weights that are higher compared to those of aperiodic structure.

From this analysis it is possible to derive a general information on the rules concerning physico-chemical properties of the structures. The β-strand and aperiodic structures have diametrically opposite trends. The β-strand has higher hydrophobicity values and, on the average, amino acids with higher molecular weights, while the aperiodic structure is less hydrophobic and is composed of amino acids having lower molecular weights. Moreover, the helix structure is intermediate, sharing a similar distribution with the β-strand structure in the case of its molecular weight, while it follows the behavior of the aperiodic structure in its hydrophobicity patterns.

More important, though, is the fact that molecular weight and hydrophobicity show strong negative correlations when each kind of secondary structure is considered separately. This negative correlation completely disappears when all the structures are considered together. Moreover, as shown in the figures, for the same hydrophobicity values, the difference among the structures is given by the molecular weight. Likewise, if the molecular weight is kept constant, the difference among the structures is given by the hydrophobicity values. In other words, if one of these two variables is kept constant, changing the other would necessarily imply a change in the secondary structure.

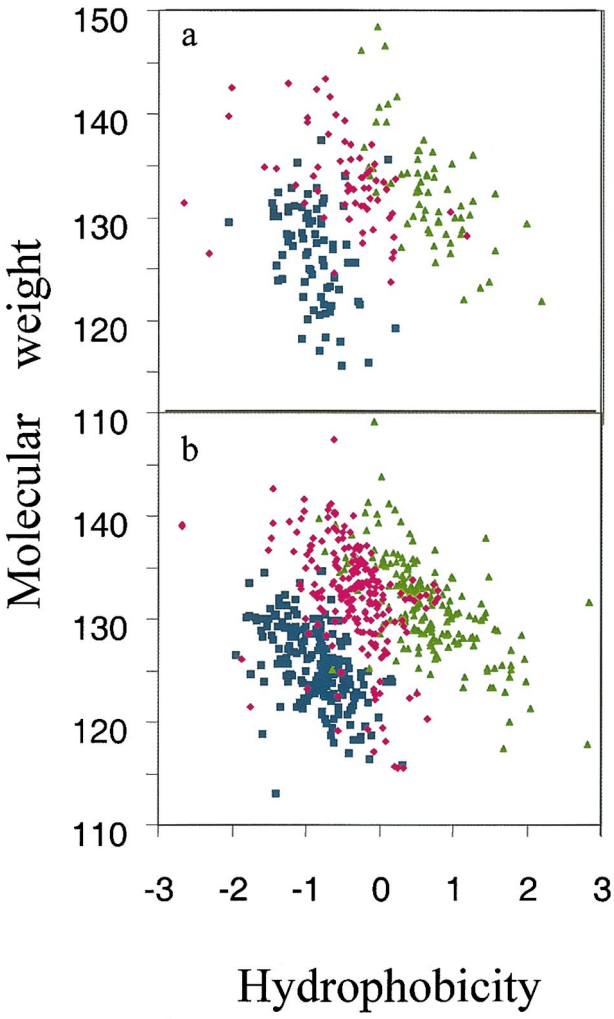

Fig. 5. Scatterplots of molecular weights versus hydrophobicity in human (a) and prokaryotic proteins (b). Symbols are as in Fig. 1.

Finally, we would like to mention that while each of these two physico-chemical variables is not powerful enough for discriminating the secondary structures when considered separately (due to the considerable overlap), the combination of the two is indeed a strong predictor.

5. Discussion

In this paper we have investigated the nucleotide frequencies in coding sequence regions corresponding to different secondary structures of the encoded proteins. Moreover, we have studied the possible implications for new prediction methods, as well as for the origin of the genetic code.

As far as the first point is concerned, we have shown that the nucleotide frequencies in the second codon position are so strikingly different that one could imagine straightforward secondary structure predictions from a nucleic acid sequence data. So far, attempts in this direction have been

only done by Lesnik and Reiss (1996, 1998) who used sequence information to predict putative transmembrane α-helix domains. The advantages of using base frequencies is that this approach would imply estimating four parameters instead of 20 (the amino acids), with the corresponding increase in the reliability of the estimates.

The main contribution of the present analyses is, however, that the physico-chemical properties of protein structures, strongly dependent on amino acid composition, are indeed correlated with well-defined choices in the second codon position of the corresponding coding sequences. This link between protein structure and the second position of the codons suggests that the organisation of the genetic code reflects somehow the secondary structures of proteins, and therefore one of the principal adaptive forces driving the organisation of the code could be protein structure. Even though this could be in part expected, there are still no clear indications that the link between the amino acid properties and the organisation of the code evolved because these properties are structural determinants (as we state here), or are involved in the codon-anticodon interactions that, according to the stereochemical hypothesis (Woese, 1967; Shimizu, 1982; Yarus, 1991), could have promoted the organisation of the genetic code (see Di Giulio, 1997, for a review). Indeed, earlier attempts to link protein secondary structures to the genetic code have been unsuccessful (Salemme et al., 1977; Goodman and Moore, 1977). More recent papers that analyze the relationship between the genetic code and the putative primitive structures have led to the conclusions that β-turn (Jurka and Smith, 1987) and β-strand structures (Di Giulio, 1996) are linked to the structure of the genetic code, suggesting that these structures could have moulded the code. However, none of these previous approaches defined a link between the organisation of the genetic code and protein structure as clearly as it was possible here. The stereochemical hypothesis suggests that the genetic code originated from the interactions between codons or anticodons and amino acids. These interactions depend on the physico-chemical properties of amino acids and anticodons. Remarkably, the same physico-chemical properties of amino acids are also determinant for the correct folding of proteins. In other words, the genetic code would have been organised via forces related to the interactions between amino acids and their codons/anticodons, the same amino acid determinants of this interaction being also linked to properties involved in the organisation of the secondary structures of the proteins.

We propose that protein secondary structures are reflected in the genetic code because secondary structures were the three-dimensional elements of the proteins that had to be preserved at the time the genetic code was organised. Thus, in order to reduce dramatic translation errors while assembling the amino acid sequence required for the determination of a given structure, amino acids with similar physico-chemical properties were grouped by identical second positions in their codons, leading to their present organisation in the columns of the standard representation of the code.

Based on this hypothesis, further details can be discussed, such as the clear separation of aperiodic structure and β-strand structure, shown by the second codon position analysis. The separation suggests that these two structures were fundamental when the genetic code was organised. The intermediary features of the helix structure demonstrates its flexibility in terms of structural requirements, but also stresses that the helix structure is not clearly distinguished by the genetic code. Actually the most frequent amino acids in helix have opposite physico-chemical properties (like leucine and glutamic acid). As a consequence the most preferred amino acids in the helix structure cannot be grouped according to common physico-chemical properties of the lateral chains, and thus they cannot be selected through common choices in the second codon positions. Alternatively, the poorer imprinting of helix in the genetic code, can be interpreted as supporting the idea that this is a late appearing structure, whose emergence could have taken place after the rules of the code were already established. Moreover, the analysis of the preferred second codon position usage in turn structure, reported here, has shown that this structure is better separated from the β-strand structure than the remaining aperiodic structures. This sharp separation could be viewed as supporting the hypothesis that the primitive structures could have been the β-turn (Jurka and Smith, 1987) and/or the β-strand structure (Brack and Orgel, 1975; Di Giulio, 1996).

Acknowledgements

M.L. Chiusano thanks the European Union for a PhD fellowship. The authors thank Dr O. Clay, Dr R. Ragone and Dr A. Facchiano for useful suggestions for improving the paper.

References

Adzhubei, I.A., Adzhubei, A.A., Neidle, S., 1998. An integrated sequence-structure database incorporating matching mRNA sequence, amino acid sequence and protein three-dimensional structure data. Nucleic Acids Res. 26, 327–331.

Bahar, I., Kaplan, M., Jernigan, R.L., 1997. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. Proteins 29, 292–308.

Bernardi, G., Bernardi, G., 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24, 1–11.

Brack, A., Orgel, L.E., 1975. Beta structures of alternating polypeptides and their possible prebiotic significance. Nature 256, 383–387.

Chirpich, T.P., 1975. Rates of protein evolution: a function of amino acid composition. Science 188, 1022–1023.

Chou, P.Y., Fasman, G.D., 1974. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13, 211–222.

Cook, D.A., 1967. The relation between amino acid sequence and protein conformation. J. Mol. Biol. 29, 167–171.

Di Giulio, M., 1996. The beta-sheets of proteins, the biosynthetic relationships between amino acids, and the origin of the genetic code. Orig. Life Evol. Biosph. 26, 589–609.

Di Giulio, M., 1997. On the origin of the genetic code. J. Theor. Biol 187, 573–581.

Epstein, C.J., 1967. Non-randomness of amino-acid changes in the evolution of homologous proteins. Nature 215, 355–359.

Goldsack, D.E., 1969. Relation of amino acid composition and the Moffitt parameters to the secondary structure of proteins. Biopolymers 7, 299–313.

Goodman, M., Moore, G.W., 1977. Use of Chou-Fasman amino acid conformational parameters to analyze the organization of the genetic code and to construct protein genealogies. J. Mol. Evol. 10, 7–47.

Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

Gupta, S.K., Majumdar, S., Bhattacharya, T.K., Ghosh, T.C., 2000. Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochem. Biophys. Res. Commun. 269, 692–696.

Guzzo, A.V., 1965. The influence of amino-acid sequence on protein structure. Biophys. J. 5, 809–822.

Havsteen, B.H., 1966. A study of the correlation between the amino acid composition and the helical content of proteins. J. Theor. Biol. 10, 1–10.

Jurka, J., Smith, T.F., 1987. Beta turns in early evolution: chirality, genetic code, and biosynthetic pathways. Cold Spring Harbor Symp. Quant. Biol. 52, 407–410.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Lesnik, T., Reiss, C., 1996. A method to detect transmembrane helical segments at the nucleotide level. Biochem. Mol. Biol. Int. 38, 937–955.

Lesnik, T., Reiss, C., 1998. Detection of transmembrane helical segments at the nucleotide level in eukaryotic membrane protein genes. Biochem. Mol. Biol. Int. 44, 471–479.

Levitt, M., 1978. Conformational preferences of amino acids in globular proteins. Biochemistry 17, 4277–4285.

Prothero, J.W., 1966. Correlation between the distribution of amino acids and alpha helices. Biophys. J. 6, 367–370.

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

Shimizu, M., 1982. Molecular basis for the genetic code. J. Mol. Evol. 18, 297–303.

Sjostrom, M., Wold, S., 1985. A multivariate study of the relationship between the genetic code and the physical-chemical properties of amino acids. J. Mol. Evol. 22, 272–277.

Szent-Gyorgyi, A.G., Cohen, C., 1957. Role of proline in polypeptide chain configuration of proteins. Science 126, 697.

Taylor, F.J., Coates, D., 1989. The code within the codons. Biosystems 22, 177–187.

Tolstrup, N., Toftgard, J., Engelbrecht, J., Brunak, S., 1994. Neural network model of the genetic code is strongly correlated to the GES scale of amino acid transfer free energies. J. Mol. Biol. 243, 816–820.

Woese, C.R., 1967. The Genetic Code. Harper & Row, New York.

Wolfenden, R.V., Cullis, P.M., Southgate, C.C., 1979. Water, protein folding, and the genetic code. Science 206, 575–577.

Yarus, M., 1991. An RNA-amino acid complex and the origin of the genetic code. New Biol. 3, 183–189.

Zhang, J., 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J. Mol. Evol. 50, 56–68.