ELSEVIER

Gene 300 (2002) 179–187

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

The base composition of the human genes is correlated with the secondary structures of the encoded proteins

Giuseppe D'Onofrio[a,*], Tapash Chandra Ghosh[b], Giorgio Bernardi[a]

[a]*Laboratorio di Evoluzione Molecolare, Stazione Zoologica A. Dohrn, 80121 Naples, Italy*
[b]*Bioinformatics Centre, Bose Institute, P 1/12, C. I. T. Scheme VII M, Kolkata 700 054, India*

Received 5 June 2002; received in revised form 18 September 2002; accepted 23 September 2002
Received by T. Gojobori

Abstract

The analysis of a non-redundant set of human proteins, for which both the crystallographic structures and the corresponding gene sequences are available, show that bases at third codon position are non-uniformly distributed along the coding sequences. Significant compositional differences are found by comparing the gene regions corresponding to the different secondary structures of the proteins. Inter- and intra-structure differences were most pronounced in the GC-richest genes. These results are not compatible with any proposed hypotheses based on a neutral process of formation/maintenance of the high GC_3 levels of the genes localized in the GC-richest isochores of the human genome.

Keywords: Base composition; Human protein; Coding sequence; GC_3; Isochore; Encoded protein

1. Introduction

Neglecting satellite DNAs, base composition is not uniform across the human genome, not only at large-scale level, as first showed by density gradient centrifugation of DNA preparations (Macaya et al., 1976; Thiery et al., 1976) and, more recently, by sliding window analysis of the draft human genome sequence (Pavliček et al., 2002), but also at the gene and at the coding sequence level. Indeed, it was reported that the GC level (the level of guanine + cytosine) of introns was 5–10% lower than the corresponding exons (Aïssani et al., 1991; Clay et al., 1996), whereas, by sliding window analysis, the GC level at third codon positions (GC_3) was found to fluctuate along the coding regions (Alvarez-Valin et al., 1998). The fluctuation of GC_3 along the coding sequences was first reported analysing a set of phage and bacterial genes, a result obtained by sliding window analyses (Wada and Suyama, 1985). The same authors also reported that, along the coding regions, the GC_3 levels were negatively correlated with the corresponding GC_{1+2} levels (whereas the average GC_3 and GC_{1+2} per coding sequence are positively correlated; Wada and Suyama, 1985; Bernardi and Bernardi, 1986; Sueoka, 1988; D'Onofrio and Bernardi, 1992). Moreover, the correlation disappeared using a sliding window shorter than five codons (Wada and Suyama, 1986). This observation stressed the fact that window length, an arbitrary parameter, can affect the final result. Therefore, we developed an alternative approach in order to analyse the base composition within gene sequences. Complete coding sequences were first aligned with the corresponding secondary structures of the proteins, and then the nucleotide regions corresponding to the same protein secondary structure elements were pooled together. In this way each nucleotide coding sequences was split in a defined number of 'nucleotide structures '. Investigations from our laboratory based on this approach using predicted secondary structures of the proteins showed that base frequencies at third codon positions, as well as synonymous and non-synonymous substitution rates, were significantly different in different secondary structures (Chiusano et al., 1999).

Here we analysed a non-redundant set of human proteins for which both the complete crystallographic structures and the corresponding gene sequences were available. First of all, we confirmed that base composition at third codons positions is different in region corresponding to different protein secondary structures, as observed in predicted

* Corresponding author. Tel.: +39-81-583-3311; fax: +39-81-746-3155.
E-mail address: donofrio@sunev.szn.it (G. D'Onofrio).

0141-933/02/$ - see front matter
PII: S0378-1119(02)01045-4

secondary structures of proteins (Chiusano et al., 1999). Second, we found systematic compositional differences not only among, but also within the secondary structures levels. To further elucidate this finding, we investigated the role played by codon usage, which was reported to be different in different secondary structures of the proteins (Chou and Zhang, 1993; Adzhubei et al., 1996; Oresic and Shalloway, 1998). The present results confirm that several synonymous codons have different frequencies in different protein secondary structures. However, codon usage accounted for only a small part of the differences in base composition at third codon positions, which were, therefore, mainly due to the frequencies of amino acids that are different in different secondary structures of proteins (Szent-Gyorgyi and Cohen, 1957; Guzzo, 1965; Havsteen, 1966; Prothero, 1966; Cook, 1967; Goldsack, 1969; Chou and Fasman, 1974; Levitt, 1978). We discuss the implications of these findings as far as the formation and maintenance of isochores are concerned.

2. Materials and methods

A non-redundant set of proteins for Homo sapiens was retrieved from http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html. We extracted only sequences whose complete or nearly complete crystal structures are available in PDB (Westbrook et al., 2002). TBLASTN was then used against the NCBI database to retrieve the corresponding gene sequences from GenBank. TBLASTN compares a query protein sequence to a DNA sequence by translating the DNA in all the six reading frames. Only protein sequences that had 100% identity with the DNA sequences were retained to avoid any ambiguity of the one-to-one correspondence between amino acids and codons. The PDB accession numbers, the gene length (bp) and the base composition of the 62 retrieved genes, as well as the corresponding length (bp%) and base composition of each structure, are listed in Table 1. The DSSP program (Kabsch and Sander, 1983) was used for the secondary structural assignments of individual residues of proteins. In the present paper, we assumed that no significant differences exist between G (3_{10}) and H (α helix) structures, which were pooled and denoted as helix. Likewise, structures E (extended strand) and B (isolated β-bridge) were pooled and denoted as strand. Those regions not annotated as G, H, E or B were pooled and defined as aperiodic. The rationale for this choice was to increase the number of nucleotides, codons and amino acids, in order to have reliable statistical analyses. A program developed in the C language was used to extract the nucleotide sequences of corresponding protein secondary structural elements assigned by DSSP. The frequencies of bases, codons and amino acids, and the Gravy scores were calculated using codonW 1.3 (J. Peden; http://molbiol.ox.ac.uk/Win95.codonW.zip).

The *t*-test for dependent samples was used to evaluate the significance of the pairwise differences in nucleotide composition in the three secondary structures.

3. Results and discussion

3.1. Compositional analysis

The GC_3 levels of strand, helix and aperiodic were 62.6% (S.E. 0.022), 56.6% (S.E. 0.021) and 54.7% (S.E. 0.019), respectively, whereas the GC_{1+2} values in the three structures were 39.8% (S.E. 0.01),45.4% (S.E. 0.008)% and 51.6% (S.E. 0.007) (Fig. 1). The GC_3 levels were significantly higher in strand than in either helix and aperiodic, the p values were $< 4.5 \times 10^{-4}$ and $< 5.1 \times 10^{-7}$, respectively; whereas in helix vs. aperiodic the p value was higher than the 5% threshold ($p < 8.4 \times 10^{-2}$). The GC_{1+2} levels were all significantly different among the secondary structures of the proteins, at least $p < 10^{-6}$. The GC level, on the contrary, was never significantly different among structures.

The Gravy scores (Kyte and Doolittle, 1982) of strand, helix and aperiodic were 0.63 (S.E. 0.07), −0.45 (S.E. 0.09) and −0.84 (S.E. 0.04), respectively. All the pairwise comparisons were statistically significant, the lowest p value was that of helix vs. aperiodic ($p < 6.0 \times 10^{-3}$). The Gravy scores of the secondary structures of the proteins showed a positive trend with GC_3, as was previously found at the protein level (D'Onofrio et al., 1999), and a negative trend with GC_{1+2}. It is clear, therefore, that the correlations between GC_3 and GC_{1+2} with hydropathy are not merely a consequence of the correlation existing among codon positions, i.e. GC_3 vs. GC_{1+2} (D'Onofrio and Bernardi, 1992), as proposed by Eyre-Walker and Hurst (2001).

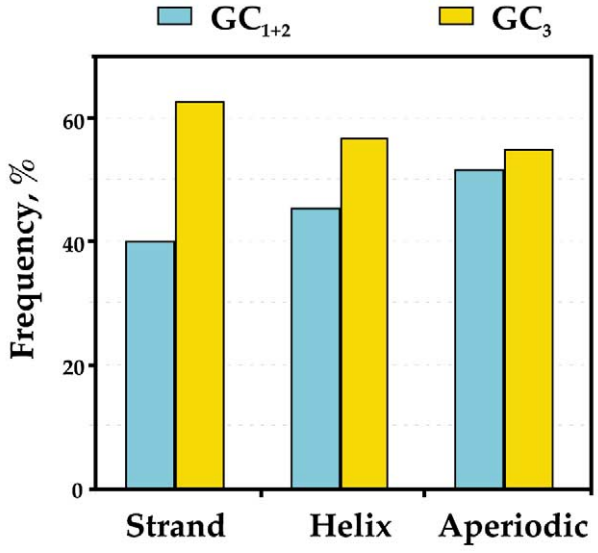

Fig. 1. Histogram of the frequencies of GC_3 and GC_{1+2} in the three secondary structures of the proteins. The structures are sorted according to decreasing Gravy scores.

The first indication that the four bases at third codon position have different frequencies in the different secondary structures of proteins was provided by an analysis of the consensus of five predictive methods run on a set of 34 human genes (Chiusano et al., 1999). The average frequencies of the four bases at third codon positions in the crystallographic secondary structure of the 62 human proteins analysed in the present work are reported in Fig. 2A, and the results of the statistical tests of the inter-structure pairwise comparisons are reported in Table 2 . C_3 and G_3 were found to be significantly different in all three secondary structures of proteins (strand, helix and aperiodic), whereas T_3 and A_3 were significantly different in two of the three structures. In spite of the fact that no method can predict the secondary structure of a protein with more than 80–85% accuracy (Frishman and Argos, 1997; Rost, 2001), the previously published data (Chiusano et al., 1999) are in good agreement with the results presented here. The discrepancies only concerned: (i) A_3, which was previously reported to be always significantly different in the three structures; (ii) G_3, which was previously reported to be not significantly different between strand and aperiodic; and (iii) T_3, reported to be not significantly different between strand and aperiodic, and significant between strand and helix. Statistical tests were also performed, within each of the three types of structure, to assess whether significant differences also occurred between G_3 and C_3, or between A_3 and T_3, at the intra-structure level. The p values reported in Table 2 showed that C_3 and G_3 were always significantly different in all the comparisons, whereas A_3 and T_3 were significantly different only in strand and aperiodic.

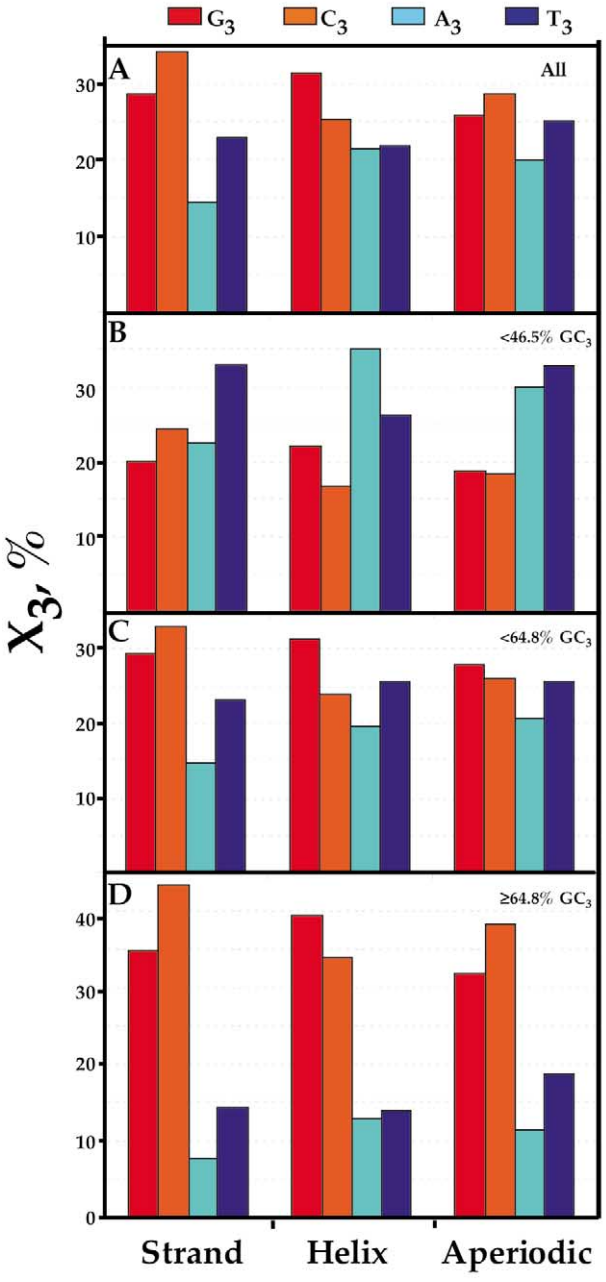

Fig. 2. Histograms of the frequencies of the bases at the third codon positions in the three secondary structures of the proteins of the all dataset (A), and of the data set partitioned according to the GC_3 boundaries defined in D'Onofrio (2002) (B–D). See also legend of Fig. 1.

3.2. Codon usage and amino acid frequency

Several authors have studied the possible relationships between the synonymous codon usage and protein secondary structural units, with different results. These conflicting conclusions were not surprising, since in none of the published papers the analyses were performed using complete crystal structure information (Chou and Zhang, 1993; Siemion and Siemion, 1994; Adzhubei et al., 1996; Oresic and Shalloway, 1998; Brunak and Engelbrecht, 1996; Tao and Dafu, 1998; Gupta et al., 2000). Indeed, some codon usage bias may be obscured when partial secondary structural information is used.

The frequency of each of the 59 synonymous codons was calculated in the three different secondary structures and a contingency χ^2 test was used to assess the significance of codon usage in pairwise comparisons. The codons found to be significantly different were: (i) **TTG** ($p < 0.01$), **CGC** ($p < 0.029$), and ACT ($p < 0.038$), in strand vs. helix (codons in bold where higher in strand); (ii) **CGT** ($p < 0.030$) and CAA ($p < 0.029$) in helix vs. aperiodic (codons in bold were higher in helix); (iii) **ATA** ($p < 0.008$), **CGC** ($p < 0.047$), ACT ($p < 0.010$), TTA ($p < 0.040$) and CAT ($p < 0.050$), in strand vs. aperiodic (codons in bold were higher in strand).

From these data it was evident that the pattern of the X-ending codons (where X can be one of the four bases) was not in agreement with the X_3 frequencies found in each structure. Indeed, only 50% of the above listed codons were in accordance with the results reported in Fig. 1 and Table 2. Moreover, when strand and helix were compared: (i) the only G-ending codon significantly different (TTG) was higher in strand, whereas G_3 was higher in helix; and (ii) A_3 was significantly different, but no single A-ending codon

Table 1
Length and nucleotide frequencies at third codon positions of the genes, as well as of the corresponding secondary structures of the proteins

		Gene					Strand					Helix					Aperiodic				
PDB no.[a]	bp	G_3	C_3	A_3	T_3	GC_3	G_3	C_3	A_3	T_3	bp, %	G_3	C_3	A_3	T_3	bp, %	G_3	C_3	A_3	T_3	bp, %
2HHM(A)	831	0.16	0.12	0.34	0.39	0.28	0.15	0.15	0.30	0.40	21.66	0.11	0.11	0.39	0.39	32.49	0.19	0.11	0.32	0.39	45.85
1FQ1(A)	636	0.13	0.18	0.40	0.30	0.31	0.04	0.25	0.33	0.38	11.32	0.17	0.22	0.38	0.23	38.68	0.10	0.13	0.43	0.33	50.00
1UCH	690	0.18	0.16	0.33	0.33	0.34	0.12	0.19	0.30	0.40	18.70	0.19	0.19	0.38	0.25	35.22	0.19	0.13	0.31	0.37	46.09
1D7Q(A)	432	0.19	0.15	0.31	0.35	0.35	0.28	0.09	0.31	0.31	22.22	0.27	0.09	0.46	0.18	7.64	0.15	0.19	0.29	0.38	70.14
1B34(A)	357	0.24	0.12	0.36	0.29	0.35	0.27	0.11	0.27	0.35	31.09	0.31	0.15	0.31	0.23	10.92	0.19	0.12	0.44	0.26	57.98
1A5R	303	0.25	0.12	0.32	0.32	0.37	0.18	0.21	0.25	0.36	27.72	0.18	0.18	0.55	0.09	10.89	0.27	0.07	0.31	0.36	61.39
1DT9(A)	1311	0.19	0.20	0.33	0.29	0.38	0.17	0.24	0.30	0.30	16.25	0.21	0.16	0.38	0.24	29.29	0.17	0.21	0.31	0.32	54.46
9ICW(A)	1005	0.20	0.19	0.30	0.30	0.40	0.18	0.33	0.16	0.33	13.43	0.20	0.17	0.32	0.31	47.16	0.21	0.18	0.33	0.28	39.40
1A6Q	1146	0.23	0.18	0.26	0.33	0.41	0.28	0.15	0.18	0.39	25.65	0.26	0.21	0.27	0.27	35.60	0.16	0.18	0.32	0.35	38.74
1HMP(A)	654	0.21	0.20	0.23	0.36	0.41	0.24	0.19	0.20	0.37	27.06	0.27	0.13	0.25	0.35	27.52	0.14	0.24	0.23	0.38	45.41
1HS6(A)	1833	0.23	0.20	0.25	0.32	0.43	0.23	0.23	0.24	0.30	22.09	0.23	0.15	0.29	0.33	11.29	0.22	0.20	0.25	0.32	66.61
1LZ1	444	0.21	0.23	0.22	0.34	0.44	0.00	0.56	0.06	0.38	10.81	0.31	0.12	0.24	0.33	34.46	0.19	0.24	0.25	0.33	54.73
1SPD(A)	462	0.23	0.22	0.25	0.31	0.45	0.33	0.25	0.10	0.31	31.17	0.11	0.00	0.56	0.33	5.84	0.18	0.23	0.30	0.30	62.99
1QIP(B)	552	0.22	0.23	0.23	0.32	0.45	0.23	0.15	0.29	0.33	28.26	0.17	0.24	0.28	0.30	25.00	0.23	0.28	0.17	0.31	46.74
1DUJ(A)	615	0.24	0.22	0.26	0.29	0.45	0.22	0.17	0.28	0.33	26.34	0.23	0.30	0.28	0.19	25.85	0.25	0.20	0.24	0.32	47.80
1CKS(B)	237	0.23	0.23	0.28	0.27	0.46	0.30	0.60	0.00	0.10	12.66	0.31	0.08	0.46	0.15	16.46	0.18	0.20	0.30	0.32	70.89
1LPB(B)	1395	0.18	0.27	0.25	0.29	0.46	0.18	0.29	0.23	0.30	27.53	0.20	0.34	0.18	0.28	22.37	0.17	0.24	0.30	0.30	50.11
1QUQ(D)	363	0.22	0.26	0.23	0.30	0.47	0.29	0.22	0.22	0.27	33.88	0.18	0.27	0.23	0.32	18.18	0.16	0.28	0.24	0.33	47.93
2ACY	297	0.27	0.22	0.30	0.20	0.50	0.20	0.29	0.34	0.17	41.41	0.38	0.13	0.17	0.33	24.24	0.27	0.21	0.38	0.15	34.34
1JDW	1269	0.23	0.27	0.22	0.28	0.50	0.24	0.18	0.21	0.38	20.09	0.25	0.20	0.25	0.31	24.82	0.21	0.35	0.21	0.24	55.08
1ULA	867	0.25	0.26	0.21	0.28	0.51	0.29	0.33	0.17	0.21	20.07	0.29	0.18	0.24	0.29	30.10	0.20	0.29	0.21	0.31	49.83
1DRF	786	0.28	0.24	0.23	0.25	0.52	0.27	0.23	0.18	0.32	22.90	0.24	0.20	0.29	0.27	15.65	0.29	0.26	0.22	0.23	61.45
1SAC(A)	669	0.27	0.27	0.18	0.29	0.54	0.25	0.27	0.21	0.28	42.15	0.27	0.27	0.18	0.27	4.93	0.28	0.28	0.14	0.30	52.91
1CRA	780	0.25	0.30	0.17	0.29	0.54	0.31	0.37	0.09	0.24	31.15	0.31	0.26	0.10	0.33	16.15	0.18	0.26	0.24	0.31	52.69
2CBA	780	0.25	0.30	0.17	0.29	0.54	0.31	0.37	0.09	0.24	31.15	0.31	0.26	0.10	0.33	16.15	0.18	0.26	0.24	0.31	52.69
1PRX(A)	672	0.25	0.31	0.17	0.27	0.55	0.26	0.44	0.04	0.26	22.32	0.21	0.29	0.26	0.24	29.46	0.25	0.26	0.19	0.31	48.21
1NDD(B)	243	0.40	0.16	0.21	0.24	0.56	0.50	0.23	0.12	0.15	32.10	0.60	0.07	0.27	0.07	18.52	0.25	0.15	0.25	0.35	49.38
3GRS	1437	0.27	0.29	0.18	0.26	0.56	0.36	0.32	0.13	0.19	24.22	0.25	0.25	0.20	0.29	32.99	0.22	0.30	0.20	0.28	42.80
1GUH(A)	666	0.28	0.28	0.23	0.22	0.56	0.47	0.20	0.13	0.20	6.76	0.25	0.29	0.25	0.22	58.56	0.29	0.27	0.21	0.23	34.68
1DG3(A)	1776	0.32	0.25	0.24	0.19	0.57	0.44	0.32	0.07	0.18	10.47	0.33	0.23	0.25	0.19	58.61	0.27	0.25	0.28	0.20	30.91
1FIT	441	0.30	0.27	0.18	0.25	0.57	0.19	0.41	0.06	0.34	21.77	0.30	0.28	0.19	0.23	29.25	0.35	0.21	0.24	0.21	48.98
1FIN(A)	894	0.28	0.29	0.15	0.27	0.57	0.28	0.28	0.19	0.26	18.12	0.27	0.32	0.13	0.28	34.56	0.28	0.28	0.16	0.27	47.32
2CPL	495	0.24	0.33	0.15	0.27	0.58	0.23	0.42	0.11	0.25	32.12	0.44	0.09	0.17	0.30	13.94	0.19	0.35	0.18	0.28	53.94
1AWR(A)	495	0.25	0.33	0.14	0.28	0.58	0.22	0.41	0.12	0.26	30.91	0.40	0.12	0.16	0.32	15.15	0.21	0.35	0.16	0.28	53.94
1URO(A)	1101	0.33	0.26	0.18	0.23	0.59	0.33	0.33	0.10	0.25	10.90	0.37	0.24	0.18	0.22	50.95	0.28	0.28	0.20	0.24	38.15
1EEM(A)	723	0.29	0.31	0.17	0.23	0.60	0.14	0.62	0.10	0.14	8.71	0.33	0.23	0.23	0.21	51.45	0.38	0.24	0.15	0.24	39.83
1CB6(A)	2130	0.34	0.27	0.18	0.22	0.60	0.38	0.29	0.11	0.22	18.59	0.31	0.31	0.17	0.22	35.21	0.24	0.33	0.18	0.25	46.20
2FKE	324	0.29	0.32	0.20	0.19	0.61	0.30	0.35	0.16	0.19	39.81	0.21	0.21	0.36	0.21	12.96	0.28	0.33	0.22	0.18	47.22
1B3O(B)	1542	0.27	0.34	0.16	0.23	0.61	0.26	0.35	0.17	0.23	12.84	0.34	0.29	0.17	0.21	24.32	0.25	0.36	0.16	0.24	62.84
2HMB	399	0.32	0.32	0.18	0.19	0.63	0.26	0.32	0.24	0.18	57.14	0.28	0.44	0.06	0.22	13.53	0.41	0.26	0.15	0.18	29.32
1HDR	732	0.33	0.31	0.15	0.21	0.64	0.29	0.35	0.18	0.18	22.54	0.35	0.32	0.09	0.25	38.52	0.33	0.27	0.20	0.20	38.93
2HGS(A)	1422	0.33	0.32	0.16	0.19	0.65	0.36	0.34	0.15	0.14	24.89	0.38	0.29	0.16	0.18	41.77	0.24	0.35	0.16	0.25	33.33
1BJ4(A)	1449	0.32	0.34	0.14	0.21	0.65	0.43	0.35	0.12	0.10	14.08	0.36	0.28	0.15	0.21	46.58	0.22	0.41	0.14	0.24	39.34

Table 1 (*continued*)

		Gene					Strand					Helix					Aperiodic				
PDB no.[a]	bp	G_3	C_3	A_3	T_3	GC_3	G_3	C_3	A_3	T_3	bp, %	G_3	C_3	A_3	T_3	bp, %	G_3	C_3	A_3	T_3	bp, %
1ILR1	429	0.25	0.41	0.16	0.18	0.66	0.26	0.43	0.15	0.16	51.75	0.20	0.40	0.20	0.20	10.49	0.26	0.37	0.17	0.20	37.76
1HTI(A)	747	0.35	0.31	0.12	0.22	0.66	0.30	0.35	0.08	0.28	16.06	0.37	0.30	0.16	0.17	41.77	0.35	0.31	0.11	0.24	42.17
1QHA(A)	2751	0.34	0.33	0.14	0.18	0.68	0.34	0.34	0.09	0.23	17.67	0.37	0.36	0.12	0.15	43.84	0.32	0.31	0.19	0.19	38.50
1DCP(C)	312	0.35	0.34	0.16	0.15	0.68	0.38	0.38	0.05	0.19	20.19	0.40	0.30	0.21	0.09	45.19	0.22	0.36	0.19	0.22	34.62
1NSK(R)	456	0.34	0.35	0.15	0.16	0.69	0.30	0.57	0.00	0.13	15.13	0.37	0.29	0.20	0.14	46.05	0.29	0.34	0.17	0.20	38.82
1AOI(C)	390	0.37	0.32	0.08	0.23	0.69	0.50	0.50	0.00	0.00	4.62	0.32	0.43	0.11	0.14	48.46	0.30	0.44	0.15	0.12	46.92
1CBS	414	0.43	0.30	0.14	0.13	0.73	0.40	0.35	0.16	0.09	54.35	0.56	0.11	0.17	0.17	13.04	0.42	0.31	0.09	0.18	32.61
1BO1(A)	1248	0.37	0.37	0.09	0.18	0.73	0.33	0.42	0.06	0.18	15.87	0.40	0.35	0.13	0.13	27.88	0.36	0.36	0.08	0.20	56.25
1ALD	1092	0.34	0.40	0.07	0.19	0.74	0.35	0.48	0.02	0.15	14.29	0.40	0.39	0.05	0.16	44.78	0.26	0.38	0.11	0.25	40.93
1A44	561	0.38	0.37	0.09	0.16	0.75	0.34	0.45	0.06	0.15	28.34	0.42	0.33	0.17	0.08	12.83	0.39	0.34	0.09	0.18	58.82
1ADS	948	0.37	0.38	0.08	0.17	0.75	0.38	0.45	0.00	0.17	14.87	0.44	0.34	0.10	0.12	38.29	0.31	0.39	0.09	0.21	46.84
1HDO(A)	618	0.38	0.38	0.13	0.12	0.76	0.35	0.47	0.14	0.05	20.87	0.44	0.32	0.10	0.14	34.47	0.35	0.38	0.14	0.13	44.66
1STF(I)	300	0.36	0.43	0.11	0.10	0.79	0.15	0.46	0.15	0.23	39.00	0.16	0.37	0.26	0.21	19.00	0.17	0.36	0.12	0.36	42.00
1A7A(A)	1296	0.36	0.43	0.05	0.16	0.79	0.35	0.44	0.06	0.15	16.67	0.40	0.43	0.05	0.12	43.75	0.32	0.44	0.05	0.20	39.58
1EFV(B)	765	0.48	0.31	0.08	0.12	0.80	0.37	0.39	0.09	0.15	30.20	0.50	0.34	0.07	0.09	30.20	0.39	0.40	0.08	0.13	39.61
1QH5(A)	780	0.42	0.37	0.08	0.12	0.80	0.49	0.35	0.07	0.09	25.00	0.52	0.23	0.12	0.13	33.08	0.44	0.35	0.07	0.15	41.92
1QGV(A)	426	0.35	0.45	0.06	0.14	0.80	0.17	0.50	0.07	0.27	21.13	0.51	0.44	0.00	0.05	28.87	0.32	0.45	0.09	0.14	50.00
1VHR(A)	555	0.32	0.50	0.07	0.11	0.83	0.30	0.61	0.04	0.04	12.43	0.30	0.46	0.09	0.15	42.70	0.34	0.53	0.05	0.08	44.86
1PIN(A)	489	0.43	0.43	0.07	0.07	0.86	0.42	0.47	0.06	0.06	22.09	0.44	0.36	0.10	0.10	23.93	0.43	0.44	0.06	0.07	53.99
Average		0.29	0.29	0.19	0.24	0.58	0.29	0.34	0.14	0.23	23.64	0.31	0.25	0.22	0.22	29.06	0.26	0.29	0.20	0.25	47.30
σ^2		0.006	0.007	0.007	0.005	0.021	0.011	0.016	0.008	0.010	121.4	0.011	0.011	0.014	0.007	188.6	0.006	0.009	0.008	0.006	90.419

[a] PDB accession numbers were sorted according to the GC_3 levels of the genes.

Table 2
p values of pairwise comparisons at inter-and intra-secondary structure of proteins

GC_3 range	X_3	S-H	H-A	S-A		S	H	A
All	G_3	3.6×10^{-2}	1.0×10^{-5}	4.5×10^{-2}	G_3-C_3	7.2×10^{-3}	9.3×10^{-4}	2.0×10^{-2}
	C_3	7.9×10^{-7}	2.0×10^{-3}	1.0×10^{-4}				
	A_3	2.7×10^{-5}	n.s.	7.7×10^{-6}	A_3-T_3	1.3×10^{-10}	n.s.	2.5×10^{-5}
	T_3	n.s.	1.0×10^{-3}	1.1×10^{-2}				
<46.5	G_3	n.s.	n.s.	n.s.	G_3-C_3	n.s.	n.s.	n.s.
	C_3	n.s.	n.s.	n.s.				
	A_3	3.0×10^{-3}	n.s.	1.0×10^{-2}	A_3-T_3	6.8×10^{-5}	3.5×10^{-2}	n.s.
	T_3	4.0×10^{-3}	5.0×10^{-3}	n.s.				
<64.8	G_3	n.s.	2.6×10^{-2}	n.s.	G_3-C_3	n.s.	2.9×10^{-2}	n.s.
	C_3	3.0×10^{-3}	n.s.	3.4×10^{-2}				
	A_3	4.3×10^{-2}	n.s.	1.0×10^{-3}	A_3-T_3	1.0×10^{-4}	1.2×10^{-2}	1.2×10^{-2}
	T_3	n.s.	n.s.	n.s.				
≥64.8	G_3	4.6×10^{-2}	9.5×10^{-5}	n.s.	G_3-C_3	6.0×10^{-3}	n.s.	5.0×10^{-3}
	C_3	2.9×10^{-6}	2.7×10^{-3}	6.0×10^{-4}				
	A_3	1.4×10^{-3}	4.6×10^{-2}	2.3×10^{-4}	A_3-T_3	1.6×10^{-3}	n.s.	3.4×10^{-5}
	T_3	n.s.	1.0×10^{-3}	5.5×10^{-3}				

S, H and A correspond to Strand, Helix and Aperiodic structures, respectively. n.s., not significant.

was found to be significantly different in the same comparison. Similarly, when helix and aperiodic were compared, the significant difference of G_3 was not supported by any G-ending codon, whereas, although a single A-ending codon was significant, A_3 was not significant. Finally, when strand and aperiodic were compared the significant difference of G_3 was not supported by any G-ending codon. It is worth stressing that 80% of the significant codons belonged to amino acids that had significantly different frequencies in the different crystallographic structures (see below).

Table 3
Average amino acid (aa) frequencies in strand (S), helix (H) and aperiodic (A) structures, and *p* values of pairwise comparisons

aa	S	H	A	S-H	H-A	S-A
Arg	4.22	6.52	4.99	6.5×10^{-3}	1.6×10^{-2}	n.s.
Ser	4.36	5.10	7.12	n.s.	2.7×10^{-4}	7.9×10^{-6}
Leu	10.59	10.48	6.72	n.s.	6.4×10^{-6}	2.4×10^{-5}
Ile	10.31	4.79	3.01	8.6×10^{-7}	2.2×10^{-4}	2.5×10^{-15}
Val	13.73	6.21	4.95	1.2×10^{-11}	1.4×10^{-2}	2.0×10^{-17}
Pro	1.76	3.01	7.28	1.8×10^{-3}	1.4×10^{-13}	5.3×10^{-16}
Thr	6.25	4.35	5.60	2.0×10^{-2}	1.2×10^{-2}	n.s.
Ala	4.69	9.07	5.97	1.0×10^{-4}	1.2×10^{-4}	4.1×10^{-2}
Gly	5.01	3.66	11.95	3.2×10^{-2}	2.1×10^{-17}	1.3×10^{-15}
Phe	5.96	3.96	3.25	3.9×10^{-4}	n.s.	4.4×10^{-6}
Tyr	4.39	2.92	2.61	2.6×10^{-5}	n.s.	1.4×10^{-2}
His	3.06	2.08	2.69	n.s.	n.s.	n.s.
Asn	2.95	3.28	4.69	n.s.	2.9×10^{-3}	1.1×10^{-2}
Lys	5.67	8.69	7.67	3.32×10^{-5}	n.s.	4.9×10^{-3}
Asp	2.88	5.28	7.66	2.8×10^{-8}	1.1×10^{-4}	7.0×10^{-13}
Glu	5.07	10.16	6.39	4.3×10^{-12}	9.8×10^{-5}	3.1×10^{-2}
Cys	2.32	1.41	1.25	3.7×10^{-3}	n.s.	2.7×10^{-3}
Gln	2.64	4.87	3.65	1.7×10^{-8}	5.4×10^{-3}	2.9×10^{-2}
Met	2.82	2.76	1.65	n.s.	6.7×10^{-3}	2.1×10^{-2}
Trp	1.26	1.50	0.93	n.s.	2.3×10^{-2}	n.s.

The analysis of the amino acid frequencies in the three secondary structures of the proteins and the corresponding *p* values of the statistical test are reported in Table 3. The results are in good agreement with those reported in the literature (Szent-Gyorgyi and Cohen, 1957; Guzzo, 1965; Havsteen, 1966; Prothero, 1966; Cook, 1967; Goldsack, 1969; Chou and Fasman, 1974; Levitt, 1978). Indeed, for example, the majority of the hydrophobic amino acids were significantly higher in strand; Ala and Glu were the most frequent amino acids in helix, in accordance with the amphipathic character of this structure; and Pro, as expected, was highly frequent in aperiodic.

In order to check how the properties of the amino acid affects the X_3 bases, the codons having the same X-ending base were assigned to three groups in hydrophobic, hydrophilic and amphipathic, according to the hydropathy scale of Kyte and Doolittle (1982). Each group was compared among structures and the corresponding differences are reported in Fig. 3.

In strand vs. helix, the significant difference in C_3 (see Table 2) was due to the combined effect of the C-ending codons of hydrophobic and amphipathic amino acids, whereas those in G_3 and A_3 were due to hydrophilic amino acids; the lack of significance in T_3 was due to a counter balance of hydrophobic amino acids on one hand and hydrophilic plus amphipathic amino acids on the other.

In the helix vs. aperiodic comparison, the significant difference in C_3 was due to the high delta of hydrophobic amino acids; that in G_3 was due to the combined effect of hydrophobic and hydrophilic, and that in T_3 was due to the amphipathic, whereas the lack of significance for A_3 was

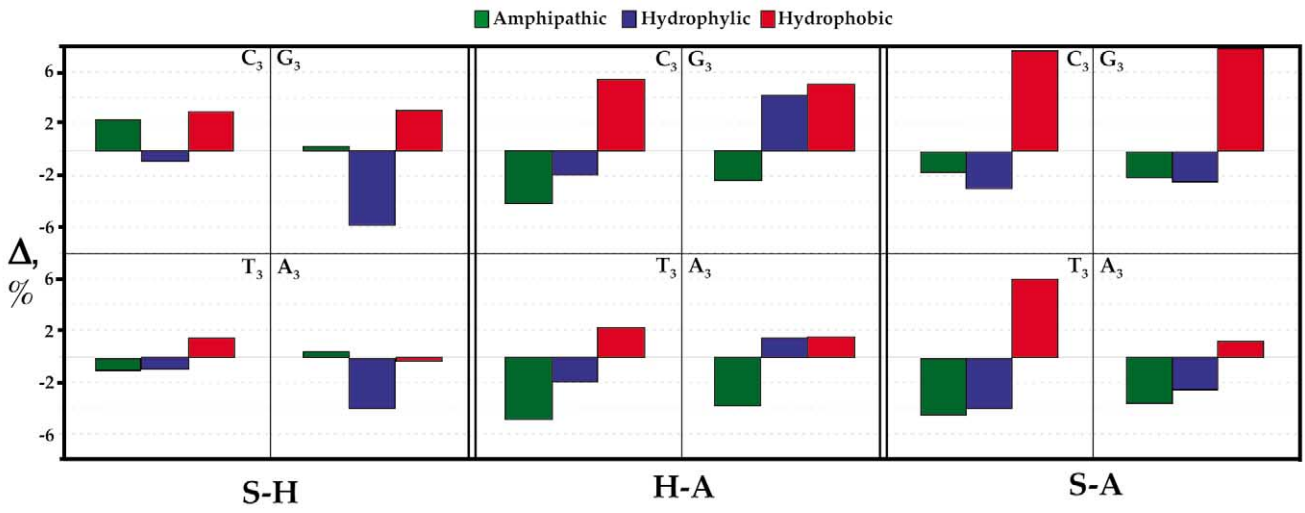

Fig. 3. Histograms of the delta of amphipathic, hydrophilic and hydrophobic amino acids in the three secondary structures of the proteins.

due to the counter balance of amphipathic versus hydrophobic plus hydrophilic amino acids.

In strand vs. aperiodic, the significant differences of C_3,G_3 and T_3 were all due to the hydrophobic amino acids, whereas that of A_3 was due to the combined effect of hydrophilic and amphipathic amino acids.

3.3. Compositional analysis after partition

The different occurrences of the four bases at the third codon positions in the different crystallographic secondary structures of the proteins prompted us to investigate the compositional properties according to the GC_3 levels of coding sequences. Using 46.5% GC_3 as the upper limit for the GC-poor genes, and 64.8% GC_3 as the lower limit for the GC-rich genes (D'Onofrio, 2002), the set of genes was split into three groups, containing 27.4, 38.7 and 33.8% of the genes, respectively. The results are reported in Fig. 2B–D, and the *p* values of the statistical tests performed at inter-and intra-structure levels are reported in Table 2.

In the $< 46.5\%$ GC_3 range, G_3 and C_3 were never significantly different, at both inter-and intra-structure levels. A_3 and T_3 were significantly different in all inter-structure comparisons except strand vs. aperiodic and helix vs. aperiodic. At intra- structure level, A_3-T_3 was significant except in aperiodic.

In the $< 64.8\%$ GC_3 range, the difference in G_3 was only significant in helix vs. aperiodic, whereas the difference C_3 and A_3 were significant in strand vs. both helix and aperiodic. T_3 turned out to be never significant. As regards the intra-structure comparisons, G_3-C_3 was significant only in helix, whereas A_3-T_3 was always significant, T_3 being always higher than A_3.

In the $< 64.8\%$ GC_3 range, the *p* values of the inter-structure comparisons were all significant, except for G_3 and T_3 in strand vs. aperiodic and helix, respectively. At the intra-structure level, helix was the only secondary structure showing non-significant differences in both G_3-C_3 and A_3-T_3 comparisons. In this range, GC_3 was significantly higher in strand (78.0%, S.E. 0.021) than aperiodic (69.9%, S.E. 0.02), $p < 5.8 \times 10^{-4}$, whereas strand vs. helix (73.0%, S.E. 0.02) and helix vs. aperiodic were at the limit of significance, $p < 5.4 \times 10^{-2}$ and $p < 6.6 \times 10^{-2}$, respectively.

Interestingly, the number of significant *p* values strongly increased from GC-poor to GC-rich genes (Table 3). Two possible explanations can be considered for such results. First, the data presented above were simply the consequence of the split done according to the GC_3 level of the genes. In other words, the lack of significance for both G_3 and C_3 in the lowest GC_3 range could be visualized as expected, since in that range C_3 and G_3 can only fluctuate narrowly, making it impossible to find significant differences. However, the fact that the same considerations do not hold for A_3 and T_3 in the GC-richest range rules out this first possibility. Second, in order to avoid misleading interpretations due to the comparison of GC-poor and GC-rich genes encoding proteins with different functions, human/*Xenopus* orthologous gene were analysed. The results obtained rule out this second possibility (Ghosh et al., in preparation).

4. Conclusions

The results reported in the present paper can be summarized as follows:

(i) the positive correlation between the GC_3 levels of the genes and the hydropathy of the encoded proteins (D'Onofrio et al., 1999) also holds at the intra-structure level; on the contrary GC_{1+2} show a negative correlation;
(ii) significant differences, at inter-structure level, in the base composition at third codon position were largely

confirmed from the analyses of the crystallographic structures, and were extended to the intra-structure level;

(iii) the number of significant differences, both at inter-and intra-structure levels, increased at increasing GC_3 levels, after partitioning the data set into three groups according to the newly defined boundaries (D'Onofrio, 2002).

The first point clearly shows that the statement that in GC-rich genes different amino acid composition and hydropathy is likely to be a consequence of the correlation between GC level of isochores and GC_{1+2} (Eyre-Walker and Hurst, 2001) is wrong. GC_3 and GC_{1+2} are, indeed, both correlated with the GC of the isochores (Bernardi et al., 1985; D'Onofrio et al., 1991), but show opposite correlations with hydropathy, also at the intra-structure level. Taking into account the alternate occurrence of strand, helix and aperiodic structures along the gene, GC_3 and GC_{1+2} are expected to fluctuate in opposite phase along the gene. The result is not in contradiction with the observed correlation of synonymous and non-synonymous substitution rates along the genes (Alvarez-Valin et al., 1998). Incidentally, the first point explains the loss of negative correlation between GC_3 and GC_{1+2} when using a sliding window less than five codons in size (Wada and Suyama, 1985, 1986). Indeed, individual strand structures generally consist of some 5–10 residues, whereas the shortest α-helix structures can be six amino acids in size (α-helix comprises 3.6 amino acids per turn and consists of four turns on the average, but can be more than a factor two longer or shorter). Therefore, using a sliding window less than five codons, a size that is lower than the inferior limit of the structure length, the compositional information of the protein structure and, consequently, the correlations between GC_3 and GC_{1+2} are lost.

The second point clearly shows that the differences in base composition at third codon positions were not due to a different codon usage, but were instead affected by the different propensity of the amino acids in the different secondary structures of the proteins. Therefore, the physico-chemical properties of the secondary structures of the proteins are the main factor affecting the base composition at third codon positions and at the second codon positions (Chiusano et al., 2000), as well as the synonymous and non-synonymous substitution rates (Chiusano et al., 1999).

The last point deserves some further comments. In the compositional transition from *Xenopus* to human, i.e. the major transition, GC-poor and GC-rich genes showed a different substitution pattern for hydrophobic, hydrophilic and amphipathic amino acids, with an increase, in the latter, of the average hydrophobicity levels of the proteins (Cruveiller et al., 1999). Taking into account that the propensity for the amino acids is different in different secondary structure of the proteins (Kohel and Levitt, 1999), it was logical to expect a non-random distribution of the amino acids substitutions along the coding sequences. Indeed, a re-analysis of *Xenopus*/human orthologous genes not only supported the expected result, but also showed that the G_3 and C_3 increments were strongly non-uniformly distributed at inter-and intra-secondary structure levels (Ghosh et al. in preparation). As a consequence, the proposed hypotheses based on a neutral process of formation/maintenance of the GC_3 levels in the GC-richest isochores of the human genome, as well as the biased gene conversion hypothesis (Eyre-Walker and Hurst, 2001), which "is basically a neutral process" (Galtier et al., 2001), hardly fit with our results.

Acknowledgements

Thanks are due to S.K. Gupta for the program developed in C language used to extract the nucleotide sequences of corresponding protein secondary structural elements. T.C.G. was supported by DBT overseas fellowship, Government of India.

References

Adzhubei, A.A., Adzhubei, I.A., Krasheninnikov, I.A., Neidle, S., 1996. Non- random usage of 'degenerate'codons is related to protein-three-dimensional structure. FEBS Lett. 399, 78–82.

Aïssani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C., Bernardi, G., 1991. The compositional properties of human genes. J. Mol. Evol. 32, 493–503.

Alvarez-Valin, F., Jabbari, K., Bernardi, G., 1998. Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. J. Mol. Evol. 46, 37–44.

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

Bernardi, G., Olofson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.

Brunak, S., Engelbrecht, J., 1996. Protein structure and sequential structure of mRNA: alpha-helix and beta-sheet signals at the nucleotide level. Proteins 25, 237–252.

Chiusano, M.L., D'Onofrio, G., Alvarez-Valin, F., Jabbari, K., Colonna, G., Bernardi, G., 1999. Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. Gene 238, 23–31.

Chiusano, M.L., Alvarez-Valin, F., Di Giulio, M., D'Onofrio, G., Ammirato, G., Colonna, G., Bernardi, G., 2000. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of genetic code. Gene 261, 63–69.

Chou, J.J., Zhang, C.T., 1993. A joint prediction of the folding types of 1490 human proteins from their genetic codons. J. Theor. Biol. 161, 251–262.

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

Clay, O., Caccio, S., Zoubak, S., Mouchiroud, D., Bernardi, G., 1996. Human coding and noncoding DNA: compositional correlations. Mol. Phylogenet. Evol. 5, 2–12.

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

Cruveiller, S., Jabbari, K., D'Onofrio, G., Bernardi, G., 1999. Different

hydrophobicities of orthologous proteins from Xenopus and human. Gene 238, 15–21.
D'Onofrio, G., 2002. Expression patterns and gene distribution in the human genome. Gene 300, 155–159.
D'Onofrio, G., Bernardi, G., 1992. A universal compositional correlation among codon position. Gene 110, 81–88.
D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., Bernardi, G., 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J. Mol. Evol. 32, 504–510.
D'Onofrio, G., Jabbari, K., Musto, H., Bernardi, G., 1999. The correlation of protein hydropathy with the base composition of coding sequences. Gene 238, 3–14.
Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. Nat. Rev. Genet. 2, 549–555.
Frishman, D., Argos, P., 1997. The future of protein secondary structure prediction accuracy. Fold Des. 2, 159–162.
Galtier, N., Piganeau, G., Mouchiroud, D., Duret, L., 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159, 907–911.
Goldsack, D.E., 1969. Relation of amino acid composition and Moffitt parameters to the secondary structure of proteins. Biopolymers 7, 299–313.
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.
Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22, 2577–2637.
Kohel, P., Levitt, M., 1999. Structure-based comformational preferences of amino acids. Proc. Natl. Acad. Sci. USA 96, 12524–12529.
Kyte, J., Doolittle, R.F., 1982. A simple method for displaying hydropathic character of a protein. J. Mol. Biol. 157, 105–132.
Levitt, M., 1978. Conformational preferences of amino acids in globular proteins. Biochemistry 17, 4277–4285.
Macaya, G., Thiery, J.P., Bernardi, G., 1976. An approach to organization of eukaryotic genomes at a macromolecular level. J. Mol. Biol. 108, 237–254.
Oresic, M., Shalloway, D., 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. J. Mol. Biol. 281, 31–48.
Pavliček, A., Paces, J., Clay, O., Bernard, G., 2002. A compact view of isochores in the draft human genome sequence. FEBS Lett. 511, 165–169.
Prothero, J.W., 1966. Correlation between the distribution of amino acids and alpha helices. Biophys. J. 6, 367–370.
Rost, B., 2001. Review: protein secondary structure prediction continues to rise. J. Struct. Biol. 134, 204–218.
Siemion, I.Z., Siemion, P.J., 1994. The informational context of the third base in amino acid codons. Biosystems 33, 139–148.
Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85, 2653–2657.
Szent-Gyorgyi, A.G., Cohen, C., 1957. Role of proline in polypeptide chain configuration of proteins. Science 126, 697.
Tao, X., Dafu, D., 1998. The relationship between synonymous codon usage and protein structure. FEBS Lett. 434, 93–96.
Thiery, J.P., Macaya, G., Bernardi, G., 1976. An analysis of eukaryotic genomes by density gradient centrifugation. J. Mol. Biol. 108, 219–235.
Wada, A., Suyama, A., 1985. Third letters in codons counterbalance the (G + C)-content of their first and second letters. FEBS Lett. 188, 291–294.
Wada, A., Suyama, A., 1986. Local stability of DNA and RNA secondary structure and its relation to biological functions. Prog. Biophys. Mol. Biol. 47, 113–157.
Westbrook, J., et al., 2002. The Protein Data Bank: unifying the archive. Nucleic Acids Res. 30, 245–248.