ELSEVIER

Gene 238 (1999) 23–31

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure

Maria Luisa Chiusano [a,b,c], Giuseppe D'Onofrio [a], Fernando Alvarez-Valin [d], Kamel Jabbari [c], Giovanni Colonna [b], Giorgio Bernardi [a,c,*]

[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, I-80138 Naples, Italy*
[c] *Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France*
[d] *Sección Biomatemática, Facultad de Ciencias, Iguá 4225, Montevideo 11400, Uruguay*

Accepted 15 June 1999; Received by T. Gojobori

Abstract

We investigated the relationships between the nucleotide substitution rates and the predicted secondary structures in the three states representation (α-helix, β-sheet, and coil). The analysis was carried out on 34 alignments, each of which comprised sequences belonging to at least four different mammalian orders. The rates of synonymous substitution were found to be significantly different in regions predicted to be α-helix, β-sheet, or coil. Likewise, the nonsynonymous rates also differ, although expectedly at a lower extent, in the three types of secondary structure, suggesting that different selective constraints associated with the different structures are affecting in a similar way the synonymous and nonsynonymous rates. Moreover, the base composition of the third codon positions is different in coding sequence regions corresponding to different secondary structures of proteins. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Evolution; Genomes; Mutational bias; Selection-protein structure

1. Introduction

As indicated in the preceding paper (Cruveiller et al., 1999), a compositional genome transition accompanied the emergence of warm-blooded from cold-blooded vertebrates, and compositional changes that took place were then maintained until the present [see Bernardi (1995) for a review]. Although, at the coding sequence level, the compositional transition can practically only be studied by comparing *Xenopus* and mammalian genes, the maintenance of the changes can be investigated by analysing the homologous sequences from different mammalian orders.

When the maintenance of the directional changes was investigated in mammalian orthologous genes, it was found that the frequencies of synonymous substitutions correlated with the frequencies of nonsynonymous substitutions (Wolfe and Sharp, 1993; Mouchiroud et al., 1995; Ohta and Ina 1995) and were gene-specific (Mouchiroud et al., 1995).

A breakthrough was subsequently made by pushing the analysis of nucleotide substitutions to the intragenic level.

In a first step, synonymous positions of quartet (fourfold degenerate) codons of orthologous coding sequences from four orders of mammals (which had been separated for about 100 million years) were divided into conserved (no change), intermediate (one change) and variable (more than one change) positions. In GC-rich genes, the frequency of the three classes of positions was shown to deviate significantly from expectation based on a stochastic process in which nucleotide substitutions accumulate at random over time, whereas this was not the case for GC-poor genes (Cacciò et al., 1995). Moreover, in GC-rich coding sequences, synonymous positions (especially conserved positions) of quartet codons exhibited significantly different base compositions compared with expectations based on a 'random' substitution process from the 'ancestral' (consensus) sequence to the present day (actual) sequences,

* Corresponding author. Tel.: +39-081-7641360; fax +39-081-7641255.
E-mail address: bernardi@alpha.szn.it (G. Bernardi)

0378-1119/99/$ – see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(99)00258-9

whereas significant differences were rare in GC- poor genes (Zoubak et al., 1995). It should be noted that the latter results provided a novel demonstration of the 'neutrality' of changes in most synonymous positions of GC-poor genes and in a number of synonymous positions of GC-rich genes.

A second step was made by using a window analysis of all synonymous and nonsynonymous positions from the same set of orthologous genes (Alvarez-Valin et al., 1998). This showed that the intragenic variability of synonymous rates of coding sequences was correlated with that of nonsynonymous rates, mainly in GC-rich genes, and that the variation in GC level (and especially in C level) of all silent positions along each gene was correlated with the variation in synonymous rate. These results indicate that synonymous and nonsynonymous rates, as well as GC levels of synonymous positions, are under some common selective constraints.

The third step, reported here, was to analyse regions of the same set of genes corresponding to different protein structures. Regions corresponding to α-helix, β-sheet and coil were analysed in terms of substitution rates and of levels of GC and of individual nucleotides in third codon positions.

2. Materials and methods

2.1. Sequence dataset

The present analysis was performed on 34 multiple alignments of orthologous mammalian coding sequences, a subset of the 48 genes previously used to study the relationship between synonymous and nonsynonymous substitution rates at the intragenic level (Alvarez-Valin et al., 1998). The 34 alignments were selected to build a dataset composed by a balanced distribution of genes that showed highly significant or positive or negative correlation between the rates of synonymous and nonsynonymous substitutions (Alvarez-Valin et al., 1998). Each alignment comprises coding sequences from at least four different mammalian orders, and in no case did the alignments include more than one sequence belonging to a species of the same order. Only six alignments did not include a sequence from the order of murids (alignments in bold in Table 1).

Table 1
Gene names, lengths and GC_3 levels of protein-coding sequence

Gene[a]		L[b]	GC_3 (%)
No.	Name		
1	Apolipoprotein E	320	0.90
2	Dipeptidase	411	0.89
3	**A1 adenosine receptor**	328	0.88
4	**Prostaglandin E**	389	0.85
5	Na–H exchange protein	823	0.84
6	Creatine kinase M	380	0.84
7	Apolipoprotein A1	267	0.84
8	CD8 alpha chain	255	0.83
9	H,K ATPase beta subunit	603	0.83
10	**Retinol Binding Protein**	294	0.83
11	Glutathione peroxidase	856	0.80
12	GMP-phosphodiesterase alpha	492	0.80
13	**TNFalpha**	236	0.79
14	Na Glucose transporter	665	0.74
15	**Erythropoietin**	195	0.74
16	TRNA ligase	477	0.73
17	CD4 antigen	476	0.73
18	Prolyl-4-hydroxylase beta	794	0.71
19	Polymeric Ig receptor	508	0.71
20	D-amino acid oxidase	347	0.66
21	Interleukin 1B receptor	272	0.65
22	Interleukin 6 receptor	213	0.62
23	Endothelin	214	0.61
24	CD3 epsilon antigen	304	0.59
25	Na–K ATPase beta-1 subunit	213	0.59
26	Prolactin Receptor	625	0.57
27	Urate oxidase	304	0.57
28	**Interleukin 1A**	275	0.52
29	Flavin-containing monooxigenase	535	0.52
30	Link protein	354	0.51
31	Apolipoprotein H	345	0.47
32	Serum Albumin	609	0.45
33	Macrophage scavenger	455	0.40
34	Ca-ATPase	1220	0.39

[a] The alignments not comprising murids are in bold.
[b] Coding sequence size is given as number of codons.

2.2. Methods and analyses

The multiple aligned nucleic acid sequences were matched with the secondary structure predicted in terms of α-helix, β-sheet, turn and coil. The predicted structures used in this work correspond to the consensus of five different predicting methods (Levin et al., 1986; Deleage and Roux, 1987; Gibrat et al., 1987; Geourjon and Deleage, 1994, 1995). The prediction was performed using the 'Consensus Secondary Structure Prediction' program, available at the WWW server http://www.ibcp.fr/predict.html. The 'consensus' in the prediction of the secondary structure is defined as at least three algorithms (out of the five used) yielding the same predicted structure. The amino acid sequence from *Homo sapiens* (or from a Hamadryas baboon in the case of the urate oxidase gene) was used as the model for the prediction.

For each alignment the α-helix, β-sheet, and coil regions were pooled into three sets. Three sub-alignments were so obtained, each one comprising only one type of secondary structure. The base composition at the third codon position was calculated for each kind of structure on every alignment. Likewise, the number of substitutions was determined separately for each kind of structure in each gene. Synonymous and nonsynonymous rates were estimated according to Nei and Gojobori (1986).

Statistical tests were performed to evaluate the significance of the pairwise differences in substitution rates in the three predicted structures, as well as the differences in nucleotide composition in third codon positions. The test used for such purpose was the *t*-test for dependent samples.

3. Results

The coding sequences investigated, their numbering, as well as their size and the average GC_3 levels of the aligned sequences, are listed in Table 1. Since no differences were observed among the three subsets of genes (i.e. those showing highly significant or positive or negative correlation between the rates of synonymous and nonsynonymous substitutions), the set was ordered according to the GC_3 levels, which range from 0.90 to 0.39 [mean value *M* and standard deviation (SD) were 0.69 and 0.15 respectively]. The length of each alignment, excluding gaps (L − g), the percentage of structure predicted, and the percentage of each selected structure, α-helix, β-sheet and coil, is reported in Table 2. No consensus of the predicted turn regions was found. This absence may be due to: (i) the generally low representation of such regions in proteins; (ii) the different sensitiveness of the predicting methods, so that a consensus is difficult to find; (iii) the exclusion of regions formed by one element only from our analysis. Thus, any information about turn structures was disregarded. The five values in italic in the β-sheet column of Table 2 were also neglected because their extensions were less than 5% of the entire predicted structure.

The average synonymous and nonsynonymous substitution rates, in regions with different predicted structures, are presented in Table 3, whereas the frequencies

Table 2
Percent of predicted structures. L − g is the length in codons without gaps

Gene no.	L − g	Structure (%)	Coil (%)	α-Helix (%)	β-Sheet[a] (%)
1	307	97.39	10.70	88.63	*0.67*
2	335	89.85	36.88	50.50	12.63
3	326	94.79	36.89	46.28	16.83
4	409	96.58	41.27	45.32	13.42
5	811	92.36	43.79	44.06	12.15
6	380	88.95	51.48	39.94	8.58
7	262	98.47	12.40	87.60	*0.00*
8	227	87.67	58.29	18.59	23.12
9	199	86.94	45.09	39.88	15.03
10	290	96.21	53.41	32.98	13.62
11	851	94.48	39.68	55.10	5.22
12	492	92.68	37.94	36.40	25.66
13	231	94.81	52.51	23.74	23.74
14	602	91.36	41.46	38.55	20.00
15	187	93.58	38.29	57.14	*4.57*
16	445	89.44	48.74	22.61	28.64
17	470	91.28	51.98	40.33	7.69
18	738	93.90	58.44	15.73	25.83
19	501	94.41	41.86	51.37	6.77
20	346	89.02	55.52	21.43	23.05
21	260	93.85	45.90	44.67	9.43
22	204	94.61	34.72	65.29	*0.00*
23	201	94.53	58.42	30.00	11.58
24	303	88.12	69.29	13.86	16.85
25	187	93.05	57.47	14.94	27.59
26	554	91.16	63.56	20.20	16.24
27	299	88.63	47.55	39.25	13.21
28	263	91.64	42.32	34.86	22.82
29	532	90.79	46.58	35.61	17.81
30	354	92.66	57.93	22.26	19.82
31	345	88.41	75.08	9.18	15.74
32	607	95.88	28.01	70.79	*1.20*
33	449	94.43	41.51	47.41	11.09
34	1176	92.35	45.30	38.21	16.48
M		92.48	46.18	39.49	14.32
SD		2.97	13.43	19.28	8.14
Max/min		98.5/86.9	75.1/10.7	88.6/9.2	28.6/0.0

[a] Values in italic were excluded from the statistical analyses (see text).

Table 3
Mean rates of synonymous (SYN) and nonsynonymous (NSY) nucleotide substitutions in regions with different predicted structures

Gene no.	Coil		α-Helix		β-Sheet	
	SYN	NSY	SYN	NSY	SYN	NSY
1	0.42	0.33	0.45	0.18	–	–
2	0.48	0.12	0.37	0.11	0.44	0.08
3	0.53	0.05	0.29	0.03	0.19	0.05
4	0.57	0.11	0.42	0.05	0.34	0.07
5	0.44	0.05	0.35	0.02	0.20	0.03
6	0.43	0.01	0.44	0.03	0.37	0.02
7	0.29	0.09	0.44	0.18	–	–
8	0.70	0.37	0.48	0.32	0.41	0.15
9	0.42	0.05	0.36	0.07	0.37	0.07
10	0.78	0.09	0.67	0.11	0.49	0.06
11	0.70	0.04	0.66	0.04	0.61	0.00
12	0.47	0.03	0.32	0.01	0.44	0.01
13	0.49	0.12	0.35	0.12	0.27	0.06
14	0.81	0.07	0.50	0.07	0.43	0.07
15	0.24	0.07	0.32	0.09	–	–
16	0.82	0.34	0.58	0.25	0.43	0.27
17	0.70	0.07	0.55	0.06	0.39	0.04
18	0.64	0.29	0.69	0.34	0.55	0.26
19	0.99	0.02	0.63	0.05	0.50	0.01
20	0.50	0.11	0.58	0.10	0.39	0.11
21	0.67	0.28	0.37	0.18	0.44	0.09
22	0.28	0.24	0.36	0.25	–	–
23	0.48	0.15	0.63	0.15	0.32	0.09
24	0.38	0.04	0.51	0.09	0.40	0.04
25	0.50	0.31	0.44	0.21	0.49	0.21
26	0.60	0.16	0.58	0.17	0.50	0.15
27	0.48	0.07	0.67	0.07	0.39	0.02
28	0.69	0.23	0.73	0.16	0.37	0.21
29	0.58	0.07	0.41	0.12	0.31	0.05
30	0.47	0.02	0.26	0.02	0.34	0.01
31	0.73	0.13	1.18	0.16	0.53	0.21
32	0.84	0.15	0.61	0.16	–	–
33	0.53	0.16	0.68	0.19	0.66	0.20
34	0.51	0.03	0.49	0.02	0.44	0.03
M	0.56	0.13	0.51	0.12	0.41	0.09
SE	0.029	0.019	0.031	0.015	0.020	0.015

of the four bases and the GC level at the third codon positions, calculated as the average level in all the aligned sequences in a coding region, are reported in Table 4. The mean *M* and the standard error (SE) values are given in Table 5, and the results of the statistical tests are reported in Table 6.

The results of Tables 1–6 can be summarized as follows:

(1) Nonsynonymous and synonymous substitution rates have average values that are higher in coil structure than in α-helix; the latter is higher than in β-sheet (Table 5). Statistical tests (Table 6) show that the different rates of both nonsynonymous and synonymous substitutions are significant in the comparisons of both coil and α-helix versus β-sheet structure; in contrast, coil versus helix comparisons do not show a significance better than 5%. However, a closer inspection of the average synonymous rates of the α-helix structure (Table 3) shows that Apolipoprotein H has an unusually high value, in fact the highest one, twice as large as the mean value. Eliminating this value from the test makes the difference statistically significant ($p<2\%$).

(2) The negative correlation between synonymous substitution rate and GC_3 content found at intragenic level (Alvarez-Valin et al., 1998) is also found at the structure level (Fig. 1). In fact, slopes are negative in all the plots of synonymous substitution rates versus GC_3 levels in each structure. However, significant correlations are only found for α-helix and β-sheet. Indeed, the slopes s and the correlation coefficients r are: $s=-0.57$, $r=0.46$ in α-helix, and $s=-0.28$ and $r=0.38$ in β-sheet. The nonsynonymous substitution rates seem to show no significant trends with the GC_3 content, even if β-sheet does show a negative slope and a correlation coefficient that are statistically significant ($s=-0.17$ and $r=0.31$).

(3) As far as the nucleotide composition of third codon position is concerned, A_3 is higher in coil than in α-helix, which is, in turn, higher than in β-sheet. T_3 is similar in coil and β-sheet, and both are higher than in α-helix. C_3 is higher in β-sheet than in coil, which is, in turn, higher than in α-helix. Moreover, the differences among the structures increase with increasing GC_3 (Fig. 2). Finally, G_3 is higher in α-helix compared with coil and β-sheet, the latter structures not showing any significant difference. To sum up, nucleotide differences between coil and α-helix, as well as between α-helix and β-sheet, are always significant (Table 6). However, in the coil versus β-sheet comparisons, only A_3 and C_3 differ significantly ($p<10^{-7}$), whereas T_3 and G_3 do not show any significant difference.

(4) The GC_3 level is very similar in α-helix and β-sheet, and both are higher than in coil. This difference is highly significant, showing, in the paired comparison test, p values lower than 10^{-7} and 10^{-8} respectively (Table 6). The lack of difference in GC_3 levels between α-helix and β-sheet is related to the opposite trend of C_3 and G_3 levels in the two structures. Indeed, G_3 is higher in α-helix than β-sheet, whereas C_3 shows the opposite trend (see above).

(5) The CpG levels in the C_3pG_1 and C_2pG_3 dinucleotides were calculated in both the entire coding regions and in the three structures. The results of these analyses are displayed in Fig. 3, which also shows that the contributions of C_2pG_3 and C_3pG_1 levels are minor, compared with total G_3 and C_3 respectively.

4. Discussion

The idea to analyse synonymous positions in relation to protein structure is not a new one. Indeed, Lipman and Wilbur (1985) discussed a possible relationship

Table 4
Nucleotide frequencies at third codon positions for predicted structures

Gene no.	Coil					α-Helix					β-Sheet				
	A_3	T_3	C_3	G_3	GC_3	A_3	T_3	C_3	G_3	GC_3	A_3	T_3	C_3	G_3	GC_3
1	0.07	0.10	0.48	0.43	0.91	0.08	0.04	0.33	0.57	0.90	–	–	–	–	–
2	0.10	0.11	0.48	0.48	0.96	0.07	0.13	0.39	0.43	0.82	0.03	0.05	0.54	0.38	0.91
3	0.06	0.11	0.51	0.33	0.83	0.03	0.06	0.50	0.41	0.90	0.03	0.04	0.69	0.24	0.94
4	0.07	0.12	0.49	0.32	0.81	0.06	0.07	0.36	0.51	0.87	0.01	0.11	0.58	0.32	0.90
5	0.10	0.11	0.49	0.30	0.79	0.06	0.09	0.45	0.41	0.86	0.03	0.08	0.54	0.35	0.89
6	0.06	0.10	0.54	0.30	0.84	0.07	0.09	0.34	0.50	0.84	0.01	0.09	0.55	0.35	0.91
7	0.16	0.09	0.53	0.27	0.80	0.08	0.08	0.34	0.50	0.84	–	–	–	–	–
8	0.15	0.10	0.39	0.42	0.81	0.09	0.06	0.39	0.55	0.95	0.07	0.11	0.59	0.23	0.83
9	0.14	0.12	0.56	0.18	0.74	0.06	0.06	0.33	0.54	0.87	0.03	0.13	0.59	0.26	0.85
10	0.11	0.10	0.51	0.29	0.79	0.09	0.06	0.44	0.42	0.85	0.03	0.13	0.57	0.27	0.84
11	0.10	0.12	0.43	0.35	0.78	0.09	0.11	0.39	0.42	0.81	0.12	0.07	0.57	0.24	0.82
12	0.12	0.13	0.44	0.31	0.75	0.05	0.08	0.40	0.47	0.87	0.03	0.18	0.52	0.27	0.79
13	0.14	0.14	0.40	0.34	0.74	0.04	0.13	0.46	0.38	0.83	0.08	0.10	0.60	0.23	0.82
14	0.19	0.21	0.41	0.28	0.69	0.10	0.20	0.43	0.34	0.77	0.04	0.24	0.45	0.35	0.80
15	0.19	0.18	0.44	0.21	0.65	0.11	0.16	0.35	0.39	0.74	–	–	–	–	–
16	0.15	0.18	0.34	0.36	0.70	0.17	0.10	0.24	0.51	0.76	0.12	0.14	0.34	0.40	0.74
17	0.11	0.20	0.46	0.24	0.69	0.14	0.12	0.32	0.43	0.74	0.03	0.14	0.67	0.18	0.85
18	0.16	0.18	0.42	0.28	0.70	0.18	0.13	0.33	0.38	0.71	0.12	0.16	0.43	0.30	0.73
19	0.14	0.18	0.41	0.27	0.68	0.13	0.13	0.30	0.44	0.74	0.04	0.24	0.53	0.19	0.72
20	0.16	0.17	0.39	0.27	0.67	0.25	0.16	0.23	0.36	0.59	0.07	0.21	0.40	0.32	0.73
21	0.22	0.19	0.38	0.21	0.60	0.19	0.19	0.32	0.34	0.66	0.10	0.09	0.46	0.36	0.82
22	0.27	0.22	0.35	0.22	0.57	0.19	0.16	0.30	0.36	0.66	–	–	–	–	–
23	0.28	0.17	0.30	0.24	0.55	0.24	0.13	0.38	0.29	0.68	0.05	0.32	0.31	0.33	0.64
24	0.19	0.24	0.31	0.26	0.57	0.17	0.16	0.29	0.39	0.68	0.13	0.27	0.31	0.29	0.60
25	0.22	0.27	0.28	0.29	0.57	0.23	0.14	0.18	0.48	0.66	0.22	0.16	0.34	0.29	0.63
26	0.28	0.26	0.31	0.23	0.54	0.21	0.31	0.25	0.38	0.63	0.16	0.24	0.31	0.32	0.62
27	0.21	0.24	0.33	0.24	0.57	0.18	0.25	0.29	0.29	0.58	0.21	0.20	0.41	0.18	0.59
28	0.28	0.29	0.25	0.20	0.46	0.23	0.23	0.27	0.30	0.56	0.18	0.26	0.32	0.24	0.56
29	0.31	0.24	0.26	0.19	0.46	0.23	0.23	0.25	0.29	0.54	0.09	0.27	0.41	0.23	0.64
30	0.22	0.31	0.32	0.15	0.47	0.15	0.26	0.23	0.36	0.59	0.14	0.30	0.29	0.27	0.56
31	0.31	0.25	0.24	0.20	0.44	0.26	0.25	0.23	0.26	0.49	0.10	0.41	0.32	0.17	0.50
32	0.28	0.28	0.24	0.21	0.45	0.25	0.30	0.23	0.22	0.45	–	–	–	–	–
33	0.33	0.35	0.19	0.14	0.33	0.32	0.25	0.18	0.26	0.44	0.21	0.33	0.21	0.25	0.46
34	0.34	0.40	0.17	0.15	0.32	0.30	0.24	0.18	0.28	0.46	0.24	0.31	0.23	0.22	0.44
M	0.18	0.19	0.38	0.27	0.65	0.15	0.15	0.32	0.40	0.72	0.09	0.19	0.45	0.28	0.73
SE	0.014	0.014	0.019	0.014	0.027	0.014	0.014	0.014	0.015	0.026	0.07	0.10	0.14	0.06	0.15

between the constraints on codon usage and those related to protein function and suggested that natural selection at the level of codon usage was responsible for higher biases in well-conserved regions compared with less-conserved regions. Chou and Zhang (1993) discussed a possible weak correlation between the folding types of human protein and the frequencies of nucleotides in third codon positions. Adzhubei et al. (1996) showed that codons overrepresented in α-helix are

Table 5
Mean values and standard deviations of base composition and synonymous and nonsynonymous substitution rates in each structure

	A_3	T_3	C_3	G_3	GC_3	SYN	NSY
Coil							
M	0.18	0.19	0.38	0.27	0.65	0.56	0.13
SE	0.014	0.014	0.019	0.014	0.027	0.029	0.019
α-Helix							
M	0.15	0.15	0.32	0.40	0.72	0.51	0.12
SE	0.014	0.014	0.014	0.015	0.026	0.031	0.015
β-Sheet							
M	0.09	0.19	0.45	0.28	0.73	0.41	0.09
SE	0.013	0.018	0.026	0.011	0.028	0.020	0.015

Table 6
Statistical values of pairwise comparisons

	Coil/α-Helix	Coil/β-Sheet	α-Helix/β-Sheet
NSY	n.s.	6.7×10^{-3}	1.8×10^{-2}
SYN	n.s.	4.4×10^{-7}	6.4×10^{-4}
A_3	4.6×10^{-5}	4.6×10^{-8}	1.3×10^{-5}
C_3	9.8×10^{-6}	9.8×10^{-6}	9.8×10^{-6}
G_3	2.3×10^{-11}	n.s.	2.0×10^{-7}
T_3	5.3×10^{-6}	n.s.	1.5×10^{-2}
GC_3	9.0×10^{-7}	2.0×10^{-8}	n.s.

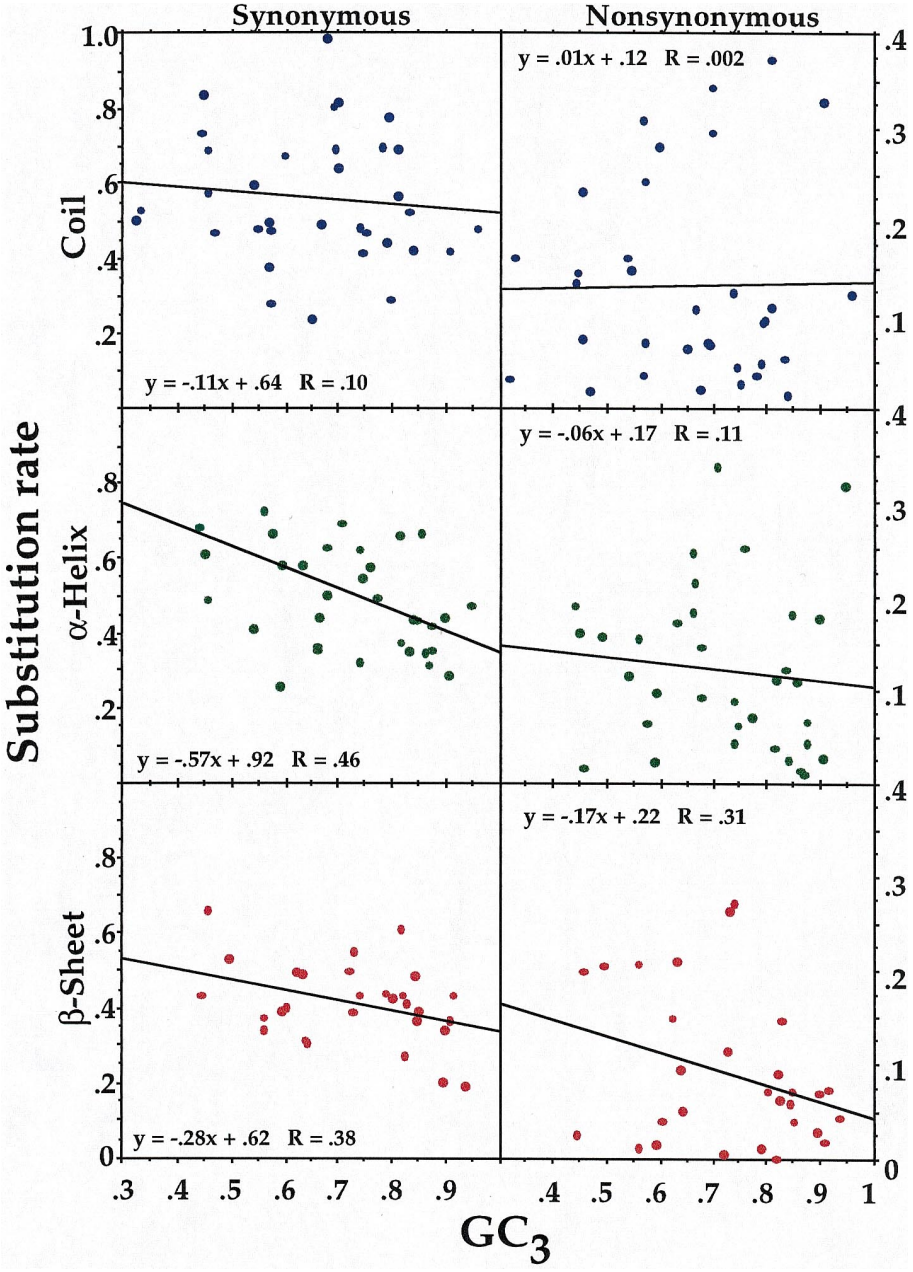

Fig. 1. Nucleotide substitution rates versus GC_3 in the predicted protein structures.

underrepresented in β-sheet and vice versa, and that the frequency of some synonymous codons depends on their position relative to secondary structure boundaries.

Since protein folding can proceed cotranslationally (Hardesty et al., 1995; Kolb et al., 1995; Netzer and Hartl, 1997), the optimization of folding may also include the tuning of codon pattern along mRNA to particular translation kinetics necessary to ensure the proper folding of a nascent peptide. In particular, codon context variations along mRNA affect speed and uniformity, either because of disparity in the rates of translation of different codons (Bonekamp et al., 1985; Wolin and Walter, 1988), or by introducing 'slow' regions of mRNA, in which a ribosome has to move over mRNA local secondary structure elements (Chaney and Morris, 1978). Synonymous codon usage might be biased towards rare codons in segments connecting domains and regular secondary structure blocks (Wolin and Walter, 1988; Krasheninnikov et al., 1989), and inter-domain regions would be characterized by lower rates of translation (Krasheninnikov et al., 1989; Thanaraj and Argos, 1996). However, Brunak and Engelbrecht (1996) found no correspondence between the distribution of rare codons and positions of structural blocks in proteins. Likewise, Tao and Dafu (1998) indicated that the correlation between synonymous

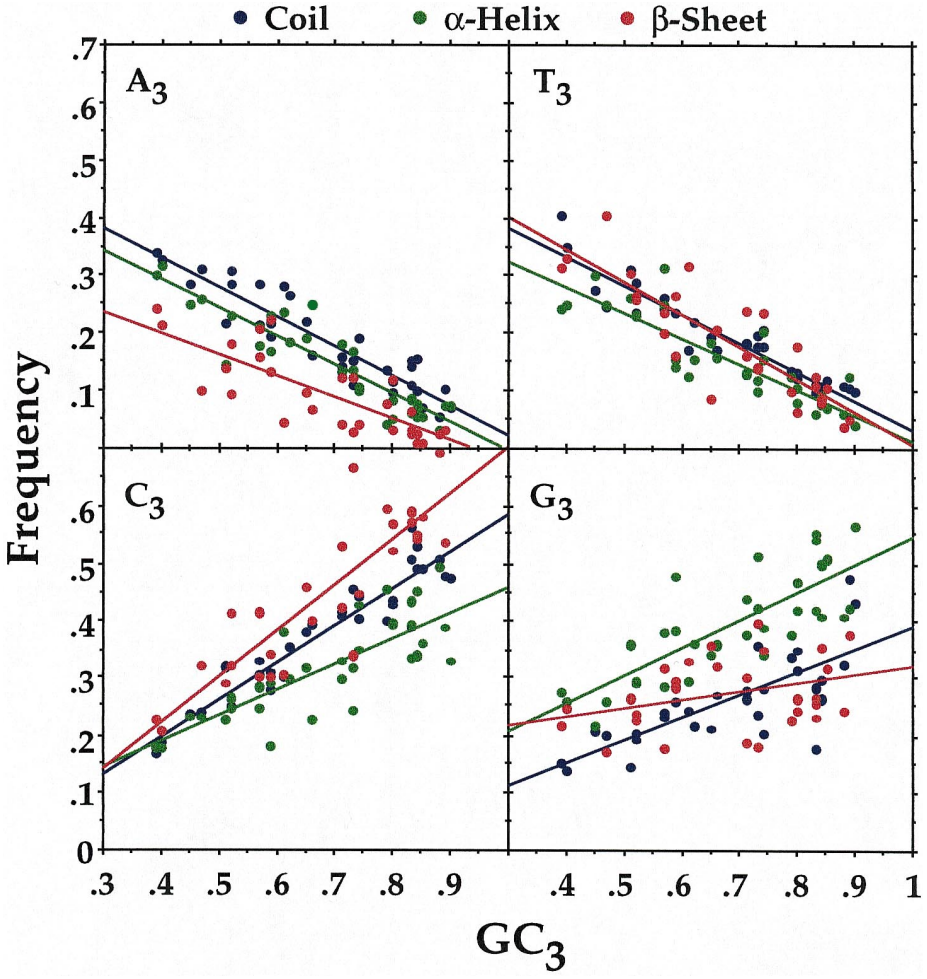

Fig. 2. Base frequencies in third codon positions of coding sequence regions corresponding to different predicted protein structures.

codon usage and protein secondary structure in *Escherichia coli* is not significant, and concluded that at least "there is not a universal codon-structure dictionary", since so many factors affect synonymous codon usage and translation rate (Osawa, 1995; Karlin and Mrazek, 1996; Solomovici et al., 1997).

Here, we have shown not only the correlations of protein structure with nucleotide substitution rates and base composition in third codon positions, but also the relevance of these results on the selectionist–neutralist controversy.

Previous work from our laboratory (see the Section 1) led to a series of observations that were difficult to reconcile with the idea that the maintenance of GC-rich coding sequences of mammalian orders could be explained by the mutational bias hypothesis (Sueoka, 1988, 1992). Indeed, at the level of intergenic comparisons, it is difficult to see how the mutational bias can explain the gene specificity of the substitution rates on the correlation between synonymous and nonsynonymous substitutions, the latter being notoriously under selection. At the level of intragenic comparisons, the conservation and the composition of a number of synonymous positions in quartet (fourfold degenerate) codons of GC-rich (but not of GC-poor genes) also raised a problem. Likewise, the correlation of nonsynonymous rates and GC level of silent positions with the variation in synonymous rates could not be accounted for by the mutational bias hypothesis. Needless to say, all these observations were perfectly compatible with an alternative explanation, namely negative selection.

The present work adds one more piece of evidence along the same line. Indeed, it shows statistically significant differences between both synonymous and nonsynonymous substitution rates in coding sequence regions corresponding to coils and β-sheet or to α-helix and β-sheet. Moreover, significant differences were also found among the nucleotides in third codon positions, with regard to coils, α-helix and β-sheet (G_3 and T_3 values being, however, not significantly different in coil and β-sheet). It is impossible to see how the random input of mutations could be so modulated in the coding sequences, other than by a selection mechanism. We conclude, therefore, that these observations complement

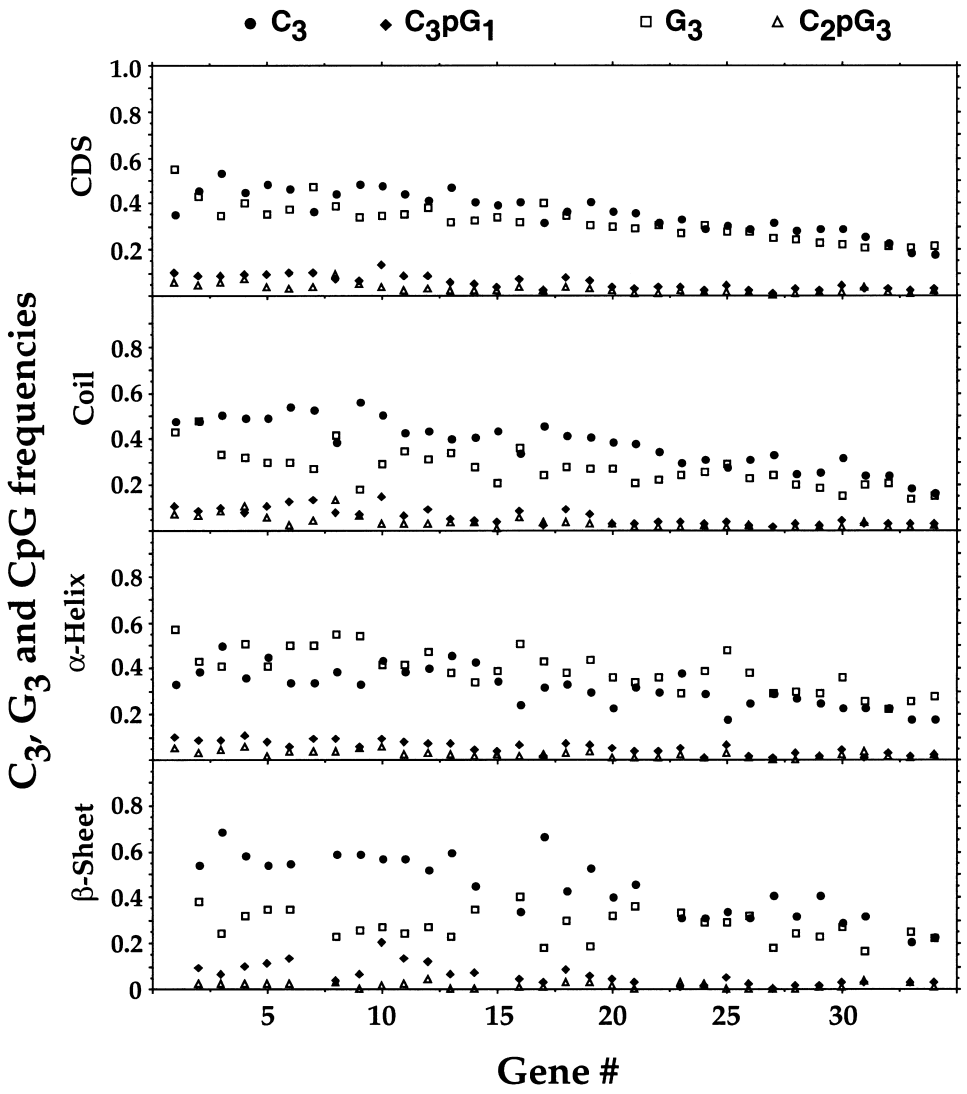

Fig. 3. C_3 and C_3pG_1, as well as G_3 and C_2pG_3, levels in the entire coding sequences and in the sequence regions corresponding to different predicted protein structures. The abscissa axis refers to the coding sequences (1–34) listed in Table 1. In the β-sheet plot, coding regions whose values were eliminated from the statistical analyses (see Table 2) are not reported.

the previous ones in supporting a selection mechanism for the conservation of GC-rich coding sequences in mammals.

Acknowledgements

M.L. Chiusano thanks the European Union for a PhD fellowship. We wish to thank Dr T. Gojobori for driving us to investigate CpG content.

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.

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., 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. 29, 445–476.

Bonekamp, F., Andersen, H.D., Christensen, T., Jensen, K.F., 1985. Codon-defined ribosomal pausing in *Escherichia coli* detected by using the pyrE attenuator to probe the coupling between transcription and translation. Nucleic Acids Res. 13, 4113–4123.

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

Cacciò, S., Zoubak, S., D'Onofrio, G., Bernardi, G., 1995. Nonrandom frequency patterns of synonymous substitution in homologous mammalian genes. J. Mol. Evol. 40, 280–292.

Chaney, W.G., Morris, A.J., 1978. Nonuniform size distribution of nascent peptides: the role of messenger RNA. Arch. Biochem. Biophys. 191, 734–741.

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.

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

hydrophobicities of orthologous proteins from *Xenopus* and man. Gene 238, 15–21.
Deleage, G., Roux, B., 1987. An algorithm for protein secondary structure prediction based on class prediction. Protein Eng. 1, 289–294.
Geourjon, C., Deleage, G., 1994. SOPM: a self optimised prediction method for protein secondary structure prediction. Protein Eng. 7, 157–164.
Geourjon, C., Deleage, G., 1995. SOPMA significant improvements in protein secondary structure prediction by prediction from multiple alignments. CABIOS 11, 681–684.
Gibrat, J.F., Garnier, J., Robson, B., 1987. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J. Mol. Biol. 198, 425–443.
Hardesty, B., Kudlicki, W., Odom, O.W., Zhang, T., McCarthy, D., Kramer, G., 1995. Cotranslational folding of nascent proteins on *Escherichia coli* ribosomes. Biochem. Cell Biol. 73, 1199–1207.
Karlin, S., Mrazek, J., 1996. What drives codon choices in human genome. J. Mol. Biol. 262, 459–472.
Krasheninnikov, I.A., Komar, A.A., Adzhubei, I.A., 1989. Role of the code redundancy in determining cotranslational protein folding. Biokhimiia 54, 187–200.
Kolb, V.A., Makeyev, E.V., Kommer, A., Spirin, A.S., 1995. Cotranslational folding of proteins. Biochem. Cell Biol. 73, 1217–1220.
Levin, J.M., Robson, B., Garnier, J., 1986. An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS Lett. 15, 303–308.
Lipman, D.J., Wilbur, W.J., 1985. Interaction of silent and replacement changes in eukaryotic coding sequences. J. Mol. Evol. 21, 161–167.
Mouchiroud, D., Gautier, C., Bernardi, G., 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. 40, 107–113.
Nei, M., Gojobori, T., 1986. Simple method for estimating the number of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.
Netzer, W.J., Hartl, F.U., 1997. Recombination of protein domains facilitate by cotranslational folding in eukaryotes. Nature 388, 343–349.
Osawa, S., 1995. Evolution of the Genetic Code. Oxford University Press, Oxford.
Ohta, T., Ina, Y., 1995. Variation in synonymous substitution rate among mammalian genes and correlation among synonymous and nonsynonymous divergences. J. Mol. Evol. 41, 717–720.
Solomovici, J., Lesnik, T., Reiss, C., 1997. Does *Escherichia coli* optimize the economics of the translation process? J. Theor. Biol. 185, 511–521.
Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85, 2653–2657.
Sueoka, N., 1992. Directional mutation pressure and molecular evolution: equilibria and asymmetric phylogenetic branching. J. Mol. Evol. 34, 95–114.
Tao, X., Dafu, D., 1998. The relationship between synonymous codon usage and protein structure. FEBS Lett. 434, 93–96.
Thanaraj, T.A., Argos, P., 1996. Ribosome-mediated translational pause and protein domain organization. Protein Sci. 5, 1594–1612.
Wolfe, K.H., Sharp, P.M., 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. 37, 441–456.
Wolin, S.L., Walter, P., 1988. Ribosome pausing and stacking during translation of a eukaryotic mRNA. EMBO J. 7, 3559–3569.
Zoubak, S., D'Onofrio, G., Cacciò, S., Bernardi, G., 1995. Specific compositional patterns of synonymous positions in homologous mammalian genes. J. Mol. Evol. 40, 293–307.