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 Università, 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 (Caccio 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,
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).

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; Gibrait 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 in each gene. Synonymous and nonsynonymous rates were estimated according to Nei and Gojobori (1986).

Table 1

<table>
<thead>
<tr>
<th>Gene name</th>
<th>L%</th>
<th>GC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apolipoprotein E</td>
<td>320</td>
<td>0.90</td>
</tr>
<tr>
<td>Dipeptidase</td>
<td>411</td>
<td>0.89</td>
</tr>
<tr>
<td>A1 adenosine receptor</td>
<td>328</td>
<td>0.88</td>
</tr>
<tr>
<td>Prostaglandin E</td>
<td>389</td>
<td>0.85</td>
</tr>
<tr>
<td>Na-H exchange protein</td>
<td>823</td>
<td>0.84</td>
</tr>
<tr>
<td>Creatine kinase M</td>
<td>380</td>
<td>0.84</td>
</tr>
<tr>
<td>Apolipoprotein A1</td>
<td>267</td>
<td>0.84</td>
</tr>
<tr>
<td>CD91 alpha chain</td>
<td>255</td>
<td>0.83</td>
</tr>
<tr>
<td>H-K ATPase beta subunit</td>
<td>603</td>
<td>0.83</td>
</tr>
<tr>
<td>Retinal Binding Protein</td>
<td>294</td>
<td>0.83</td>
</tr>
<tr>
<td>Glutathione peroxidase</td>
<td>856</td>
<td>0.80</td>
</tr>
<tr>
<td>GMP-phosphodiesterase alpha</td>
<td>492</td>
<td>0.80</td>
</tr>
<tr>
<td>TNFalpha</td>
<td>236</td>
<td>0.79</td>
</tr>
<tr>
<td>Na Glucose transporter</td>
<td>665</td>
<td>0.74</td>
</tr>
<tr>
<td>Erythropoietin</td>
<td>195</td>
<td>0.74</td>
</tr>
<tr>
<td>TRNA ligase</td>
<td>477</td>
<td>0.73</td>
</tr>
<tr>
<td>CD4 antigen</td>
<td>476</td>
<td>0.73</td>
</tr>
<tr>
<td>Prolyl-4-hydroxylase beta</td>
<td>794</td>
<td>0.71</td>
</tr>
<tr>
<td>Polymeric Ig receptor</td>
<td>508</td>
<td>0.71</td>
</tr>
<tr>
<td>D-amino acid oxidase</td>
<td>347</td>
<td>0.66</td>
</tr>
<tr>
<td>Interleukin 1B receptor</td>
<td>272</td>
<td>0.65</td>
</tr>
<tr>
<td>Interleukin 6 receptor</td>
<td>213</td>
<td>0.62</td>
</tr>
<tr>
<td>Endothelin</td>
<td>214</td>
<td>0.61</td>
</tr>
<tr>
<td>CD9 epsilon antigen</td>
<td>304</td>
<td>0.59</td>
</tr>
<tr>
<td>Na-K ATPase beta-1 subunit</td>
<td>213</td>
<td>0.59</td>
</tr>
<tr>
<td>Proslatin Receptor</td>
<td>623</td>
<td>0.57</td>
</tr>
<tr>
<td>Urate oxidase</td>
<td>304</td>
<td>0.57</td>
</tr>
<tr>
<td>Interleukin 1A</td>
<td>275</td>
<td>0.52</td>
</tr>
<tr>
<td>Flavin-containing monooxygenase</td>
<td>535</td>
<td>0.52</td>
</tr>
<tr>
<td>Link protein</td>
<td>354</td>
<td>0.51</td>
</tr>
<tr>
<td>Apolipoprotein H</td>
<td>345</td>
<td>0.47</td>
</tr>
<tr>
<td>Serum Albumin</td>
<td>609</td>
<td>0.45</td>
</tr>
<tr>
<td>Macrophage scavenger</td>
<td>455</td>
<td>0.40</td>
</tr>
<tr>
<td>Ca-ATPase</td>
<td>1220</td>
<td>0.39</td>
</tr>
</tbody>
</table>

The alignments not comprising murids are in bold.

Coding sequence size is given as number of codons.
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, $\alpha$-helix, $\beta$-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 $\beta$-sheet column of Table 2 were also neglected because their extensions were less than 3% 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>
<thead>
<tr>
<th>Gene no.</th>
<th>L $-$ g</th>
<th>Structure (%)</th>
<th>Coil (%)</th>
<th>$\alpha$-Helix (%)</th>
<th>$\beta$-Sheet* (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>307</td>
<td>97.39</td>
<td>10.70</td>
<td>88.63</td>
<td>0.67</td>
</tr>
<tr>
<td>2</td>
<td>355</td>
<td>89.85</td>
<td>36.88</td>
<td>50.50</td>
<td>12.63</td>
</tr>
<tr>
<td>3</td>
<td>326</td>
<td>94.79</td>
<td>36.89</td>
<td>46.28</td>
<td>16.83</td>
</tr>
<tr>
<td>4</td>
<td>409</td>
<td>96.58</td>
<td>41.27</td>
<td>45.32</td>
<td>13.42</td>
</tr>
<tr>
<td>5</td>
<td>811</td>
<td>92.36</td>
<td>43.79</td>
<td>44.06</td>
<td>12.15</td>
</tr>
<tr>
<td>6</td>
<td>380</td>
<td>88.95</td>
<td>51.48</td>
<td>39.94</td>
<td>8.58</td>
</tr>
<tr>
<td>7</td>
<td>262</td>
<td>98.47</td>
<td>12.40</td>
<td>87.60</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>227</td>
<td>87.67</td>
<td>58.29</td>
<td>18.59</td>
<td>23.12</td>
</tr>
<tr>
<td>9</td>
<td>199</td>
<td>86.94</td>
<td>45.09</td>
<td>39.88</td>
<td>15.03</td>
</tr>
<tr>
<td>10</td>
<td>290</td>
<td>96.21</td>
<td>53.41</td>
<td>32.98</td>
<td>13.62</td>
</tr>
<tr>
<td>11</td>
<td>851</td>
<td>94.48</td>
<td>39.68</td>
<td>55.10</td>
<td>5.22</td>
</tr>
<tr>
<td>12</td>
<td>492</td>
<td>92.68</td>
<td>37.94</td>
<td>36.40</td>
<td>25.66</td>
</tr>
<tr>
<td>13</td>
<td>231</td>
<td>94.81</td>
<td>52.51</td>
<td>23.74</td>
<td>23.74</td>
</tr>
<tr>
<td>14</td>
<td>602</td>
<td>91.36</td>
<td>41.46</td>
<td>38.55</td>
<td>20.00</td>
</tr>
<tr>
<td>15</td>
<td>187</td>
<td>93.58</td>
<td>36.29</td>
<td>57.14</td>
<td>4.57</td>
</tr>
<tr>
<td>16</td>
<td>445</td>
<td>89.44</td>
<td>48.74</td>
<td>22.61</td>
<td>20.64</td>
</tr>
<tr>
<td>17</td>
<td>470</td>
<td>91.28</td>
<td>51.98</td>
<td>40.33</td>
<td>7.69</td>
</tr>
<tr>
<td>18</td>
<td>718</td>
<td>93.90</td>
<td>56.44</td>
<td>15.73</td>
<td>25.83</td>
</tr>
<tr>
<td>19</td>
<td>501</td>
<td>94.41</td>
<td>41.06</td>
<td>51.37</td>
<td>6.77</td>
</tr>
<tr>
<td>20</td>
<td>346</td>
<td>89.02</td>
<td>55.52</td>
<td>21.43</td>
<td>25.05</td>
</tr>
<tr>
<td>21</td>
<td>260</td>
<td>93.85</td>
<td>45.59</td>
<td>44.67</td>
<td>9.43</td>
</tr>
<tr>
<td>22</td>
<td>204</td>
<td>94.61</td>
<td>34.72</td>
<td>65.29</td>
<td>0.00</td>
</tr>
<tr>
<td>23</td>
<td>201</td>
<td>94.53</td>
<td>34.82</td>
<td>30.00</td>
<td>11.38</td>
</tr>
<tr>
<td>24</td>
<td>303</td>
<td>88.12</td>
<td>49.29</td>
<td>13.86</td>
<td>16.85</td>
</tr>
<tr>
<td>25</td>
<td>187</td>
<td>93.05</td>
<td>57.47</td>
<td>14.94</td>
<td>27.59</td>
</tr>
<tr>
<td>26</td>
<td>554</td>
<td>91.16</td>
<td>63.56</td>
<td>20.20</td>
<td>16.24</td>
</tr>
<tr>
<td>27</td>
<td>299</td>
<td>88.83</td>
<td>47.55</td>
<td>39.25</td>
<td>13.21</td>
</tr>
<tr>
<td>28</td>
<td>263</td>
<td>91.64</td>
<td>42.32</td>
<td>34.86</td>
<td>22.82</td>
</tr>
<tr>
<td>29</td>
<td>532</td>
<td>90.79</td>
<td>46.51</td>
<td>35.61</td>
<td>17.81</td>
</tr>
<tr>
<td>30</td>
<td>354</td>
<td>92.66</td>
<td>57.93</td>
<td>22.26</td>
<td>19.82</td>
</tr>
<tr>
<td>31</td>
<td>345</td>
<td>88.41</td>
<td>75.08</td>
<td>9.18</td>
<td>15.74</td>
</tr>
<tr>
<td>32</td>
<td>607</td>
<td>95.88</td>
<td>28.01</td>
<td>70.79</td>
<td>1.20</td>
</tr>
<tr>
<td>33</td>
<td>449</td>
<td>94.43</td>
<td>41.51</td>
<td>47.41</td>
<td>11.09</td>
</tr>
<tr>
<td>34</td>
<td>1176</td>
<td>92.35</td>
<td>45.30</td>
<td>38.21</td>
<td>16.48</td>
</tr>
<tr>
<td>$M$</td>
<td>92.48</td>
<td></td>
<td>46.18</td>
<td>39.49</td>
<td>14.32</td>
</tr>
<tr>
<td>SD</td>
<td>2.97</td>
<td></td>
<td>13.45</td>
<td>19.28</td>
<td>8.14</td>
</tr>
<tr>
<td>Max/min</td>
<td>98.5/98.9</td>
<td></td>
<td>75.1/10.7</td>
<td>88.6/9.2</td>
<td>20.6/0.0</td>
</tr>
</tbody>
</table>

* Values in italic were excluded from the statistical analyses (see text).
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 < 0.05).

(2) The negative correlation between synonymous substitution rate and GC3 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 GC3 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 GC3 content, even if β-sheet does show a negative slope and a correlation coefficient that is statistically significant (s = −0.17 and r = 0.31).

(3) As far as the nucleotide composition of third codon position is concerned, A3 is higher in coil than in α-helix, which is, in turn, higher than in β-sheet. T3 is similar in coil and β-sheet, and both are higher than in α-helix. C3 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 GC3 (Fig. 2). Finally, G3 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 A3 and C3 differ significantly (p < 10−5), whereas T3 and G3 do not show any significant difference.

(4) The GC3 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 GC3 levels between α-helix and β-sheet is related to the opposite trend of C3 and G3 levels in the two structures. Indeed, G3 is higher in α-helix than β-sheet, whereas C3 shows the opposite trend (see above).

(5) The CpG levels in the CpG3 and CpG2 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 CpG3 and CpG2 levels are minor, compared with total G3 and C3 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

<table>
<thead>
<tr>
<th>Gene no.</th>
<th>coil</th>
<th>n-Helix</th>
<th>β-Sheet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A₁</td>
<td>T₁</td>
<td>C₁</td>
</tr>
<tr>
<td>1</td>
<td>0.07 0.10 0.48 0.43 0.91 0.08 0.04 0.33 0.57 0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.10 0.11 0.46 0.45 0.96 0.07 0.13 0.29 0.34 0.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.06 0.11 0.51 0.33 0.83 0.03 0.06 0.50 0.41 0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.07 0.12 0.49 0.32 0.81 0.06 0.07 0.56 0.51 0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.10 0.11 0.49 0.30 0.79 0.06 0.09 0.43 0.41 0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.06 0.10 0.54 0.30 0.84 0.07 0.09 0.34 0.50 0.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.16 0.09 0.53 0.27 0.80 0.00 0.00 0.54 0.50 0.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.15 0.10 0.39 0.42 0.81 0.09 0.06 0.39 0.55 0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.14 0.12 0.56 0.18 0.74 0.06 0.06 0.33 0.54 0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.11 0.10 0.51 0.29 0.79 0.09 0.06 0.44 0.42 0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.10 0.12 0.43 0.35 0.78 0.09 0.11 0.39 0.42 0.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.12 0.13 0.44 0.31 0.75 0.03 0.08 0.40 0.40 0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.14 0.14 0.46 0.34 0.74 0.04 0.13 0.46 0.36 0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.19 0.21 0.41 0.28 0.69 0.09 0.20 0.43 0.34 0.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.19 0.18 0.44 0.21 0.65 0.11 0.16 0.53 0.39 0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.15 0.18 0.34 0.36 0.70 0.17 0.10 0.24 0.51 0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.11 0.20 0.46 0.24 0.69 0.14 0.12 0.32 0.43 0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.16 0.18 0.42 0.20 0.70 0.18 0.13 0.33 0.36 0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.14 0.18 0.41 0.27 0.68 0.13 0.13 0.40 0.44 0.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.16 0.17 0.39 0.27 0.67 0.23 0.16 0.23 0.36 0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0.22 0.19 0.38 0.21 0.60 0.19 0.19 0.32 0.34 0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.27 0.22 0.35 0.22 0.57 0.19 0.16 0.30 0.36 0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.26 0.17 0.30 0.24 0.55 0.24 0.13 0.30 0.29 0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.19 0.24 0.31 0.26 0.57 0.17 0.16 0.29 0.39 0.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.22 0.27 0.28 0.29 0.57 0.23 0.14 0.18 0.48 0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>0.28 0.26 0.31 0.23 0.54 0.21 0.31 0.23 0.38 0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>0.21 0.24 0.33 0.24 0.57 0.18 0.25 0.29 0.29 0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0.28 0.25 0.23 0.20 0.46 0.23 0.23 0.27 0.30 0.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0.31 0.24 0.26 0.19 0.46 0.23 0.23 0.29 0.34 0.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.22 0.31 0.32 0.15 0.47 0.15 0.26 0.23 0.36 0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0.31 0.25 0.24 0.20 0.44 0.26 0.25 0.23 0.26 0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.26 0.28 0.24 0.21 0.45 0.25 0.30 0.22 0.22 0.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.33 0.35 0.19 0.14 0.33 0.32 0.23 0.18 0.26 0.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>0.34 0.40 0.17 0.15 0.32 0.30 0.24 0.18 0.28 0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>0.18 0.19 0.38 0.27 0.65 0.56 0.13 0.08 0.04 0.33 0.57 0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.014 0.014 0.019 0.014 0.027 0.029 0.019 0.014 0.014 0.019 0.027 0.029 0.019</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>coil</th>
<th>n-Helix</th>
<th>β-Sheet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A₁</td>
<td>T₁</td>
</tr>
<tr>
<td>M</td>
<td>0.18</td>
<td>0.19</td>
</tr>
<tr>
<td>SE</td>
<td>0.014 0.014 0.019 0.014 0.027 0.029 0.019</td>
<td></td>
</tr>
</tbody>
</table>

Table 6
Statistical values of pairwise comparisons

<table>
<thead>
<tr>
<th>coil</th>
<th>n-Helix</th>
<th>β-Sheet</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>6.7 x 10⁻⁵</td>
<td>1.8 x 10⁻⁵</td>
</tr>
<tr>
<td>SYN</td>
<td>4.6 x 10⁻⁷</td>
<td>6.4 x 10⁻⁷</td>
</tr>
<tr>
<td>C₁</td>
<td>9.8 x 10⁻⁷</td>
<td>9.8 x 10⁻⁷</td>
</tr>
<tr>
<td>GC₁</td>
<td>2.3 x 10⁻¹</td>
<td>n.s.</td>
</tr>
<tr>
<td>T₁</td>
<td>5.3 x 10⁻⁹</td>
<td>1.5 x 10⁻⁹</td>
</tr>
<tr>
<td>GC₁</td>
<td>9.0 x 10⁻⁹</td>
<td>2.0 x 10⁻⁹</td>
</tr>
</tbody>
</table>

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 n-Helix are
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
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 (G3 and T3 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. 2. Base frequencies in third codon positions of coding sequence regions corresponding to different predicted protein structures](image-url)
Fig. 3. $C_3$ and $C_{pG_1}$, as well as $G_3$ and $C_{pG_3}$, 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.

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


Craveiller, S., Jabbari, K., D’Onofrio, G., Bernardi, G., 1999. Different
hydrophobicities of orthologous proteins from *Xenopus* and man. Gene 238, 15–21.


