J Mol Evol (1989) 28:286–298

B 14

Journal of Molecular Evolution

Nucleic Acid Composition, Codon Usage, and the Rate of Synonymous Substitution in Protein-Coding Genes

Aharon Ticher and Dan Graur

Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel

Summary. Based on the rates of synonymous substitution in 42 protein-coding gene pairs from rat and human, a correlation is shown to exist between the frequency of the nucleotides in all positions of the codon and the synonymous substitution rate. The correlation coefficients were positive for A and T and negative for C and G. This means that AT-rich genes accumulate more synonymous substitutions than GC-rich genes. Biased patterns of mutation could not account for this phenomenon. Thus, the variation in synonymous substitution rates and the resulting unequal codon usage must be the consequence of selection against A and T in synonymous positions. Most of the variation in rates of synonymous substitution can be explained by the nucleotide composition in synonymous positions. Codon–anticodon interactions, dinucleotide frequencies, and contextual factors influence neither the rates of synonymous substitution nor codon usage. Interestingly, the nucleotide in the second position of codons (always a nonsynonymous position) was found to affect the rate of synonymous substitution. This finding links the rate of nonsynonymous substitution with the synonymous rate. Consequently, highly conservative proteins are expected to be encoded by genes that evolve slowly in terms of synonymous substitutions, and are consequently highly biased in their codon usage.

Key words: Rate of synonymous substitution — Codon usage — Nucleotide composition — Codon–anticodon interaction

Offprint requests to: D. Graur

Introduction

The rate of synonymous substitutions is generally much higher than that of amino acid replacements (Kimura 1977; Jukes and King 1979; Miyata et al. 1980; Li et al. 1985a; Sharp and Li 1987a), but lower than the rate of substitution in pseudogenes (Li et al. 1981; Miyata and Yasunaga 1981). This means that synonymous substitutions are not completely free of selective constraint, and that some selection operates against synonymous changes (Miyata and Hayashida 1981). It has been suggested (Miyata et al. 1980) that the rate of synonymous substitution is similar among genes. More recently, however, Li et al. (1985a) found that the rate of synonymous substitution varies greatly from gene to gene, albeit not nearly as much as the nonsynonymous rate. They concluded that this variation is too large to reflect random fluctuation only.

According to the neutral theory of molecular evolution (Kimura 1983), biased codon usage and diminished rates of synonymous substitution are both results of purifying selection. Indeed, Sharp and Li (1987a) found a negative correlation between the bias in codon usage and the rate of synonymous substitution. This result indicates that not only are certain codons from among each isoacceptor group selected against, but that the magnitude of selection varies between genes. In the literature (see review by Li et al. 1985a) there have been suggestions that the rate of synonymous substitution, and consequently the pattern of codon usage, are affected by such factors as tRNA availability (Ikemura 1980, 1981a,b, 1982, 1985; Ikemura and Ozeki 1983; Sharp and Li 1987b), the strength of the hydrogen bond between the codon and its tRNA anticodon

(Grosjean et al. 1978; Grosjean and Fiers 1982), contextual constraints (Lipman and Wilbur 1983), dinucleotide preferences (Nussinov 1981), and overall GC content (Nichols et al. 1980; Yanofsky and van Cleemput 1982). Eight rules affecting codon usage are listed in Ikemura (1985). The rules, however, are not applicable to either all genes (Ikemura and Ozeki 1983) or all organisms (Chen et al. 1986). One of the aims of this study was to check some of these hypotheses and to look for other factors that may underlie the pattern and rate of synonymous substitutions in different genes.

Several findings in the literature suggest that a connection exists between the rate of synonymous substitution and the rate of amino acid replacement. In particular, the rate of synonymous substitution is positively correlated with the nonsynonymous rate (Graur 1985). In addition, poorly conserved regions at the protein level were found to have less biased codon usages than conserved regions (Lipman and Wilbur 1983). In this study we shall look for compositional factors that might explain the connection between synonymous rates and biased codon usage, on the one hand, and nonsynonymous rates and constraints at the amino acid level, on the other. Because Graur (1985) found that nonsynonymous substitution rates are influenced by compositional factors, we investigate here compositional parameters that might either predict or influence the synonymous substitution rate. By implication, we look for the level of genetic information transfer (e.g., replication, transcription, or translation) at which selection against synonymous changes is exerted.

Data and Methods

Nucleotide sequence data were obtained either from the GenBank library or directly from the literature. Forty-two complete human and rat protein-coding genes were selected. Introns and the initiation and termination codons were excluded from the analysis. Also excluded were gaps in the sequences that came out of alignment by the algorithm of Wilbur and Lipman (1985).

Synonymous and nonsynonymous substitution rates (K_s and K_a) were calculated according to Li et al. (1985b). We note that the results reported here are not affected by the choice of method for estimating K_s and K_a. Because we do not want to assume any particular time in regard to the rat–human divergence event, the rates of substitution are expressed as nucleotide substitutions per site for the combined rat–human lineages. By using only rat–human comparisons we also avoid complications related to the fact that DNA sequences evolve at different rates in different organisms (Wu and Li 1985; Li and Tanimura 1987; Li and Wu 1987; Li et al. 1987; Graur et al. 1988). The genes and their sizes and rates of synonymous and nonsynonymous substitution are listed in Table 1, together with the literature sources.

The following compositional parameters were calculated for each of the genes: (1) The frequencies of the four nucleotides according to codon position. These were denoted as f(m, i), where m stands for the nucleotide (m = A, C, G, T), and i for the codon position (i = 1, 2, 3). For example, f(A, 2) is the frequency of A in the second position. The values of f(m, i) for the 42 genes are shown in Table 2. Table 2 also contains the frequencies of the four nucleotides pooled over all positions. (2) Dinucleotide frequencies, denoted as f(d, j), where d stands for the dinucleotide (d = ApA, ApT, ApC, . . .), and j for the location relative to the reading frame (j = 1p2, 2p3, 3p1), and (3) trinucleotide frequencies, denoted as f(t, k), where t stands for the trinucleotide (t = ApApA, ApApT, ApApC, . . .) and k represents the frame of reference (k = 1p2p3, 2p3p1, 3p1p2). For reasons of brevity we do not present the dinucleotide and trinucleotide data.

Correlation coefficients (r) were calculated between K_s and f(m, i), f(d, j), and f(t, k).

Results

The Rate of Synonymous Substitution

The rate of synonymous substitution was found to vary in the present sample of genes by a factor of about 4, from 0.289 substitutions/site in metallothionein I, to 1.074 in serum albumin. The mean rate was 0.643 substitutions per synonymous site in the combined human and rat lineages (Table 1). In comparison, the nonsynonymous substitution rate in the same sample varies over a range of about 350. A positive correlation exists between K_s and K_a (r = 0.513), similar to Graur (1985).

Nucleotide Frequencies

From Table 2 we see that the variation in nucleotide usage in each of the codon positions is quite large. The largest variation in nucleotide usage was seen in the third position, e.g., for f(C, 3) the maximum value was 0.725 and the minimum was 0.184, spanning a range of 0.541. By using the methods of Wright (1969, p. 26) and Tajima and Nei (1982), the equilibrium frequencies of the four nucleotides due to mutational biases were calculated by Gojobori et al. (1982) and Li et al. (1984). They found that the equilibrium frequencies for A, T, C, and G were 0.34, 0.34, 0.16, and 0.16, respectively. Interestingly, the frequencies of A and T in the third position are smaller than 0.34, and the frequencies of C and G are larger than 0.16. The only exceptions are α-amylase (39.5% T), metallothionein II (14.2% G), and metallothionein I (15.0% G), but these deviations from the equilibrium frequencies were not statistically significant. This finding repudiates the notion that consistent mutational biases throughout the genome may be responsible for the variation in K_s.

We found significant correlation coefficients between the frequencies of the four nucleotides in the different positions of codons and the synonymous substitution rates (Table 3). Whenever statistically significant, the correlation coefficients were positive for A and T, and negative for C and G. Thus, we

Table 1. Rates of nucleotide substitution in various genes

Gene no.*	Genes	Gene size	Synonymous rate	Nonsynonymous rate
1	Serum albumin	1821	1.074	0.149
2	Prolactin	672	1.018	0.221
3	Relaxin	471	0.943	0.326
4	Thyrotropin β	411	0.927	0.078
5	α-fibrinogen	1305	0.908	0.097
6	α-fetoprotein	1842	0.894	0.204
7	Parathyroid hormone	342	0.871	0.160
8	Ornithine aminotransferase	1314	0.852	0.048
9	Lactate dehydrogenase	993	0.837	0.036
10	α-lactalbumin	423	0.833	0.177
11	Glycoprotein hormone α	345	0.802	0.153
12	β-tubulin	1324	0.761	0.029
13	Apolipoprotein A-II	231	0.727	0.363
14	Apolipoprotein A-IV	1113	0.719	0.232
15	Growth hormone	642	0.705	0.201
16	Apolipoprotein A-I	774	0.694	0.249
17	Insulin	327	0.691	0.102
18	γ-crystallin	519	0.660	0.162
19	α-amylase	1059	0.659	0.097
20	α-actin (cardiac)	1128	0.591	0.001
21	ATPase β	906	0.576	0.028
22	*Thy*-1 antigen	480	0.573	0.199
23	Aldolase A	1089	0.572	0.013
24	Transthyretin	438	0.571	0.115
25	Atrial natriuretic factor	447	0.569	0.093
26	Apolipoprotein E	925	0.564	0.179
27	Creatine kinase M	1140	0.549	0.027
28	Apoferritin (light subunit)	447	0.547	0.078
29	Glucagon	534	0.534	0.041
30	Proopiomelanocortin	687	0.529	0.115
31	Aldolase B	540	0.526	0.030
32	Lutenizing hormone β	420	0.526	0.164
33	Cholecystokinin	342	0.523	0.131
34	Glyceraldehyde-3-P-dehydrogenase	996	0.477	0.033
35	β-actin (cytoplasmic)	1047	0.451	0.004
36	Somatostatin	345	0.449	0.016
37	Liver glycogen phosphorylase	243	0.436	0.034
38	Insulin like growth factor II	537	0.409	0.077
39	Oxytocin-neurophysin I	372	0.405	0.055
40	α-tubulin	435	0.388	0.006
41	Metallothionein II	180	0.367	0.052
42	Metallothionein I	180	0.289	0.021

* References: (1) H (human): Dugaiczyk et al. 1982; R (rat): Sargent et al. 1981; (2) H: Truong et al. 1984; R: Cooke and Baxter 1982; (3) H: Hudson et al. 1984; R: Hudson et al. 1981; (4) H: Hayashizaki et al. 1985; R: Godine et al. 1982; (5) H: Rixon et al. 1983; R: Crabtree et al. 1985; (6) H: Morinaga et al. 1983; R: Jagodzinski et al. 1981; (7) H: Vasicek et al. 1983; R: Heinrich et al. 1984; (8) H: Inana et al. 1986; R: Mueckler and Pitot 1985; (9) H: Tsujibo et al. 1985; R: Li et al. 1983; (10) H: Hall et al. 1982; R: Qasba and Safaya 1984; (1) H: Fiddes and Goodman 1981; R: Godine et al. 1982; (12) H: Wilde et al. 1982; R: Sullivan et al. 1984; (13) H: Luo et al. 1986; R: Luo et al. 1986; (14) H: Karathanasis 1985; R: Boguski et al. 1985; (15) H: Roskam and Rougeon 1979; R: Page et al. 1981; (16) H: Law and Brewer 1984; R: Boguski et al. 1985; (17) H: Ullrich et al. 1980; R: Soares et al. 1985; (18) H: Meakin et al. 1985; R: den Dunnen et al. 1986; (19) H: Nakamura 1984; R: MacDonald et al. 1980; (20) H: Hamada et al. 1982; R: Mayer et al. 1984; (21) H: Kawakami et al. 1986; R: Young et al. 1987; (22) H: van Rijs et al. 1985; R: Moriuchi et al. 1982; (23) H: Sakakibara et al. 1985; R: Joh et al. 1985; (24) H: Wallace et al. 1985; R: Sundelin et al. 1985; (25) H: Greenberg et al. 1984; R: Argentin et al. 1985; (26) H: McLean et al. 1984; R: McLean et al. 1983; (27) H: Perryman et al. 1986; R: Benfield et al. 1984; (28) H: Santoro et al. 1986; R: Leibold and Munro 1987; (29) H: White and Saunders 1986; R: Lopez et al. 1983; (30) H: Chang et al. 1980; R: Drouin and Goodman 1980; (31) H: Rottmann et al. 1984; R: Tsutsumi et al. 1983; (32) H: Chin et al. 1983; R: Chin et al. 1983; (33) H: Takahashi et al. 1985; R: Deschenes et al. 1985; (34) H: Hanauer and Mandel 1984; R: Fort et al. 1985; (35) H: Ng et al. 1985; R: Nudel et al. 1983; (36) H: Shen et al. 1982; R: Argos et al. 1983; (37) H: Newgard et al. 1986; R: Osawa et al. 1986; (38) H: Graeme et al. 1984; R: Dull et al. 1984; (39) H: Sausville et al. 1985; R: Heinrich et al. 1984; (40) H: Cowan et al. 1983; R: Ginzburg et al. 1981; (41) H: Karin and Richards 1982; R: Andersen et al. 1986; (42) H: Varshney and Gedamu 1984; R: Andersen et al. 1983

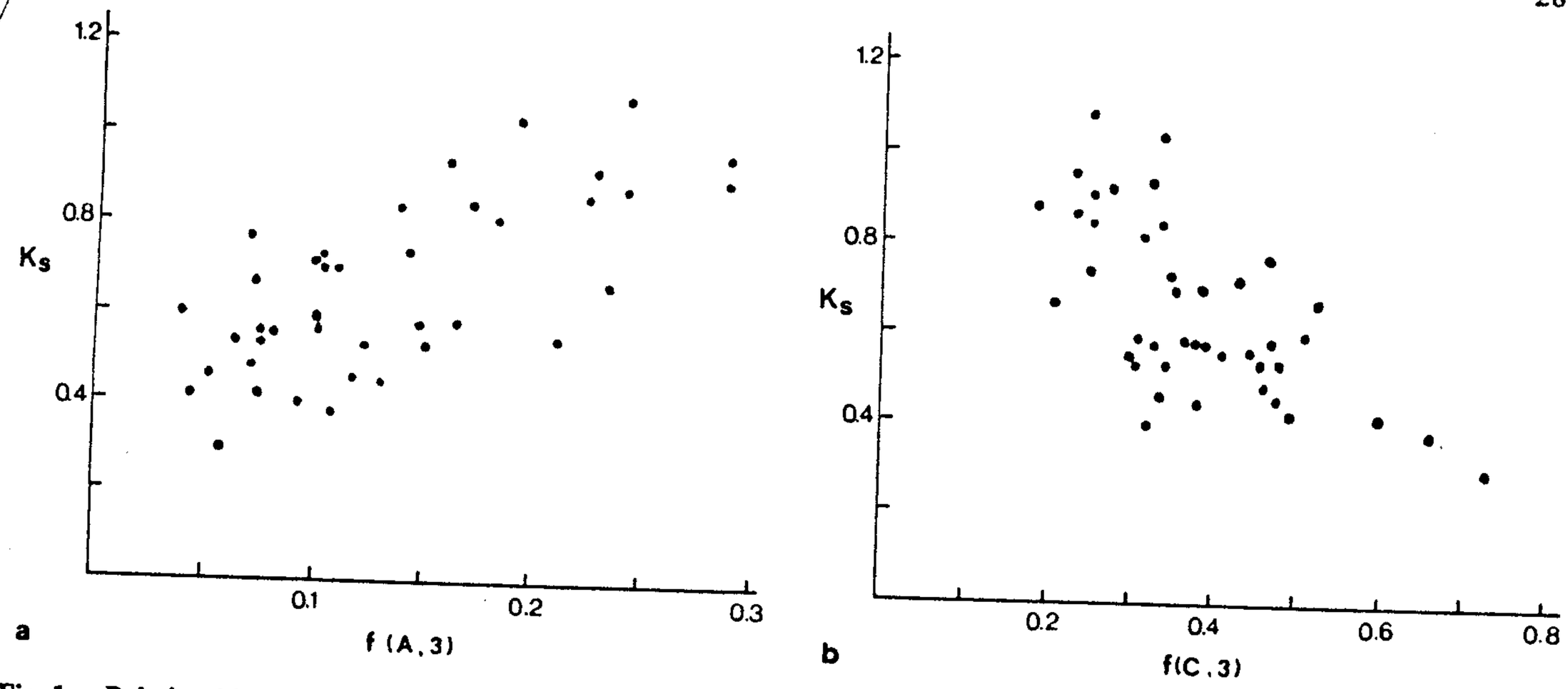

Fig. 1. Relationship between frequencies f(A, 3) **a** and f(C, 3) **b** and rates of synonymous substitutions (K_s)

conclude that AT-rich genes accumulate more synonymous substitutions than GC-rich genes. The highest absolute values were found for A (Fig. 1a), C (Fig. 1b), and T in the third position (0.696, −0.652, and 0.490, respectively), and for G in the second position (−0.490). All four nucleotides were correlated significantly with the synonymous substitution rates only in the second position of codons. Despite the fact that the second position is always nonsynonymous, we find that it greatly affects the rate of synonymous substitution. The overall nucleotide composition of genes was also correlated with K_s.

Table 4 shows how much of the variation in K_s can be explained by the different factors of nucleotide frequency composition. We found, for instance, that by using f(A, 3) and f(G, 3), we could explain more than 52% of the variation in the synonymous substitution rate. Similarly, by using f(A, 3) and f(A, 2), we could explain about 54% of the variation. The conclusions from these results are: (1) that the nucleic acid composition at the third position predicts the rate of synonymous substitution better than the second, and the second position better than the first, and (2) that in terms of nucleotide composition in the three positions of codons, A explained more of the variation than C, C more than G, and G more than T. Again, we see that the second (nonsynonymous) position influences the rate of synonymous substitution more than the first position (~5% synonymous).

Dinucleotide and Trinucleotide Frequencies

In Table 5, we list the statistically significant correlation coefficients between the frequencies of the 16 dinucleotides in the different positions of codons and the rate of synonymous substitution. We find that the dinucleotides behave according to the rules established for the mononucleotides. For example, if we look at f(CpA, 2p3), the correlation coefficient is positive (0.588). On the other hand, for f(CpA, 3p1), r is negative (−0.376). This is expected because the third position was shown to have a greater influence on the rate of synonymous substitution than other codon positions. The fact that dinucleotide frequencies do not affect K_s independently of their mononucleotide constituents rules out any explanation (e.g., Nussinov 1981) that attributes importance to dinucleotide preferences in the determination of the rate of synonymous substitution.

The absence of G in the third position (Tables 3 and 5) clearly stands out. Furthermore, when A, C, or T is in the third position, the r values that were obtained were statistically significant, regardless of the nucleotide in the first position of the next codon. The only exception was f(TpG, 3p1). This means that the synonymous substitution rate is not affected by interactions between neighboring codons. This finding agrees with Ayer and Yarus (1986). We cannot, however, rule out more distant effects (Shpaer 1986). By using a multiple regression equation with the frequencies of six dinucleotides in codon positions 2 and 3 we can explain up to 69% of the variation in K_s (Appendix 1).

In Table 6 we show the statistically significant correlation coefficients between the frequencies of the 64 trinucleotides in the reading frame and the rate of synonymous substitution. We find that the trinucleotides also obey the rules established for the mononucleotides. Interestingly, by using the frequency of only one triplet, TCC (serine), we can already explain about 38% of the variation in the rate of synonymous substitution. Figure 2 shows the

Table 2. The frequencies of the four nucleotides in different codon positions

Gene no.*	Position 1				Position 2				Position 3			
	A	C	G	T	A	C	G	T	A	C	G	T
1	0.250	0.198	0.348	0.204	0.374	0.231	0.146	0.250	0.242	0.249	0.231	0.278
2	0.263	0.292	0.268	0.176	0.353	0.192	0.147	0.308	0.192	0.333	0.243	0.232
3	0.258	0.226	0.315	0.201	0.319	0.245	0.198	0.239	0.287	0.229	0.242	0.242
4	0.296	0.171	0.252	0.281	0.285	0.252	0.197	0.266	0.161	0.321	0.183	0.336
5	0.318	0.171	0.294	0.216	0.378	0.208	0.189	0.225	0.228	0.271	0.208	0.293
6	0.287	0.206	0.300	0.207	0.358	0.219	0.158	0.265	0.286	0.249	0.219	0.246
7	0.338	0.202	0.364	0.097	0.355	0.162	0.167	0.316	0.241	0.184	0.276	0.298
8	0.266	0.204	0.365	0.164	0.300	0.232	0.176	0.292	0.225	0.232	0.220	0.323
9	0.311	0.205	0.347	0.136	0.317	0.165	0.186	0.332	0.171	0.249	0.293	0.287
10	0.312	0.177	0.287	0.223	0.351	0.163	0.188	0.298	0.138	0.333	0.273	0.255
11	0.291	0.187	0.243	0.278	0.304	0.274	0.178	0.243	0.183	0.309	0.261	0.248
12	0.270	0.229	0.352	0.149	0.326	0.210	0.187	0.276	0.070	0.459	0.377	0.094
13	0.292	0.240	0.305	0.162	0.409	0.240	0.104	0.247	0.143	0.247	0.416	0.195
14	0.266	0.327	0.330	0.077	0.447	0.171	0.121	0.260	0.102	0.344	0.460	0.094
15	0.238	0.299	0.269	0.194	0.315	0.220	0.173	0.292	0.101	0.425	0.322	0.152
16	0.244	0.291	0.359	0.107	0.407	0.178	0.157	0.258	0.103	0.349	0.430	0.118
17	0.119	0.367	0.339	0.174	0.275	0.197	0.229	0.298	0.110	0.381	0.381	0.128
18	0.211	0.292	0.254	0.243	0.353	0.127	0.289	0.231	0.072	0.520	0.306	0.101
19	0.311	0.154	0.334	0.201	0.326	0.181	0.240	0.254	0.234	0.205	0.167	0.395
20	0.295	0.210	0.340	0.154	0.313	0.250	0.160	0.278	0.040	0.505	0.327	0.128
21	0.306	0.215	0.290	0.189	0.384	0.149	0.197	0.270	0.164	0.306	0.288	0.242
22	0.306	0.306	0.219	0.169	0.278	0.219	0.172	0.331	0.100	0.463	0.294	0.144
23	0.255	0.252	0.362	0.131	0.321	0.266	0.174	0.240	0.101	0.369	0.313	0.218
24	0.223	0.202	0.401	0.175	0.264	0.301	0.171	0.264	0.164	0.363	0.229	0.243
25	0.238	0.265	0.336	0.161	0.235	0.245	0.255	0.265	0.148	0.383	0.299	0.171
26	0.168	0.355	0.393	0.085	0.337	0.200	0.217	0.246	0.101	0.324	0.510	0.065
27	0.267	0.264	0.341	0.128	0.379	0.170	0.172	0.279	0.075	0.443	0.359	0.122
28	0.201	0.282	0.356	0.161	0.389	0.188	0.141	0.282	0.080	0.406	0.332	0.181
29	0.301	0.191	0.337	0.171	0.365	0.205	0.194	0.236	0.211	0.295	0.230	0.264
30	0.251	0.264	0.325	0.159	0.314	0.214	0.271	0.201	0.063	0.452	0.415	0.070
31	0.242	0.231	0.367	0.161	0.303	0.325	0.147	0.225	0.150	0.300	0.239	0.311
32	0.171	0.368	0.275	0.186	0.136	0.271	0.282	0.311	0.075	0.475	0.286	0.164
33	0.180	0.325	0.382	0.114	0.259	0.246	0.254	0.241	0.123	0.338	0.412	0.127
34	0.310	0.154	0.392	0.145	0.312	0.244	0.157	0.288	0.071	0.456	0.249	0.224
35	0.295	0.229	0.318	0.158	0.301	0.252	0.161	0.287	0.052	0.471	0.317	0.161
36	0.187	0.291	0.330	0.191	0.296	0.317	0.152	0.235	0.117	0.330	0.387	0.165
37	0.309	0.222	0.296	0.173	0.358	0.198	0.185	0.259	0.130	0.377	0.296	0.198
38	0.204	0.293	0.310	0.193	0.215	0.260	0.268	0.257	0.075	0.486	0.299	0.140
39	0.145	0.274	0.351	0.230	0.202	0.258	0.363	0.177	0.044	0.597	0.286	0.073
40	0.214	0.162	0.472	0.152	0.352	0.203	0.197	0.248	0.093	0.317	0.307	0.283
41	0.233	0.058	0.258	0.450	0.242	0.283	0.442	0.033	0.108	0.658	0.142	0.092
42	0.258	0.050	0.217	0.475	0.200	0.300	0.467	0.033	0.058	0.725	0.150	0.067
Mean	0.255	0.236	0.324	0.186	0.317	0.225	0.205	0.253	0.134	0.375	0.297	0.194
SD	0.052	0.071	0.052	0.077	0.063	0.047	0.075	0.059	0.067	0.119	0.083	0.085
Maximum	0.338	0.368	0.472	0.475	0.447	0.325	0.467	0.332	0.287	0.725	0.510	0.395
Minimum	0.119	0.050	0.217	0.077	0.136	0.127	0.104	0.033	0.040	0.184	0.142	0.065
Range	0.219	0.318	0.255	0.398	0.311	0.198	0.363	0.299	0.247	0.541	0.368	0.330

* The names of the genes are listed in Table 1

percentage of variation in K_s that is explained by progressively increasing the number of trinucleotides in a stepwise multiple regression analysis. By using the frequencies of 9 codons, we can explain in a statistically significant manner up to 88% of the variation in K_s (Appendix 2). Again, we find no evidence for interaction between codons in the determination of synonymous substitution rates. The data for the two nonreading frames (not shown) indicate that whatever effects there are of the trinucleotide frequencies on K_s, these are restricted to the reading frame.

Discrepancy Tests

In order to separate the effects of position and composition, and to ascertain the hierarchy of effects on synonymous substitution rates, we constructed discrepancy tables for the trinucleotides in the different frames. Following our findings concerning the

Table 2. Extended

All positions			
A	C	G	T
0.289	0.226	0.242	0.244
0.269	0.272	0.220	0.239
0.288	0.234	0.252	0.227
0.247	0.248	0.211	0.294
0.308	0.217	0.230	0.245
0.310	0.225	0.225	0.239
0.311	0.183	0.269	0.237
0.264	0.223	0.254	0.260
0.266	0.206	0.275	0.252
0.267	0.225	0.249	0.259
0.259	0.257	0.228	0.257
0.222	0.299	0.306	0.173
0.281	0.242	0.275	0.201
0.272	0.281	0.304	0.144
0.218	0.315	0.255	0.213
0.251	0.273	0.315	0.161
0.168	0.315	0.317	0.200
0.212	0.313	0.283	0.192
0.290	0.180	0.247	0.283
0.216	0.322	0.276	0.187
0.285	0.223	0.258	0.233
0.228	0.329	0.228	0.215
0.225	0.296	0.283	0.196
0.217	0.289	0.267	0.227
0.207	0.298	0.296	0.199
0.202	0.293	0.373	0.132
0.240	0.292	0.291	0.176
0.224	0.292	0.276	0.208
0.292	0.230	0.254	0.224
0.210	0.310	0.337	0.143
0.231	0.285	0.251	0.232
0.127	0.371	0.281	0.220
0.187	0.303	0.349	0.161
0.231	0.285	0.266	0.219
0.216	0.318	0.265	0.201
0.200	0.313	0.290	0.197
0.265	0.265	0.259	0.210
0.165	0.346	0.292	0.197
0.130	0.376	0.333	0.160
0.220	0.228	0.325	0.228
0.194	0.333	0.281	0.192
0.172	0.358	0.278	0.192
0.235	0.278	0.275	0.211
0.046	0.049	0.036	0.037
0.311	0.376	0.373	0.294
0.127	0.180	0.211	0.132
0.184	0.196	0.182	0.162

mononucleotides, we assumed that C and G are negatively correlated with K_s, whereas A and T are positively correlated. Discrepancies between the expected effect of a nucleotide at a given position were recorded. Thus, for instance, if we observe that f(CpGpT, 1p2p3) is negatively correlated with K_s, we assign a plus (+) sign in the first position, a plus (+) in the second position, and a minus (−) in the third position, i.e., the sign of the correlation coefficient is as expected given the nucleotides occu-

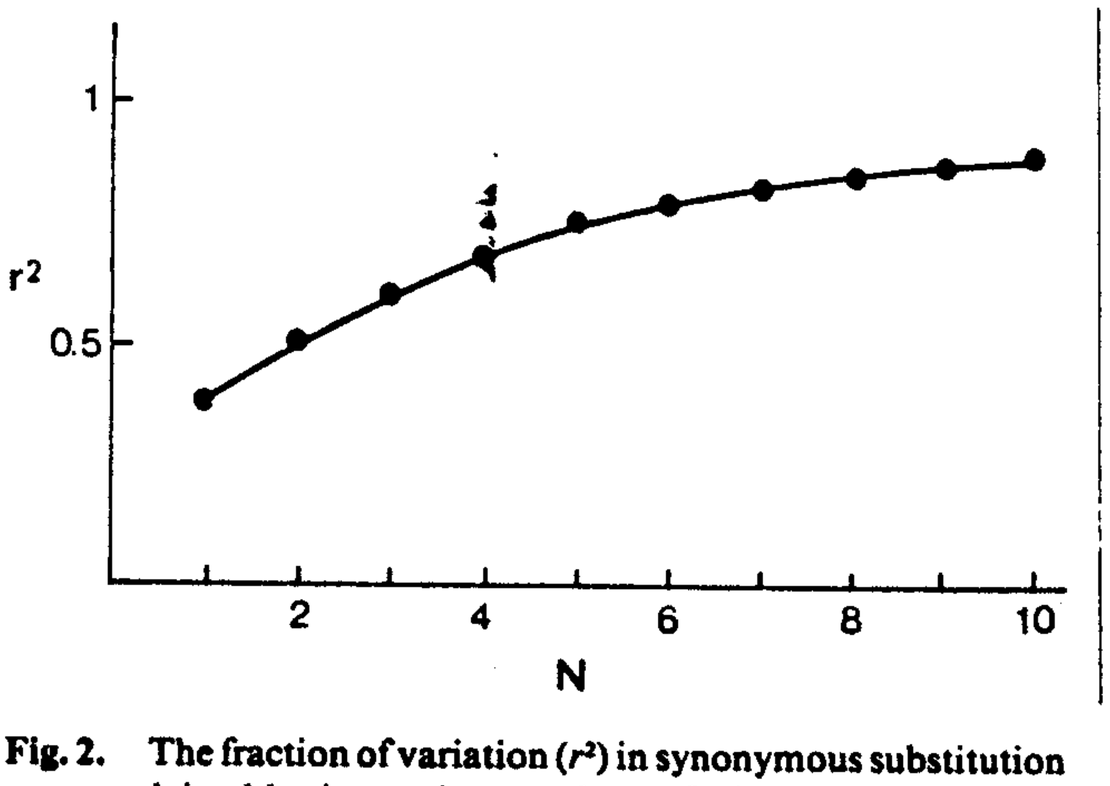

Fig. 2. The fraction of variation (r^2) in synonymous substitution rates explained by increasing numbers of trinucleotide frequencies (N)

Table 3. Statistically significant correlation coefficients between frequencies of nucleotides and the synonymous substitution rate

Nucleotide	Position of codon			
	1	2	3	1 + 2 + 3
T	ns*	0.442**	0.490***	0.494***
C	ns	−0.347*	−0.652***	−0.622***
A	0.355*	0.427**	0.696***	0.663***
G	ns	−0.490***	ns	−0.503***

* ns = $P > 0.05$; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$

Table 4. The variation in synonymous substitution rates that can be explained by A) nucleotide composition in each codon position and B) frequencies of each nucleotide in the three codon positions

A) Position	Variable	Simple *r*	Multiple *r*	r^2	*P*
i = 1	f(A, 1)	0.355	0.355	0.126	0.021
	f(C, 1)	0.025	0.443	0.196	0.014
i = 2	f(G, 2)	−0.490	0.489	0.239	0.001
	f(C, 2)	−0.346	0.544	0.296	0.001
i = 3	f(A, 3)	0.696	0.696	0.484	0.000
	f(G, 3)	−0.128	0.724	0.524	0.000
i = 1 + 2 + 3	f(A, all)	0.663	0.663	0.440	0.000
	f(T, all)	0.494	0.688	0.474	0.000
B) Nucleotide	**Variable**	**Simple *r***	**Multiple *r***	**r^2**	***P***
m = A	f(A, 3)	0.696	0.696	0.484	0.000
	f(A, 2)	0.427	0.735	0.540	0.000
m = C	f(C, 3)	−0.652	0.652	0.425	0.000
	f(C, 2)	−0.347	0.673	0.452	0.000
m = G	f(G, 2)	−0.490	0.489	0.239	0.001
	f(G, 3)	−0.128	0.575	0.330	0.000
	f(G, 1)	−0.131	0.618	0.382	0.000
m = T	f(T, 3)	0.490	0.490	0.241	0.001
	f(T, 2)	0.442	0.571	0.326	0.000
	f(T, 1)	−0.174	0.589	0.346	0.001

Table 5. Statistically significant correlation coefficients between frequencies of dinucleotides and the synonymous substitution rate

Dinucleotide	Position relative to codon		
	12*	*23	**3,1**
TT	0.456***a	0.581***	0.570***
TC	ns	ns	0.549***
TA	ns	0.663***	0.587***
TG	−0.314*	ns	ns
CT	ns	ns	−0.449**
CC	ns	−0.688***	−0.362*
CA	0.378*	0.588***	−0.376*
CG	−0.357*	ns	−0.393**
AT	ns	0.553***	0.419**
AC	ns	ns	0.582***
AA	ns	0.575***	0.707***
AG	ns	ns	0.490***
GT	ns	ns	ns
GC	ns	−0.547***	ns
GA	ns	0.428**	ns
GG	−0.420**	ns	ns

[a] ns = $P > 0.05$; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$

pying the first and second positions, but opposite to what is expected from the nucleotide in position 3. A 0% value would mean no discrepancy between the observations and expectations, 50% would mean a random pattern, i.e., no compositional effect at this position, and 100% would mean that the rules governing mononucleotide effects on the rate of synonymous substitution are exactly the opposite to what had been assumed in the premises of the analysis. The percentages of discrepancy for each of the positions, according to the reading frame, are shown in Table 7. The results are essentially the same regardless of the frame used, i.e., the effects on K_s are determined by the reading frame. Thus, we may conclude that whatever constraints there are that affect the rate of synonymous substitution, they are mostly exercised at the level of genetic information transfer at which the reading frame is taken into account, i.e., translation. The third position of the reading frame had the lowest discrepancy value (0.008). The first position has no effect on K_s.

Table 7. Percent discrepancy for position effects on rates of synonymous substitution in three frames

Frame	Codon position		
	1	2	3
1p2p3 (27)[a]	0.488	0.161	0.025
2p3p1 (34)	0.475	0.306	0.000
3p1p2 (28)	0.415	0.246	0.000
Mean (89)	0.460	0.243	0.008

[a] The number of codons significantly correlated with K_s (maximum possible number = 61)

Strength of the Codon–Anticodon Bond

Grosjean et al. (1978) and Grosjean and Fiers (1982) suggested that codons with an intermediate GC content, and hence with an intermediate codon–anticodon binding energy, are optimal for translation. They proposed two rules. The first was that codons in which the first two nucleotides are T and/or A (each forming two hydrogen bonds with either DNA or RNA) should stabilize their interaction with the anticodon by using a C (with three hydrogen bonds) in the wobble position. They noted, for instance, that in bacteriophage MS2 there was a systematic preference for TTC and AAC codons over TTT and AAT. For example, the ratio TTT/TTC (both coding for phenylalanine) was 21/32. In contrast, we find that in some genes, like serum albumin, the opposite situation prevails. In this gene the ratio

Table 6. Statistically significant correlation coefficients between frequencies of trinucleotides in the reading frame and the synonymous substitution rate

TTT	0.594***[a]	TCT	ns	TAT	0.463**	TGT	ns
TTC	ns	TCC	−0.408**	TAC	ns	TGC	−0.378*
TTA	0.571***	TCA	0.618***				
TTG	ns	TCG	−0.360*			TGG	ns
CTT	0.600***	CCT	ns	CAT	0.330*	CGT	−0.377*
CTC	ns	CCC	−0.444**	CAC	ns	CGC	−0.369*
CTA	0.403**	CCA	ns	CAA	0.608***	CGA	ns
CTG	ns	CCG	ns	CAG	ns	CGG	ns
ATT	0.318*	ACT	0.339*	AAT	0.411**	AGT	ns
ATC	ns	ACC	−0.382*	AAC	ns	AGC	ns
ATA	0.493***	ACA	0.455**	AAA	ns	AGA	0.453**
ATG	ns	ACG	ns	AAG	ns	AGG	ns
GTT	0.419**	GCT	ns	GAT	0.354*	GGT	ns
GTC	ns	GCC	−0.577***	GAC	ns	GGC	−0.602***
GTA	ns	GCA	0.391**	GAA	0.572***	GGA	ns
GTG	ns	GCG	ns	GAG	ns	GGG	ns

[a] ns = $P > 0.05$; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$

TTT/TTC is 55/42. In fact, a significant positive correlation exists between the ratio of TTT/TTC and K_s ($r = 0.413$). The same is true for the other WWT/WWC ratios, where W stands for A or T. The second rule is the complementary one, i.e., that in cases where the first two nucleotides of the codon are C and/or G, a preference would exist for T in the third position to avoid too strong a codon–anticodon interaction. In MS2, CCT and GGT are systematically preferred over synonymous codons ending in C. In contrast, we find that in some genes a preference exists for SSC codons over SST ones, where S stands for C or G. Again, a significant positive correlation exists between the ratio CCT/CCC and K_s ($r = 0.410$). Similar results were obtained for the other SST/SSC ratios.

The hypothesis of Grosjean and colleagues can also be tested by asking whether or not a correlation exists between the rate of synonymous substitution and the frequency of certain types of codons. According to their theory, codons with seven or eight hydrogen bonds are expected to have an optimal codon–anticodon binding energy, and should therefore be more conservative. The opposite is expected in regard to codons with either six or nine hydrogen bonds. We found no significant correlation between the frequency of medium-strength codons and K_s ($r = 0.149$). Similarly, no correlation exists between the frequencies of suboptimal codons (six and nine hydrogen bonds) that are expected to be unstable, and K_s. In contrast, codons that contain only C or G (nine or eight hydrogen bonds, $r = -0.611$ and $r = -0.591$, respectively) are conserved more than codons containing A or T (six or seven hydrogen bonds, $r = 0.589$ and $r = 0.654$, respectively). Thus, we find no evidence for selection operating against extreme codon–anticodon interaction strengths. The choice at the third position is independent of the combined strength of the hydrogen bonds between the nucleotides at the first and second positions and the anticodon.

Discussion

Because pseudogenes are presumably subject to no selective constraint, it is possible to infer the pattern of spontaneous mutation from the pattern of substitution in pseudogenes. The mutation pattern was found to be nonrandom (Gojobori et al. 1982; Li et al. 1984). This nonrandomness in the mutation pattern is expected to result in an AT-rich composition (Li et al. 1985a). Indeed, pseudogenes are known to accumulate A and T. Thus, the accumulation of C and G in the third position of codons in protein-coding genes cannot be explained by a bias in the direction of mutation (e.g., Sueoka 1988), unless the pattern of mutation is different in genes than in pseudogenes, a somewhat teleological assumption. Mutational biases can be ruled out for two additional reasons: (1) Among mutations, there is a preponderance of transitions that affect the GC content over transversions that do not. Thus, at equilibrium, GC-rich and AT-rich genes should exhibit higher substitution rates than genes with intermediate GC compositions. This is not the case. (2) A mutational bias, because it affects both strands of DNA, cannot explain the consistent asymmetry we find in this study between f(C, 3) that affects K_s, and f(G, 3) that does not. We propose that in protein-coding genes, synonymous mutations giving rise to A and T are selected against, resulting in diminished rates of synonymous substitution in comparison to the rate of substitution in pseudogenes. Selection against A and T operates on all positions, but, in the absence of superimposed functional constraints (coding for amino acids), it is especially pronounced in synonymous positions. This pattern of selection on synonymous mutations brings about an accumulation of C in the third position. The more stringent the selection, the higher levels of C in the third position. Consequently, the bias in codon usage reflects the magnitude of selection against A and T. Therefore, genes in which the rate of synonymous substitutions is low are more biased in their codon usage.

Our main finding is that the rate of synonymous substitutions can be predicted from considering the nucleic acid composition of the genes. In particular, the nucleic acid composition in the third position seems to determine the rate of synonymous substitutions. Codons ending in A and T seem to have a distinct disadvantage, and are selected against. We can thus see that selection and mutation work in opposite directions, with mutation producing A and T codon termini, and selection removing them. Consequently, every gene is at any given time at a distance from the optimal situation, i.e., containing a preponderance of C in the third position. The larger the distance from this optimum, the higher the proportion of neutral or advantageous mutations, and consequently the faster the substitution rate. Much of the variation in rates of synonymous substitution can be explained by compositional factors, without recourse to external factors, such as degree of expressivity and adaptation to tRNA frequencies.

Interestingly, the strength of selection against certain synonymous changes seems to be influenced by the nucleic acid composition in the second, nonsynonymous site, in particular by the frequencies of C and G in this position. Thus, the composition in the third position will depend on the nucleotide in the second position. We find, for instance, that the frequency of G in the second position is strongly

correlated with the frequency of C in the third (r = 0.695). Indeed, in all six amino acids with fourfold codon degeneracy that contain either C and G in the second position (serine^{TCN}, proline, threonine, alanine, arginine^{CGN}, and glycine) the most common codon is the one ending with C. In comparison, the two amino acids with fourfold codon degeneracy that contain T in the second position (leucine^{CTN} and valine) use mostly codons ending in G (data from Aota et al. 1988). Moreover, because C and G in the second position encode for some of the most conservative amino acids (e.g., glycine, proline, cysteine, and tryptophan, for a discussion see: Graur 1985), and are thus highly immutable, the effect of the second position on the synonymous substitution rate is such that conserved proteins are expected to also exhibit lower rates of synonymous substitution and consequently high biases toward codons ending in C. This dependency can explain several findings that were previously considered as disparate phenomena: (1) the positive correlation between synonymous and nonsynonymous rates (Graur 1985), (2) the fact that highly expressed genes that are probably subject to stringent selection at the amino acid level are highly biased in codon usage (Bennetzen and Hall 1982; Gouy and Gautier 1982), and (3) that poorly conserved regions at the protein level have less biased codon usages than the conserved regions (Lipman and Wilbur 1983).

Let us now discuss the level at which selection might operate against synonymous mutations. In principle, selection against certain codons can be exerted at one or more of the three levels of information transfer, i.e., DNA replication, DNA to RNA transcription, and translation. There are indications that codon usage may be influenced by factors that are independent of translation. For instance, the nucleotide composition in the third position of codons is linearly dependent on the nucleic acid composition of flanking regions and introns (Aota and Ikemura 1986; Mouchiroud 1986; Bulmer 1987a). At the level of DNA–DNA replication there are some clues indicating that GC-rich segments of DNA are subject to more errors in replication than AT-rich regions (Chang 1973; Loeb and Kunkel 1982; Modrich 1987). Similarly, there are fewer errors in the transcription of A and T than of G and C. However, if reducing the number of errors in replication and transcription would have been important in the evolution of genes, selection should have favored A and T in synonymous positions. In reality the opposite occurs. This leaves us with translation as the level at which most purifying selection against synonymous changes is exerted. Indeed, most of the selection against A and T is restricted to the reading frame and influenced by compositional factors that are frame-dependent.

One of the major factors implicated in codon usage bias is tRNA abundance. Various authors have recognized the mechanistical disadvantages of using a codon that is recognized only by a rare tRNA cognate (Zuckerkandl 1965; Zuckerkandl and Pauling 1965; Varenne et al. 1984). However, Bulmer (1987b, 1988) suggested that the selection pressure is of the same order of magnitude as mutation, and thus would affect codon usage only in organisms with extremely large effective population sizes. Moreover, adaptation of the codon usage to tRNA availability is expected to be species-specific, i.e., in each organism a different bias will exist. Such a state of affairs seems to prevail in unicellular organisms (Chen et al. 1986). In higher eukaryotes, on the other hand, there are indications that the codon usage does not reflect phylogenetic affinities (Hatfield and Rice 1986; Wells et al. 1986). Moreover, in tetrapods there seems to be no tissue-specific coadaptation of codon usage patterns and tRNA abundance (Hastings and Emerson 1983; Ouenzar et al. 1988). Thus, tRNA availability may not be an important factor in determining the pattern of codon usage in multicellular organisms (Wain-Hobson et al. 1981).

To date, the only indications we have on the nature of selection at the translation level involve fidelity in translation. It is known, for instance, that the frequency of errors varies for different codons in the same position and for the same codon in different positions (Rosenberger and Hilton 1983). In prokaryotes, for instance, codons ending in T are misread four to nine times more frequently than those ending in C (Precup and Parker 1987). Thus, reducing errors in translation may be a selective force affecting the rate of synonymous substitution and the pattern of codon usage.

Acknowledgments. We thank Paul M. Sharp for kindly disputing every statement in this note, thus making the presentation somewhat clearer. Ken Wolfe mercifully identified some embarrassing typos and omissions. This work was supported in part by grants from the Foundation for Basic Research, Tel Aviv University, and the Hertz Foundation.

References

Andersen RD, Birren BW, Ganz T, Piletz JE, Herschman HR (1983) Molecular cloning of the rat metallothionein 1 (*mt-1*) mRNA sequence. DNA 2:15–22

Andersen RD, Birren BW, Taplitz SJ, Herschman HR (1986) Rat metallothionein-1 structural gene and three pseudogenes, one of which contains 5′ regulatory sequences. Mol Cell Biol 6:302–314

Aota S, Ikemura T (1986) Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res 14:6345–6355 + Erratum 14:8702

Aota S, Gojobori T, Maruyama T, Ikemura T (1988) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res 16:r315–r402

Argentin S, Nemer M, Drouin J, Scott GK, Kennedy BP, Davies PL (1985) The gene for rat atrial natriuretic factor. J Biol Chem 260:4568–4571

Argos P, Taylor WL, Minth CD, Dixon JE (1983) Nucleotide and amino acid sequence comparisons of preprosomatostatins. J Biol Chem 258:8788–8793

Ayer D, Yarus M (1986) The context effect does not require a fourth base pair. Science 231:393–395

Benfield PA, Zivin RA, Miller LS, Sowder R, Smythers GW, Henderson L, Oroszlan S, Pearson ML (1984) Isolation and sequence analysis of cDNA clones coding for rat skeletal muscle creatine kinase. J Biol Chem 259:14979–14984

Bennetzen JL, Hall BD (1982) Codon selection in yeast. J Biol Chem 257:3026–3031

Boguski MS, Elshourbagy NA, Taylor JM, Gordon JI (1985) Comparative analysis of repeated sequences in rat apolipoproteins A-I, A-IV and E. Proc Natl Acad Sci USA 82:992–996

Bulmer M (1987a) A statistical analysis of nucleotide sequences of introns and exons in human genes. Mol Biol Evol 4:395–405

Bulmer M (1987b) Coevolution of codon usage and transfer RNA abundance. Nature 325:728–730

Bulmer M (1988) Are codon usage patterns in unicellular organisms determined by selection-mutation balance? J Evol Biol 1:15–26

Chang ACY, Cochet M, Cohen SN (1980) Structural organization of human genomic DNA encoding the pro-opiomelanocortin peptide. Proc Natl Acad Sci USA 77:4890–4894

Chang LMS (1973) Low molecular weight deoxyribonucleic acid polymerase from calf thymus chromatin. J Biol Chem 248:6983–6992

Chen KCK, Chen JS, Johnson JL (1986) Structural features of multiple *mifH*-like sequences and very biased codon usage in nitrogenase genes of *Clostridium pasteurianum*. J Bacteriol 166:162–172

Chin WW, Godine JE, Klein DR, Chang AS, Tan LK, Habener JF (1983) Nucleotide sequence of the cDNA encoding the precursor of the β subunit of rat leutropin. Proc Natl Acad Sci USA 80:4649–4653

Cooke NE, Baxter JD (1982) Structural analysis of the prolactin gene suggests a separate origin for its 5′ end. Nature 297:603–606

Cowan NJ, Dobner PR, Fuchs EV, Cleveland DW (1983) Expression of human α-tubulin genes: interspecies conservation of 3′ untranslated regions. Mol Cell Biol 3:1738–1745

Crabtree GR, Comeau CM, Fowlkes DM, Fornace AJ, Malley JD, Kant JA (1985) Evolution and structure of the fibrinogen genes: random insertion of introns or selective loss? J Mol Biol 185:1–19

den Dunnen JT, Moormann RJM, Lubsen NH, Schoenmakers JGG (1986) Intron insertions and deletions in the β/γ-crystallin gene family: the rat βB1 gene. Proc Natl Acad Sci USA 83:2855–2859

Deschenes RJ, Haun RS, Funckes CL, Dixon JE (1985) A gene encoding rat cholecystokinin: isolation, nucleotide sequence and promoter activity. J Biol Chem 260:1280–1286

Drouin J, Goodman HM (1980) Most of the coding region of rat ACTH-β-LPH precursor gene lacks intervening sequences. Nature 288:610–613

Dugaiczyk A, Law SW, Dennison OE (1982) Nucleotide sequence and the encoded amino acids of human serum albumin mRNA. Proc Natl Acad Sci USA 79:71–75

Dull TJ, Gray A, Hayflick JS, Ullrich A (1984) Insulin-like growth factor II precursor gene organization in relation to insulin gene family. Nature 310:777–781

Fiddes JC, Goodman HM (1981) The gene encoding the common α subunit of the four glycoprotein hormones. J Mol Appl Genet 1:3–18

Fort PH, Marty L, Piechaczyk M, el Sabrouty S, Dani C, Jeanteur P, Blanchard JM (1985) Various rat adult tissues express only one major mRNA species from the glyceraldehyde-3-phosphate-dehydrogenase multigenic family. Nucleic Acids Res 13:1431–1442

Ginzburg I, Behar L, Givol D, Littauer UZ (1981) The nucleotide sequence of rat α-tubulin: 3′-end characteristics, and evolutionary conservation. Nucleic Acids Res 9:2691–2697

Godine JE, Chin WW, Habener JF (1982) α subunit of rat pituitary glycoprotein hormones. J Biol Chem 257:8368–8371

Gojobori T, Li WH, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expression. Nucleic Acids Res 10:7055–7074

Graeme IB, Merryweather JP, Sanchez-Pescador RP, Stempien MM, Priestley L, Scott J, Rall LB (1984) Sequence of a cDNA clone encoding human preproinsulin-like growth factor II. Nature 310:775–777

Graur D (1985) Amino acid composition and the evolutionary rates of protein-coding genes. J Mol Evol 22:53–62

Graur D, Shuali Y, Li WC (1988) Deletions in processed pseudogenes accumulate faster in rodents than in humans. J Mol Evol (in press)

Greenberg BD, Bencen GH, Seilhamer JJ, Lewicki JA, Fiddes JC (1984) Nucleotide sequence of the gene encoding human atrial natriuretic factor precursor. Nature 312:656–658

Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon–anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18:199–209

Grosjean H, Sankoff D, Min Jou W, Fiers W, Cedergren RJ (1978) Bacteriophage *MS2* RNA: a correlation between the stability of codon–anticodon interaction and the choice of codon words. J Mol Evol 12:113–119

Hall L, Craig RK, Edbrooke MR, Campbell PN (1982) Comparison of the nucleotide sequence of cloned human and guinea-pig pre-α lactalbumin cDNA with that of chick pre-lysozyme cDNA suggests evolution from a common ancestral gene. Nucleic Acids Res 10:3503–3515

Hamada H, Petrino MG, Kakunaga T (1982) Molecular structure and evolutionary origin of human cardiac muscle actin gene. Proc Natl Acad Sci USA 79:5901–5905

Hanauer A, Mandel JL (1984) The glyceraldehyde-3-phosphate dehydrogenase gene family: structure of a human cDNA and of an X chromosome linked pseudogene; amazing complexity of the gene family in the mouse. EMBO J 3:2627–2633

Hastings KEM, Emerson CP (1983) Codon usage in muscle genes and liver genes. J Mol Evol 19:214–218

Hatfield D, Rice M (1986) Aminoacyl-tRNA (anticodon): codon adaptation in human and rabbit reticulocytes. Biochem Int 13:835–842

Hayashizaki Y, Miyai K, Kato K, Matsubara K (1985) Molecular cloning of human thyrotropin-β subunit gene. FEBS Lett 188:394–400

Heinrich G, Kronenberg HM, Potts JT, Habener JF (1984) Gene encoding parathyroid hormone: nucleotide sequence of the rat gene deduced amino acid sequence of rat preproparathyroid hormone. J Biol Chem 259:3320–3329

Hudson P, Haley J, Cronk M, Shine J, Niall H (1981) Molecular cloning and characterization of cDNA sequences coding for rat relaxin. Nature 291:127–131

Hudson P, John M, Crawford R, Haralambidis J, Scanlon D, Gorman J, Tregear G, Shine J, Niall H (1984) Relaxin gene expression in human ovaries and the predicted structure of

human preprorelaxin by analysis of cDNA clones. EMBO J 3:2333–2339

Ikemura T (1980) The frequency of codon usage in *E. coli* genes: correlation with the abundance of cognate tRNA. In: Osawa S, Ozeki H, Uchida H, Yura T (eds) Genetics and evolution of RNA polymerase, tRNA and ribosomes. University of Tokyo Press, Tokyo, pp 519–523

Ikemura T (1981a) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the repetitive codon in protein genes. J Mol Biol 146:1–21

Ikemura T (1981b) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the repetitive codon in protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151:389–409

Ikemura T (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the repetitive codon in protein genes: differences in synonymous codon choice patterns of yeast and *Escherichia coli* with references to the abundance of isoacceptor transfer RNAs. J Mol Biol 158:573–579

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Ikemura T, Ozeki H (1983) Codon usage and transfer RNA contents: organism-specific codon choice patterns in reference to isoacceptor contents. Cold Spring Harbor Symp Quant Biol 42:1087–1097

Inana G, Totsuka S, Redmond M, Dougherty T, Nagle J, Shiono T, Ohura T, Kominami E, Katunuma N (1986) Molecular cloning of human ornithine aminotransferase mRNA. Proc Natl Acad Sci USA 83:1203–1207

Jagodzinski LL, Sargent TD, Yang M, Glackin C, Bonner J (1981) Sequence homology between RNAs encoding rat α-fetoprotein and rat serum albumin. Proc Natl Acad Sci USA 78: 3521–3525

Joh K, Mukai T, Yatsuki H, Hori K (1985) Rat aldolase-A messenger RNA: the nucleotide sequence and multiple mRNA species with different 5′-terminal regions. Gene 39:17–24

Jukes TH, King JL (1979) Evolutionary nucleotide replacements in DNA. Nature 281:605–606

Karathanasis SK (1985) Apolipoprotein multigene family: tandem organization of human apolipoprotein AI, CIII, and AIV genes. Proc Natl Acad Sci USA 82:6374–6378

Karin M, Richards RI (1982) Human metallothionein gene—primary structure of the metallothionein-II gene and a related processed gene. Nature 299:797–802

Kawakami K, Nojima H, Ohta T, Nagano K (1986) Molecular cloning and sequence analysis of human Na^+,K^+-ATPase β-subunit. Nucleic Acids Res 14:2833–2844

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267:275–276

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Law SW, Brewer HB (1984) Nucleotide sequence and the encoded amino acids of human apolipoprotein A-I mRNA. Proc Natl Acad Sci USA 81:66–70

Leibold EA, Munro HN (1987) Characterization and evolution of the expressed rat ferritin light subunit gene and its pseudogene family. J Biol Chem 262:7335–7341

Li SSL, Fitch WM, Pan YCE, Sharief FS (1983) Evolutionary relationship of vertebrate lactate dehydrogenase isozymes A_4 (muscle), B_4 (heart), and C_4 (testis). J Biol Chem 258:7029–7032

Li WH, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. Nature 326:93–96

Li WH, Wu CI (1987) Rates of nucleotide substitution are evidently higher in rodents than in man. Mol Biol Evol 4: 74–77

Li WH, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239

Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions and its evolutionary implications. J Mol Evol 21:58–71

Li WH, Luo CC, Wu CI (1985a) Evolution of DNA sequences. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum, New York, pp 1–94

Li WH, Wu CI, Luo CC (1985b) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Li WH, Tanimura M, Sharp PM (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330–342

Lipman DJ, Wilbur WJ (1983) Contexual constraints on synonymous codon choice. J Mol Biol 163:363–376

Loeb LA, Kunkel TA (1982) Fidelity of DNA synthesis. Annu Rev Biochem 52:429–457

Lopez LC, Farzier ML, Su CJ, Kumar A, Saunders GF (1983) Mammalian pancreatic preproglucagon contains three glucagon-related peptides. Proc Natl Acad Sci USA 80:5484–5489

Luo CC, Li WH, Moore MN, Chan L (1986) Structure and evolution of the apolipoprotein multigene family. J Mol Biol 187:325–340

MacDonald RJ, Crerar MM, Swain WF, Pictet RL, Thomas G, Rutter WJ (1980) Structure of a family of rat amylase genes. Nature 287:117–122

Mayer Y, Czosnek H, Zeelon PE, Yaffe D, Nudel U (1984) Expression of the genes coding for the skeletal muscle and cardiac actins in the heart. Nucleic Acids Res 12:1087–1100

McLean JW, Fukazawa C, Taylor JM (1983) Rat apolipoprotein E mRNA: cloning and sequencing of double-stranded cDNA. J Biol Chem 258:8993–9000

McLean JW, Elshourbagy NA, Chang DJ, Mahley RW, Taylor JM (1984) Human apolipoprotein E mRNA: cDNA cloning and nucleotide sequence of a new variant. J Biol Chem 259: 6498–6504

Meakin SO, Breitman ML, Tsui LC (1985) Structural and evolutionary relationships among five members of the human γ-crystallin gene family. Mol Cell Biol 5:1408–1414

Miyata T, Hayashida H (1981) Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. Proc Natl Acad Sci USA 78:5739–5743

Miyata T, Yasunaga T (1981) Rapidly evolving mouse α-globin-related pseudogene and its evolutionary history. Proc Natl Acad Sci USA 78:450–453

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328–7332

Modrich P (1987) DNA mismatch correction. Annu Rev Biochem 56:435–466

Morinaga T, Sakai M, Wegmann TG, Tamaoki T (1983) Primary structures of human α-fetoprotein and its mRNA. Proc Natl Acad Sci USA 80:4604–4608

Moriuchi T, Chang HC, Denome R, Silver J (1982) *Thy-1* cDNA sequence suggests a novel regulatory mechanism. Nature 301:80–82

Mouchiroud D (1986) Relation entre la composition en base de l'ADN non codant du gene et la composition en codon. CR Acad Sci Paris 303:743–748

Mueckler MM, Pitot HC (1985) Sequence of the precursor to the rat ornithine aminotransferase deduced from a cDNA clone. J Biol Chem 250:12993–12997

Nakamura Y, Ogawa M, Nishide T, Emi M, Kosaki G, Himeno S, Matsubara K (1984) Sequences of cDNAs for human salivary and pancreatic α-amylases. Gene 28:263–270

Newgard CB, Nakano K, Hwang PK, Fletterick RJ (1986) Sequence analysis of the complementary DNA encoding human

liver glycogen phosphorylase reveals tissue-specific codon usage. Proc Natl Acad Sci USA 83:8132–8136

Ng SY, Gunning P, Eddy R, Ponte P, Leavitt J, Shows T, Kedes L (1985) Evolution of the functional human β-actin gene and its multipseudogene family: conservation of noncoding regions and chromosomal dispersion of pseudogenes. Mol Cell Biol 5:2720–2732

Nichols BP, Miozzari GF, van Cleemput M, Bennett GM, Yanofsky C (1980) Nucleotide sequences of the *trpG* regions of *Escherichia coli, Shigella dysenteriae, Salmonella typhimurium* and *Serratia marcescens*. J Mol Biol 142:503

Nie NH, Hadlaihull C, Jenkins JG, Steinbrenner K, Bent DH (1975) SPSS. McGraw-Hill, New York

Nudel U, Zakut R, Shani M, Neuman S, Levy Z, Yaffe D (1983) The nucleotide sequence of the rat cytoplasmic β-actin gene. Nucleic Acids Res 11:1759–1771

Nussinov R (1981) Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. J Mol Biol 149:125–131

Osawa S, Chiu RH, McDonough A, Miller TB, Johnson GL (1986) Isolation of partial cDNAs for rat liver and muscle glycogen phosphorylase isozymes. FEBS Lett 202:282–288

Ouenzar B, Agoutin B, Reinisch F, Weill D, Perin F, Keith G, Heyman T (1988) Distribution of isoaccepting tRNAs and codons for proline and glycine in collagenous and noncollagenous chicken tissues. Biochem Biophys Res Commun 150: 148–155

Page GS, Smith S, Goodman GM (1981) DNA sequence of the rat growth hormone gene: location of the 5′ terminus of the growth hormone mRNA and identification of an internal transposon-like element. Nucleic Acids Res 9:2087–2104

Perryman MB, Kerner SA, Bohlmeyer TJ, Roberts R (1986) Isolation and sequence analysis of a full-length cDNA for human M creatine kinase. Biochem Biophys Res Commun 140:981–989

Precup J, Parker J (1987) Missense misreading of asparagine codons as a function of codon identity and context. J Biol Chem 262:11351–11355

Qasba PK, Safaya SK (1984) Similarity of the nucleotide sequences of rat α-lactalbumin and chicken lysozyme gene. Nature 303:377–380

Rixon MW, Chan WY, Davie EW, Chung DW (1983) Characterization of a complementary deoxyribonucleic acid coding for the α-chain of human fibrinogen. Biochemistry 22: 3237–3244

Rosenberger RF, Hilton J (1983) The frequency of transcriptional and translational errors at nonsense codons in the *lacZ* gene of *Escherichia coli*. Mol Gen Genet 191:207–212

Roskam WG, Rougeon F (1979) Molecular cloning and nucleotide sequence of the human growth hormone structural gene. Nucleic Acids Res 7:305–320

Rottmann WH, Tolan DR, Penhoet EE (1984) Complete amino acid sequence for human aldolase B derived from cDNA and genomic clones. Proc Natl Acad Sci USA 81:2738–2742

Sakakibara M, Mukai T, Hori K (1985) Nucleotide sequence of a cDNA clone for human aldolase: a messenger RNA in the liver. Biochem Biophys Res Commun 131:413–420

Santoro C, Marone M, Ferrone M, Costanzo F, Colombo M, Minganti C, Cortese R, Silengo L (1986) Cloning of the gene coding for human L apoferritin. Nucleic Acids Res 14:721–735

Sargent TD, Yang M, Booner J (1981) Nucleotide sequence of cloned rat serum albumin messenger RNA. Proc Natl Acad Sci USA 78:243–246

Sausville E, Carney D, Battey J (1985) The human vasopressin gene is linked to the oxytocin gene and is selectively expressed in a cultured lung cancer cell line. J Biol Chem 260:10236–10241

Sharp PM, Li WH (1987a) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

Sharp PM, Li WH (1987b) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1294

Shen LP, Pictet RL, Rutter WJ (1982) Human somatostatin sequence of the cDNA. Proc Natl Acad Sci USA 79:4575–4579

Shpaer EG (1986) Constraint on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation. J Mol Biol 188:555–564

Soares MB, Schon E, Henderson A, Karathanasis SK, Cate R, Zeitlin S, Chirgwin J, Efstratiadis A (1985) RNA-mediated duplication: the rat preproinsulin I gene is a functional retroposon. Mol Cell Biol 5:2090–2103

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Sullivan KF, Havercroft JC, Cleveland DW (1984) Primary structure and expression of a vertebrate β-tubulin gene family. In: Dorisy GG, Cleveland DW, Murphy DW (eds) Molecular biology of the cytoskeleton. Cold Spring Harbor Press, Cold Spring Harbor NY, pp 321–332

Sundelin J, Melhus H, Das S, Eriksson U, Lind P, Traegardh L, Peterson PA, Rask L (1985) The primary structure of rabbit and rat prealbumin and a comparison with the tertiary structure of human prealbumin. J Biol Chem 260:6481–6487

Tajimi F, Nei M (1982) Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J Mol Evol 18:115–120

Takahashi Y, Kato K, Hayashizaki Y, Wakabayash T, Ohtsuka E, Matsuki S, Ikehara M, Matsubara K (1985) Molecular cloning of the human cholecystokinin gene by use of synthetic probe containing deoxyinosine. Proc Natl Acad Sci USA 82: 1931–1935

Truong AT, Duez C, Belayew A, Renard A, Pictet R, Bell GI, Martial JA (1984) Isolation and characterization of human prolactin gene. EMBO J 3:429–437

Tsujibo H, Tiano HF, Li SSL (1985) Nucleotide sequence of the cDNA and an intronless pseudogene for human lactate dehydrogenase-A isozyme. Eur J Biochem 147:9–15

Tsutsumi KI, Mukai T, Hidaka S, Miyahara H, Tsutsumi R, Tanaka T, Hori K, Ishikawa K (1983) Rat aldolase isozyme gene: cloning and characterization of cDNA for aldolase B messenger RNA. J Biol Chem 258:6537–6542

Ullrich A, Dull TJ, Gray A, Brosius J, Sures I (1980) Genetic variation in the human insulin gene. Science 209:612–615

van Rijs J, Giruere V, Hurst J, van Agthoven T, van Kessel AD, Goyert S, Grosveld F (1985) Chromosomal localization of human *thy-1* gene. Proc Natl Acad Sci USA 82:5832–5835

Varenne S, Buc J, Lloubes R, Lazdunski C (1984) Translation in a nonuniform process: effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J Mol Biol 180:549–576

Varshney U, Gedamu L (1984) Human metallothionein MT-I and MT-II processed genes. Gene 31:135–145

Vasicek TJ, McCevitt BE, Freeman MW, Fennick BJ, Hendy GN, Potts JT, Rich A, Kronenberg HM (1983) Nucleotide sequence of the human parathyroid hormone gene. Proc Natl Acad Sci USA 80:2127–2131

Wain-Hobson S, Nussinov R, Brown RJ, Sussman JL (1981) Preferential codon usage in genes. Gene 13:335–364

Wallace MR, Naylor SL, Kluve-Beckerman B, Long GL, McDonald L, Shows TB, Benson MD (1985) Localization of the human prealbumin gene to chromosome 18. Biochem Biophys Res Commun 129:753–758

Wells D, Bauis W, Kedes L (1986) Codon usage in histone gene families of higher eukaryotes reflects functional rather than phylogenetic relationships. J Mol Evol 23:224–241

White JW, Saunders GF (1986) Structure of the human glucagon gene. Nucleic Acids Res 14:4719–4730

Wilbur WJ, Lipman DJ (1985) Rapid similarity searches of nucleic acid and protein data banks. Proc Natl Acad Sci USA 80:726–730

Wilde CD, Crowther CE, Cripe TP, Gwo-Shu Lee M, Cowan NJ (1982) Evidence that a human β-tubulin pseudogene is derived from its corresponding mRNA. Nature 297:83–84

Wright S (1969) Evolution and genetics of populations, vol 2. University of Chicago Press, Chicago

Wu CI, Li WH (1985) Evidence for higher rates in rodents than in man. Proc Natl Acad Sci USA 82:1741–1745

Yanofsky C, van Cleemput M (1982) Nucleotide sequence of trpE of *Salmonella typhimurium* and its homology with the corresponding sequence of *Escherichia coli*. J Mol Biol 155: 235–246

Young RM, Shull GE, Lingrel JB (1987) Multiple mRNAs from rat kidney and brain encode a single Na^+,K^+-ATPase β-subunit protein. J Biol Chem 262:4905–4910

Zuckerkandl E (1965) Remarques sur l'évolution des polynucléotides comparée à celle des polypeptides. Bull Soc Chim Biol 47:1729–1730

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357–366

Received April 4, 1988/Revised and accepted September 22, 1988

Appendix 1

The equations for predicting the rate of synonymous substitution in a gene from its dinucleotide composition in positions 2 and 3 were calculated by a forward (stepwise) inclusion approach. The independent variables (the frequencies of the dinucleotides in positions 2 and 3) were entered one by one into the multiple regression function, the order of inclusion being determined by their respective added contribution to the explained variance in rates of synonymous substitution (Nie et al. 1975, pp. 321–367). The formulae for the first six predictors, denoted as Y_n, where n stands for number of independent variables, are as follows:

$$Y_1 = 0.923 - 3.051f_{CpC}$$
$$Y_2 = 0.781 - 2.410f_{CpC} + 2.163f_{ApA}$$
$$Y_3 = 0.674 - 1.946f_{CpC} + 1.905f_{ApA} + 1.902f_{CpA}$$
$$Y_4 = 0.515 - 1.744f_{CpC} + 2.459f_{ApA} + 2.185f_{CpA} + 1.200f_{TpC}$$
$$Y_5 = 0.621 - 2.067f_{CpC} + 2.274f_{ApA} + 1.920f_{CpA} + 1.143f_{TpC} - 1.692f_{GpG}$$
$$Y_6 = 0.682 - 2.193f_{CpC} + 2.214f_{ApA} + 2.664f_{CpA} + 1.543f_{TpC} - 1.971f_{GpG} + 1.229f_{CpT}$$

Appendix 2

The equations for predicting the rate of synonymous substitution in a gene from its trinucleotide composition in the reading frame were calculated as in Appendix 1. The formulae for the first 9 predictors, denoted as Z_n, where n stands for the number of independent variables (frequencies of codons), are as follows:

$$Z_1 = 0.498 + 12.025f_{CTT}$$
$$Z_2 = 0.440 + 9.374f_{CTT} + 9.513f_{CAA}$$
$$Z_3 = 0.517 + 9.171f_{CTT} + 8.953f_{CAA} - 8.749f_{CGT}$$
$$Z_4 = 0.597 + 7.3860_{CTT} + 8.254f_{CAA} - 11.323f_{CGT} - 0.965f_{TGC}$$
$$Z_5 = 0.584 + 5.448f_{CTT} + 8.351f_{CAA} - 10.644f_{CGT} - 1.547f_{TGC} + 3.751f_{TGT}$$
$$Z_6 = 0.455 + 7.430f_{CTT} + 8.784f_{CAA} - 11.633f_{CGT} - 1.280f_{TGC} + 4.477f_{TGT} + 2.015f_{GAG}$$
$$Z_7 = 0.508 + 6.680f_{CTT} + 9.610f_{CAA} - 11.313f_{CGT} - 1.462f_{TGC} + 4.273f_{TGT} + 1.834f_{GAG} - 3.080f_{TGG}$$
$$Z_8 = 0.468 + 6.989f_{CTT} + 10.345f_{CAA} - 12.521f_{CGT} - 1.443f_{TGC} + 4.212f_{TGT} + 2.119f_{GAG} - 4.447f_{TGG} + 4.200f_{GGT}$$
$$Z_9 = -0.429 + 7.665f_{CTT} + 10.897f_{CAA} - 13.017f_{CGT} - 1.406f_{TGC} + 4.447f_{TGT} + 2.174f_{GAG} - 5.910f_{TGG} + 5.684f_{GGT} + 3.418f_{CGG}$$