J Mol Evol (2004) 59:400–415
DOI: 10.1007/s00239-004-2634-7

JOURNAL OF MOLECULAR EVOLUTION

Selection on Codon Usage for Error Minimization at the Protein Level

Marco Archetti

Département de Biologie, Section Ecologie et Evolution, Université de Fribourg, Chemin du Musée 10, 1700, Fribourg, Switzerland

Received: 30 September 2003 / Accepted: 12 April 2004 [Reviewing Editor: Dr. Massimo Di Giulio]

Abstract. Given the structure of the genetic code, synonymous codons differ in their capacity to minimize the effects of errors due to mutation or mistranslation. I suggest that this may lead, in protein-coding genes, to a preference for codons that minimize the impact of errors at the protein level. I develop a theoretical measure of error minimization for each codon, based on amino acid similarity. This measure is used to calculate the degree of error minimization for 82 genes of *Drosophila melanogaster* and 432 rodent genes and to study its relationship with CG content, the degree of codon usage bias, and the rate of nucleotide substitution. I show that (i) *Drosophila* and rodent genes tend to prefer codons that minimize errors; (ii) this cannot be merely the effect of mutation bias; (iii) the degree of error minimization is correlated with the degree of codon usage bias; (iv) the amino acids that contribute more to codon usage bias are the ones for which synonymous codons differ more in the capacity to minimize errors; and (v) the degree of error minimization is correlated with the rate of nonsynonymous substitution. These results suggest that natural selection for error minimization at the protein level plays a role in the evolution of coding sequences in *Drosophila* and rodents.

Key words: Genetic code — Error minimization — Codon usage — Nucleotide substitution — *Drosophila melanogaster* — Rodents — Amino acid similarity

Correspondence to: Marco Archetti; *email:* marco.archetti@unifr.ch

Introduction

Studies on the origin of the genetic code have shown that the code is arranged in a way that reduces errors due to mutation or mistranslation; that is, amino acids with similar chemical properties are coded by similar codons (Woese 1965; Epstein 1966). Therefore it is possible (Haig and Hurst 1991; Freeland and Hurst 1998; Knight et al. 1999; Freeland et al. 2000) that the main force that shaped the genetic code is selection for minimization of the chemical distances between amino acids, that is, error minimization at the protein level (though this theory is still debated [see Di Giulio 2000a, 2000b; Archetti 2004]).

Since the genetic code is degenerate (there are 64 codons coding for only 20 amino acids and a termination signal), most amino acids are encoded by several synonymous codons, and as Grantham et al. (1980) first demonstrated, and as further data from a number of organisms eventually corroborated, codon usage is not random: some synonymous codons are more used than others. The theory of error minimization for the evolution of the genetic codes postulates that the codons are *arranged* in the code in a way that reduces errors. The scope of this paper is to test whether error minimization at the protein level plays a role also in the evolution of codon usage in protein-coding genes. The hypothesis is that the *preferred* codons are the ones that, after mutation or mistranslation, keep on coding for the same amino acid or for amino acids with similar chemical properties.

Following the same logic used to measure the optimization of the genetic code, it is possible to calculate a measure of the capacity to minimize errors due to mutation or mistranslation for each codon. This measure can then be used to calculate the degree

of error minimization for individual genes or genomes.

If the hypothesis of error minimization for codon usage bias is correct, then not only should the preferred codons be the ones that minimize errors, but also it is expected that there is a correlation between the degree of error minimization and the rate of nonsynonymous nucleotide substitution; that is, highly conserved genes should prefer synonymous codons that minimize the effect of errors. Moreover, the degree of codon usage bias should be correlated with the degree of error minimization, and the amino acids that contribute more to codon usage bias should be the ones whose synonymous codons have a greater variance in the capacity to minimize errors.

I first develop a theoretical, quantitative measure of error minimization for synonymous codons, based on similarities between amino acids; then I use this measure to calculate the degree of error minimization for 82 genes of *Drosophila melanogaster* and for 432 rodent genes; finally, these values are used to study the relationship of error minimization with the degree of codon usage bias, with the CG content, and with the rates of nucleotide substitution.

Methods

I developed a computer program that measures the capacity of each codon to minimize the deleterious effects of errors due to mutation or mistranslation. The basic concept is to calculate the mean dissimilarity between the amino acid coded by the original codon and the amino acids coded by its possible mutants; this measure depends only on the structure of the genetic code and on the similarities between amino acids. For the synonymous codons of each amino acid, then, these values of error minimization are correlated with codon usage. The mean value of the correlations is taken as a measure of error minimization for the gene. The method is described in detail in the following sections.

Amino Acid Similarity

The impact of mutation or mistranslation can be deduced from amino acid similarity matrices. In a similarity scoring matrix, higher values are assigned to more similar pairs of amino acids (George et al. 1990). Throughout this paper I use McLachlan's (1971) classical matrix based on chemical properties, but at the end I compare the results obtained with other matrices.

For each pair of amino acids, I derive the measure $D_{AA/AA^*} = \omega_{AA/AA} - \omega_{AA/AA^*}$ from the matrix, where $\omega_{AA/AA}$ is the similarity of amino acid AA to itself (this value is usually the same for all amino acids, but not in all matrices: in McLachlan's it is either 8 or 9) and ω_{AA/AA^*} is the similarity of AA to the mutant amino acid AA* obtained after an error at one of the three positions of the original codon. Hence, D_{AA/AA^*} is the distance (dissimilarity) between the original (AA) and the mutant (AA*) amino acid.

Similarities between amino acids and termination signals ($\omega_{AA/STOP}$) have, of course, no meaning and are not tabulated in similarity scoring matrices; however, a measure of the damage produced by mutations to termination codons must be considered. I use diverse scores for this that are less than or equal to the lowest similarity score of the matrix (0 in McLachlan's matrix).

Error Minimization of Synonymous Codons: The MD Value

Since $\omega_{AA/AA} \geq \omega_{AA/AA^*}$ for every amino acid, D_{AA/AA^*} is always positive, and the measures of D_{AA/AA^*} for the possible mutant codons arising by point mutation are positive. Their mean value is taken as a measure of distance (dissimilarity) between the original codon and its possible mutants. I call this measure MD (mean distance). Optimal codons are predicted to have small MD values.

The same procedure can be used to measure the MD between one codon and the possible codons that arise by point mutation after n mutation events. This is important because some synonymous codons have the same MD value after one mutation (Fig. 1). These include most twofold degenerate amino acids and some others such as Pro, because its similarity score with His and Gln is the same (in McLachlan's matrix). In these cases selection for optimal codons will not operate after one mutation but after successive mutation events. Actually, this process can be seen as a differential survival of lineages originating from genomes with different codon usage patterns. For example, consider the part of the progeny that inherited Gly instead of Glu after the first mutation. In the first generation, there is no differential selection between the two lineages with Gly–GGA and Gly–GGG. These two lineages will, however, produce some mutant progeny with the termination signal and Trp, respectively.

I use the following procedure to take into account the importance of each mutation event. I measure MD for each codon and variance of MD for synonymous codons of each amino acid, for every mutation event. For example, the variance after one mutation for Pro is zero, while there is a high variance for the six codons of Leu (Fig. 2). In general, the higher the variance, the higher the intensity of selection. The intensity of selection (σ) for each amino acid (AA) for each mutation event (n) is

1st base	2nd T	2nd C	2nd A	2nd G	3rd base
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Fig. 1. The standard genetic code. The codons in light gray are the mutant codons originated from the original codon (in dark gray) after one point mutation.

$$\sigma_{AA(n)} = \frac{v_{AA(n)}}{\sum_{m=1}^{10} v_{AA(m)}}$$

where $v_{AA(n)}$ is the variance of MD values for synonymous codons coding for amino acid AA after the nth mutation. I consider mutations up to the 10th: although this is an arbitrary limit, MD values clearly converge to the same value, for all codons, as n grows (Fig. 2) because there are 9^n possible mutants but only 64 codons; therefore after many mutation events the differences of MD values for different codons become negligible.

The importance of mutations also depends on the order. Further mutations are likely to be less important if selection had already occurred in previous generations. I use the following simple method (a decision on how to weight successive mutations is likely to be arbitrary; however, different methods gave very similar results): for each amino acid (AA), the weight (Ω) of the first mutation is equal to its intensity of selection [$\Omega_{AA(1)} = \sigma_{AA(1)}$], while for the successive mutations, at the nth mutation, the weight is

$$\Omega_{AA(n)} = \left[1 - \sum_2^n \sigma_{AA(n-1)}\right]\sigma_{AA(n)}$$

Imagine, for example, that $\sigma_{AA(1)} = 0.2$, $\sigma_{AA(2)} = 0.7$, $\sigma_{AA(3)} = 0.1$, and $\sigma_{AA(n)} = 0$ for all $n > 3$. The corresponding weights will be $\Omega_{AA(1)} = 0.2$, $\Omega_{AA(2)} = 0.56$, and $\Omega_{AA(3)} = 0.01$. The corresponding MD values are multiplied by these weights and then summed (up to the 10th mutation, unless otherwise stated in the text).

Mutation Bias

The values taken from the matrix can be weighted also by the probability that such mutations occur.

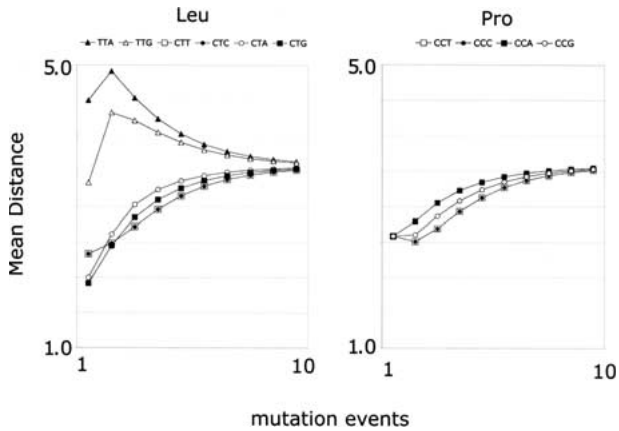

Fig. 2. MD values for the codons of Leu and Pro after further mutation events (no transition/transversion bias; CG/AT mutation ratio = 1; $\omega_{AA/STOP} = -10$).

Transitions (C↔T, A↔G) and transversions (C,T↔A,G) are not equally likely to occur (Topal and Fresco 1976). Moreover, it is also possible that accuracy varies during translation according to base position within a codon (Woese 1965), since the translation machinery acts upon mRNA in a specific reading frame, reading bases in triplets. Therefore, before calculating MD, D_{AA/AA^*} can be multiplied by the probability of a change from the original base to the mutant one (allowing for a consideration of whether it is a transition or a transversion) and by the probability of mistranslation based on its position on the codon (1st, 2nd, or 3rd).

I have followed the same assumptions used by Freeland and Hurst (1998) to test the optimization of the genetic code, namely, that mistranslation of the second position is much less frequent than mistranslation of the first or third position; that mistranslation of the first position is less frequent than mistranslation of the third codon position; that mistranslations at the second position are almost exclusively transitional; that mistranslation at the first position is biased towards transitional errors; and that there is very little transition bias at the third position. The precise values used to weight each mutation are summarized in Table 1.

As with translation, different rates for transition and transversion must be considered for mutation but in this case there is no reason to assume that these rates vary depending upon base position (see Table 1).

Moreover, I consider the possibility of different mutation rates for CG and AT, because C and G, which have three hydrogen bonds, may be more stable than A or T.

Error Minimization for Individual Genes

Since MD is a measure of dissimilarity, lower values of MD correspond to optimal codons (codons that

Table 1. Relative frequencies of mutation and mistranslation used in the model

	Mistranslation						
	Frequency[a]			T/T ratio[b]			Mutation (all bases) T/T ratio
	1st	2nd	3rd	1st	2nd	3rd	
A	0.5	0.1	1	2	5	1	—
B	—	—	—	—	—	—	1.5
C	0.5	0.1	1	2	5	1	1.5

[a]Frequency of mistranslation for the first, second, or third base of the codon.
[b]Transition/transversion ratio (relative to a transversion rate = 1) for the first, second, or third base of the codon.

minimize the effects of errors). MD measures can be compared only within each synonymous family (amino acid) and not between different amino acids, because each position in the sequence of a protein must be occupied by a certain amino acid, and codons for that amino acid can be chosen only among synonymous codons. The sample size is too small ($n = 2$ to 6) to yield a significant correlation within each amino acid, but the importance of error minimization can be measured by the sum of the correlations between MD and synonymous codon frequencies for all the 18 degenerate amino acids. This measure (R) ranges between $-N$ and $+N$, where N is the number of amino acids on which the correlations are measured (for the standard genetic code, usually $N = 18$, but $N < 18$ if for some amino acid there is no variance in the MD values or in the frequency of its synonymous codons). A normalized version of R (R_N: from -1 to $+1$) can be obtained by dividing R by N; both measures can be transformed weighting the correlation of each amino acid by its frequency (wR and wR_N). R and wR values close to $-N$ (or R_N and wR_N values close to -1) will mean a strong tendency to minimize the effects of errors.

Genes and Other Measures Used in the Analysis

Codon usage and the degree of codon usage bias for individual genes are calculated from the coding sequences of 82 *D. melanogaster* genes and 432 mouse–rat homologs. Codon usage for the entire genome of *D. melanogaster* is taken from the Codon Usage Database (www://kazusa.or.jp/codon; described by Nakamura et al. 2000). The measure used for the degree of codon usage bias is the effective number of codons (ENC; Wright 1990); ENC values range from 20 (high bias—only one codon per amino acid is used) to 61 (low bias—all codons are used equally).

For *Drosophila*, the genes are the same used by Moriyama and Powell (1997), while the substitution rates are taken from Akashi (1994) for a limited set of genes. Genes and substitution rates for rodents are the ones used in Smith and Hurst (1997). K_a and K_s are the rates of synonymous and nonsynonymous substitution, K_4 is the rate of nonsynonymous substitution at fourfold sites (K_4 is more reliable than K_s because it does not have to combine rates of sites with different degeneracies).

Random Sequences

To study the importance of CG content, I also use random sequences (82 for *Drosophila*, 432 for rodents) with the same length and the same CG content as the real genes (for rodents I use the CG content of the mouse genes). I also use random sequences with the same CG content (as the real genes) at each of the three positions. Both these kinds of sequences maintain the same CG content, but not the same degree of codon usage bias (ENC).

As a null model in the study of the degree of codon usage bias I use, instead, sequences (82 for *Drosophila*, 432 for rodents) with the same length and the same ENC values as the real genes, produced by switching at random the frequencies of synonymous codons. These sequences maintain the same ENC of the real genes but not the same CG content.

Results

MD Values

Table 2 shows MD values obtained with different values of the parameters. With no transition/transversion bias and with the same mutation rate for CG and AT, some amino acids show an ambiguity about which codons have the lowest MD value. This reflects the arrangement of the codons in the code (Fig. 1): in the standard code with no bias, for example, all twofold degenerate amino acids except Glu, Lys, and Gln have no variance in MD values for their synonymous codons because each couple of codons can mutate only to new couples of codons that are synonymous as well (and if only the first mutation event is considered, only codons for Lys have different MD values because one of them can mutate to Met and the other one to Ile—but, incidentally, in McLach-

Table 2. MD values

		No bias,[a] 1 mutation only		No bias				
		$\chi^b = 0.9$	$\chi = 1$	$\chi = 0.9$	$\chi = 1$	A[a] ($\chi = 0.9$)	B[a] ($\chi = 0.9$)	C[a] ($\chi = 0.9$)
Arg								
	CGA	2.775	3.083	1.947	2.351	1.117	7.356	2.814
	CGC	2.250	2.500	1.552	1.898	1.002	6.477	2.429
	CGG	1.800	2.000	1.411	1.662	0.754	6.450	2.339
	CGT	2.250	2.500	1.601	1.898	1.009	7.105	2.584
	AGA	3.708	3.833	2.422	2.781	2.076	8.520	3.603
	AGG	2.558	2.750	1.805	2.099	1.258	7.423	2.885
Leu								
	CTA	1.950	2.000	1.567	1.696	0.684	5.938	2.522
	CTC	2.258	2.333	1.524	1.719	0.762	5.319	2.343
	CTG	**1.875**	**1.917**	**1.438**	**1.609**	**0.631**	5.326	**2.291**
	CTT	2.258	2.333	1.577	1.719	0.770	5.789	2.459
	TTA	4.500	4.500	3.023	3.172	1.745	8.271	5.124
	TTG	3.283	3.333	2.445	2.652	1.315	7.177	4.067
Ser								
	TCA	4.150	4.500	2.799	3.172	1.816	7.929	4.464
	TCC	2.275	2.417	1.745	1.989	1.045	6.565	2.744
	TCG	3.175	3.417	2.243	2.578	1.414	6.899	3.491
	TCT	2.275	2.417	1.798	1.989	1.048	7.173	2.889
	AGC	2.750	2.917	1.881	2.119	1.660	**6.424**	**2.681**
	AGT	2.817	2.917	1.957	2.119	1.717	7.019	2.823
Thr								
	ACA	2.208	2.333	1.705	1.998	0.717	10.290	3.565
	ACC	***2.058***	***2.167***	**1.457**	***1.785***	**0.607**	**9.154**	**3.044**
	ACG	2.208	2.333	1.535	1.911	0.651	9.209	3.158
	ACT	2.058	2.167	1.566	1.785	0.635	10.107	3.314
Pro								
	CCA	*2.325*	*2.583*	1.558	2.071	0.518	9.459	3.306
	CCC	*2.325*	*2.583*	**1.337**	***1.866***	**0.451**	**8.432**	**2.887**
	CCG	*2.325*	*2.583*	1.391	1.958	0.465	8.469	2.941
	CCT	*2.325*	*2.583*	**1.444**	1.866	0.481	9.316	3.162
Ala								
	GCA	2.025	2.250	1.552	2.035	0.631	9.339	3.287
	GCC	2.100	2.333	**1.335**	**1.838**	**0.531**	**8.312**	**2.829**
	GCG	2.025	2.250	1.391	1.934	0.563	8.357	2.911
	GCT	2.100	2.333	1.439	1.838	0.559	9.178	3.084
Gly								
	GGA	3.300	3.667	2.153	2.637	1.737	7.989	3.226
	GGC	***2.475***	***2.750***	**1.654**	***2.043***	1.137	**6.976**	**2.610**
	GGG	2.475	2.750	1.677	2.068	1.134	6.993	2.611
	GGT	2.475	2.750	1.711	2.043	1.145	7.659	2.775
Val								
	GTA	2.092	2.167	1.686	1.928	0.756	10.467	3.684
	GTC	2.325	2.417	1.493	1.856	0.849	9.329	3.212
	GTG	2.167	2.250	1.526	1.878	0.761	9.370	3.272
	GTT	2.325	2.417	1.607	1.856	0.856	10.302	3.502
Ile								
	ATA	2.583	2.583	1.941	2.167	0.809	11.176	4.041
	ATC	2.642	2.667	**1.674**	***1.990***	**0.783**	**9.941**	**3.465**
	ATT	2.667	2.667	1.809	1.990	0.816	10.972	3.773
Lys								
	AAA	4.083	*4.083*	2.220	2.671	1.247	11.913	4.478
	AAG	**4.017**	*4.083*	**2.005**	**2.626**	**1.176**	**10.676**	**3.999**
Asn								
	AAC	**2.933**	*3.000*	**1.841**	*2.250*	**1.009**	**11.314**	**4.045**
	AAT	3.000	*3.000*	2.007	*2.250*	1.082	12.510	4.451
Gln								
	CAA	3.625	*3.833*	1.992	2.632	1.003	10.693	3.906
	CAG	**3.558**	*3.833*	**1.800**	**2.562**	**0.947**	9.585	**3.497**
His								
	CAC	**2.667**	*2.833*	**1.650**	*2.203*	**0.826**	**10.144**	**3.560**
	CAT	2.733	*2.833*	1.798	*2.203*	0.880	11.217	3.918

continued

Table 2. Continued

		No bias,[a] 1 mutation only		No bias		A[a]	B[a]	C[a]
		χ[b] = 0.9	χ = 1	χ = 0.9	χ = 1	(χ = 0.9)	(χ = 0.9)	(χ = 0.9)
Glu								
	GAA	3.625	*3.833*	2.001	2.639	1.082	10.715	4.025
	GAG	**3.575**	*3.833*	**1.807**	**2.584**	**1.016**	**9.602**	**3.595**
Asp								
	GAC	2.917	*3.083*	1.666	*2.244*	0.884	10.196	3.650
	GAT	2.967	*3.083*	1.816	*2.244*	0.948	11.273	4.015
Tyr								
	TAC	**5.933**	*6.250*	**2.755**	*3.087*	**2.885**	**11.133**	**4.440**
	TAT	6.250	*6.250*	2.975	*3.087*	3.091	12.295	4.860
Cys								
	TGC	**5.667**	*6.083*	**2.366**	*2.969*	**2.320**	**10.066**	**3.935**
	TGT	5.883	*6.083*	2.552	*2.969*	2.464	11.119	4.310
Phe								
	TTC	**3.517**	*3.583*	**1.956**	*2.415*	**0.968**	**11.368**	**4.290**
	TTT	3.583	*3.583*	2.130	*2.415*	1.027	12.567	4.715

Note. The most used codons for *Drosophila melanogaster* are in boldface; the less used codons are underlined. MD values are typed in boldface/underlined when they correspond to the most/less used codon and in italics when the case is ambiguous. $\omega_{AA/STOP} = -10$.
[a]A, mistranslation only; B, mutation only; C, mutation and mistraslation (parameters as in Table 1).
[b]χ = CG/AT mutation ratio.

lan's matrix $\omega_{Lys/Ile} = \omega_{Lys/Met}$, so even Lys has no variance for MD after one mutation). Introducing a bias in the mutation rate for CG versus AT, or in the transition/transversion ratio, produces a higher variability in the MD values.

The measures of MD also depend on the "similarity" with the termination signal ($\omega_{AA/STOP}$). The relative differences of MD values for synonymous codons depend slightly on $\omega_{AA/STOP}$ (Fig. 3). Therefore, in the analysis I use $\omega_{AA/STOP}$ values ranging from 0 to −50 (0 is the lowest value in McLachlan's matrix, in which scores range from 0 to 9).

Error Minimization

It can already be noted (Table 2, Fig. 4) that in the genome of *D. melanogster*, in most cases the optimal codons (the ones with the lowest MD values) correspond to the most used codons and the worst codons (the ones with the highest MD values) correspond to the less used codons. The main exception is apparently Ser, for which AGC is the most frequent codon, but the MD value for AGC is not the lowest (it is the lowest, however, if the transition/transversion bias is considered). Also, the preferred codons for Arg and Val do not correspond to the lowest MD values, the less used codons in Gly and Pro do not correspond to the highest MD values, and the values are inverted in Asp.

Codon usage frequencies (obtained from the Codon Usage Database) are quite rough measures of codon preference in the whole genome. Moriyama and Powell (1997) use, as a measure of codon preference in *Drosophila*, the correlation between the degree of bias for more than 1000 genes and the frequency of T-, C-, A-, or G-ending codons within each gene (positive values indicate codons that are increasingly used as the codon usage bias for that amino acid increase). Moreover, to avoid stochastic fluctuations, Moriyama and Powell (1997) examine only genes longer than 200 bp. These values should therefore give more reliable information about the preference of synonymous codons than mere usage percentages. When these measures are used there is a difference with Ser: in this case the preferred codon is TCC, which indeed corresponds to the lowest MD value with no transition/transversion bias, while the lowest value in Moriyama and Powell is for TCA, which corresponds to the highest MD value. There are also minor differences between codon frequency (from the Codon Usage Database) and Moriyama and Powell measures for Arg and Gly.

The pattern for rodents is quite similar. For a quantitative measure of error minimization from now on I use mainly the wR_N values (described in the Methods section), but results obtained with R_N are quite similar. If error minimization based on the similarity between amino acids plays no role in the evolution of the genes analyzed, then wR_N (or R_N) values should have a Gaussian distribution centered in zero. A deviation from this distribution will indicate a tendency toward error minimization (if wR_N values are more frequent toward −1) or maximization (toward +1).

Figure 5 shows that there is a strong prevalence of low wR_N values, that is, the genes analyzed prefer codons that reduce the impact of errors at the protein level, both for *Drosophila* and for rodents. A very

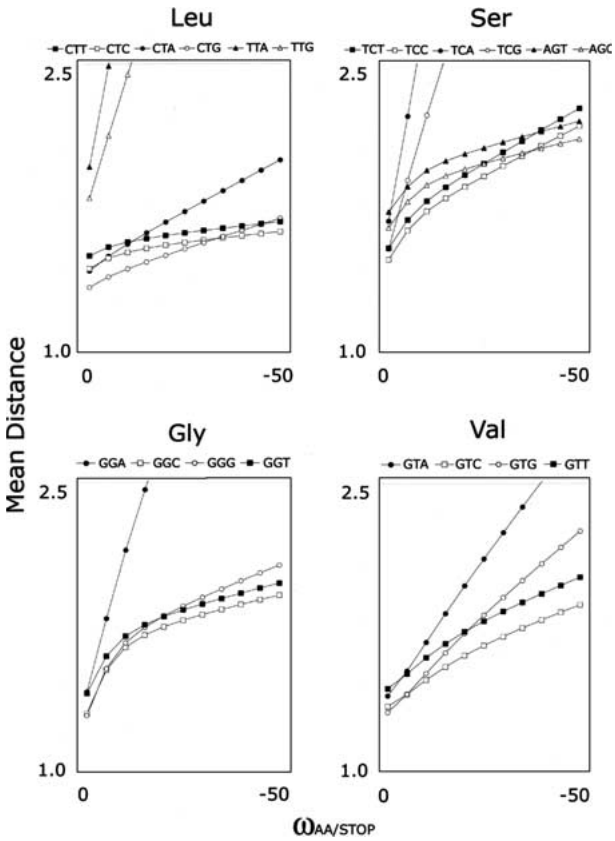

Fig. 3. MD values for the codons of Leu, Ser, Gly, and Val, for different values of $\omega_{AA/STOP}$ (no transition/transversion bias; CG/AT mutation ratio = 0.9).

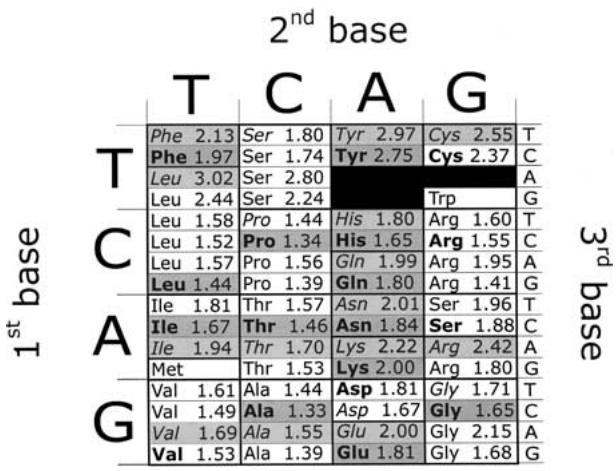

Fig. 4. The standard genetic code with MD values ($\omega_{AA/STOP} = -10$; CG/AT mutation ratio = 0.9; no transition/transversion bias). Values at the right in each cell are the MD values for the corresponding codon. Codons in boldface/italics are the most/less used in *Drosophila melanogaster*. Cases where the lowest/higest MD values correspond to the most/less used codons are marked in dark/light gray.

similar pattern is obtained with R_N values (not shown; see Appendix for *Drosophila* wR_N values; values for rodents are available on request). Note that even if the frequency of very low wR_N values (close to −1) is not high, this does not necessarily mean that error minimization is weak; the comparison must be done with the expected random distribution or, better, with the distribution obtained with random sequences with the same CG content (see next section).

These results do not change much when a bias in the transition/transversion rate is considered for mutation, mistranslation, or both or different values of the similarity score with the termination signal ($\omega_{AA/STOP}$), ranging between 0 and −50, are used. The CG/AT mutation ratio also does not seem to have a strong influence, especially for rodents (Fig. 6). Finally, when only one mutation event is considered, the distribution of wR_N values is less biased toward error minimization (Fig. 6). This might mean that multiple mutation events, that is, selection on mutations rather than on mistranslation, are important in error minimization at the protein level.

In general, it seems that both *Drosophila* and rodent genes prefer codons that reduce the impact of errors.

Error Minimization and CG Content

It is possible, in principle, that wR_N (or R_N) values are correlated with CG content, and a bias in their values might simply be the by-product of a mutational bias. To investigate this possibility, I produce random sequences (82 for *Drosophila*,432 for rodents), 300 codons long, with exactly the same CG content as the real genes analyzed here (for rodents I use the CG content of the mouse genes, but results are virtually identical for rat genes). If CG content, irrespective of the specific codons used, was the main cause of the pattern observed, then the distribution of wR_N (or R_N) values for these random sequences would be similar to the distribution of wR_N (or R_N) values for the real genes.

Figure 5 shows that wR_N values for sequences with "random" codons are almost Gaussian, not biased and very different from the wR_N values of the real genes. A similar result is obtained when using position-specific CG content.

Moreover, CG content is not correlated with wR_N in the real genes (see Table 3). These results do not change much with a bias in the transition/transversion or CG/AT mutation rate or with different values of the similarity score with the termination signal ($\omega_{AA/STOP}$), ranging between 0 and −50 (Table 3).

Since wR_N values are obtained weighting the frequencies of amino acids, this may lead to an underestimation of the importance of some amino acids and an overestimation of some others; it may be thought, therefore, that these results depend on amino acid frequencies, and that the random sequences fail to show error minimization because they have a deficit in some amino acid. However, though I have discussed mainly results obtained with wR_N

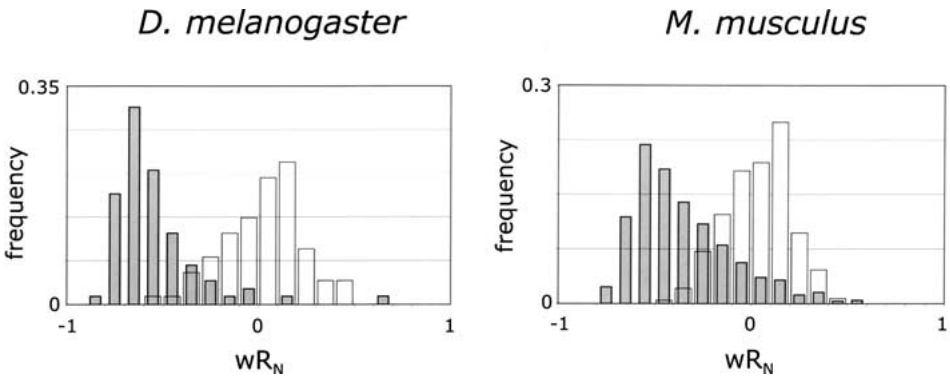

Fig. 5. Distribution of the wR_N values for the real genes (gray) and for random sequences with the same CG content (white).

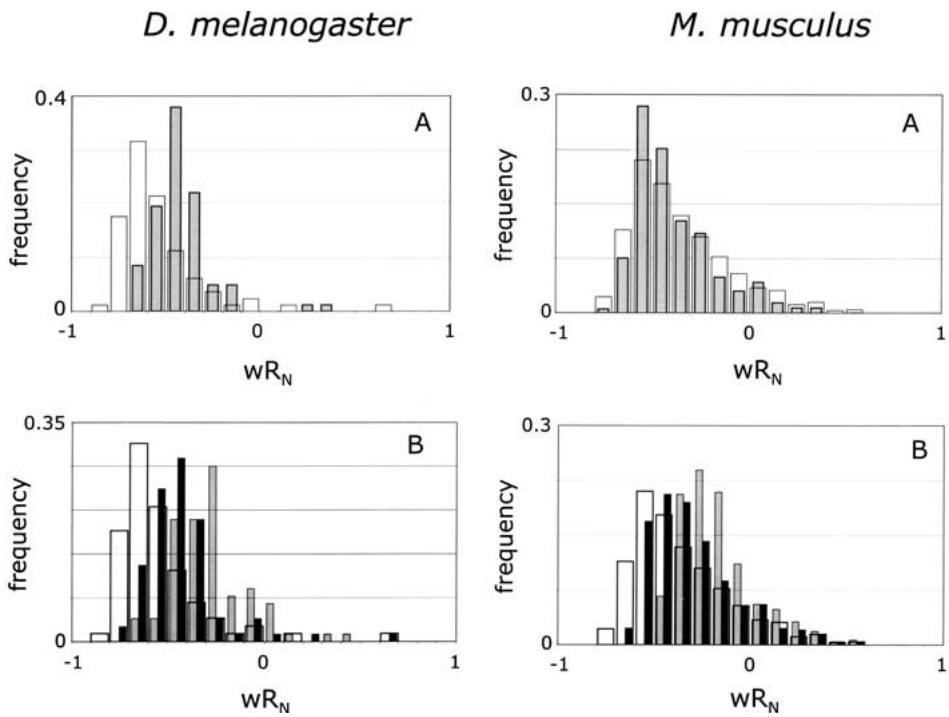

Fig. 6. Distribution of the wR_N values for the real genes obtained with different CG/AT mutation ratios (**A**; white = 0.9, gray = 1) or with a different number of mutation events (**B**; white = 10, black = 2, gray = 1) ($\omega_{AA/STOP} = -10$; no transition/transversion bias; CG/AT mutation ratio = 0.9).

values, the results obtained with R_N values are very similar. R_N values do not depend on the frequency of amino acids, as all amino acids are weighted the same. The use of R_N may also be questionable, however, because the more frequent an amino acid, the more likely it is to influence selection on that protein. In any case, the results obtained using wR_N and R_N are very similar.

Therefore, it seems that it is the specific choice of codons, and not solely CG content on amino acid composition, that determines error minimization. That is, error minimization is not due to mutation bias.

Error Minimization and Codon Usage Bias

If codon usage bias is influenced by the degree of error minimization, then it should be possible to find a correlation between the variance of MD values for synonymous codons and the contribution of each individual amino acid to the overall codon usage bias. Moreover, it should be possible to find a correlation between the degree of error minimization and the degree of codon usage bias for individual genes. I analyze these two aspects in turn.

Variance of MD and Contribution to the Bias. Moriyama and Powell (1997) report the contribution to the bias of each amino acid for *Drosophila*. A correlation between the variance of MD values for synonymous codons (the measure developed here) and the contribution of the corresponding amino acid to codon usage bias (taken from Moriyama and Powell 1997) is expected because natural selection for error minimization, if it actually

Table 3a. Correlation between wR_N and CG content, ENC, K_a, and K_s for *Drosophila*

	Correlation of wR_N with[a]			
Parameters[b]	%CG	ENC	K_s	K_a
$\omega_{AA/STOP} = 0$	−0.02	**0.51****	**0.44****	**0.54*****
$\omega_{AA/STOP} = -10$	−0.11	**0.60****	**0.56*****	**0.43***
$\omega_{AA/STOP} = -20$	−0.18	**0.61****	**0.58*****	**0.38***
$\omega_{AA/STOP} = -50$	−0.21	**0.60****	**0.57*****	**0.33***
Bias = A[c]	−0.12	**0.56****	**0.51*****	**0.35***
Bias = B[c]	−0.02	**0.51****	**0.54*****	**0.45****
Bias = C[c]	−0.06	**0.56****	**0.50****	**0.47****
$\chi = 1$	−0.21	**0.48****	**0.55*****	**0.44****
1 mutation	0.21	**0.32***	0.24	**0.44****
2 mutations	−0.14	**0.51****	**0.53*****	**0.40***

Table 3b. Correlation between wR_N and CG content, ENC, K_a, and K_s, for the 432 rodent genes

	Correlation of wR_N with[a]			
Parameters[b]	%CG	ENC	K_s	K_a
$\omega_{AA/STOP} = 0$	0.04	**0.33****	−0.02	**0.19*****
$\omega_{AA/STOP} = -10$	−0.03	**0.32****	−0.01	**0.20*****
$\omega_{AA/STOP} = -20$	−0.06	**0.29****	−0.01	**0.20*****
$\omega_{AA/STOP} = -50$	−0.08	**0.28****	0.00	**0.20*****
Bias = A[c]	−0.04	**0.34****	0.00	**0.21*****
Bias = B[c]	0.04	**0.34****	−0.01	**0.19*****
Bias = C[c]	−0.03	**0.32****	−0.01	**0.20*****
$\chi = 1$	**− 0.18*****	**0.24****	−0.00	**0.23*****
1 mutation	0.04	**0.31****	0.01	**0.10***
2 mutations	−0.05	**0.32****	0.01	**0.21*****

Note. The correlation with %CG and with ENC for *Drosophila* is calculated for all 82 genes; the correlation with K_s, and K_a, only for a subset of 38 genes (see Appendix). Significant correlations are in boldface (*$p < 0.05$; **$p < 0.005$; ***$p < 0.001$; ****$p < 0.0001$).
[a]%CG, CG content; ENC, effective number of codons; K_s and K_a, rate of synonymous and nonsynonymous substitution.
[b]Unless otherwise stated, $\omega_{AA/STOP} = -10$; χ (=CG/AT mutation ratio) = 0.9; no bias in the transition/transversion ratio; MD values calculated up to the 10th mutation event.
[c]Parameters as in Table 1.

plays a role in the evolution of coding sequences, will be stronger for amino acids whose codons have a high variance in the capacity to minimize errors.

Table 4 shows that there is a positive correlation between the variance of MD values and the contribution of each amino acid to the overall codon usage bias. There is a weak, but significant, correlation if one considers all the 18 degenerate amino acids; the correlation is stronger if one does not consider twofold degenerate amino acids. The transition/transversion ratio and the $\omega_{AA/STOP}$ value do not seem to affect the results significantly. A higher mutation rate for AT versus CG is necessary for a significant correlation if one considers all 18 degenerate amino acids. The reason why the correlation is more significant when we do not consider twofold degenerate amino acids is probably because variance for twofold degenerate amino acids is usually very low or zero.

Degree of Error Minimization and Degree of Codon Usage Bias. If codon usage is influenced by selection for error minimization at the protein level, then a positive correlation is expected between the degree of codon usage bias and the degree of error minimization. That is, genes with a highly biased codon usage should be those with a strong preference for codons that minimize the effect of errors.

For "random" sequences with the same CG content as the 82 *Drosophila* genes, the degree of error minimization is not correlated with the degree of codon usage bias (ENC values) ($r = -0.07$, $p = 0.70$, for wR_N values; $r = 0.09$, $p = 0.62$, for R_N values) ($\omega_{AA/STOP} = -10$; CG/AT mutation bias = 0.9; no transition/transversion bias). On the other hand, there is a positive correlation (Fig. 7, Table 3a) between wR_N (or R_N) values and the degree of codon usage bias (ENC) for the real 82 *Drosophila* genes considered.

A correlation between the degree of codon usage bias (ENC) and wR_N is also not found in 432 random sequences with the same CG content as the rodent genes ($r = 0.01$, $p = 0.79$) ($\omega_{AA/STOP} = -10$; CG/AT mutation bias = 0.9; no transition/transversion bias). For the real 432 mouse (or rat) genes, there is instead a positive correlation (Fig. 7, Table 3b) be-

Table 4. Correlation between variance of MD values for each amino acid and its contribution to codon usage bias

		Twofold degenerate amino acids included		Twofold degenerate amino acids excluded	
Bias[a]	$\omega_{AA/STOP}$	χ[b] = 0.9	χ = 1	χ = 0.9	χ = 1
No					
	0	0.42	0.47	**0.81****	**0.86*****
	−10	**0.47***	0.46	**0.83*****	**0.78***
	−50	**0.53***	0.41	**0.85*****	**0.71***
C					
	0	**0.52***	0.43	**0.81****	**0.79****
	−10	**0.47***	0.41	**0.83*****	**0.72***
	−50	**0.51***	0.40	**0.84*****	**0.68***

Note. Only degenerate amino acids are included. *$p < 0.05$; **$p < 0.01$; ***$p < 0.005$ (the p values differ because the number of amino acids on which the correlation is calculated is different).
[a]Bias in the transition/transversion ratio: parameters as in Table 1.
[b]χ = CG/AT mutation ratio.

tween the degree of codon usage bias (ENC) and wR_N (or R_N).

The result does not change drastically when "random" sequences with the position-specific CG content are used or when different values of $\omega_{AA/STOP}$ or for the transition /transversion ratio are considered. When only one or two mutation events are considered, the correlation with ENC still holds (Table 3).

The existence of a positive correlation between ENC and wR_N means that as codon usage bias increases, the preference for optimal codons increases too. This suggests that codon usage bias is due, at least in part, to selection for error minimization at the protein level.

It may be thought that the difference between the random sequences and the real sequences is that the random sequences have no codon usage bias, and that error minimization is intrinsic in codon usage bias. To investigate this possibility, I also analyze sequences (82 for *Drosophila*, 432 for rodents—I use the mouse sequences, but results are virtually identical for rat) in which the frequency of synonymous codons is switched at random. In this case ENC values are exactly the same as in the real genes, but the relative frequencies of synonymous codons is modified.

If ENC was intrinsic in the measure of wR_N (that is, if biased wR_N values were simply due to biased codon usage), then the distribution of wR_N (or R_N) values for these sequences would be similar to the distribution of wR_N (or R_N) values of the real genes. On the contrary, the distribution of wR_N values for these sequences with rearranged codon frequencies is not biased toward error minimization (it is similar to a Gaussian distribution; not shown), and ENC is not correlated with the degree of error minimization (*Drosophila*: $r = 0.03$, $p = 0.75$, for wR_N; $r = -0.01$, $p = 0.95$, for R_N; mouse: $r = -0.04$, $p = 0.42$, for wR_N; $r = 0.01$, $p = 0.84$, for R_N).

Error Minimization and Rates of Nucleotide Substitution

If error minimization plays a role in the evolution of coding sequences, then the rate of nonsynonymous substitutions (K_a) could be correlated with the degree of error minimization (wR_N or R_N), while this would not necessarily be the case for the rate of synonymous substitution (K_s), because selection for error minimization at the protein level would occur only in the event of a nonsynonymous substitution. This means, in other words, that highly conserved genes are expected to prefer strongly codons that minimize the impact of errors due to mutation or mistranslation, while poorly conserved genes are expected to have more relaxed preferences.

I analyzed 38 of the *Drosophila* genes used for the previous analysis (the ones for which K_a and K_s measures were available—see Appendix) and found a positive correlation between R_N or wR_N values (Fig. 8) and the rates of nucleotide substitution, both synonymous and nonsynonymous. Changing the parameters does not change the results drastically (Table 3a). However, when only one mutation event is considered, the correlation with K_s disappears. Note that ENC is correlated with K_s ($r = 0.69$, $p < 0.0001$) but not with K_a ($r = 0.10$, $p = 0.55$) and that K_a and K_s are not correlated ($r = 0.19$, $p = 0.24$).

The observed correlation with K_s in *Drosophila* is rather unexpected. In rodents, however, for the 432 real genes analyzed here, K_a and wR_N are correlated but K_s and wR_N are not (for both mouse and rat; see Fig. 9 and Table 3b); even when K_4 is used instead of K_s there is no correlation with wR_N (for both mouse and rat, $r = -0.01$, $p = 0.91$). Different values for the transition/transversion bias or for the $\omega_{AA/STOP}$ score have negligible effects on the results (Table 3b). When only one or two mutation events are consid-

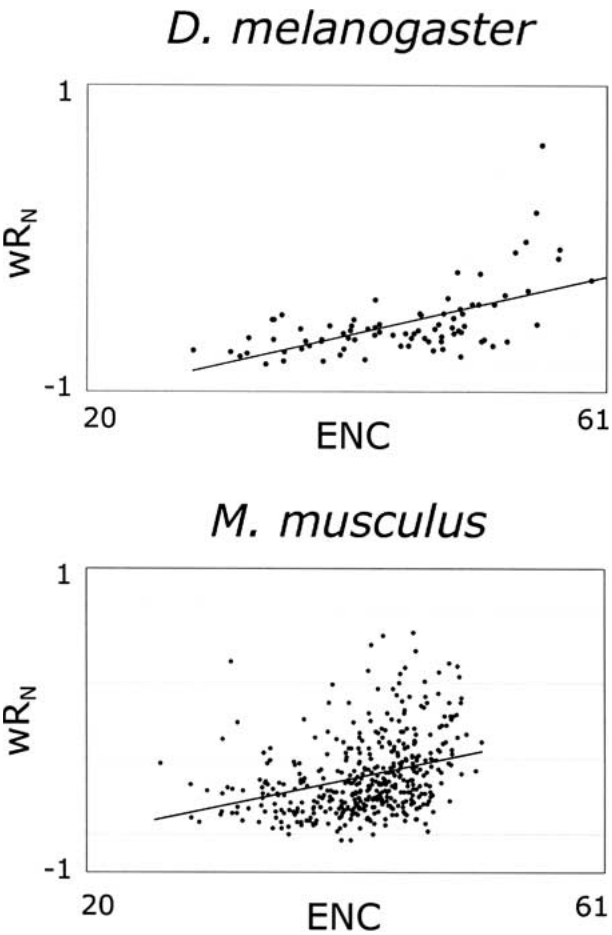

Fig. 7. wR_N plotted against ENC for the 82 *Drosophila* and for the 432 mouse genes analyzed. The linear regression line is shown. $\omega_{AA/STOP} = -10$; no transition/transversion bias; CG/AT mutation ratio = 0.9.

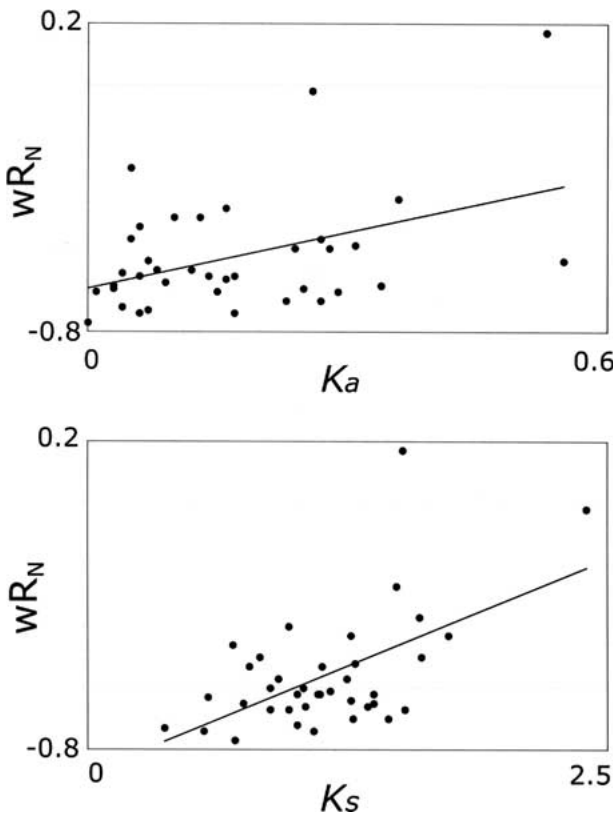

Fig. 8. wR_N plotted against K_a and K_s for the 38 *Drosophila* genes analyzed. The linear regression line is shown. $\omega_{AA/STOP} = -10$; no transition/transversion bias; CG/AT mutation ratio = 0.9.

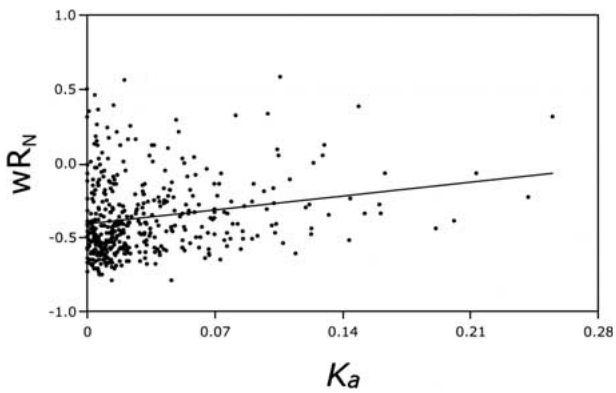

Fig. 9. wR_N against K_a for the 432 mouse genes analyzed. The linear regression line is shown. $\omega_{AA/STOP} = -10$; no transition/transversion bias; CG/AT mutation ratio = 0.9.

ered, the correlation with K_a still holds, though with only one mutation it is weaker (Table 3b).

These results suggest that error minimization at the protein level plays a role in the evolution of coding sequences. An analysis of the single genes reveals that, for example, actin genes (known to be among the most conserved genes because of their fundamental role in the cytoskeleton and in the processes of muscle motility) have some of the lowest wR_N values (for example, $wR_N = -0.79$ for the mouse α1-actin); the highest wR_N value in mouse, on the contrary, is scored by the ST2 gene ($wR_N = +0.58$), which has a high similarity to the interleukin-1 receptor, known to be among the most variable genes.

Importance of Amino Acid Similarity Matrices

For all the previous analyses I have used an amino acid similarity matrix based on chemical properties (McLachlan 1971). However, using different matrices may lead to very different results when analyzing the level of error minimization of the genetic code (Haig and Hurst 1991; Archetti 2004). Therefore, I have checked the robustness of the results obtained here with McLachlan's matrix, by using different similarity matrices under different sets of parameters.

The main results do not change (Table 5); the two main differences are a weak correlation between wR_N and CG content when a matrix based on hydrophobicity is used and a slightly weaker correlation between wR_N and K_a when the PAM 74-100 matrix is used. All in all it seems that using different matrices does not affect the main results.

Table 5. Correlation between wR_N and CG content, ENC, K_a and K_s for the 432 mouse genes when different similarity matrices are used

Matrix[a]	Correlation of wR_N with[b]			
	%CG	ENC	K_s	K_a
DAYM780301 (−20)	−0.08	**0.29******	0.00	**0.20******
BENS940103 (−1)	0.03	**0.33******	−0.02	**0.15****
OVEJ920101 (−20)	−0.04	**0.31******	0.00	0.20****
RIER950101 (−100)	−0.01	**0.32******	0.00	**0.19******
RISJ880101 (−8)	0.02	**0.32******	0.00	**0.18******
GEOD900101 (−10)	**−0.12***	**0.27******	0.00	**0.20******

Note. Significant correlations are in boldface (*$p < 0.05$; *$p < 0.005$; ****$p < 0.0001$).

[a]The label corresponds to the AAindex accession number (http://www.genome.ad.jp/dbget/AAindex/list_of_matrices), followed (in parentheses) by the value of the similarity score with the stop codon ($\omega_{AA/STOP}$) used here. DAYM780301: log-odds matrix for 250 PAMs (Dayhoff et al. 1978); BENS940103: log-odds scoring matrix collected in 74-100 PAM (Benner et al. 1994); OVEJ920101: STR matrix from structure-based alignments (Overington et al. 1992); RIER950101: hydrophobicity scoring matrix (Riek et al. 1995); RISJ880101: scoring matrix (Risler et al. 1988); GEOD900101: hydrophobicity scoring matrix (George et al. 1990).

[b]%CG, CG content; ENC, effective number of codons; K_s and K_a, rates of synonymous and nonsynonymous substitution. CG/AT mutation ratio = 0.9; no bias in the transition/transversion ratio; MD values calculated up to the 10th mutation event.

Discussion

The hypothesis presented here is based on the concept of error minimization at the protein level, a concept that has already been used to study the evolution of the genetic code. The idea discussed in this paper does not deal with the evolution of the genetic code itself but predicts that, given the genetic code (in which codons are *arranged* to minimize errors), within each synonymous family, the most used codons are the ones that, after mutation or mistranslation, keep on coding for the same amino acid or for amino acids with similar properties; that is, that codons are *used* in a way that minimizes errors.

To my knowledge this is the first time this hypothesis has been tested, though some hints in this direction have been put forward by Modiano et al. (1980) and McPherson (1988). McPherson (1988) suggested that codon preference may reflect mistranslational constraints, while Modiano et al. (1980) observed that codons that can mutate to termination codons (''pretermination'' codons) by point mutation are avoided in human α and β-globin genes and suggested that it be so to avoid mutations with drastic effects. This hypothesis, however, has not received much attention. In his book on the neutral theory of natural selection, Kimura (1983) mentions it only to dismiss it in few words: ''It seems to me that the selective advantage coming from such strategy [avoiding pre-termination codons] is too small (presumably the order of the mutation rate) to be effective in the actual course of evolution.'' However, the work eventually done on codon usage evolution has shown that selection is in fact able to operate at a very fine scale; as Duret (2002) puts it, selection is able to act on codon usage even if it leads to a synonymous substitution on a single gene in a whole genome. Moreover, if selection shaped the genetic code, it is not inconceivable that it also shaped codon usage, though selection for codon usage may have occurred after a genetic code had been fixed (or perhaps coevolved).

I have developed a theoretical measure of the capacity of each codon to minimize the effects of errors due to mutation and mistranslation. I have then used these values to measure the degree of error minimization for individual genes of *Drosophila melanogaster* and rodents.

The main result is (i) that the distribution of these values is not random: both *Drosophila melanogaster* and rodents prefer codons that lead to the minimization of errors at the protein level, that is, codons that, after mutation or mistranslation, keep on coding for the same amino acid or for amino acids with similar properties. Therefore, error minimization at the protein level may affect not only the evolution of the genetic code itself, as has long been discussed, but also the frequency of synonymous codons.

I have also shown (ii) that it is possible to exclude that GC content is solely responsible for the pattern observed, as the degree of error minimization is not correlated with CG content, and random sequences with the same CG content do not lead to error minimization. This suggests that selection, and not mutation bias, is responsible for the observed pattern of error minimization.

I have then shown (iii) that the higher the bias in codon usage, the stronger the tendency to use codons that minimize errors, and that (iv) the higher the contribution of an amino acid to codon usage bias, the higher the variance of the capacity to minimize errors for its synonymous codons. This suggests that codon usage bias can be explained by the preferential usage of codons that minimize errors. Other evidence (McVean and Vieira 2001) suggests that Leu (the amino acid shown here to have the highest variance

for MD values) is the amino acid on which the influence of selection is strongest.

Classically, two models have been proposed to explain codon usage bias: the selective model, which postulates that there is a coadaptation of codon usage and tRNA abundance; and the neutral model, which postulates that the bias results from biases in the mutational process (see recent review by Duret 2002). In unicellular organisms, codon usage bias is usually thought to be due to selection for translation efficiency: there is preferential usage of codons that bind more efficiently to the corresponding tRNA or that are recognized by a more abundant tRNA (Post et al. 1979; Ikemura 1981, 1982, 1985, 1992). In this way proteins are produced more rapidly and with fewer errors. In mammals codon usage may be related to isochores, segments several kilobases long of similar CG content but with substantial differences in CG content between them (Bernardi et al. 1985). In *Drosophila* it has been suggested that codon usage bias is caused by selection on synonymous sites for translational accuracy (Akashi 1994; Powell and Moriyama 1997), but there is evidence that other factors, for example, recombination and compositional bias, influence codon usage in *Drosophila* (Marais et al. 2001; Hey and Kliman 2002). In general, therefore, it is thought that codon usage reflects a combination of different forces. The results presented here show that it is possible that selection on error minimization plays a role also in the evolution of codon usage bias.

The results seem rather robust, not depending drastically on the transition/transversion ratio or on the CG/AT mutation ratio. I have also shown that they do not depend on the amino acid similarity matrix used. Matrices based on observed substitutions are usually considered better than chemical similarity matrices, but they are derived from sequences that include themselves codon usage bias; for this reason these matrices may not provide a fair measure (though it is not demonstrated that this is the case) to study the relationship between error minimization and codon usage bias. Di Giulio (2001) used a similar argument to critique the use of the PAM 74-100 matrix in studies of the genetic code origin. Note, however, that there is a difference: in the study of the origin of the genetic code, the similarity matrix used must be independent from the structure of the genetic code, otherwise the analysis would be tautologous (*genetic code structure* explained with a matrix that depends on the *genetic code structure*). In this case, instead, even if the similarity matrix depended on the genetic code structure, this would not affect the analysis; indeed the hypothesis I want to test *is* that codon usage depends on the similarity matrix and on the arrangement of codons in the genetic code (*codon usage bias* explained with a matrix that may depend on the *genetic code structure*).

Error minimization seems to be weaker when only one mutation is considered instead of multiple mutation events. This might suggest that mutation is more important than mistranslation in natural selection for the minimization of errors at the protein level. If selection during mistranslation is important, there might be a correlation between error minimization and gene length or gene expression. Studies on the evolution of the genetic code have been rather vague as to whether the main mechanism of selection for error minimization acts during translation or mutation. The analysis presented here also does not lead to a conclusive answer concerning the relative importance of mutation and mistranslation for error minimization due to codon usage. The observation that error minimization is stronger with multiple mutation events and other preliminary results of mine (M.A., unpublished) suggest that mutation rather than mistranslation is the driving force of error minimization at the protein level. If mistranslation is important in selection for error minimization, then gene expression levels should be correlated with the degree of error minimization, as highly expressed genes should be the ones under stronger selective pressure to reduce errors. This is a point that would be worth investigating.

Finally I have shown (v) that the degree of error minimization varies according to the rate of evolution of the gene. That is, genes with a low rate of evolution tend to use more codons that minimize the effect of errors. There is, however, a difference between rodents and *Drosophila*. In rodents only the rate of nonsynonymous substitution is correlated with the degree of error minimization, while in *Drosophila* the rate of synonymous substitution is also correlated with the degree of error minimization. Possibly this difference is due to the different methods used to calculate the substitution rates. For the 432 rodent genes used here, K_a and K_s have been obtained by Smith and Hurst (1997) by a maximum-likelihood approach (ML), while for the *Drosophila* genes K_a and K_s had been obtained by Akashi (1994) using the approximate method of Nei and Gojobori (1986) (NG). It has been shown (Yang and Nielsen 1998, 2000; Bielawski et al. 2000) that ML methods are superior to approximate methods and that the NG method may lead to biased estimates of substitution rates. It is possible that the correlation between wR_N and K_s observed in *Drosophila* is due to the usage of approximate methods in the calculation of the substitution rates. It seems that it cannot be just an artifact of a correlation of K_s with K_a, because in that case (Akashi 1994) K_s and K_a are not correlated. The alternative possibility is that wR_N and K_s are correlated in *Drosophila* but not in mammals. To decide about these two alternatives a reexamination of *Drosophila* genes, with K_s and K_a, derived by ML methods would be useful.

What can we deduce from this correlation? Highly conserved genes are expected to have a high degree of error minimization if the minimization of errors at the protein level is important, because any change in these genes would result in a change in the function of the protein and, as a consequence, if the change is deleterious, a reduction in fitness. Less conserved genes, instead, can have more relaxed preferences on codon usage because changes at the protein level will not have a drastic impact on fitness. On the other hand, it could be possible to argue that genes evolving faster should show error minimization even more clearly, because they are more likely to have evolved to their current state by means of small mutations. This is a basic question in molecular evolution. Amino acid replacement changes should be expected to have greater influence on genes whose structure and function is "important." Unfortunately, as Kreitman (1996) points out, the problem is not so simple: constraining selection alone cannot explain protein evolution because the observed differences in proteins must, by definition, be ones that have escaped constraining selection. The study presented here does not help understanding this problem, but further studies on error minimization might provide some useful insights. The method itself could be used to detect selection from single DNA sequences.

The degree of error minimization could also depend on gene function. Genes of parasitic organisms that code for proteins involved in host–parasite interactions, for example, in some cases prefer codons that increase the effect of errors to escape the host defenses more easily (M.A., upublished). This might lead to different optimal codons for different species, according to their life history.

Indeed, the explanation of codon usage bias presented here, based on the structure of the genetic code, should predict a unique set of optimal codons for all the species that use the same code. I have shown here that results for *Drosophila* are consistent with those for rodents. A comprehensive analysis of many more species will be presented in another paper, with a more precise method to detect selection from codon usage analysis of single nucleotide sequences (M.A., unpublished).

Appendix

Table AI. wR_N values for different values of the parameters for 82 *Drosophila melanogaster* genes

Gene	$wR_{N\text{-}0}$	$wR_{N\text{-}10}$	$wR_{N\text{-}20}$	$wR_{N\text{-}50}$	$wR_{N\text{-}10(A)}$	$wR_{N\text{-}10(B)}$	$wR_{N\text{-}10(C)}$
*ade3	−0.42	−0.43	−0.46	−0.43	−0.38	−0.51	−0.49
*Adh	−0.67	−0.74	−0.73	−0.70	−0.61	−0.70	−0.72
*Adhr	−0.24	−0.27	−0.25	−0.21	−0.09	−0.27	−0.27
*Amy−d	−0.73	−0.73	−0.69	−0.65	−0.62	−0.78	−0.73
*Antp	−0.76	−0.72	−0.68	−0.64	−0.62	−0.81	−0.75
*Aprt	−0.55	−0.64	−0.62	−0.60	−0.49	−0.63	−0.63
*bcd	−0.73	−0.60	−0.54	−0.49	−0.56	−0.74	−0.65
*boss	−0.42	−0.43	−0.41	−0.38	−0.37	−0.43	−0.43
bw	−0.80	−0.76	−0.70	−0.65	−0.68	−0.86	−0.78
cdc37	−0.77	−0.79	−0.77	−0.74	−0.72	−0.81	−0.79
*Cp15	−0.37	−0.57	−0.60	−0.61	−0.43	−0.49	−0.54
*Cp16	−0.44	−0.53	−0.55	−0.55	−0.47	−0.46	−0.50
*Cp18	−0.38	−0.50	−0.51	−0.50	−0.44	−0.43	−0.46
*Cp19	−0.49	−0.65	−0.66	−0.65	−0.55	−0.60	−0.63
*Cp36	−0.58	−0.63	−0.60	−0.56	−0.56	−0.63	−0.63
csw	−0.77	−0.63	−0.62	−0.57	−0.60	−0.82	−0.66
cybt-b5	−0.48	−0.66	−0.50	−0.48	−0.51	−0.54	−0.53
Ddx1	−0.74	−0.77	−0.74	−0.71	−0.65	−0.79	−0.77
dpp	−0.76	−0.70	−0.63	−0.56	−0.58	−0.83	−0.74
e(r)	−0.57	−0.59	−0.56	−0.53	−0.51	−0.63	−0.61
*elav	−0.75	−0.65	−0.57	−0.50	−0.57	−0.77	−0.68
*en	−0.60	−0.62	−0.59	−0.55	−0.53	−0.70	−0.64
esc	−0.61	−0.59	−0.55	−0.52	−0.51	−0.67	−0.61
*Est-6	0.01	−0.02	−0.02	0.00	0.05	−0.04	−0.03
exu	−0.50	−0.49	−0.46	−0.43	−0.46	−0.51	−0.48
Fbp2	−0.75	−0.80	−0.78	−0.75	−0.62	−0.82	−0.80
*Fmrf	−0.55	−0.53	−0.50	−0.47	−0.51	−0.59	−0.55
fu	−0.79	−0.73	−0.68	−0.63	−0.67	−0.82	−0.75
fz	−0.72	−0.67	−0.62	−0.57	−0.52	−0.74	−0.69
gl	−0.65	−0.65	−0.60	−0.56	−0.54	−0.75	−0.67
*Gld	−0.62	−0.62	−0.58	−0.53	−0.54	−0.65	−0.63
*Gpdh	−0.66	−0.66	−0.63	−0.60	−0.58	−0.67	−0.66
*h	−0.58	−0.60	−0.58	−0.55	−0.51	−0.66	−0.61

continued

Table AI. Continued

Gene	$wR_{N\text{-}0}$	$wR_{N\text{-}10}$	$wR_{N\text{-}20}$	$wR_{N\text{-}50}$	$wR_{N\text{-}10(A)}$	$wR_{N\text{-}10(B)}$	$wR_{N\text{-}10(C)}$
*hb	−0.58	−0.62	−0.54	−0.55	−0.47	−0.70	−0.65
His1	−0.21	−0.23	−0.25	−0.27	−0.31	−0.12	−0.19
*Hsp83	−0.76	−0.77	−0.80	−0.66	−0.74	−0.78	−0.79
janA	−0.31	−0.22	−0.19	−0.15	−0.15	−0.30	−0.24
janB	0.05	−0.07	−0.11	−0.14	0.03	0.01	−0.04
kni	−0.65	−0.65	−0.61	−0.57	−0.59	−0.73	−0.67
Kr	−0.36	−0.35	−0.35	−0.34	−0.34	−0.37	−0.35
ksr	−0.67	−0.67	−0.57	−0.58	−0.58	−0.77	−0.70
l(2)gl	0.69	0.61	0.49	0.50	0.61	0.69	0.62
l(2)not	−0.78	−0.72	−0.67	−0.63	−0.68	−0.81	−0.74
l(2)tid	−0.76	−0.67	−0.60	−0.55	−0.58	−0.79	−0.70
lama	−0.47	−0.49	−0.46	−0.43	−0.44	−0.50	−0.49
* *mam*	−0.43	−0.67	−0.63	−0.59	−0.61	−0.71	−0.69
Mlc1	−0.46	−0.53	−0.54	−0.53	−0.42	−0.50	−0.51
neur	−0.55	−0.58	−0.56	−0.53	−0.52	−0.55	−0.58
ninaE	−0.75	−0.70	−0.65	−0.61	−0.62	−0.77	−0.71
nos	−0.63	−0.56	−0.50	−0.45	−0.48	−0.67	−0.61
osk	−0.48	−0.43	−0.37	−0.32	−0.40	−0.54	−0.47
para	−0.41	−0.34	−0.29	−0.23	−0.31	−0.43	−0.37
* *Pcp*	−0.47	−0.40	−0.36	−0.31	−0.37	−0.47	−0.43
pdm2	−0.59	−0.56	−0.51	−0.46	−0.44	−0.71	−0.62
* *per*	−0.76	−0.66	−0.58	−0.52	−0.52	−0.83	−0.71
Rh2	−0.59	−0.58	−0.55	−0.51	−0.49	−0.61	−0.59
Rh3	−0.68	−0.57	−0.50	−0.43	−0.56	−0.65	−0.59
* *Rh4*	−0.62	−0.50	−0.44	−0.40	−0.46	−0.60	−0.52
**ro*	−0.68	−0.67	−0.62	−0.58	−0.55	−0.77	−0.70
run	−0.67	−0.59	−0.52	−0.46	−0.51	−0.73	−0.63
* *ry*	−0.57	−0.57	−0.54	−0.50	−0.52	−0.55	−0.57
salm	−0.63	−0.61	−0.57	−0.52	−0.57	−0.67	−0.63
**sev*	−0.51	−0.52	−0.48	−0.41	−0.42	−0.53	−0.55
* *sina*	−0.78	−0.67	−0.60	−0.55	−0.61	−0.77	−0.68
* *slbo*	−0.79	−0.74	−0.69	−0.65	−0.64	−0.85	−0.77
Sod	−0.60	−0.72	−0.72	−0.70	−0.48	−0.70	−0.72
sry-a	−0.56	−0.60	−0.58	−0.56	−0.50	−0.61	−0.60
sry-b	−0.86	−0.80	−0.75	−0.71	−0.77	−0.88	−0.82
sry-d	−0.87	−0.82	−0.77	−0.73	−0.78	−0.88	−0.82
* *su(Hw)*	−0.68	−0.70	−0.69	−0.67	−0.66	−0.72	−0.70
su(s)	−0.60	−0.56	−0.54	−0.52	−0.50	−0.64	−0.59
su(var)	−0.44	−0.39	−0.36	−0.33	−0.37	−0.45	−0.40
Tgfb-60A	−0.77	−0.66	−0.69	−0.65	−0.66	−0.82	−0.76
Tl	−0.55	−0.49	−0.44	−0.38	−0.45	−0.58	−0.52
* *tll*	−0.65	−0.61	−0.58	−0.55	−0.59	−0.66	−0.61
* *tra*	0.15	0.17	0.20	0.22	0.29	0.02	0.09
trx	−0.12	−0.13	−0.12	−0.11	−0.13	−0.05	−0.11
* *tub*	−0.37	−0.37	−0.35	−0.33	−0.30	−0.43	−0.40
* *Ubx*	−0.51	−0.46	−0.40	−0.34	−0.36	−0.57	−0.50
* *Uro*	−0.65	−0.62	−0.59	−0.55	−0.52	−0.67	−0.63
Yp1	−0.64	−0.75	−0.75	−0.74	−0.62	−0.64	−0.74
* z	−0.70	−0.70	−0.67	−0.63	−0.63	−0.75	−0.71

Note. $wR_{N\text{-}X(Y)}$ is the wR_N value calculated with $\omega_{\mathrm{AA/STOP}} = X$ and with values of the transition/transversion ratio corresponding to Y (see Table 1). Accession numbers for all the genes are given by Moriyama and Powell (1997). An asterisk precedes genes used in the analysis of correlation with K_s and K_a (taken from Akashi 1994).

Acknowledgments. Nick Smith provided the substitution rates used by Smith and Hurst (1997).

References

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster:* Natural selection and translational accuracy. Genetics 136:927–935

Archetti M (2004) Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. J Mol Evol (in press)

Benner SA, Cohen MA, Gonnet GH (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng 7(11):1323–1332

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

Bielawski JP, Dunn KA, Yang Z (2000) Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. Genetics 156:1299–1308

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins.. In: Dayhoff MO (ed) Atlas of protein sequence and structure, Vol 5, Supp1 3. National Biomedical Research Foundation, Washington, DC, p 352

Di Giulio M (2000a) Genetic code origin and the strength of natural selection. J Theor Biol 205:659–661

Di Giulio M (2000b) The origin of the genetic code. Trends Biochem Sci 25(2):44

Di Giulio M (2001) The origin of the genetic code cannot be studied using measurements based on the PAM matrix because this matrix reflects the code itself, making any such analyses tautologous. J Theor Biol 208:141–144

Duret L (2002) Evolution of synonymous codon usage in metazoans. Curr Opin Gen Dev 12:640–649

Epstein CJ (1966) Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. Nature 210:25–28

Freeland SJ, Hurst LD (1998) The genetic code is one in a million. J Mol Evol 47:238–248

Freeland SJ, Knight RD, Landweber LF (2000) Measuring adaptation within the genetic code. Trends Biochem Sci 25(2):44–45

George DG, Barker WC, Hunt LT (1990) Mutation data matrix and its uses. Methods Enzymol 183:333–351

Grantham R, Gautier C, Gouy M, Mercier R, Pavè A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acid Res 8:r49–r62

Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. J Mol Evol 33(5):412–417

Hey J, Kliman RM (2002) Interactions between natural selection, recombination and gene density in the genes of *Drosopila*. Genetics 160:595–608

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its 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 respective codons in protein genes Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158:573–597

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

Ikemura T (1992) Correlation between codon usage and tRNA content in microorganisms. In: Hatfield DL, Lee BJ, Pirtle RM (eds) Transfer RNA in protein synthesis. CRC Press, Boca Raton, pp 87–111

Kimura M (1983) The Neutral Theory of Natural Selection. Cambridge University Press, Cambridge

Knight RD, Freeland SJ, Landweber LF (1999) Selection, history and chemistry: The three faces of the genetic code. Trends Biochem Sci 24(6):241–247

Kreitman M (1996) The neutral theory is dead. Long live the neutral theory. BioEssays 18:678–682

Marais G, Mouchiroud D, Duret L (2001) Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci USA 98:5688–5692

McLachlan AD (1971) Tests for comparing related amino-acid sequences cytochrome c and cytochrome c 551. J Mol Biol 61:409–424

McPherson DT (1988) Codon preference reflects mistranslational constraints: A proposal. Nucleic Acid Res 16:4111–4120

McVean GAT, Vieira J (2001) Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics 157(1):245–257

Modiano G, Battistuzzi G, Motulsky AG (1981) Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha- and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? Proc Natl Acad Sci USA 78:1110–1114

Moriyama EN, Powell JR (1997) Codon usage bias and tRNA abundance in *Drosophila*. J Mol Evol 45:514–523

Nakamura Y, Gojobori T, Ikemura T (2000) Codon usage tabulated from the the international DNA sequence databases: Status for the year 2000. Nucleic Acids Res 28:292

Nei M, Gojobori T (1986) Simple methods for estimating the number of synonymous and non-synonymous nucleotide substitution. Mol Biol Evol 3:418–426

Overington J, Donnelly D, Johnson MS, et al. (1992) Environment-specific amino-acid substitution tables—Tertiary templates and prediction of protein folds. Protein Sci 1(2):216–226

Post LE, Strycharz GD, Nomura M, Lewis H, Dennis PP (1979) Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in *Escherichia coli*. Proc Natl Acad Sci USA 76:1697–1701

Powell JR, Moriyama EN (1997) Evolution of codon usage bias in *Drosophila*. Proc Natl Acad Sci USA 94:7784–7790

Riek RP, Handschumacher MD, Sung SS, Tan M, Glynias MJ, Schluchter MD, Novotny J, Graham RM (1995) Evolutionary conservation of both the hydrophilic and hydrophobic nature of transmembrane residues. J Theor Biol 172(3):245–258

Risler JL, Delorme MO, Delacroix H, Henaut A (1988) Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J Mol Biol 204(4):1019–1029

Smith NG, Hurst LD (1997) The effect of tandem substitutions on the correlation between synonymous and non-synonymous rates in rodents. Genetics 153:1395–1402

Topal MD, Fresco JR (1976) Complementary base pairing and the origin of substitution mutations. Nature 263:285–289

Woese CR (1965) On the evolution of the genetic code. Proc Natl Acad Sci USA 54:1546–1552

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23–29

Yang Z, Nielsen R (2000) Estimating synonymous and non-synonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409–418

Yang Z, Nielsen R (1998) Synonymous and non-synonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17(1):32–43