Evidence for translational selection in codon usage in *Echinococcus* spp.

V. FERNÁNDEZ[1], A. ZAVALA[2] *and* H. MUSTO[2]*

[1] *Cátedra de Inmunología, Facultad de Química, Montevideo, Uruguay*
[2] *Laboratorio de Organización y Evolución del Genoma, Sección Bioquímica, Facultad de Ciencias, Iguá 4225, Montevideo 11400, Uruguay*

(*Received 3 January 2001; revised 28 February 2001; accepted 28 February 2001*)

SUMMARY

We analysed the intragenomic variation in codon usage in *Echinococcus* spp. by correspondence analysis. This approach detected a trend among genes which was correlated with expression levels. Among the (presumed) highly expressed sequences we found an increased usage of a subset of codons, almost all of them G- or C- ending. Since an increase in these bases at the synonymous sites is against the mutational bias (these genomes are slightly A+T- rich), we conclude that codon usage in *Echinococcus* is the result of an equilibrium between compositional pressure and selection, the latter acting at the level of translation, mainly on highly expressed genes. This is the first report where translational selection for codon usage is detected among Platyhelminthes.

Key words: *Echinococcus* spp., codon usage, translational selection, optimal codons.

INTRODUCTION

Several studies have demonstrated that the synonymous codon usage is far from random. This unequal usage was first explained by Grantham *et al.* (1981) who proposed the 'genome hypothesis', stating that the biases are species specific. Subsequently it was shown that in several bacteria, such as *Escherichia coli* (Ikemura, 1981; Gouy & Gautier 1982) and *Bacillus subtilis* (Shields & Sharp, 1987), highly expressed genes display a more biased pattern of codon preferences than less expressed sequences. This was explained as the result of two main factors: mutational biases and natural selection acting at the level of translation, the latter being more evident in highly expressed sequences. Since the direction (towards G+C or A+T) and strength of these two factors vary among genomes, different patterns of preferences result among genes from different organisms (for reviews see Anderson & Kurland, 1990; Sharp & Matassi, 1994; Sharp *et al.* 1995). When the complete genome of several prokaryotes became available, this paradigm was reinforced (see for instance de Miranda *et al.* 2000; Romero, Zavala & Musto, 2000*a*; Lafay, Atherton & Sharp, 2000). Among unicellular eukaryotes, such as *Saccharomyces cerevisiae*, kinetoplastids, *Plasmodium falciparum* and *Entamoeba histolytica*, the same factors seem to shape codon usage (Sharp, Touhy & Mosurski, 1986; Alvarez, Robello & Vignali, 1994; Musto *et al.* 1999*a*; Romero, Zavala & Musto, 2000*b*).

In multicellular organisms different patterns have been reported. For example, in *Caenorhabditis elegans* and *Drosophila melanogaster*, similar to unicellular species, the factors which govern the codon choices have been attributed to a balance between the mutational biases and natural selection (Shields *et al.* 1988; Sharp & Li, 1989; Moriyama & Gojobori, 1992; Carulli *et al.* 1993; Akashi, 1994, 1997; Stenico, Lloyd & Sharp, 1994; Moriyama & Powell, 1997; Powell & Moriyama, 1997). Furthermore, translational selection at silent sites has also been reported to be the main factor shaping codon usage in plants like *Zea mays* (Fennoy & Bailey-Serres, 1993) and *Arabidopsis thaliana* (Chiapello *et al.* 1998). For vertebrates, it is generally accepted that the most important factor shaping codon choices is the localization of each gene. Indeed, these genomes are a mosaic of isochores (which are long compositionally homogeneous DNA segments which can be subdivided into a small number of families characterized by different GC levels) in which strong compositional correlations hold, especially between GC3 (G+C content at silent sites) and the segment harbouring the gene (Bernardi *et al.* 1985; Aota & Ikemura, 1986; D'Onofrio *et al.* 1991; Clay *et al.* 1996; Musto *et al.* 1999*b*). Therefore, for these species it has been argued that codon usage is mainly the reflection of the physical localization of each gene in the corresponding isochore (Eyre-Walker, 1991; Sharp & Matassi, 1994).

For Platyhelminthes the number of studies is few.

* Corresponding author: Laboratorio de Organización y Evolución del Genoma, Sección Bioquímica, Facultad de Ciencias, Iguá 4225, Montevideo 11400, Uruguay.
Tel: +598 2 525 20 95. Fax: +598 2 525 86 17.
E-mail: hmusto@fcien.edu.uy

DOI: 10.1017/S0031182001008150 Printed in the United Kingdom

Table 1. *Echinococcus* spp. gene sequences

(Genes are listed in order of their position on the first axis of the COA of RSCU. Acc. no. is the accession number; Sp. is the species of origin of the gene; Eg is *E. granulosus*; Em is *E. multilocularis*; (P) denotes if the sequence is partial; L is the length of the gene in codons, GC3s is the G+C content at third codon positions in synonymous sites, and Nc is the effective number of codons.)

Acc. no.	Sp.	Description	L	GC3s	Nc
U26448	Eg	Heat shock 70 kDa	665	0·71	35
L23315	Em	Ubiquitin	76	0·59	42
AF143813	Eg	Antigen B8/1	81	0·60	46
AF034959	Eg (P)	Thioredoxin peroxidase	185	0·59	61
J04664	Eg (P)	Cyclophilin	161	0·57	46
Z21787	Eg	Paramyosin	863	0·59	53
Z29075	Eg	Myophilin	190	0·57	50
AF246979	Eg (P)	Translation Elongation Factor	244	0·52	59
M59323	Em (P)	Antigen	354	0·55	57
X65947	Eg	Fatty acid binding protein	133	0·64	56
Z29489	Eg	Eg10	559	0·57	56
Z31712	Eg	Ferritin	173	0·56	52
L07773	Eg	Actin 1	375	0·57	49
L08894	Eg	Malate dehydrogenase	332	0·62	49
L07774	Eg	Actin 2	376	0·50	54
U19101	Em	Glyceraldehyde-3-P dehydrogenase	336	0·60	45
AJ249550	Em (P)	Tubulin	443	0·57	57
AF011923	Eg (P)	Tropomyosin-like protein	148	0·49	53
X66817	Eg (P)	Hbx1	368	0·45	61
X90928	Eg (P)	EG95	153	0·44	57
AF101269	Eg	Glutathione S-transferase	219	0·55	47
M63605	Eg	Glucose regulated protein	651	0·47	52
X66819	Eg (P)	Hbx3	167	0·52	60
M96564	Eg	Antigen	426	0·55	60
L33460	Eg	Laminin-binding protein	268	0·48	52
AF207904	Eg	14-3-3 protein	244	0·54	58
AF012071	Eg	Unknown protein	236	0·45	61
AF078931	Eg	Antigen	238	0·52	58
U63410	Em (P)	Antigen	260	0·49	52
AF034637	Eg	Thioredoxin	107	0·55	41
AF252859	Eg	Antigen B subunit precursor	89	0·51	47
M55441	Eg (P)	Antigen S epitope	52	0·51	53
AF067807	Eg	Paramyosin related protein	601	0·30	49
L34050	Eg (P)	Severin	111	0·59	55
L48620	Eg	Antigen B8/2	90	0·41	46

For the trematode *Schistosoma mansoni* it has been shown that there is a clear tendency to use A- and T-ending codons (Ellis & Morrison, 1995). Recently, it has been reported that this genome seems to be composed of isochore-like structures and, although natural selection might operate on certain sequences, the main factor shaping codon usage is the mutational pressure (Musto, Romero & Rodríguez-Maseda, 1998). For Cestodes, the only genus studied is *Echinococcus* spp. (*E. granulosus* and *E. multilocularis*), and the pattern found indicates that there is a strong bias towards C- and G- ending codons (Alvarez *et al*. 1993; Kalinna & McManus, 1994; Ellis, Morrison, & Kalinna, 1995). However, in the first 2 reports the intragenomic variability of codon choices was not analysed, and in the third one the number of sequences studied was very low (only 10). In this paper we report the pattern of codon usage in *Echinococcus* spp. through multivariate analysis, taking advantage of the increased amount of sequences now available.

MATERIALS AND METHODS

Sequences

DNA sequences from *E. granulosus* and *E. multilocularis* were taken from GenBank (September, 2000), and pooled given the strong identity of orthologous sequences at synonymous and at non-synonymous sites. In the cases of orthologous genes from the two species, the sequence from *E. granulosus* was chosen since this species is more represented in the GenBank. The data set comprised complete sequences (e.g. including initiation and stop codons), and incomplete genes but longer than 100 amino

https://doi.org/10.1017/S0031182001008150 Published online by Cambridge University Press

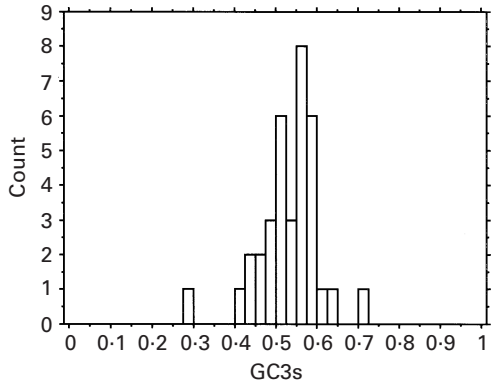

Fig. 1. Distribution of GC3s level (G+C content at the third codon position) in the genes from *Echinococcus* spp. analysed in this paper.

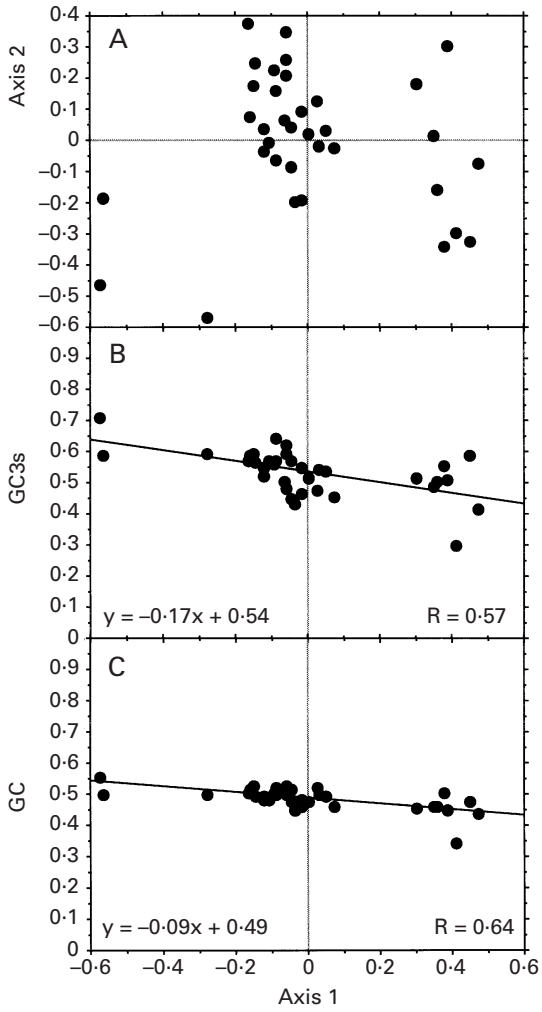

Fig. 2. (A) Distribution of the genes from *Echinococcus* spp. on the plane defined by the two main axes of the correspondence analysis. (B) Correlation between the position of each gene along the first axis and the GC3s content of the respective sequences. (C) Correlation between the position of each gene along the first axis and the G+C content of the respective exons. The equations of the regression lines and the regression coefficients (*R*) are shown.

acids. A total of 35 non-redundant genes were analysed, and are listed in Table 1.

Analyses

Codon usage, correspondence analysis (COA), frequency of codons ending in C or G, excluding Met, Trp and stop codons (GC3s), 'effective number of codons' (Nc, Wright, 1990) and relative synonymous codon usage (RSCU, Sharp *et al.* 1986) were calculated using the program CodonW 1.3 (written by John Peden and obtained from ftp://molbiol.ox.ac.uk/Win95.codonW.zip). Nc is a measure of the bias in synonymous codon usage, and it is independent of amino acid composition and codon number. Nc values can range from 20, when only 1 codon is used per amino acid, and 61 when all codons are used equally. The expected value for Nc under random (except for the influence of GC content) codon usage is approximately given by:

$$Nc = 2 + s + \{29/[s + (1 - s)^2]\},$$

where s = GC3s. RSCU is the observed frequency of a codon divided by the frequency expected if all synonyms coding for that amino acid are used equally; therefore RSCU values close to 1·0 indicate a lack of bias for that codon. To investigate the major trends in codon usage among genes, a COA was performed.

RESULTS

In order to understand if there is some variation in the pattern of codon usage in *Echinococcus* spp., we made 2 complementary analyses. First, we studied the GC3s content of all coding sequences, and in Fig. 1 it can be seen that there is a huge variation among the genes, since the extreme values are 30% and 71%. The second approach was to analyse the effective number of codons (Nc) of each gene. This parameter is a measure of codon bias, and generally highly expressed sequences display lowest values than sequences expressed at low or very low levels (Wright, 1990). In Table 1 it can be seen that the Nc values range from 35 to 61. Therefore, we concluded that there is a large variation in codon usage among the sequences.

In order to understand the causes of this variation, we conducted a COA for all the genes from *Echinococcus* spp. COA has been extensively used to study the intragenomic variation in synonymous codon usage patterns (see for instance Sharp & Devine, 1989; Alvarez *et al.* 1994; Ellis *et al.* 1995; Musto *et al.* 1998; Musto *et al.* 1999*a*; Romero *et al.* 2000*a*, *b*). With this multivariate statistical approach the data (sequences) are plotted in a multidimensional space of 59 axes. Subsequently, the axes which represent the most prominent factors contributing to the variation among the data set are

https://doi.org/10.1017/S0031182001008150 Published online by Cambridge University Press

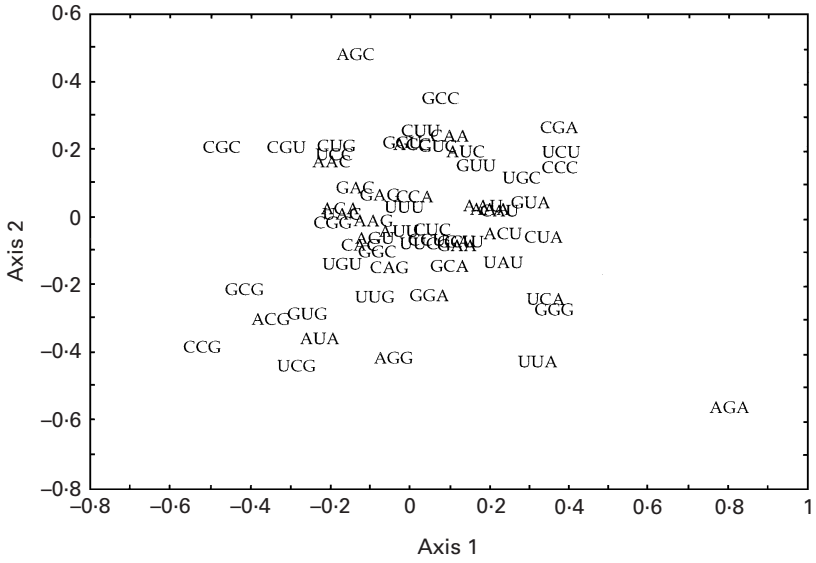

Fig. 3. Distribution of the codons from *Echinococcus* spp. on the plane defined by the two main axes of the correspondence analysis.

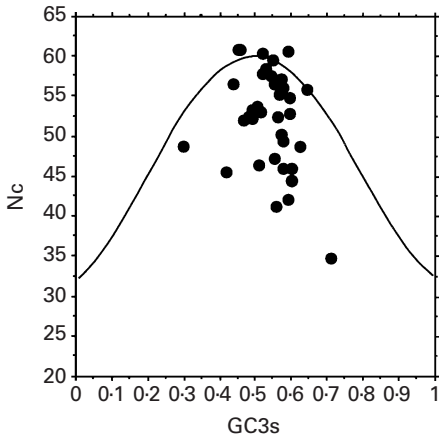

Fig. 4. Nc plot computed for 35 sequences from *Echinococcus* spp. The continuous curve in the figure represents the relationship between Nc and GC3s under random codon usage (except for the influence of GC content).

identified. In this report we performed the study on the RSCU data (excluding Met, Trp, and stop codons) in order to minimize the effects of amino acid composition. In Fig. 2A the position of the genes on the plane defined by the first (horizontal) and second (vertical) axes is shown. The axes represented 14·3 % and 11·0 % of the total variation, respectively.

As can be seen, the first axis splits the genes into 3 groups, and is significantly correlated with the GC3s level of each gene (Fig. 2B) and with the global GC content of the whole translated sequence (Fig. 2C). The second axis strongly correlates with the purine (or pyrimidine) content at the silent sites of each gene ($R = 0{\cdot}59$, $P < 0{\cdot}0005$). The position of each codon on these two main axes is represented in Fig. 3, which shows that there is a trend towards increasing GC from negative to positive values on the axis 1, and from purine-towards pyrimidine-ending codons with increasing co-ordinates on the second axis.

In Table 1 the genes are sorted according to their position along the first axis of the COA. In spite of the scarcity of data concerning expression levels in *Echinococcus* spp. (and especially, comparative expression levels among different sequences), there appears to be a tendency for the genes to be grouped according to that biological trait, since in the upper half of the Table the sequences which are expected to be highly expressed are clustered. This is the case of HSP 70, ubiquitin, actins, tubulin, glyceraldehyde-3-phosphate dehydrogenase, etc. This is reinforced by the fact that the antigen B8/1 is more heavily expressed than B8/2 (G. González, personal communication), and it occupies a clearly higher position (see Table 1). Therefore, the first axis of the COA seems to be related to 2 biological features: GC3s and expression levels. The Nc plot displayed in Fig. 4 supports this hypothesis since it is clear that several genes display an effective number of codons that is not determined exclusively by the mutational bias.

Our next step was to investigate the codon usage pattern in the genes displaying the most extreme values at both ends of the first axis of the COA (5 genes each), assuming, as noted above, that codon usage biases at each end of the distribution might be representative of the biases characteristic of highly and lowly expressed sequences. A χ^2 test was carried out and the result of this analysis is shown in Table 2. We could detect 14 putative translational optimal triplets coding for 12 amino acids, and 13 of them are either C- or G- ending. A careful inspection of these incremented codons shows that they tend to follow 2 rules: (1) for quartets, the G- ending triplet is the translational optimal; and (2) for duets, either the C- or the G- ending codon are the preferred among the

https://doi.org/10.1017/S0031182001008150 Published online by Cambridge University Press

Table 2. Codon usage in putatively highly- and lowly-expressed genes in *Echinococcus* spp.

(RSCU of putatively highly[a] and lowly[b] expressed genes. Each group was constructed by summing the appearences of each codon (N) in the 5 sequences at either extreme of the first axis determined by the COA. Codons occurring significantly more often in the 'high' group are in bold ($P < 0{\cdot}01$) or underlined ($P < 0{\cdot}05$).)

AA	Codon	RSCU[a]	N[a]	RSCU[b]	N[b]
Phe	UUU	0·63	18	0·40	5
	UUC	1·37	39	1·60	20
Tyr	UAU	0·67	7	1·29	9
	UAC	1·33	14	0·71	5
His	CAU	0·35	3	1·41	12
	CAC	1·65	14	0·59	5
Asn	AAU	1·05	22	1·46	41
	AAC	0·95	20	0·54	15
Asp	GAU	1·00	40	1·46	46
	GAC	1·00	40	0·54	17
Cys	UGU	0·88	7	0·73	4
	UGC	1·13	9	1·27	7
Gln	CAA	0·51	11	0·88	23
	CAG	1·49	32	1·12	29
Lys	AAA	0·18	8	0·97	65
	AAG	1·82	81	1·03	69
Glu	GAA	0·38	16	1·20	80
	GAG	1·62	68	0·80	53
Ile	AUU	0·80	19	1·30	30
	AUC	1·06	25	0·65	15
	AUA	1·14	27	1·04	24
Val	GUU	0·54	11	1·19	16
	GUC	0·49	10	1·11	15
	GUA	0·29	6	0·74	10
	GUG	2·68	55	0·96	13
Pro	CCU	0·30	3	1·33	7
	CCC	0·50	5	1·71	9
	CCA	1·10	11	0·76	4
	CCG	2·10	21	0·19	1
Thr	ACU	0·57	10	1·71	27
	ACC	0·86	15	0·83	13
	ACA	0·51	9	1·02	16
	ACG	2·06	36	0·44	7
Ala	GCU	0·71	14	1·83	33
	GCC	0·61	12	0·89	16
	GCA	0·35	7	0·89	16
	GCG	2·33	46	0·39	7
Gly	GGU	1·91	55	1·25	10
	GGC	0·97	28	1·00	8
	GGA	0·66	19	1·13	9
	GGG	0·45	13	0·63	5
Leu	UUA	0·06	1	1·20	24
	UUG	2·77	43	1·40	28
	CUU	0·65	10	1·00	20
	CUC	0·90	14	1·05	21
	CUA	0·06	1	0·85	17
	CUG	1·55	24	0·50	10
Ser	UCU	0·25	3	2·00	17
	UCC	0·58	7	0·35	3
	UCA	0·41	5	1·06	9
	UCG	2·63	32	0·71	6
	AGU	1·23	15	1·53	13
	AGC	0·90	11	0·35	3
Arg	**CGU**	2·10	22	0·78	6
	CGC	1·33	14	0·13	1
	CGA	0·57	6	1·17	9
	CGG	0·67	7	0·13	1
	AGA	0·48	5	3·00	23
	AGG	0·86	9	0·78	6
Met	AUG	1·00	28	1·00	24
Trp	UGG	1·00	5	1·00	4
TER	UAA	1·20	2	0·75	1
	UAG	0·60	1	0·75	1
	UGA	1·20	2	1·50	2

highly expressed genes. Of course, this is in agreement with the correlation between the first axis of the COA and GC3s (Fig. 2B)

DISCUSSION

Several previous papers indicate that codon usage in *Echinococcus* spp. is biased towards G- and C-ending triplets (Alvarez *et al.* 1993; Kalinna & McManus, 1994; Ellis, Morrison & Kalinna, 1995). However, 2 measures of codon usage bias suggest that there is some intragenomic variability: the histogram of the distribution of GC3s and the Nc values. Therefore, it seems evident that there is a large variation in codon usage among the sequences.

The first axis generated by the COA seems to be related to biological features: GC and expression levels. As happens in several species, this result strongly suggests that in *Echinococcus* codon usage is the result of an equilibrium between the mutational bias and natural selection acting mainly on highly expressed sequences (for reviews, see Sharp & Matassi, 1994; Sharp *et al.* 1995; Akashi & Eyre-Walker, 1998). This conclusion is supported by the Nc plot displayed in Fig. 4, since it is clear that several genes display an effective number of codons that is not determined exclusively by the mutational bias.

The statistical comparison of codon usage of the genes displaying the most extreme values along the first axis, in other words, the codon frequencies of putatively highly and lowly expressed sequences, allowed us to detect several synonymous triplets that are more frequent among the highly expressed genes. As can be seen, most of these triplets are either C- or G- ending. This is an interesting result, since such an increment is probably against the mutational bias characteristic of *Echinococcus*. Indeed, the genomic GC content reported for these species is 44% (Ellis *et al.* 1995), a value very close to the mean GC content of the available introns (42%, our result, data not shown). Therefore, an incremented level of

https://doi.org/10.1017/S0031182001008150 Published online by Cambridge University Press

GC3s in certain genes (and remarkably, most of them presumably highly expressed) is probably due to natural selection acting at the synonymous sites incrementing the frequency of optimal codons in highly expressed genes.

Summarizing, in this work we have shown that there is a group of codons in *Echinococcus* spp. which are very probably translationally optimal. While in trematodes the main factor shaping the codon usage seems to be the mutational bias (Musto *et al*. 1998), among cestodes natural selection acting at the level of translation might be operative. Therefore, the factors shaping codon usage do not seem to be identical for all Platyhelminthes.

REFERENCES

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

AKASHI, H. (1997). Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene* **205**, 269–278.

AKASHI, H. & EYRE-WALKER, A. (1998). Translational selection and molecular evolution. *Current Opinion in Genetics and Development* **8**, 688–693.

ALVAREZ, F., GARAT, B., MUSTO, H., PICON, M. & EHRLICH, R. (1993). Tendencies in *Echinococcus* sp. codon usage. *Memorias do Instituto Oswaldo Cruz* **88**, 345–346.

ALVAREZ, F., ROBELLO, C. & VIGNALI, M. (1994). Evolution of codon usage and base contents in kinetoplastid protozoans. *Molecular Biology and Evolution* **11**, 790–802.

ANDERSSON, S. G. & KURLAND, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiological Reviews* **54**, 198–210.

AOTA, S. & IKEMURA, T. (1986). Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Research* **14**, 6345–6355.

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.

CARULLI, J. P., KRANE, D. E., HARTL, D. L. & OCHMAN, H. (1993). Compositional heterogeneity and patterns of molecular evolution in the *Drosophila* genome. *Genetics* **134**, 837–845.

CHIAPELLO, H., LISACEK, F., CABOCHE, M. & HENAUT, A. (1998). Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene* **209**, GC1–GC38.

CLAY, O., CACCIO, S., ZOUBAK, S., MOUCHIROUD, D. & BERNARDI, G. (1996). Human coding and noncoding DNA: compositional correlations. *Molecular Phylogenetics and Evolution* **5**, 2–12.

DE MIRANDA, A. B., ALVAREZ-VALIN, F., JABBARI, K., DEGRAVE, W. M. & BERNARDI, G. (2000). Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. *Journal of Molecular Evolution* **50**, 45–55.

D'ONOFRIO, G., MOUCHIROUD, D., AISSANI, B., GAUTIER, C. & BERNARDI, G. (1991). Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *Journal of Molecular Evolution* **32**, 504–510.

ELLIS, J. T. & MORRISON, D. A. (1995). *Schistosoma mansoni*: patterns of codon usage and bias. *Parasitology* **110**, 53–60.

ELLIS, J., MORRISON, D. A. & KALINNA, B. (1995). Comparison of the patterns of codon usage and bias between *Brugia*, *Echinococcus*, *Onchocerca* and *Schistosoma* species. *Parasitology Research* **81**, 388–393.

EYRE-WALKER, A. C. (1991). An analysis of codon usage in mammals: selection or mutation bias? *Journal of Molecular Evolution* **33**, 442–449.

FENNOY, S. L. & BAILEY-SERRES, J. (1993). Synonymous codon usage in *Zea mays L.* nuclear genes is varied by levels of C and G- ending codons. *Nucleic Acids Research* **21**, 5294–5300.

GOUY, M. & GAUTIER, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research* **10**, 7055–7074.

GRANTHAM, R., GAUTIER, C., GOUY, M., JACOBZONE, M. & MERCIER, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research* **9**, r43–r74.

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. *Journal of Molecular Biology* **151**, 389–409.

KALINNA, B. H. & MCMANUS, D. P. (1994). Codon usage in *Echinococcus*. *Experimental Parasitology* **79**, 72–76.

LAFAY, B., ATHERTON, J. C. & SHARP, P. M. (2000). Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology* **146**, 851–860.

MORIYAMA, E. N. & GOJOBORI, T. (1992). Rates of synonymous substitution and base composition of nuclear genes in *Drosophila*. *Genetics* **130**, 855–864.

MORIYAMA, E. N. & POWELL, J. R. (1997). Codon usage bias and tRNA abundance in *Drosophila*. *Journal of Molecular Evolution* **45**, 514–523.

MUSTO, H., ROMERO, H. & RODRIGUEZ-MASEDA, H. (1998). Heterogeneity in codon usage in the flatworm *Schistosoma mansoni*. *Journal of Molecular Evolution* **46**, 159–167.

MUSTO, H., ROMERO, H., ZAVALA, A., JABBARI, K. & BERNARDI, G. (1999*a*). Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection. *Journal of Molecular Evolution* **49**, 27–35.

MUSTO, H., ROMERO, H., ZAVALA, A. & BERNARDI, G. (1999*b*). Compositional correlations in the chicken genome. *Journal of Molecular Evolution* **49**, 325–329.

POWELL, J. R. & MORIYAMA, E. N. (1997). Evolution of codon usage bias in *Drosophila*. *Proceedings of the National Academy of Sciences, USA* **94**, 7784–7790.

ROMERO, H., ZAVALA, A. & MUSTO, H. (2000*a*). Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Research* **28**, 2084–2090.

ROMERO, H., ZAVALA, A. & MUSTO, H. (2000*b*). Compositional pressure and translational selection determine codon usage in the extremely GC-poor unicellular eukaryote *Entamoeba histolytica*. *Gene* **242**, 307–311.

SHARP, P. M. & DEVINE, K. M. (1989). Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons. *Nucleic Acids Research* **17**, 5029–5039.

SHARP, P. M. & LI, W. H. (1989). On the rate of DNA sequence evolution in *Drosophila*. *Journal of Molecular Evolution* **28**, 398–402.

SHARP, P. M. & MATASSI, G. (1994). Codon usage and genome evolution. *Current opinion in Genetics and Development* **4**, 851–860.

SHARP, P. M., AVEROF, M., LLOYD, A. T., MATASSI, G. & PEDEN, J. F. (1995). DNA sequence evolution: the sounds of silence. *Philosophical Transactions of the Royal Society of London*, *B* **349**, 241–247.

SHARP, P. M., TUOHY, T. M. & MOSURSKI, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* **14**, 5125–5143.

SHIELDS, D. C. & SHARP, P. M. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Research* **15**, 8023–8040.

SHIELDS, D. C., SHARP, P. M., HIGGINS, D. G. & WRIGHT, F. (1988). 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution* **5**, 704–716.

STENICO, M., LLOYD, A. T. & SHARP, P. M. (1994). Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Research* **22**, 2437–2446.

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