Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 317 (2003) 141–147
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

The influence of translational selection on codon usage in fishes from the family Cyprinidae

Héctor Romero[a,b,c], Alejandro Zavala[b], Héctor Musto[a,b,*], Giorgio Bernardi[a]

[a] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, Naples 80121, Italy*
[b] *Laboratorio de Organización y Evolución del Genoma, Departamento de Biología Celular y Molecular, Facultad de Ciencias, Igua 4225, Montevideo 11400, Uruguay*
[c] *Escuela de Tecnología Médica, Facultad de Medicina, Hospital de Clínicas, Avda Italia s/n, Montevideo 11600, Uruguay*

Received 10 February 2003; received in revised form 28 February 2003; accepted 12 May 2003

Abstract

In this paper, the main factors shaping codon usage in three species of fishes that belong to the family Cyprinidae (namely *Brachidanio rerio*, *Cyprinus carpio*, and *Carassius auratus*) are reported. Correspondence analysis (COA), a commonly used multivariate statistical approach, was used to analyze codon usage bias. Our results show that the main trend is strongly correlated with the GC_3 content at silent sites of each sequence. On the other hand, the second axis discriminates between presumed highly and lowly expressed genes, a result that is confirmed by the distribution of matching expressed sequence tags (ESTs) along that axis. Translational selection appears, therefore, to influence synonymous codon usage in these fishes. The comparison of codon usages of the sequences displaying the extreme values on the second axis indicates that several codons are significantly incremented among the heavily expressed sequences. Interestingly, several of these triplets are not only shared by the three fishes but also by *Xenopus laevis*, another cold-blooded vertebrate in which translational selection influences codon choices. We postulate that natural selection was operative for codon usage in the last common ancestor of these fishes and *Xenopus*, and will probably be detected in cold-blooded vertebrates in general. Finally, we raise the possibility that the same phenomena will be found among warm-blooded vertebrates.

Keywords: Codon usage; Translational selection; Isochores; Vertebrates; Correspondence analysis

1. Introduction

Among prokaryotes, it is generally agreed that the non-random usage of synonymous codons (Grantham et al., 1980) is the result of the balance between compositional constraints and natural selection acting at the level of translation, the latter effect being detectable only if it is strong enough to overcome the effect of random genetic drift (Sharp and Li, 1986a; Bulmer, 1991; Akashi and Eyre-Walker, 1998). This paradigm has been successfully applied to several eukaryotes, *Saccharomyces cerevisiae* (Sharp et al., 1986), *Drosophila melanogaster* (Shields et al., 1988), *Caenorhabditis elegans* (Stenico et al., 1994), *Plasmodium falciparum* (Musto et al., 1999), and *Chlamydomonas reinhardtii* (Naya et al., 2001).

On the other hand, in vertebrates, which are characterized by smaller effective population sizes and by larger generation times compared to unicellular species, it is generally accepted that translational selection is unable to overcome the effect of genetic drift. Furthermore (and more importantly), these genomes (especially, those of mammals and birds) are compositionally compartmentalized and this feature has a very great influence on codon usage. Indeed, it is well known that strong compositional correlations exist between the different codon positions and the isochore in which each gene is embedded (for a review, see Bernardi, 2003). Hence, a coding sequence located in a GC-poor isochore will display low GC_3

Abbreviations: GC_3, G+C content at synonymous third positions; COA, correspondence analysis; RSCU, relative synonymous codon usage; Nc, effective number of codons; CAI, codon adaptation index; EST, expressed sequence tag; Y, pyrimidine.

* Corresponding author. Laboratorio de Organización y Evolución del Genoma, Departamento de Biología Celular y Molecular, Facultad de Ciencias, Igua 4225, Montevideo 11400, Uruguay. Tel.: +598-2-525-2095; fax: +598-2-525-8617.
E-mail address: hmusto@fcien.edu.uy (H. Musto).

0378-1119/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0378-1119(03)00701-7

values (consequently, a codon usage biased towards A and T at 'silent' sites) while the reverse is true for a gene embedded in a GC-rich isochore. One then should conclude that the strong variation in codon usage that is always found in vertebrates is the result of the isochore structure characteristic of these genomes. This idea was reinforced by the finding that in human and *Xenopus laevis*, there are no differences in codon usage between genes encoding ribosomal proteins (that are certainly heavily expressed) and other genes (Kanaya et al., 2001). In line with this result, no correlation was found between (a) GC_3 and codon usage with expression levels for mammalian genes (Duret, 2002; Duret and Mouchiroud, 2000), and (b) between the rate of synonymous substitutions with either the expression level or the tissue specificity of genes in a mouse/rat comparison (Wolfe and Sharp, 1993). All these results strongly suggest that synonymous codon usage is not constrained by translational selection in mammals.

One of the most popular statistical approaches used to detect the pressures that shape codon usage, in particular among unicellular species, is multivariate analysis (for review, see Ermolaeva, 2001; Perriere and Thioulouse, 2002). When applied to vertebrates, these analyses reveal a single major trend that is, not surprisingly, always strongly correlated with the GC content at the third codon position, and does not discriminate any aspect of gene function, including expression.

We have recently found, however, that when this analysis is applied to the coding sequences of *X. laevis* (which displays a much narrower distribution of isochores and consequently of GC_3 than mammals and birds), it reveals two major trends. While the first one is strongly correlated with GC_3, the second is correlated with gene expression, as measured with expressed sequence tag (ESTs). In other words, highly and lowly expressed sequences display a different pattern of codon usage, which we interpreted as the result of natural selection acting at the level of translation (Musto et al., 2001). To see whether this also applies to other cold-blooded vertebrates (which are generally characterized, like *Xenopus*, by a narrower compositional distribution of isochores compared to warm-blooded vertebrates), we applied multivariate analysis to the coding sequences of three fishes that belong to the family Cyprinidae, namely *Brachidanio rerio*, *Cyprinus carpio*, and *Carassius auratus*. Our results show that after the effect of isochore structure is eliminated, the effect of translational selection can be detected. Therefore, we conclude that at least among several cold-blooded vertebrates, codon usage is shaped not only by compositional constraints but also by translational selection.

2. Materials and methods

Complete coding sequences (CDS) from *C. carpio* and *C. auratus* were retrieved from GenBank using the ACNUC retrieval system (Gouy et al., 1985). Redundancies and partial genes were removed. The final data set included 219 and 170 sequences for each species, respectively. For *B. rerio*, the sequences were retrieved from UniGene build no. 49 (ftp://ncbi.nlm.nih.gov/repository/UniGene/Dr.seq.uniq.gz). Only the putative complete coding sequences were considered, and the final data set was of 1899 sequences.

Codon usage, correspondence analysis (COA), GC_3 (the frequency of codons ending in G or C, excluding Met, Trp, and stop codons), the 'effective number of codons' (Nc) (Wright, 1990), the relative synonymous codon usage (RSCU) (Sharp and Li, 1986b), and the codon adaptation index (CAI) (Sharp and Li, 1987a) were calculated using the program CodonW 1.3 (J. Peden; http://molbiol.ox.ac.uk/Win95.codonW.zip).

The orthologous sequences were identified by running a BLAST query of the whole set of proteins of each species against the other two using the stand-alone package of Altschul et al. (1997). Only those pairs of sequences displaying a minimal value of 50% identity and a maximal length difference of 20% at the amino acid level were considered. The final data set included 63 pairs of orthologous genes in the case of *C. carpio–B. rerio*, 53 for *C. auratus–B. rerio*, and 27 for *C. carpio–C. auratus*. Using the same approach, we found 337 pairs of orthologous sequences between *B. rerio* and *X. laevis*. Expression levels were estimated using the data generated by UniGene system using the Sugano SJD adult male library, dbEST Library ID.9968 (http://www.ncbi.nlm.nih.gov:80/UniGene/library.cgi?ORG=Dr&LID=9968).

The estimated number of synonymous substitutions (Ks) was calculated according to the method of Li (1993) as implemented in the JaDis package (Goncalves et al., 1999).

3. Results and discussion

3.1. Compositional properties of coding sequences

The three species analyzed in this paper belong to the same family and, therefore, it is expected that their coding sequences should display very similar compositional features. This was investigated in two different ways. First, we analyzed the distribution of GC_3 (G+C content at silent sites of coding sequences) for each species, and the results are shown in Fig. 1 and Table 1. As it can be seen, in spite of the different numbers of coding sequences available, the shapes of the distributions are very similar, cover an equivalent range, and display nearly identical mean values and standard deviations. This is in complete agreement with the results obtained by the ultracentrifugation of genomic DNAs, which show very similar CsCl profiles for the three species (Bernardi and Bernardi, 1990; Bucciarelli et al., 2002). Second, we plotted the GC_3 values of orthologous genes, and we found that there are strong and significant correlations in the three comparisons: *B. rerio* against *C. auratus*

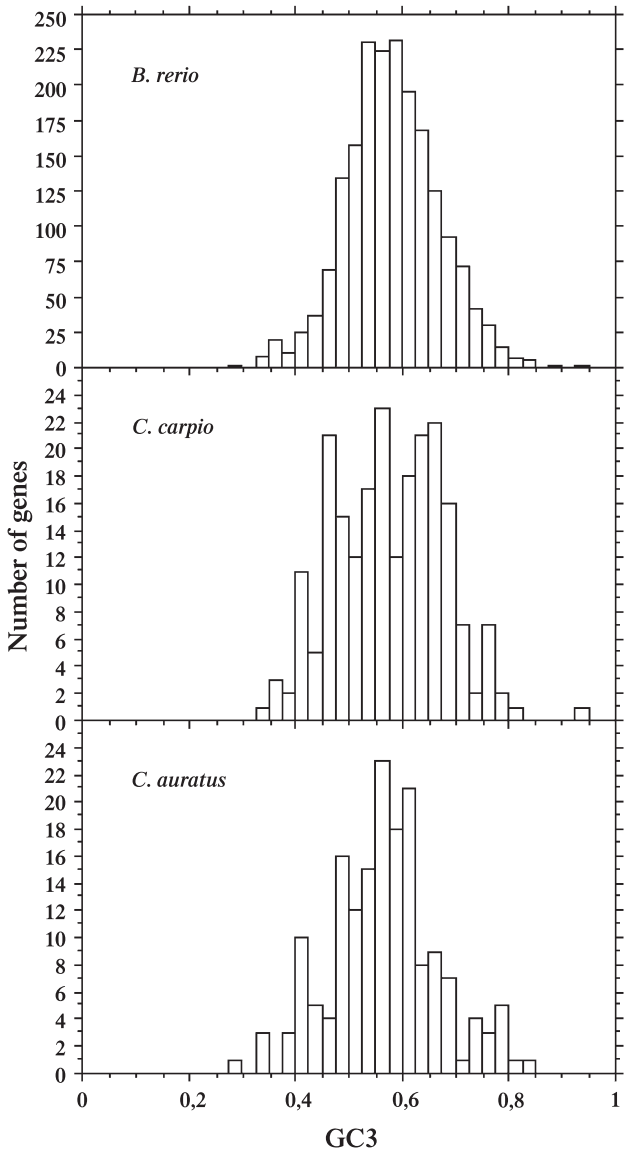

Fig. 1. Distribution of the GC_3 levels of the species analyzed in this paper.

(R=0.73, P<0.0001) (we should stress that in this comparison, there are two outliers: a member of the histone H2A family and a member of the cytochrome *P*450 family; if these genes are removed, the R value increases to 0.89), *B. rerio* against *C. carpio* (R=0.85, P<0.0001), and *C. carpio* against *C. auratus* (R=0.94, P<0.0001). Therefore, we conclude that, from a compositional point of view, the three species are almost identical. From the histograms, it can be seen that the distributions tend to be unimodal and symmetrical (specially in *B. rerio*, which is, by far, the species with the larger number of available sequences within Cyprinidae), a finding again in line with the results of Bernardi and Bernardi (1990) and Bucciarelli et al. (2002). We should stress that these compositional features are generally found both at the DNA and GC_3 levels among cold-blooded vertebrates (see Bernardi, 2003 for a review).

In spite of the rather low heterogeneity at the DNA level (compared to mammals and birds), it can be seen in Fig. 1 that there are several sequences in the three distributions that display GC_3 levels rather apart from the mean value, a point also apparent from the standard deviations and minimum and maximum values shown in Table 1. This is a clear indication that in the three species, the pattern of codon usage is not uniform among their genes.

3.2. Global patterns of codon usage

In order to better understand this variation, we first analyzed the average codon usage for the three fishes, which is displayed in Table 2. It appears that, in spite of some minor differences, the global pattern is very similar for the three species, which again confirms that there have not been significant compositional changes since these species diverged from their last common ancestor. Another conclusion derived from the inspection of Table 2 is that among fourfold degenerate codons, the "compositionally equivalent" bases (i.e., G and C, and A and T) are not used at similar frequencies. This can be taken as an evidence that the genome composition, although the main factor, cannot explain per se the whole codon usage patterns. However, we should stress that since significant correlations hold between the GC_3 level of each gene and its flanking regions and introns (Romero et al., in preparation), codon usage is strongly influenced by the physical location of each gene in the different isochores that make up the genome of these species (Bernardi and Bernardi, 1990; Bucciarelli et al., 2002). Another evidence of the variation in codon usage among these fishes comes from the analysis of the effective number of codons (Wright, 1990) for each gene. Nc is a measure of the bias in codon usage of the genes, and in genomes where translational selection contributes to codon usage, highly expressed sequences tend to display lower values compared with lowly expressed genes. This index shows a relative broad range in the three species, displaying a minimum of 33, 29, and 31 for *B. rerio*, *C. carpio*, and *C. auratus*, respectively, and a maximum of 61 in all the cases (not shown). All these features, taken together, suggest that apart from the effect of compositional constraints, there is some variation in codon usage among the sequences.

Table 1
G+C content at silent sites for *B. rerio*, *C. carpio*, and *C. auratus*

Organism	*n*	GC_3	Min	Max
B. rerio	1159	0.57 (0.09)	0.26	0.92
C. carpio	219	0.58 (0.10)	0.35	0.93
C. auratus	170	0.57 (0.10)	0.29	0.83

n=is the number of sequences; GC_3=the G+C content at silent sites; min=the minimum of GC_3 found for each species; max=the maximum of GC_3 found for each species. Standard deviations are given in parentheses.

Table 2
Codon usage and putative preferred codons in *B. rerio*, *C. carpio*, and *C. auratus*

aa	Codon	*B.r.*	*C.c.*	*C.a.*
Phe	TTT	0.87	0.85	0.83
	TTC	1.13**	1.15*	1.17**
Tyr	TAT	0.79	0.78	0.81
	TAC	1.21*	1.22	1.19
His	*CAT*	0.78	0.86	0.86
	CAC	1.22	1.14	1.14
Asn	AAT	0.73	0.81	0.78
	AAC	1.27	1.19	1.22
Asp	*GAT*	0.88**	0.96	0.94
	GAC	1.12	1.04	1.06
Cys	TGT	0.93	1.04	0.95
	TGC	1.07	0.96**	1.05
Gln	CAA	0.51	0.55	0.54
	CAG	1.49**	1.45**	1.46
Lys	AAA	0.94	0.92	0.92
	AAG	1.06**	1.08**	1.08**
Glu	GAA	0.69	0.74	0.70
	GAG	1.31**	1.26**	1.30**
Val	*GTT*	0.83	0.89**	0.87
	GTC	0.96**	1.00**	0.97*
	GTA	0.40	0.35	0.38
	GTG	1.81	1.76	1.78
Pro	*CCT*	1.18**	1.22*	1.27**
	CCC	1.04**	1.05**	1.05
	CCA	1.09	1.21	1.12
	CCG	0.69	0.52	0.56
Thr	*ACT*	0.98**	1.15**	1.07**
	ACC	1.30**	1.24**	1.29**
	ACA	1.13	1.15	1.10
	ACG	0.59	0.46	0.54
Ala	*GCT*	1.21**	1.40**	1.30**
	GCC	1.27**	1.24**	1.24**
	GCA	0.95	0.93	0.96
	GCG	0.57	0.43	0.50
Gly	*GGT*	0.86**	1.03**	1.02**
	GGC	1.17	1.09	1.06
	GGA	1.33	1.32	1.26
	GGG	0.63	0.55	0.66
Leu	TTA	0.39	0.35	0.33
	TTG	0.77	0.76	0.79
	CTT	0.79**	0.82	0.90**
	CTC	1.17	1.15	1.13
	CTA	0.39	0.35	0.40
	CTG	2.49**	2.57**	2.45
Ser	AGT	0.88	0.95	0.86
	AGC	1.43	1.40	1.36
	TCT	1.15**	1.23	1.20
	TCC	1.18**	1.12**	1.21**
	TCA	0.89	0.97	0.96
	TCG	0.46	0.32	0.41
Arg	AGA	1.45	1.66	1.40
	AGG	1.09	1.27**	1.26
	CGT	0.77**	0.90**	0.91**
	CGC	1.17**	0.91	1.19
	CGA	0.74	0.64	0.59
	CGG	0.79	0.62	0.65
Ile	*ATT*	1.00	1.01	1.01
	ATC	1.59**	1.55	1.58*
	ATA	0.42	0.44	0.41
Ter	TAA	1.02	1.36	1.27
	TAG	0.50	0.62	0.56
	TGA	1.48	1.03	1.16

3.3. Correspondence analysis on codon usage

In order to understand the causes of this variation, we first conducted a correspondence analysis of the RSCU values for all the genes in each of the species. The proportions of the total variance accounted for by the two principal axes of the COA were 17.4% and 7.5% in *B. rerio*, 15.1% and 9.7% in *C. carpio*, and 17.3% and 8.4% in *C. auratus*. The position of each sequence on the plane defined by the first two axes for each species is displayed in Fig. 2. Therefore, this analysis detects for each species a single major source of variation in the data set which is, in the three cases, strongly correlated with the GC_3 content of the respective coding sequences, with R values ranging from 0.95 to 0.97. We note that no other biological feature apart from composition at silent sites could be associated with this major trend.

On the other hand, the position of the sequences along the second axes is strongly correlated with the pyrimidine (Y) content of the genes at the third codon positions (R=−0.41, −0.38, and −0.60) for *B. rerio*, *C. carpio*, and *C. auratus*, respectively. However, a more important result emerged when the genes were sorted according to their position on the second axes, since we detected for the three species a cluster of highly expressed genes at one end of the distribution. Indeed, genes encoding ribosomal proteins, elongation factors, alpha- and beta-actins, and heat shock proteins (namely, heavily expressed housekeeping genes) clustered together with highly expressed tissue-specific sequences like several alpha- and beta-globins genes. Some regulatory enzymes and homeoboxes were localized at the other end of the distribution, where no presumed highly expressed sequence could be found. Then, we conclude that in these three fishes, the second main axis of COA is related to the expression level of each gene, the highly expressed sequences being Y-rich at third codon positions.

It is very interesting to note that these results are identical to those previously reported in *X. laevis*. Indeed, the analysis of 1303 sequences of *Xenopus* showed a main trend (20.3% of the total variance) that is strongly correlated (R=0.98) with the GC_3 content of each gene, while the second trend was (a) xnegatively correlated with the Y content at third codon positions (R=−0.37), and b) discriminated between highly and lowly expressed sequences (Musto et al., 2001). This correlation with expressivity was confirmed by the signifi-

Notes to Table 2:
aa=amino acid; *B.r.*, *C.c.*, and *C.a.*=*B. rerio*, *C. carpio*, and *C. auratus*, respectively. For comparative purposes, only the RSCU values are shown for each codon. Italicized codons are those proposed as translationally optimal in *X. laevis* (Musto et al., 2001).

* Statistically more frequent among highly expressed sequences in each fish (P<0.05).

** Statistically more frequent among highly expressed sequences in each fish (P<0.01).

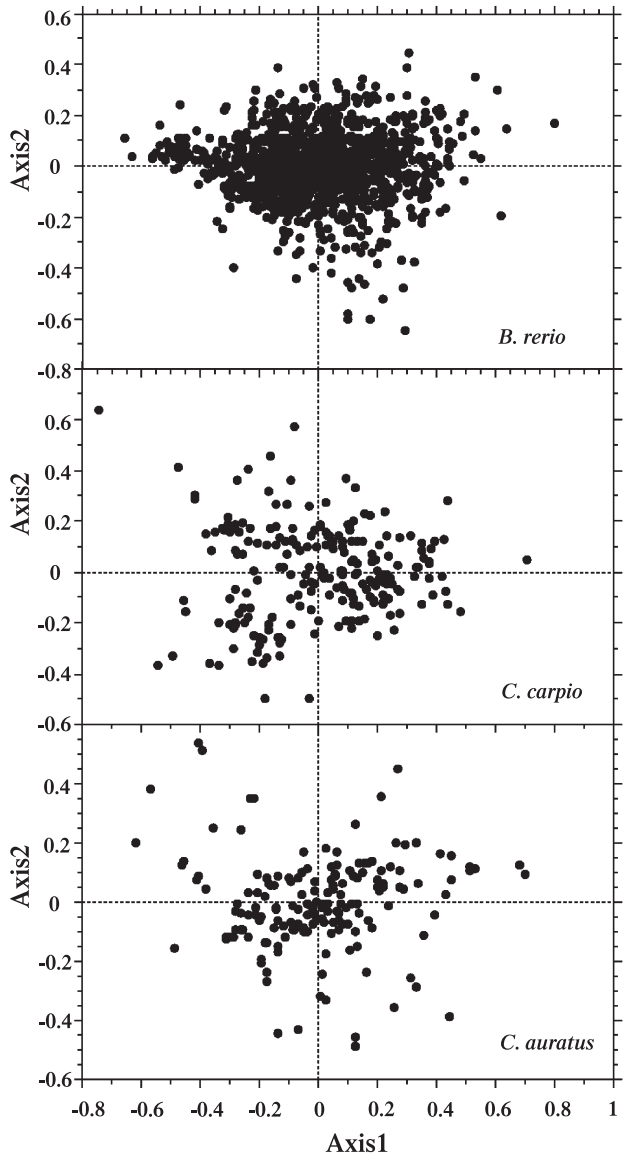

Fig. 2. Plot of the two first axes generated by the COA of RSCU values for each species.

cant correlation found between the position of each gene along the second axis and the respective CAI (Sharp and Li, 1987b), and, more importantly, by the distribution of ESTs along that axis. In other words, after the effect of the isochore structure in *Xenopus* (axis 1) is eliminated, the following most prominent source of variation (axis 2) is correlated with gene expression, which implies that highly and lowly expressed sequences display a different pattern of codon usage. We interpreted these results in terms of natural selection acting at the level of translation (Musto et al., 2001). Then, it immediately follows that the results described above for the three fishes studied here, which are almost identical to those reported for *X. laevis*, can be interpreted in the same way: translational selection contributes to the variability in codon usage of Cyprinidae.

3.4. Translational selection affects codon usage in Cyprinidae

If translational selection affects codon usage in Cyprinidae, two results should be expected. First, the position of orthologous genes along the axis that discriminates expression levels in each species (axis 2) should be similar. The rationale of this idea comes from the similarities in average codon usage among the three species (Table 2). Moreover, given the similarities among the three species, it is very probable that the expression levels of orthologous genes are comparable and, therefore, their relative positions along axis two should be similar. We found that this is indeed the case: the *R* values are 0.78, 0.80, and 0.72 for the comparisons of *B. rerio* vs. *C. carpio*, *B. rerio* vs. *C. auratus*, and *C. carpio* vs. *C. auratus*, respectively. Second, CAI values should correlate with the respective position along axis 2. In order to investigate this possibility, we calculated the CAI value for each sequence, taking as a reference the codon usage of genes encoding ribosomal proteins from *B. rerio*, since in this species, the number of sequences is higher. We found that for *B. rerio*, the *R* value was 0.66, while for *C. auratus* and *C. carpio*, the values were 0.49 and 0.55, respectively. We should mention that, as expected, the putatively highly expressed genes displayed in the three cases the highest CAI values. Then, it seems clear that axis 2 of COA in the three species is related to the expression level of each sequence.

With the purpose of confirming this interpretation and obtaining an estimation of the expression levels of the genes from *B. rerio* (which is, by far, the Cyprinidae member most represented in databases), we counted the number of matching ESTs for each sequence and their distribution along axis 2. Of the 3611 UniGene clusters reported for the dbEST Library ID.9968, we found that 874 of them where present in our curated database (46%). This analysis (Fig. 3) showed that when all these matching ESTs are considered, more than 70% of them match with the genes placed in the first and second deciles of the distribution along axis 2 (49% and

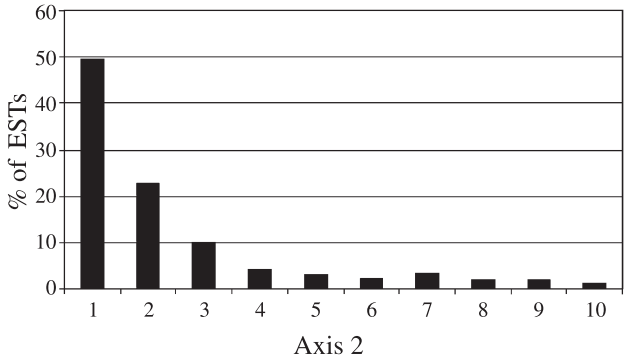

Fig. 3. Histogram of the distribution of expressed sequence tags along axis 2. The axis was divided into 10 bins, each of them containing an equal number of genes. In each bin, the total number of matching ESTs was calculated and expressed as percentage of the total matches.

23%, respectively), where the majority of highly expressed sequences are placed. This analysis clearly demonstrates that the second axis of COA indeed discriminates expression levels, as it does in *Xenopus* (Musto et al., 2001). In other words, the second most important source of variation in codon usage among these cold-blooded vertebrates is related to expression. From this, it can be immediately deduced that, within Cyprinidae, highly and lowly expressed sequences display differences in codon usage and, therefore, translationally optimal codons might exist. We compared, therefore, the codon usage patterns of the genes displaying the extreme values at both ends of the second axis (10% in the case of *C. carpio* and *C. auratus*, and 5% for *B. rerio*), and the differences between the two groups were evaluated with a χ^2 test. Two results concerning this analysis (Table 2) are interesting. First, 13 presumed optimal codons (coding for 10 amino acids) are shared by the three species, which confirms the similar pattern of codon usage mentioned above. Second, and very significantly, there are nine triplets (coding for nine different amino acids) incremented among the highly expressed genes in the three fishes that are also incremented in *Xenopus* (Musto et al., 2001). These are TTC (Phe), AAG (Lys), GAG (Glu), CCT (Pro), ACT (Thr), GCT (Ala), GGT (Gly), TCC (Ser), and CGT (Arg). Furthermore, there are another seven triplets that are significantly incremented in *Xenopus* and *B. rerio*: TAC (Tyr), GAT (Asp), CTT and CTG (Leu), TCT (Ser), CGC (Arg), and ATC (Ile). Given that in the other two fishes these codons display the same trend (see Table 2), it seems reasonable to assume that they did not reach the minimum level of statistical significance because of the rather low number of genes available. If this is the case, there are 16 common preferred codons among highly expressed sequences in *Xenopus* and Cyprinidae, which might imply that these triplets were translationally optimal in the last common ancestor of these two highly diverged *taxa*. Incidentally, this similarity is reinforced by two results. First, there is a highly significant correlation (R=0.70) between the position of orthologous genes on the respective second axes generated by the COA of *Xenopus* and *B. rerio*. Second, in spite of the enormous distance between these two species, which led to an accumulation of synonymous and nonsynonymous substitutions, a significant correlation holds (R=0.42, P<0.0001) between Ks and the position of the sequences along the second axis of the COA of *B. rerio* (when only sequences displaying Ks<2 are considered). Of course, the analysis of the pattern of codon usage and expression levels of more species is needed to confirm this hypothesis.

4. Conclusions

Codon usage in fishes from the family Cyprinidae, although mainly shaped by compositional constraints, is influenced by translational selection. Indeed, we have shown that a COA conducted on RSCU values from the three species detects two major trends. The first is strongly correlated with the GC_3 content of each sequence, while the second discriminates between presumed lowly and highly expressed genes. This result is confirmed by the analysis of matching ESTs that cluster at the end of the second trend, where the highly expressed sequences are located. Furthermore, we found that several putative preferred codons of these fishes are the same, as were previously postulated to be preferred in *X. laevis* (Musto et al., 2001). These results, obtained in cold-blooded vertebrates, might imply that (1) translational selection for codon usage was operative in the last common ancestor of these fishes and *Xenopus*, (2) will probably be found in more cold-blooded species, and (3) raises the possibility of detecting its presence among warm-blooded vertebrates.

Acknowledgements

This work was partially supported by award 7094 from the ‘Fondo Clemente Estable,’ Uruguay.

References

Akashi, H., Eyre-Walker, A., 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. 8, 688–693.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Bernardi, G., 2003. Structural and Evolutionary Genomics. Elsevier, Amsterdam (in press).

Bernardi, G., Bernardi, G., 1990. Compositional patterns in the nuclear genome of cold-blooded vertebrates. J. Mol. Evol. 31, 265–281.

Bucciarelli, G., Bernardi, G., Bernardi, G., 2002. An ultracentrifugation analysis of two hundred fish genomes. Gene 295, 153–162.

Bulmer, M., 1991. The selection–mutation–drift theory of synonymous codon usage. Genetics 129, 897–907.

Duret, L., 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12, 640–649.

Duret, L., Mouchiroud, D., 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17, 68–74.

Ermolaeva, M., 2001. Synonymous codon usage in bacteria. Curr. Issues Mol. Biol. 3, 91–97.

Goncalves, I., Robinson, M., Perriere, G., Mouchiroud, D., 1999. JaDis: computing distances between nucleic acid sequences. Bioinformatics 15, 424–425.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., di Paola, G., 1985. ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput. Appl. Biosci. 1, 167–172.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pave, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, r49–r62.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., Ikemura, T., 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. 53, 290–298.

Li, W.H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96–99.

Musto, H., Romero, H., Zavala, A., Jabbari, K., Bernardi, G., 1999. Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection. J. Mol. Evol. 49, 27–35.

Musto, H., Cruveiller, S., D'Onofrio, G., Romero, H., Bernardi, G., 2001. Translational selection on codon usage in *Xenopus laevis*. Mol. Biol. Evol. 18, 1703–1707.

Naya, H., Romero, H., Carels, N., Zavala, A., Musto, H., 2001. Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii*. FEBS Lett. 501, 127–130.

Perriere, G., Thioulouse, J., 2002. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res. 30, 4548–4555.

Sharp, P.M., Li, W.H., 1986a. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res. 14, 7737–7749.

Sharp, P.M., Li, W.H., 1986b. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.

Sharp, P.M., Li, W.H., 1987a. The codon adaptation index: a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.

Sharp, P.M., Li, W.H., 1987b. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4, 222–230.

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 Res. 14, 5125–5143.

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. Mol. Biol. Evol. 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 Res. 22, 2437–2446.

Wolfe, K.H., Sharp, P.M., 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. 37, 441–456.

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