ELSEVIER

Gene 238 (1999) 15–21

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Different hydrophobicities of orthologous proteins from *Xenopus* and human

S. Cruveiller [a,b], K. Jabbari [b], G. D'Onofrio [a], G. Bernardi [a,b,*]

[a] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121, Naples, Italy*
[b] *Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, F-75005 Paris, France*

Accepted 15 June 1999; Received by T. Gojobori

Abstract

A compositional transition was previously detected by comparing orthologous coding sequences from cold- and warm-blooded vertebrates (see Bernardi, G., Hughes, S., Mouchiroud, D., 1997. The major compositional transitions in the vertebrate genome. J. Mol. Evol. 44, S44–S51 for a review). The transition is characterized by higher GC levels (GC is the molar ratio of guanine+cytosine in DNA) and, especially, by higher GC_3 levels (GC_3 is the GC level of third codon positions) in coding sequences from warm-blooded vertebrates. This transition essentially affects GC-rich genes, although the nucleotide substitution rate is of the same order of magnitude in both GC-poor and GC-rich genes.

In order to understand the evolutionary basis of the changes, we have compared the hydrophobicity of orthologous proteins from *Xenopus* and human. Although the differences are small in proteins encoded by coding sequences ranging from 0 to 65% in GC_3, they are large in the proteins encoded by sequences characterized by GC_3 values higher than 65%. The latter proteins are more hydrophobic in human than in *Xenopus*.

Keywords: Isochores; Mutational bias; Natural selection

1. Introduction

A compositional analysis performed on high-molecular-weight DNA (Thiery et al., 1976) showed that the nuclear genomes of warm-blooded vertebrates are characterized by a remarkable base composition heterogeneity, whereas those of cold-blooded vertebrates are much more homogeneous. The major difference between these two classes of genomes is that the former attain much higher GC levels (GC is the molar ratio of guanine+cytosine in DNA) than the latter. This finding was confirmed by detailed studies on the genomes of cold-blooded vertebrates (Hudson et al., 1980; Pizon et al., 1984; Bernardi and Bernardi, 1990a,b; Aïssani and Bernardi, 1991). The difference between warm- and cold-blooded vertebrates was also seen by comparing total and homologous coding sequences (Perrin and Bernardi, 1987; Bernardi et al., 1988, 1997; Bernardi and Bernardi, 1991), showing that whereas, in GC-poor genes, GC_3 values (the GC levels of third codon positions) were relatively close, in GC-rich genes, GC_3 values are higher in warm-blooded vertebrates. In fact, the compositional transition (i.e. the difference in GC content affecting coding and non-coding sequences) essentially concerned the GC-rich coding sequences of cold-blooded vertebrates, which became much more GC-rich. At this point, it should be recalled that the gene concentration is low in the GC-poor regions of all vertebrate genomes and high in the GC-rich regions [see Bernardi (1995) for a review]. Interestingly, the increased GC levels of warm-blooded vertebrates were maintained since the compositional transition took place.

The major transition affecting the GC-rich minority of the vertebrate genome has been viewed as the result of either selection for some directional changes favoring GC increase (Bernardi and Bernardi, 1986; Bernardi et al., 1988), or mutational bias (Sueoka, 1988, 1992). In the present paper, we have further investigated this problem by analyzing the hydropathic changes at the proteins level resulting from compositional changes in the corresponding coding sequences. The rationale of this approach is posed on the observations that the

* Corresponding author. Tel.: +39-08-1247-808; fax: +39-08-1245-807.
E-mail address: bernardi@alpha.s2n.it (G. Bernardi)

0378-1119/99/$ – see front matter
PII: S0378-1119(99)00259-0

frequencies of the amino acids change as the GC content of the coding sequences changes (D'Onofrio et al., 1991, 1999b; Lobry, 1997), and that an increased hydrophobicity is one of the forces contributing to protein stability [see Pace et al. (1996) for a review].

Investigations on the maintenance of the compositional patterns of mammalian coding sequences will be presented in the following paper (Chiusano et al., 1999).

2. Materials and methods

2.1. Data retrieval

Due to the lack of reptilian sequences, the compositional transition from cold- to warm-blooded vertebrates can only be studied by comparing *Xenopus* and mammalian genes. Homologous sequences from human (*Homo sapiens*) and *Xenopus* (*Xenopus laevis*) were retrieved from the HOVERGEN Database (Duret et al., 1994, release 31 July 1998), using a new computer program written in C language. A first dataset of 664 pairs of homologous genes was checked in order to remove paralogous and partial genes. Paralogous sequences were identified on the basis of the distance $D=\ln(1-p-0.2p^2)$ (Kimura, 1983), where p is the proportion of amino acids that differ between two sequences, using the PROTDIST software (Felsenstein, 1989). Finally, only proteins presenting a size difference minor than, or equal to, 30 residues were used, creating a dataset with 460 pairs of orthologous genes. A subdataset of 72 orthologous genes (hereafter referred to as the three-species dataset) shared by human (*Homo sapiens*), calf (*Bos taurus*) and *Xenopus* (*Xenopus laevis*), was retrieved using the same approach.

2.2. Analysis of amino acid frequencies

Orthologous sequences were aligned with CLUSTALW (version 1.7, Thompson et al., 1994), and a C language program was written to compute the substitution matrix based on the alignments. This substitution matrix reports changes from one amino acid to another in absolute frequency. Each matrix was then handled in order to determine: (1) the total number of substitutions; (2) the frequency difference (Δaa) of each amino acid; (3) the number of conserved/variable sites and insertions or deletions (indels).

A substitution matrix was created for a GC-poor ($GC_3<45\%$), a GC-intermediate ($45\%<GC_3<65\%$) and a GC-rich group ($GC_3>65\%$) of coding sequences. The same criteria were used to divide the three-species dataset, and in each group, a substitution matrix was made for both human/calf and human/*Xenopus*.

2.3. Analysis of hydrophobicity in proteins

The 460 pairs of human/*Xenopus* orthologous genes were first sorted according to GC_3 levels in human genes, and then the dataset was divided into 20 groups of equal size to perform reliable statistical tests. For each group, a substitution matrix was used in order to calculate: (1) the frequency difference (Δaa) of each amino acid; (2) the hydrophobicity difference (ΔH), determined using the Δaa values and the hydropathy scale of Kyte and Doolittle (1982).

3. Results

3.1. Compositional distributions of coding sequences from human and Xenopus

The compositional distributions of the third codon positions in human and *Xenopus* are reported in Fig. 1. Each distribution shows the typical pattern of warm- and cold-blooded vertebrates, respectively [see Bernardi (1995) for a review]. Indeed, the GC_3 content in human ranged from 30 to 95%, and the gene distribution was strongly skewed toward high GC_3 levels. In contrast, the GC_3 content in *Xenopus* only ranged from 30 to 75%. The gene distribution was narrower, centered around 45%, and only a few genes (3.5%) exhibited a GC_3 level higher than 70%.

The present set of orthologous genes only represents less than 8 and 20% of the sequences that are retrievable at present in databanks for human and *Xenopus*, respectively. It is, however, representative of the entire set of coding sequences in that it shows the same compositional properties. As expected, the correlation coefficient ($R=0.46$, $p<10^{-4}$) and the slope ($s=2.88$) of the orthogonal regression line in the human versus *Xenopus* GC_3 plot (Fig. 2) are close to those previously reported, $R=0.57$ and $s=2.47$ (Bernardi et al., 1997).

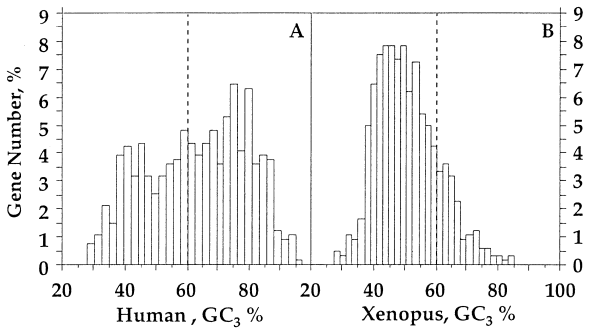

Fig. 1. Compositional distribution of third codon positions (GC_3) from human (**A**) and *Xenopus* (**B**). The number of genes analyzed for each organism is 460.

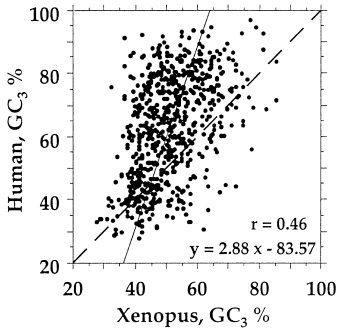

Fig. 2. Correlation between GC_3 values of orthologous genes from human and *Xenopus*. The number of gene pairs analyzed is 460. The equation of the orthogonal regression line and the regression coefficient are indicated at the bottom of the figure. The broken line is the diagonal.

3.2. Amino acid frequencies

The results concerning substitution matrices are reported in Fig. 3. Panels (A), (B) and (C), were arranged from top to bottom by increasing GC_3 levels, and in each panel, amino acids were sorted from GC-rich (black bars) to GC-poor (white bars), according to the GC content in the first and second codon positions (D'Onofrio et al., 1991).

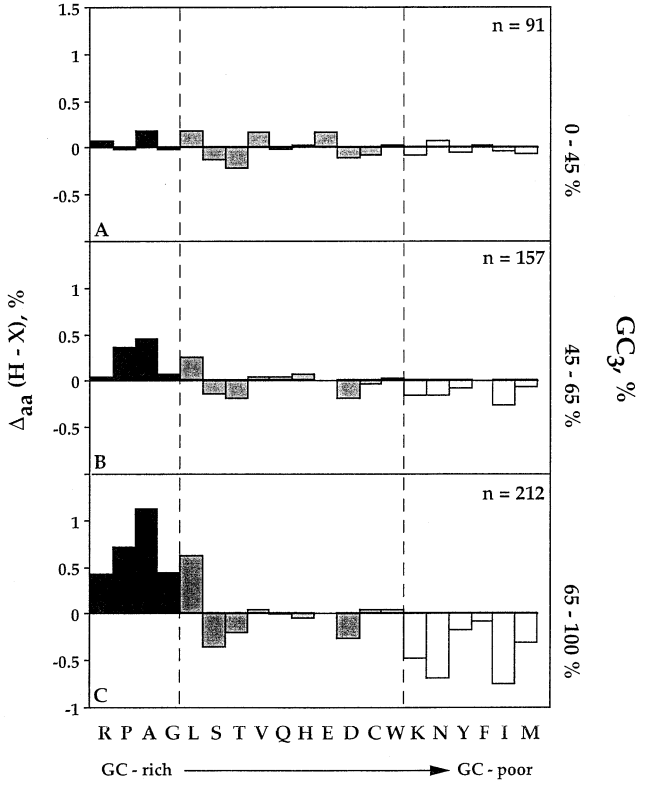

Fig. 3. Histograms showing amino acid frequency differences between orthologous protein from human and *Xenopus*. Proteins were partitioned in three groups according to the GC_3 values of the corresponding human coding sequences. *N* is the number of genes in each group. Amino acids encoded by GC-rich, GC-intermediate and GC-poor codons are shown in black, gray and white, respectively.

Table 1
Amino acid frequency diffrences obtained from the comparison of 460 orthologous genes from human and *Xenopus*[a]

	I[b]	II[b]	III[b]	$\Delta_{(III-I)}$	$\Delta_{(III-II)}$
Arg	0.063	0.033	0.417	0.354	0.384
Pro	−0.027	0.356	0.703	0.731	0.347
Ala	0.173	0.449	1.125	0.952	0.675
Gly	−0.030	0.065	0.437	0.467	0.372
Leu	0.175	0.244	0.612	0.436	0.367
Ser	−0.124	−0.154	−0.358	−0.233	−0.204
Thr	−0.221	−0.193	−0.203	0.018	−0.010
Val	0.162	0.042	0.036	−0.127	−0.006
Gln	−0.017	0.035	−0.011	0.006	−0.046
His	0.025	0.062	−0.057	−0.082	−0.119
Glu	0.156	0.005	0.011	−0.145	0.007
Asp	−0.110	−0.188	−0.273	−0.163	−0.084
Cys	−0.084	−0.036	0.036	0.120	0.072
Try	0.019	0.027	0.035	0.016	0.008
Lys	−0.091	−0.169	−0.482	−0.391	−0.313
Asn	0.063	−0.163	−0.700	−0.763	−0.537
Tyr	−0.051	−0.089	−0.173	−0.123	−0.084
Phe	0.025	0.014	−0.082	−0.107	−0.095
Ile	−0.040	−0.270	−0.761	−0.721	−0.491
Met	−0.067	−0.069	−0.311	−0.243	−0.241

[a] The last two columns report the difference between the third class and the two other.

[b] I, II and III represent the class of composition (respectively GC-poor, GC-intermediate and GC-rich class).

In the 0–45% GC_3 group, the Δaa values were very small in the three classes of amino acids. Indeed, the mean and standard deviations of the Δaa absolute values ($\langle|\Delta aa|\rangle$) were 0.073 ± 0.07 and 0.06 ± 0.02, for the GC-rich and the GC-poor classes, respectively, and only slightly higher in the intermediate class (0.11 ± 0.07).

In the 45–65% GC_3 group, $\langle|\Delta aa|\rangle$ was higher in the GC-rich and GC-poor classes (0.23 ± 0.21 and 0.13 ± 0.09, respectively), whereas it remained the same in the intermediate class, 0.10 ± 0.09.

In the 65–100% GC_3 group, the $\langle|\Delta aa|\rangle$ values of the two extreme amino acid compositional classes are much higher, reaching the maximum mean values of 0.67 ± 0.33 and 0.42 ± 0.28 for the GC-rich and the GC-poor class, respectively. The $\langle|\Delta aa|\rangle$ values of the intermediate class also showed a small increment, reaching a mean value of 0.16 ± 0.20.

The $\langle|\Delta aa|\rangle$ increase accompanying the increasing GC_3 levels is not to be ascribed to only a few amino acids. Indeed, using the absolute value of the highest frequency difference (that of Thr) in the $<45\%$ GC_3 group as a reference, 11 amino acids were above this value in the $>65\%$ GC_3 range, namely Arg, Pro, Ala, Gly, Leu, Ser, Asp, Lys, Asn, Ile and Met (see Table 1).

Interestingly, the substitution rate was rather constant in the three GC_3 groups (27, 23 and 26%, from the low

to the high GC_3 group), confirming the lack of dependence of rate on composition previously reported (Bernardi et al., 1993).

3.3. Orthologous genes in human, calf and Xenopus

In order to estimate the fluctuation values of amino acid frequencies in the absence and in the presence of a compositional transition, the three-species dataset of orthologous genes was analyzed, the human/calf Δaa values providing references for those of human/*Xenopus*. The three-species dataset of orthologous genes was divided as previously described (see Section 2). The histograms of the human/calf and human/*Xenopus* comparisons are reported in Fig. 4A–C and A′–C′ (see also Table 2). The range of Δaa variability in each GC_3 group of the human/calf comparison, delimited by broken lines, is reported on the corresponding human/*Xenopus* comparison.

In the human/*Xenopus* comparison, the number of Δaa values that lie outside the broken lines increases as the GC_3 levels increase. Indeed, in the lowest GC_3 group, only one Δaa value, that of Ala, is above the range, whereas in the highest GC_3 group, ten Δaa values are outside the broken lines. More precisely, the Δaa values of the GC-rich amino acids become positive, thus increasing in human, whereas Δaa values of the GC-poor amino acids become negative. Unexpectedly, the GC levels of the genes also influence the Δaa values of the intermediate class since an increase of Leu and a decrease of Asp and Ser were observed.

The *t*-test for paired comparisons shows that the Δaa values in the $<45\%$ GC_3 range for the human/calf comparison are not much different from those of the human/*Xenopus* comparison. They become increasingly significant in the 45–65% and $>65\%$ GC_3 ranges, with $p<5\times10^{-3}$ and $p<2\times10^{-3}$, respectively.

3.4. Hydrophobicity

The Δaa increase accompanying the GC_3 increase, prompted us to check the hydropathy of the proteins affected by the compositional transition *Xenopus*/human. The set of 460 orthologous genes was divided into 20 protein sub sets, and for each group, the hydrophobicity difference (ΔH) was calculated.

The ΔH values are positively correlated with the ΔGC_3 values (Fig. 5), with a correlation coefficient, $R=0.59$, that is statistically significant, $p<6.5\times10^{-3}$.

The amino acids were then partitioned into three sets (hydrophobic, amphipathic or hydrophilic) according to the hydropathy scale of Kyte and Doolittle (1982).

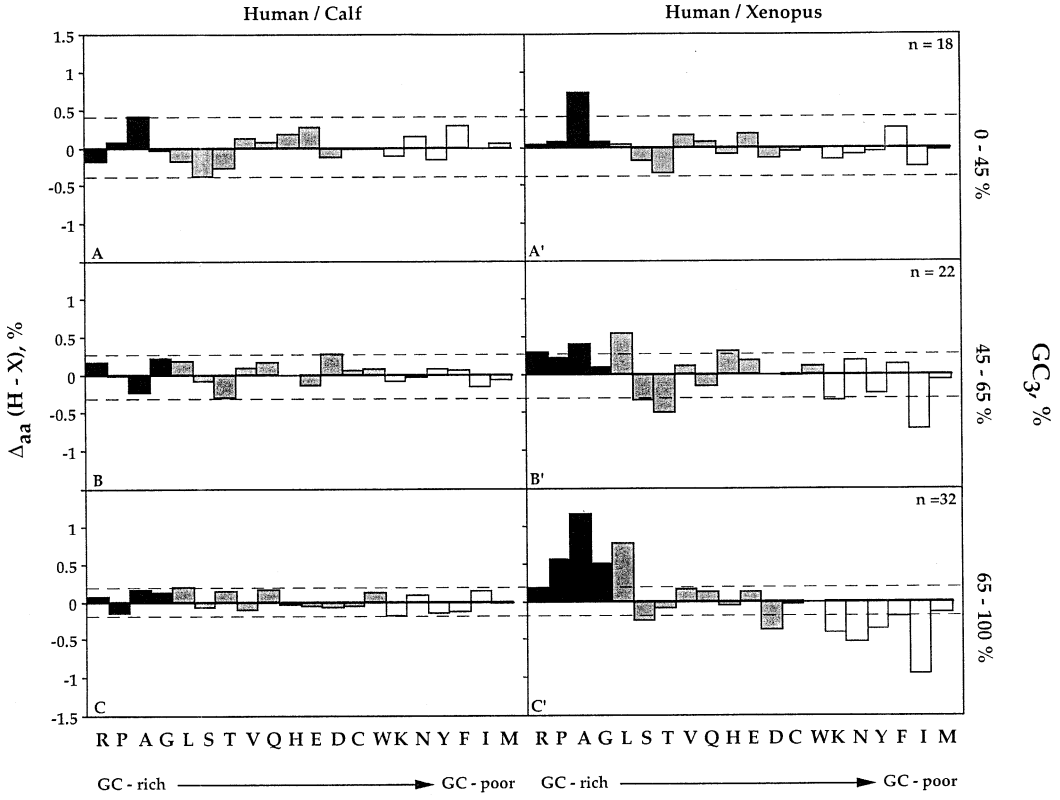

Fig. 4. Histograms showing Δaa between orthologous proteins from human and calf (left) shared by *Xenopus* (right) for each group of composition. *N* is the number of proteins for the GC_3 level considered.

Table 2
Amino acid frequency diffrences obtained from the comparison of the 72 orthologous genes from human, *Xenopus* (right), shared by calf (left)

	Human/calf			Human/*Xenopus*		
	I[a]	II[a]	III[a]	I[a]	II[a]	III[a]
Arg	−0.191	0.147	0.069	0.027	0.282	0.172
Pro	0.055	−0.027	−0.146	0.068	0.215	0.549
Ala	0.409	−0.241	0.155	0.709	0.390	1.159
Gly	−0.041	0.200	0.121	0.055	0.081	0.489
Leu	−0.191	0.174	0.181	0.027	0.537	0.764
Ser	−0.382	−0.107	−0.086	−0.191	−0.349	−0.258
Thr	−0.273	−0.307	0.138	−0.354	−0.510	−0.094
Val	0.123	0.080	−0.112	0.150	0.107	0.146
Glu	0.068	0.160	0.146	0.068	−0.175	0.120
His	0.177	−0.013	−0.052	−0.095	0.296	−0.069
Gln	0.259	−0.147	−0.060	0.164	0.175	0.112
Asp	−0.136	0.254	−0.086	−0.150	−0.013	−0.395
Cys	−0.027	0.040	−0.069	−0.068	−0.027	−0.043
Try	−0.027	0.067	0.112	−0.027	0.107	−0.009
Lys	−0.109	−0.094	−0.189	−0.177	−0.349	−0.421
Asn	0.136	−0.040	0.077	−0.095	0.175	−0.541
Tyr	−0.177	0.067	−0.155	−0.055	−0.255	−0.369
Phe	0.286	0.040	−0.138	0.245	0.121	−0.206
Ile	0.000	−0.174	0.129	−0.259	−0.725	−0.953
Met	0.041	−0.080	−0.034	−0.041	−0.081	−0.155

[a] I, II and III represent the class of composition (respectively GC-poor, GC-intermediateand GC-rich class).

Fig. 6 shows the exchange frequency differences observed among the three groups of amino acids. Hydrophilic amino acids are strongly substituted by hydrophobic amino acids in the GC-richest group, but barely so in the GC-poorest group. Hydrophilic amino acids are also very strongly substituted by amphipathic amino acids in the GC-richest group, whereas the trend is opposite and modest in the GC-poorest group. Finally, amphipathic amino acids are equally substituted by hydrophobic amino acids in both groups.

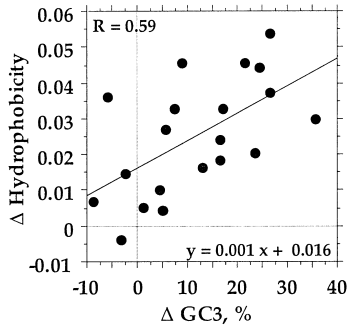

Fig. 5. Correlation between ΔGC_3 and Δhydrophobicity of orthologous proteins from human and *Xenopus*. Proteins were first sorted according to the GC_3 values of the corresponding human coding sequences, and then the entire dataset was divided in 20 groups, each comprising 23 proteins. The equation of the linear regression line and the regression coefficient are indicated at the bottom of the plot. The significance of the correlation is $p = 0.0065$.

4. Discussion

The comparison of 460 orthologous genes from human and *Xenopus* corroborates the observation that the majority of human genes underwent an increase in GC level (Bernardi, 1989, 1995; Bernardi et al., 1997). Indeed, as much as 77% of the genes showed a higher GC_3 level in human (Fig. 2). This compositional transition is not related to an increase in the substitution rate affecting the GC-rich genes since the substitution rate (25% on the average) in the GC-poor genes was as high as that in the GC-rich genes. The compositional transition must then be ascribed to the directional changes

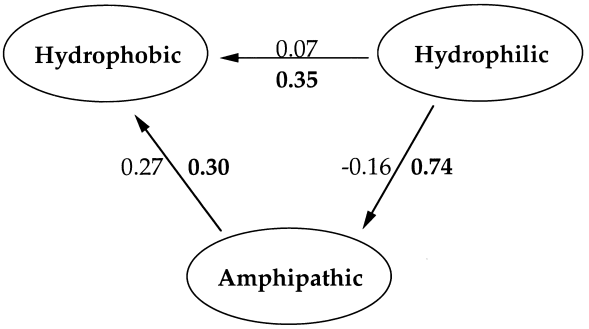

Fig. 6. Substitution frequency differences among hydrophobic, amphipathic, hydrophilic amino acids from *Xenopus* and human. Values in normal type refer to the 0–45% GC_3 group of proteins, whereas values in bold refer to the >65% GC_3 group of proteins.

that the GC-richest human genes underwent during the major transition.

The increase of the Δaa values, as the GC_3 level of the human genes increases, clearly supports the last point. Indeed, in the lowest GC_3 group (<45%), all Δaa values were located in a much smaller range than that of the highest GC_3 group (>65%).

The directionality of the changes was even more evident in a set of orthologous genes shared by human, calf and *Xenopus*. The amphibian/mammalian transition was compared with an intra-mammalian comparison, where no compositional changes took place (Mouchiroud and Bernardi, 1993; Mouchiroud et al., 1995). In the >65% GC_3 range, 11 amino acids in the human/*Xenopus* comparison were out of the range of variability of the human/calf comparison, compared to only one, Ala, in the lowest GC_3 range.

Despite an average substitution rate that is very close in the three groups of proteins, the type of amino acid (hydrophobic, amphipathic, hydrophilic) involved in changes in the GC-rich human genes indicates that the corresponding proteins have undergone strong changes in hydropathy. These changes could be expected from the results obtained in previous research on a very large number of genes from prokaryotes and eukaryotes (D'Onofrio et al., 1999a,b). Also, as expected, the GC_3 differences and hydropathy differences in the human/*Xenopus* comparison were the same as previously reported.

The difference in hydrophobicity of the human and *Xenopus* proteins is due to two phenomena, which have additive effects. First, human proteins have more hydrophobic amino acids compared to orthologous *Xenopus* proteins. Second, they also have fewer hydrophilic amino acids. This overall increase in hydrophobicity should lead to a better protein stability at the higher body temperature of warm-blooded vertebrates.

Body temperature may be the selective advantage underlying the changes under discussion, a concept supported by the increased hydrophilicity of proteins from psychrophilic organisms (Genicot et al., 1996; and papers quoted therein). Indeed, the finding that human proteins encoded by GC-rich genes are more hydrophobic than their *Xenopus* orthologues has a counterpart in the comparison between the trypsin from an Antarctic fish (*Paranotothenia magellanica*) and its homolog from calf (Genicot et al., 1996). In that case, the Antarctic fish protein exhibits a lower level of hydrophobic amino acids and a higher level of hydrophilic amino acids compared to the calf protein.

Acknowledgements

We would like to thank our colleagues Dr O. Clay and Dr F. Alvarez-Valin for their comments and valuable advice on computer programming.

References

Aïssani, B., Bernardi, G., 1991. CpG islands: features and distribution in the genome of vertebrates. Gene 106, 173–183.

Bernardi, G., Bernardi, G., 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24, 1–11.

Bernardi, G., Mouchiroud, D., Gautier, C., Bernardi, G., 1988. Compositional patterns in vertebrate genomes: conservation and change in evolution. J. Mol. Evol. 28, 7–18.

Bernardi, G., 1989. The isochore organization of the human genome. Annu. Rev. Genet. 23, 637–661.

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

Bernardi, G., Bernardi, G., 1990b. Compositional transitions in the nuclear genomes of cold-blooded vertebrates. J. Mol. Evol. 31, 282–293.

Bernardi, G., Bernardi, G., 1991. Compositional properties of nuclear genes from cold-blooded vertebrates. J. Mol. Evol. 31, 57–67.

Bernardi, G., Mouchiroud, D., Gautier, C., 1993. Silent substitutions in mammalian genomes and their evolutionary implications. J. Mol. Evol. 37, 583–589.

Bernardi, G., 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. 29, 445–476.

Bernardi, G., Hughes, S., Mouchiroud, D., 1997. The major compositional transitions in the vertebrate genome. J. Mol. Evol. 44, S44–S51.

Chiusano, M.L., D'Onofrio, G., Alvarez-Valin, F., Jabbari, K., Bernardi, G., 1999. Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. Gene 238, 23–31.

D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., Bernardi, G., 1991. Correlations between the compositional properties of human genes, codon usage and amino acid composition of proteins. J. Mol. Evol. 32, 504–510.

D'Onofrio, G., Jabbari, K., Musto, H., Alvarez-Valin, F., Cruveiller, S., Bernardi, G., 1999a. Evolutionary genomics of vertebrates and its implications. In: Caporale, L.H., Arber, W. (Eds.), Molecular Strategies in Biological Evolution. Ann. NY Acad. Sci., New York, pp. 1–14.

D'Onofrio, G., Jabbari, K., Musto, H., Bernardi, G., 1999b. The correlation of protein hydropathy with the composition of coding sequences. Gene. in press

Duret, L., Mouchiroud, D., Gouy, M., 1994. HOVERGEN: Homologous Vertebrate Genes data base. Nucleic Acids Res. 22, 2360–2363.

Felsenstein, J., 1989. Phylogeny Inference Package (Version 3.527). Cladistics 5, 164–166.

Genicot, S., Rentier-Delrue, F., Edwards, D., VanBeeumen, J., Gerday, C., 1996. Trypsin and trypsinogen from an Antarctic fish: molecular basis of cold adaptation. Biochim. Biophys. Acta 1298, 45–57.

Lobry, J.R., 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205, 309–316.

Hudson, A.P., Cuny, G., Cortadas, J., Haschemeyer, A.E.V., Bernardi, G., 1980. An analysis of fish genomes by density gradient centrifugation. Eur. J. Biochem. 112, 203–210.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Mouchiroud, D., Bernardi, G., 1993. Compositional properties of coding sequences and mammalian phylogeny. J. Mol. Evol. 37, 109–116.

Mouchiroud, D., Gautier, C., Bernardi, G., 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of non-synonymous substitutions. J. Mol. Evol. 40, 107–113.

Pace, C.N., Shirley, B.A., McNutt, M., Gajiwala, K., 1996. Forces contributing to the conformational stability of proteins. FASEB J. 10, 75–83.

Perrin, P., Bernardi, G., 1987. Directional fixation of mutations in vertebrate evolution. J. Mol. Evol. 26, 301–310.

Pizon, V., Cuny, G., Bernardi, G., 1984. Nucleotide sequence organization in the very small genome of a tetraodontid fish, *Arothron diadematus*. Eur. J. Biochem. 140, 25–30.

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

Sueoka, N., 1992. Directional mutation pressure and molecular evolution: equilibria and asymmetric phylogenetic branching. J. Mol. Evol. 34, 95–114.

Thiery, J.P., Macaya, G., Bernardi, G., 1976. An analysis of eukaryotic genomes by density gradient centrifugation. J. Mol. Biol. 108, 219–235.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.