COMPOSITIONAL PATTERNS IN VERTEBRATE GENOMES:

CONSERVATION AND CHANGE IN EVOLUTION

Giorgio Bernardi (1), Dominique Mouchiroud (2), Christian Gautier (2) and Giacomo Bernardi (1)

(1) Laboratoire de Génétique Moléculaire
Institut Jacques Monod, 2 Place Jussieu
75005 Paris (France)

(2) Laboratoire de Biométrie, U.A. 243
Université Claude Bernard Lyon I
69622 Lyon (France)

INTRODUCTION

Three approaches have recently provided new insights into the organization and evolution of the nuclear genomes of vertebrates. They are all based on the compositional properties of genome segments ranging in size from about 1 Kb, for coding sequences, to 100 Kb, for DNA fragments. The rationale for these approaches is the fact that vertebrate genomes are mosaics of isochores, which are evolutionarily relevant structures (see the following paragraph).

In the first approach, large DNA fragments (in the 30-100 Kb size range) were fractionated according to their GC levels.This allowed the study of their compositional distribution (by plotting the relative amounts of DNA fractions against their GC levels), and the localization of specific sequences to these fractions. Such investigations (see Bernardi et al., 1985, for a review) revealed that : (1) vertebrate genomes are made up of very long DNA segments (estimated to be larger than 200-300 Kb), the isochores, that are compositionally fairly homogeneous (at least above sizes of 3 Kb), and belong to a small number of classes characterized by different GC levels; (2) GC-rich isochores represent about one third of the genome of warm-blooded vertebrates, whereas they are absent from, or poorly represented in, most cold-blooded vertebrates; lesser, but highly significant differences in the compositional distribution of isochores were found between mammals and birds, and also, to a smaller extent yet, between most mammals and murids; (3) in warm-blooded vertebrates, GC-rich isochores are located in Giemsa light bands (or in Reverse bands) of metaphase chromosomes and replicate early in the cell cycle, whereas GC-poor isochores are located in Giemsa dark bands and replicate late in the cell cycle; in cold-

blooded vertebrates, metaphase chromosomes show poor Giemsa banding or no banding at all (see Medrano et al., 1988); (4) GC levels of coding sequences and their different codon positions are linearly correlated with those of the corresponding introns and of the intergenic non-coding sequences in which they are embedded; (5) the CpG doublets (the potential methylation sites) are very low in GC-poor, but almost "statistically abundant" (i.e., not under-represented compared to other doublets) in GC-rich coding sequences (as is also the case for the genomes of mammalian and avian viruses; Bernardi, 1985; Bernardi and Bernardi, 1986); (6) genes are mainly concentrated in the GC-richest isochores of the genomes of warm-blooded vertebrates.

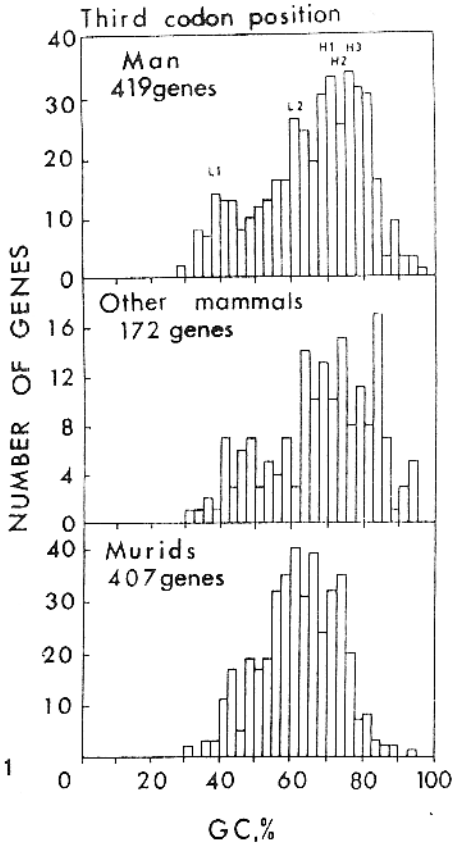
In the second approach, data from sequence banks were used (Bernardi et al., 1985; Bernardi and Bernardi, 1985, 1986; Mouchiroud et al., 1987, 1988; Perrin and Bernardi, 1987) to analyze the GC levels of individual coding sequences and the average GC levels of first, second and third codon positions within such sequences (see Fig. 1 for an example of compositional distribution of third codon positions). These investigations have shown that (1) in cold-blooded vertebrates, the compositional distribution curves of coding sequences and of DNA fragments are roughly symmetrical; (2) in warm-blooded vertebrates, GC-rich coding sequences represent the majority of coding sequences, whereas GC-rich DNA fragments correspond to a minor part of the genome; (3) the compositional distributions of coding sequences are different in mammals and birds (chicken), and also, but less so, in most mammals on the one hand, and murids on the other.

In a third approach, we have compared the GC levels of pairs of homologous coding sequences from vertebrates and of their different codon positions (Perrin and Bernardi, 1987; Mouchiroud and Gautier, 1988; Mouchiroud et al., 1988). These comparisons have provided information on the evolutionary conservation of regional compositional patterns (as we shall call here the compositional distributions of DNA fragments, of coding sequences and of their different codon positions), and shown how such patterns can shift in evolution.

Here, we shall investigate in more detail the conservation and the shifts of compositional patterns of vertebrate genomes using more recent sets of data than those previously analyzed. We shall then discuss the mechanisms and the causes of these phenomena. A more detailed presentation of this work will appear elsewhere (Bernardi et al., 1988).

Two types of compositional patterns in mammals

We have recently identified two types of compositional distributions in the available mammalian coding sequences. These types will be called here the "general" distribution and the "murid" distribution. Indeed, when coding sequences are examined (as a whole, or in their different codon positions), those of several mammals (mainly man, artiodactyls and rabbit) show similar broader distributions, and higher average GC values compared to those of murids (Mouchiroud et al., 1987, 1988; data for third codon positions are shown in Fig. 1).



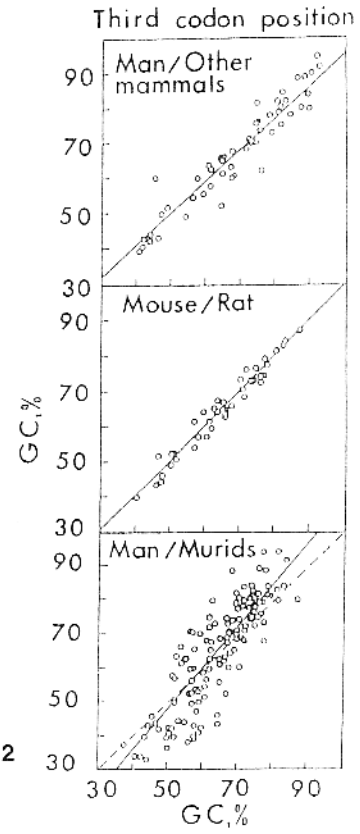


Fig. 1. (left) Compositional distribution of third codon positions for genes (i) from man; (ii) from mammals other than man, murids and hamster; and (iii) from murids (rat and mouse). A 2.5% GC window was used. Tentative identifications of different compositional classes of coding sequences corresponding to the compartments of the human genome (L1,L2, H1, H2 and H3) are indicated following Mouchiroud et al.(1987, 1988).

Fig. 2. (right) Relationships between GC contents of third codon position for pairs of homologous genes (i) from man (ordinate) and other mammals (except murids and hamster; abscissa); (ii) from mouse (ordinate) and rat (abscissa); and (iii) from man (ordinate) and murids (abscissa).

These similarities and differences are confirmed by comparisons of GC levels of homologous coding sequences, and of their different codon positions. Indeed, these GC levels show linear relationships, passing through the origin and showing unity slopes, in the case of both the "general" and the "murid" distribution (see Fig. 2 for plots of third codon positions). When homologous sequences from the "general" and from the "murid" distributions are compared with each other, linear relationships with high correlation coefficients are still found, but points show a larger scatter and the slope is significantly different from unity in the case of third codon positions (Mouchiroud et al., 1988; see Fig. 2). This difference is due to deviations mainly affecting the two opposite ends of distributions. In murids, GC-poor and GC-rich coding sequences (and their different codon positions) are less GC-poor and less GC-rich, respectively, than in other mammals. It should be stressed that "minor" shifts correspond to orderly changes in the genome, since the order of coding sequences by increasing GC content is very largely conserved in the two distributions (Mouchiroud et al., 1988).

Parallel investigations have shown that the compositional distribution of large DNA fragments (in the 30-100 Kb size range) of mouse is narrower (Salinas et al., 1986; Zerial et al., 1986) and, in contrast with that of coding sequences, centered on a slightly higher GC level, compared to man. When the distribution of DNA fragments from rat and mouse, or from man and mouse, are plotted against each other, the results are very similar to those obtained in Fig. 2. In the first case, the distributions are identical; in the second, the distribution of human DNA fragments starts at lower GC values and ends at higher ones, compared to mouse.

The causes of the conservation of mammalian compositional patterns : compositional conservation of the base substitution process and negative selection at the isochore level

The extraordinary conservation of compositional patterns just described can only be due to two factors, the base substitution process itself and/or selection. It is conceivable that, when averaged over large time intervals, base substitutions exhibit a certain degree of compositional conservation. A complete absence of compositional biases in mutations is, however, most unlikely, in view of the demonstrated influence of nearest-base composition, particularly in coding sequences where composition is averaged over a few hundred bases only. Under these circumstances, the additional intervention of the second factor, selection, appears to be inescapable, as already suggested (Bernardi and Bernardi, 1986).

Selection of individual point mutations, however, can hardly be the explanation for the findings under consideration, because this would require assigning a significant advantage or disadvantage to any event affecting one base pair out of about 3.10^{9}. Even if this is conceivable for a very small number of sites, where base substitutions change critical amino acids or critical

nucleotides in tRNAs, rRNAs and in signal sequences, it cannot be the general rule in vertebrate genomes, where more than 90% of DNA is noncoding, because this would cause an unbearable mutational load. In contrast, a negative, "stabilizing", selection process, acting at a regional level, and eliminating deviations from a narrow range of values, presumably corresponding to functionally optimal regional compositional patterns, appears to be the only plausible explanation for our results. This regional level can be identified with isochores, since isochores are, in fact, the genome segments exhibiting compositional conservation.

It should be stressed that isochores correspond to individual or contiguous chromosomal domains, the "chromatin loops", since these are each estimated to comprise 30-300 Kb of supercoiled DNA in mammals. In turn, chromatin loops are supposed to correspond to replicons and to contiguous transcription units.

Compositional shifts in the evolution of vertebrate genomes

Two "major" shifts in compositional patterns occurred in the evolution of vertebrate genomes. Indeed, massive changes in extended genome segments led (i) to the formation, in mammals and birds, of GC-rich isochores that are absent or scarcely represented in the vast majority of cold-blooded vertebrates (Bernardi et al., 1985); and (ii) to changes in the compositional distributions of coding sequences (Bernardi et al., 1985; Mouchiroud et al., 1987; see Fig. 3). These shifts were completely independent from each other, since the paleontological record indicates that mammals derived from therapsids over 200 million years ago, birds from dinosaurs about 150 million years ago. In agreement with this conclusion, compositional patterns of birds and mammals are different from each other (see Fig.3 and below) and from those of reptiles. Even if the precise durations of the shifts are not known, they certainly were much shorter than the time of existence of mammals and birds.

The main mechanism by which the "major" compositional shifts were achieved was a directional fixation of point mutations (Perrin and Bernardi, 1987). Indeed, when homologous coding sequences from 21 genes of cold-blooded vertebrates and from 41 genes of warm-blooded vertebrates were compared, most of them showed GC increases mainly in third codon positions, but also in first and second positions (implying, therefore, amino acid changes). A smaller number of homologous genes showed no change in GC levels and corresponded to genes located in GC-poor isochores of warm-blooded vertebrates. (Three exceptional cases, in which GC levels were lower in warm-blooded vertebrates, corresponded, in all likelihood, to genes located in the very scarce GC-rich isochores that arose in some cold-blooded vertebrates).

The situation just described can also be seen in comparisons of homologous coding sequences from man and Xenopus. These show that most third codon positions of man are higher in GC, a second smaller set of positions is equal, and just one position is lower (Fig. 4). Expectedly, no significant linear correlations were found in this case between GC levels of codon positions from homologous genes (contrary to what was found for mammalian genes; Fig. 2).

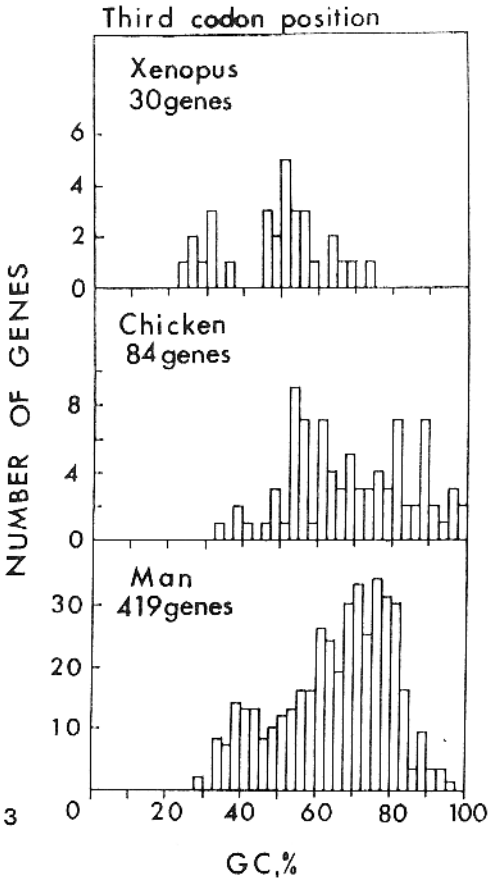
Several additional phenomena accompanied major shifts: (i) gene translocations led to the very high gene concentrations found in the GC-richest compartments of genomes from warm-blooded vertebrates; an example of this phenomenon is that of and globin genes, which are clustered in Xenopus, but underwent translocations to different chromosomes in both mammals and birds; this process was accompanied, or followed, by GC increases in the gene (in mammals), or in both and genes (in birds); (ii) large changes in chromosome structure took place, as indicated by the appearance of Giemsa and Reverse bands; moreover, the Giemsa light (or Reverse) bands apparently acquired a more complex structure relative to Giemsa dark bands, since their DNA is a mosaic of the different GC-rich isochores and also seems to have a less dense packing (G. Bernardi, unpublished observations); (iii) non-methylated CpG-rich islands, associated with genes (Bird, 1987), that are absent, or very poorly represented, in fishes and amphibians, became abundant and concentrated in GC-rich isochores of warm-blooded vertebrates (B. Aissani, G. Bernardi and G. Bernardi, unpublished observations).

A "minor" compositional shift, involving more moderate changes (towards both GC increases and GC decreases) in less extended genome sections, separated murids (as well as cricetids and spalacids) from most other mammals (Salinas et al., 1986; Zerial et al., 1986; Mouchiroud et al., 1987, 1988; Mouchiroud and Gautier, 1988). Amplification and insertion of GC-rich interspersed repeats, like Alu sequences (Bernardi et al., 1985; Zerial et al., 1986) apparently also played a role in this shift. In contrast with "major" shifts, in "minor" shifts correlations are still found between GC levels of third codon positions from homologous genes (compare Fig. 2 and 3), and the order of coding sequences by increasing GC is largely preserved.

The causes of compositional shifts in genome evolution : negative and positive selection at the isochore level

An explanation for the compositional shifts (Bernardi and Bernardi, 1986) is that they are mainly due to both negative and positive, "directional", selection acting at the isochore level, (positive selection at individual nucleotides is likely to be so rare that it will be neglected in the present discussion).

In the case of compositional shifts, negative selection should be visualized as eliminating only compositional deviations directed towards lower GC contents instead of any compositional deviation(as in the conservative evolutionary process discussed before). On the other hand, positive selection is the only explanation that can also account for the diverse events accompanying the major shifts in compositional patterns (gene translocations into GC-rich isochores, chromosome restructuring, formation or increase of unmethylated CpG islands). Moreover, positive selection is an explanation that rests on demonstrated functional advantages.



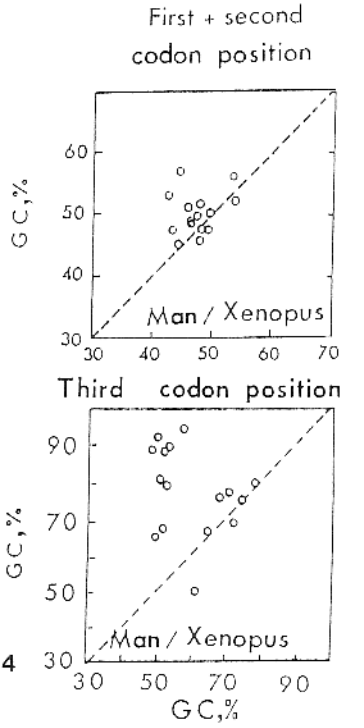


Fig. 3. (left) Compositional distribution of third codon positions for genes from Xenopus, chicken and man. Other indications as in Figure 1.

Fig. 4. (right) GC contents of first + second and third codon positions of pairs of homologous genes are plotted against each other for man and Xenopus (16 genes). Data for man correspond to the ordinate, data for Xenopus to the abscissa.

Indeed, as already pointed out (Bernardi and Bernardi, 1986), the increase in body temperature that accompanied these shifts, is associated with GC increases that have advantageous consequences : (i) GC increases in first and second codon positions lead to amino acid changes that confer (Argos et al., 1979) thermal stability to proteins; indeed, GC increases in coding sequences of vertebrates have been shown to be accompanied by increases in stabilizing amino acids (like alanine and arginine) and by decreases in destabilizing amino acids (like serine and lysine) in the encoded proteins (Bernardi and Bernardi, 1986); interestingly, reports on choices of thermally stabilizing amino acids and of GC-rich codons in thermophilic organisms are rapidly accumulating in the literature; (ii) GC increases in introns, in third codon positions and in DNA segments corresponding to untranslated regions contribute, in addition, to the thermal stability (Wada and Suyama, 1986) of primary transcripts and mRNAs; (iii) GC increases in intergenic, noncoding sequences can also conceivably help in stabilizing DNA structures, possibly through changes in DNA-protein interactions.

It should be stressed that, if our discussion was centered so far on temperature as the main selection factor responsible for the major compositional shifts of vertebrate genomes, it is only because precise selective advantages can be identified in this case. Our general suggestion is, however, that regional negative and positive selection is due to the functional advantages associated with the changes in compositional patterns. These advantages may be of a very different nature, and may be elusive because of the interplay of many factors. Thus, we do not have yet, for example, an explanation for the minor shifts leading to the murid pattern.

The selectionist-neutralist controversy

At this point, it may be interesting to reconsider the selectionist-neutralist controversy which has continued for the past twenty years (Kimura, 1968, 1983), in the light of the present results.

First of all, it should be stressed that previous investigations (including those on which the neutral theory was built) practically only dealt with the evolution of mammalian proteins and genes; they only concerned, therefore, the accumulation of mutations in the conservative mode of evolution (we neglect here the differences between the general compositional pattern of mammals and the murid pattern). In other words, these investigations missed (i) the compositional conservation of homologous coding sequences (and isochores), that characterize the conservative mode of evolution; and (ii) the compositional shifts of homologous coding sequences (and of isochores), that took place at the transition between cold-blooded and warm-blooded vertebrates.

As a consequence, the selectionist-neutralist controversy was only considered at the level of individual base substitutions. On this basis, the conclusions were drawn (Kimura, 1983) that (i) "the great majority of evolutionary changes at the molecular level are caused not by Darwinian selection acting on advantageous mutations, but by

random fixation of selectively neutral or nearly neutral mutants"; and (ii) "only a minute fraction of DNA changes in evolution are adaptive in nature". It is obvious that these conclusions cannot be considered any longer as having a general validity in genome evolution, since they could not take into account some aspects of the problem.

Indeed, positive selection is not only the very rare process operating on individual base substitutions, but also a regional process largely underlying compositional shifts, that appear to have an adaptive value. It should be noted that the transitional mode of genome evolution is rare and concerns only a very small part of the genome in mammals. It is, however, increasingly more frequent and more extensive when moving from warm-blooded to cold-blooded vertebrates, to invertebrates, plants, unicellular eukaryotes and prokaryotes, as indicated by the increasing spread of genome compositions of these organisms. This phenomenon may be related to the increasingly variable environmental conditions to which these genomes are submitted.

Moreover, even in the conservative mode of evolution, there are compositional constraints that affect the fixation of mutations. These particular "selective molecular constraints" not only contradict the "randomness in the pattern of substitutions" predicted by the neutral theory (Kimura, 1983), but also are so pervasive that the definition of a neutral mutation rate (namely, of a substitution rate reaching the maximum value set by the mutation rate) has proven elusive so far (see Zuckerkandl, 1986). A number of individual substitutions occurring in the conservative mode of evolution (which comprises most changes in the evolution of vertebrates) may, however, conceivably approach neutral, nearly neutral or slightly deleterious mutations, as described by the neutral theory.

Under these circumstances, while the overall process of genome evolution can be fully understood only within a selectionist framework, the neutralist view and the selectionist view (as presented here) appear to be more complementary than contradictory.

REFERENCES

Argos, P., Rossmann, M.G., Grau, U.M., Zuber, A., Franck, G. and Tratschin, J.D., 1979, Thermal stability and protein structure, Biochemistry, 18:5698.

Bernardi, G., 1985, The organization of the vertebrate genome and the problem of the CpG shortage, in: "Biochemistry and Biology of DNA methylation", G.L. Cantoni and A. Razin, eds., p. 3, Alan Liss, New York.

Bernardi, G. and Bernardi, G., 1985, Codon usage and genome composition, J. Mol. Evol., 22:363.

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

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

Bernardi, G. and Bernardi, G., 1986, Compositional constraints and genome evolution, J. Mol. Evol., 24:1.
Bird, A.P., 1987, CpG islands as gene markers in the vertebrate nucleus, Trends in Genetics, 3:342.
Kimura, M., 1968, Evolutionary rate at the molecular level, Nature, 217:624.
Kimura, M., 1983, The neutral theory of molecular evolution, Cambridge University Press, Cambridge, England.
Medrano, L., Bernardi, G., Couturier, J., Dutrillaux, B. and Bernardi, G., 1988, Chromosome banding and genome compartmentalization in fishes, Chromosoma, 96:178.
Mouchiroud, D., Fichant, G. and Bernardi, G. 1987, Compositional compartmentalization and gene composition in the genomes of vertebrates, J. Mol. Evol., 26:198.
Mouchiroud, D. and Gautier, C., 1988, High codon-usage changes in mammalian genes, Mol. Biol. Evol., 5:192.
Mouchiroud, D., Gautier, C. and Bernardi, G., 1988, The compositionaldistribution of coding sequences and DNA molecules in man and murids, J. Mol. Evol., in press.
Perrin, P. and Bernardi, G. 1987, Directional fixation of mutations in vertebrate evolution, J. Mol. Evol., 26:301.
Salinas, J., Zerial, M., Filipski, J. and Bernardi, G., 1986, Gene distribution and nucleotide sequence organization of the mouse genome, Eur. J. Biochem., 160:469.
Wada, A. and Suyama, A. 1986, Local stability of DNA and RNA secondary structure and its relation to biological function, Prog. Biophys. Mol. Biol., 47:113.
Zerial, M., Salinas, J., Filipski, J. and Bernardi, G., 1986, Gene distribution and nucleotide sequence organization in the human genome, Eur. J. Biochem., 160:479.
Zuckerkandl, E. 1986, Polite DNA: functional density and functional compatibility in genomes, J. Mol. Evol., 24:12.