THE VERTEBRATE GENOME: ISOCHORES AND 'EVOLUTION

GIORGIO BERNARDI

Laboratoire de Génétique Moleculaire Institut Jacques Monod,
2 Place Jussieu, 75005 Paris, France.
Tel. (33-1) 43 29 58 24 Fax (33-1) 44 27 79 77

ABSTRACT

Vertebrate genomes are mosaics of isochores, namely of long (>300 Kb), compositionally homogeneous DNA segments that can be subdivided into a small number of families characterized by different GC levels. In the human genome (which is representative of a number of mammalian genomes, and, more broadly, of the genomes of warm-blooded vertebrates), the compositional range of isochores is 30% to 60% GC, and five families of isochores have been identified: two GC- poor families, L1 and L2, representing together 62% of the genome, and three GC-rich families, H1, H2 and H3, representing 22%, 9% and 3%, respectively; (the remaining 4% of the genome are formed by satellite and ribosomal DNA). Gene concentration is strikingly non-uniform, being highest in the H3 isochore family, lowest in the L1 + L2 families, and intermediate in the H1 + H2 families. The H3 family corresponds to T(elomeric) bands of metaphase chromosomes, and the L1 + L2 families to G(iemsa) bands, whereas R(everse) bands comprise both GC-poor and GC-rich isochores.

The compositional distributions of large genome fragments, of exons (and their codon positions), and introns are correlated with each other. They represent compositional patterns and are very different between the genomes of cold- and warm-blooded vertebrates, mainly in that the former are much less heterogeneous in base composition, are generally GC-poorer, and never reach the highest GC levels attained by the latter. Only relatively small compositional differences are found among the genomes of either cold- or warm-blooded vertebrates.

Compositional patterns allow to define two modes in genome evolution: a conservative mode, with no compositional change, and a transitional (or shifting) mode, with compositional changes.

The conservative mode can be observed among either cold- or warm- blooded vertebrates. The transitional mode comprises both major and minor compositional changes. In vertebrate genomes, the major changes are associated with the appearance of GC-rich and very GC-rich isochores in mammalian and avian genomes. Mutational biases play a role in both modes of compositional evolution. According to one viewpoint, mutational biases are responsible for the transitional mode of evolution of bacterial genomes; in the conservative mode of evolution of vertebrates, they accomplish their role in conjunction with differences either in chromatin structures that modulate replication errors or in chromatin transcriptional activities that may lead to various extents of repair DNA synthesis.

According to another viewpoint, defended here, selection controls, at the isochore level, mutational biases both in the conservative and in the transitional mode of evolution.

Keywords: Genome composition - Mutational bias - Selection

Abbreviations: GC, % G+C; Kb, kilobase(s); Mb, Megabase(s)

INTRODUCTION

Vertebrate genomes are mosaics of *isochores*, namely of long, compositionally homogeneous DNA segments which can be subdivided into a small number of families characterized by different GC levels. Figure 1 displays a scheme of the isochores forming the human genome, which is representative of a number of mammalian genomes and, more broadly, of the genomes of warm-blooded vertebrates. The compositional heterogeneity of high molecular weight bovine DNA was discovered twenty years ago (Filipski et al., 1973) using Cs_2SO_4 preparative density gradient centrifugation in the presence of a sequence-specific DNA ligand, Ag^+. This heterogeneity concerned the so-called "main band" DNA, and was different from the heterogeneity previously detected in analytical CsCl gradient (Sueoka, 1959), which was essentially due to the eight GC-rich satellite DNAs that form 23% of the bovine genome (Filipski et al., 1973; Cortadas et al., 1977; Macaya et al., 1978; Kopecka et al., 1978). The basic properties of isochores (as they were called by Cunny et al., 1981) from the vertebrate genomes were defined by Thiery et al. (1976) and Macaya et al. (1976) and were later studied in further detail (see Bernardi, 1989, for a review).

Isochores from the mouse genome were estimated to exceed 300 Kb in size (Macaya et al., 1976). Recent work has shown that, in the

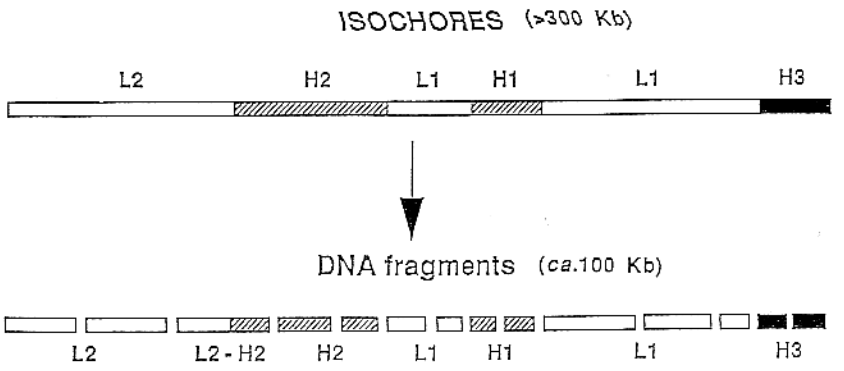


Fig. 1. Scheme of the isochore organization of the human genome. The genome is a mosaic of large DNA segments (> 300 Kb), the isochores, that are compositionally homogeneous and belong to a small number of families, GC-poor (L1 and L2), GC-rich (H1 and H2) and very GC- rich (H3). Physical and enzymatic degradation occurring during DNA preparation generates large DNA fragments, currently around 100 Kb in size. (Modified from Bernardi et al., 1985).

human genome, isochore size ranges between 0.36 and >0.7 Mb in the dystrophin gene (Bettecken et al., 1992) and is larger than 1 Mb in the cystic fibrosis locus (Krane et al., 1991). These sizes (also investigated by Ikemura and Aota, 1988; and by Ikemura et al., 1990) are intermediate between those of genes, or gene clusters, and those of chromosomal bands; they cover the most interesting range for physical and compositional mapping.

In the human genome, the compositional spectrum of isochores ranges between 30% and 60% GC, and five families of isochores have been identified, two GC-poor families, L1 and L2, representing together 62% of the genome, and three GC-rich families, H1, H2 and H3, representing 22%, 9% and 3%, respectively. The remaining 4% of the genome consist of satellite and ribosomal DNAs, which can also be visualized as isochores because of their homogeneous base composition (Bernardi, 1989).

Compositional patterns and compositional correlations

The *compositional distribution* of large (*ca.* 100 Kb) DNA fragments (such as those forming current DNA preparations) represents a *compostional pattern* that reflects the *isochore pattern* (see Figure 2).

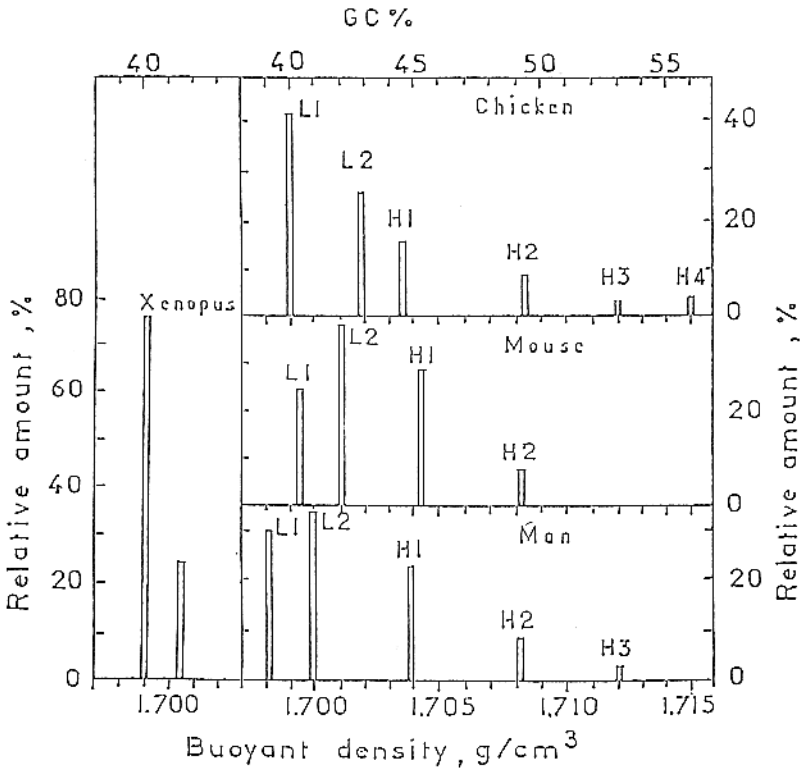


Fig. 2. Histograms showing the relative amounts, modal buoyant densities and GC levels of the "major DNA components" from Xenopus, chicken, mouse and man. The major DNA components are the large DNA fragments (see Figure 1) derived from the isochores families. Satellite and minor DNA components (like ribosomal DNA) are not shown in these histograms. (From Bernardi, 1989; see this reference for further details).

Other compositional patterns are represented by the compositional distributions of exons (and of their codon positions; see Figure 3) and of introns. These compositional patterns characterize *genome phenotypes* (Bernardi & Bernardi, 1986), which are very different in cold- and warm-blooded vertebrates. The main differences are that the former are much less heterogeneous in composition, are generally GC-poorer, and never attain very high GC levels compared to the latter (Bernardi & Bernardi, 1990a,b; 1991). Smaller compositional differences exist among the genomes of either cold-blooded or warm-blooded vertebrates.

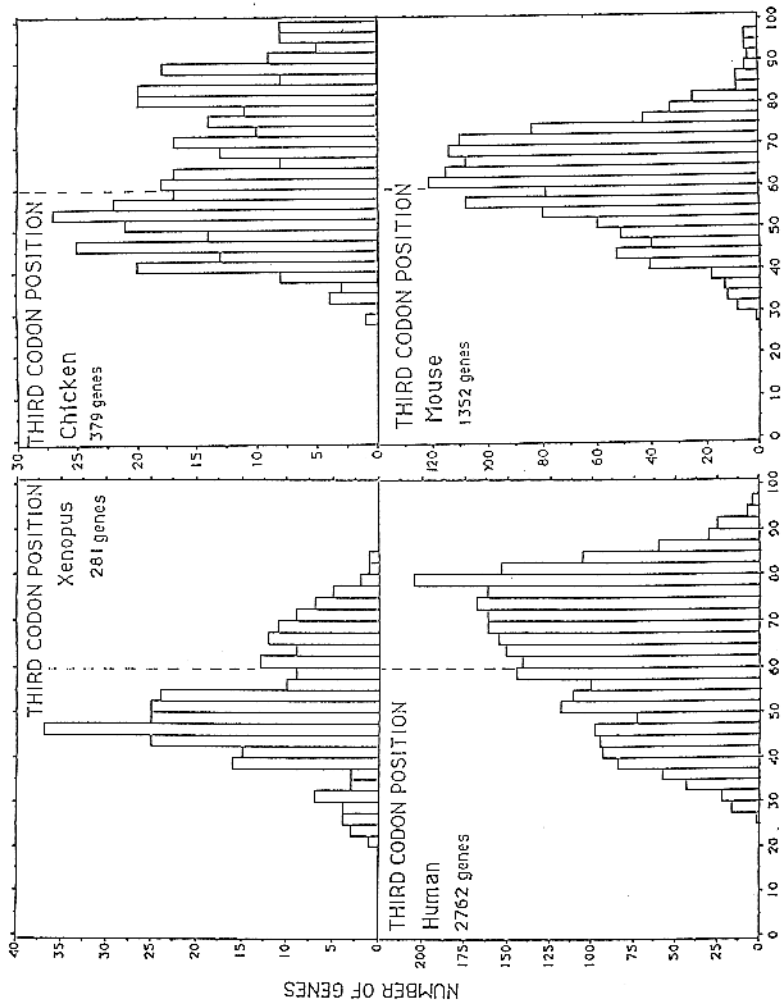


Fig. 3. Compositional distribution of third codon positions of genes from Xenopus, chicken, mouse and man. The broken line corresponds to the 60% GC level in all histograms (From Mouchiroud et al., 1992).

Compositional correlations hold between exons (and their codon positions) and the isochores in which they are embedded (Figure 4A), as well as between exons and the corresponding introns (Bernardi et al., 1985; Bernardi & Bernardi, 1985; 1986; Bernardi, 1989); Aïssani et al., 1991; D'Onofrio et al., 1991; Mouchiroud et al., 1991; see also Ikemura, 1985, and Aota & Ikemura, 1986). These compositional correlations link, in a linear fashion, the coding sequences and the non-coding sequences which surround them, or are contained in them. Moreover, a *universal correlation* among codon positions (Bernardi & Bernardi, 1985; 1986; Wada et al., 1991; D'Onofrio & Bernardi, 1992; Wada, 1992) was found to hold for all genomes (Figure 4B).

The correlations between coding and non-coding sequences and those among codon positions define a *genomic code* (Bernardi, 1990; Bernardi & Bernardi, 1991; D'Onofrio & Bernardi, 1992), which indicates that *compositional constraints* (Bernardi & Bernardi, 1986; also called AT pressure and GC pressure by Jukes & Bhushan, 1986) operate in the same direction, but not to the same extent, on all codon positions and on coding as well as on non-coding sequences.

The gene distribution in the human genome

The compositional correlation which links GC levels of third codon positions of human genes with the GC levels of the extended sequences in which the genes are located (Figure 4A) can be used in order to assess gene distribution in the different isochore families and to quantify the finding (Bernardi et al., 1985) that gene distribution in the human genome is strikingly non-uniform. This approach (Figure 5) has shown that, while 34% of all genes currently present in gene banks are contained in isochore families L1 and L2, 38% are contained in H1 and H2, and 28% in H3 (Mouchiroud et al., 1991). If the gene sample used is representative of all human genes, and if account is taken of the different relative amounts of isochore families (see Introduction), gene concentration in H3 would be 16 times higher than in L1+L2 and 8 times higher than in H1+H2. These ratios are, however, probably underestimated, because housekeeping genes, which are likely to be more abundant in H3 than in other isochore families (see next section), are currently very much under- represented in gene banks (Mouchiroud et al., 1991).

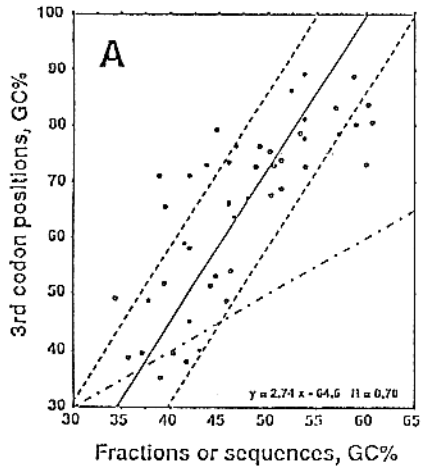
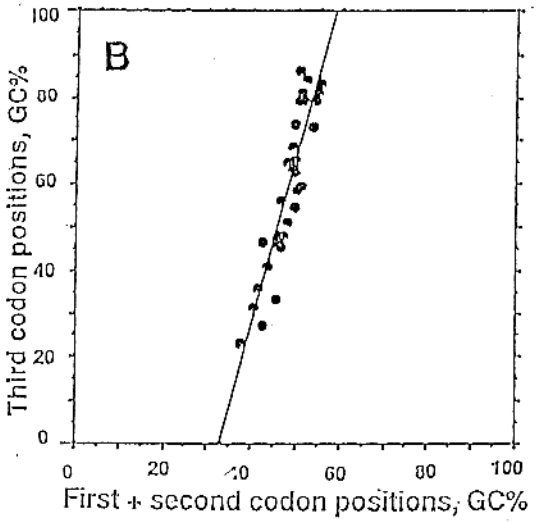


Fig. 4 A. GC levels of third codon positions from human genes are plotted against the GC levels of DNA fractions (solid circles) or extended sequences (open circles) in which the genes are located. The correlation coefficient and the slope are indicated. The dash-and-point line is the diagonal line (slope = 1). GC of third codon positions should fall on this line if they were identical to GC levels of surrounding DNA. The broken lines indicate a ± 5% GC range around the slope (From Mouchiroud et al., 1991).



B. Plot of GC levels of third codon positions against GC levels of first + second positions of prokaryotic and eukaryotic genomes. All values are averaged per genome (or genome compartment, in the case of compositionally compartmentalized genomes). (From D'Onofrio & Bernardi, 1992).

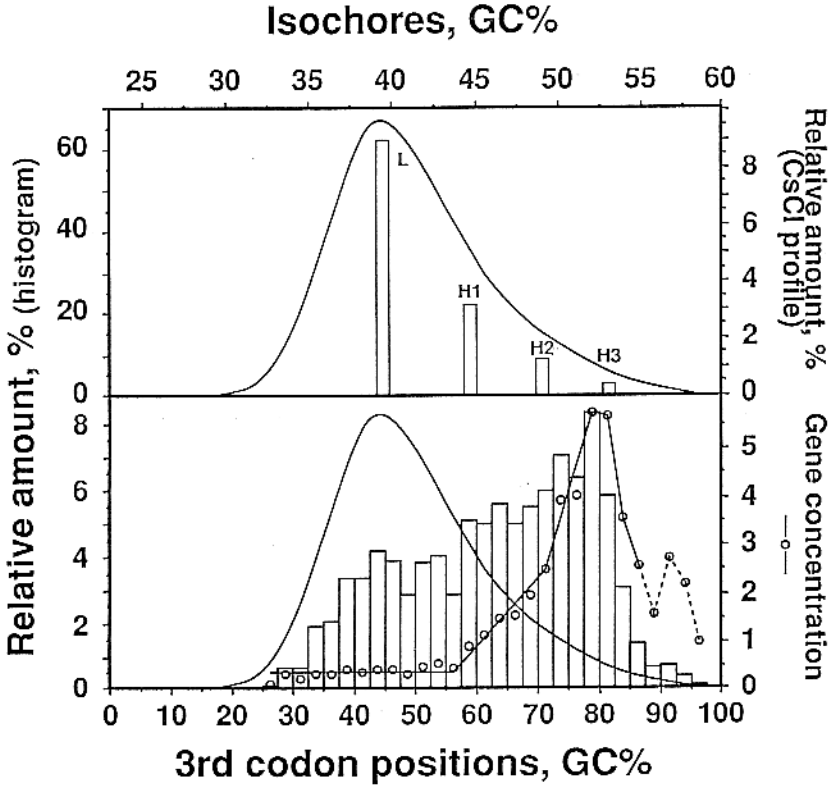


Fig. 5.A. A histogram of relative amounts of isochore (or "major DNA components") L (L1 + L2), H1, H2, H3 from the human genome. The upper scale concerns the GC levels of the isochores, as well as those of a CsCl profile of human DNA, and was obtained from the lower scale through the correlation of Figure 4A.
B. Histogram of relative amounts of human genes divided in classes according to GC levels of third codon positions. The profile of gene concentration in the human genome is also shown (open circles). The profile was obtained by dividing the relative amounts of genes in each 2.5% GC interval of the histogram by the corresponding relative amounts of DNA, as deduced from the CsCl profile (Modified from Mouchiroud et al., 1991).

The results just described (i) indicate that increasing gene concentrations are accompanied by increasing GC levels in the genome of warm-blooded vertebrates; the evolutionary process underlying this phenomenon will be discussed later; and (ii) independently confirm the classification of isochore families, which was originally based on purely compositional grounds; for instance, gene concentration is low and constant over isochore families L1 and L2, and is highest in isochore family H3. Very

interestingly, the gradient of gene concentration is paralleled by a series of changes in a number of properties which have functional significance. This will be illustrated by describing the extreme situation found in the GC-richest isochore family H3.

The human genome core

The isochore family H3 corresponds to a genome compartment endowed with very remarkable properties. This family has not only the highest GC level and the highest gene concentrations, but also the highest concentrations of CpG doublets (Bernardi, 1985), the only potential sites of methylation in vertebrates, and the highest concentrations of CpG islands (Aïssani & Bernardi, 1991a,b), which are very GC-rich sequences characterized by abundant, unmethylated CpG doublets (Bird, 1986). Since CpG islands (which are located in the 5'flanking sequences of genes) are preferentially associated with housekeeping genes (Gardiner- Garden & Frommer, 1987), the latter should be more abundant in H3 than in the other isochore families.

The coding sequences of H3 isochores are much higher in GC level than their genomic environment, compared with those from other isochore families, especially from GC-poor isochores (Aïssani et al., 1991). Moreover, these genes and their associated CpG islands are characterized by a particular chromatin structure, with nucleosome-free regions, absence or scarcity of histone H1, and acetylation of histones H3 and H4 (Tazi & Bird, 1990; see also Aïssani & Bernardi, 1991a,b). These properties make these chromosomal regions more "open", as also indicated by their sensitivity to nuclease attack (Kerem et al., 1984).

The H3 isochore family presumably has the highest level of transcription because of its very high concentration of genes, and especially of housekeeping genes. It also has the highest recombination rate, possibly because of its "open" chromatin structure and because of the abundance of repetitive sequences, like Alu sequences and minisatellites. The very high recombination rate of H3 isochores may also be largely responsible for the much higher rate of karyotypic rearrangements (and speciation) shown by mammals compared to cold-blooded vertebrates (Bernardi, 1992).

Indications exist that the H3 isochores may be the main integration regions for the majority of (GC-rich) retroviral sequences (see Rynditch et al., 1991; Zoubak et al., 1992; and the section "Two modes of genome evolution: the conservative mode").

The H3 isochore family has an extremely biased codon usage, a number of codons being absent or very scarce because of the very high GC levels in third codon positions, and an extreme amino acid utilization, which favors aminoacids corresponding to codons having only G and/or C in the first two codon positions (D'Onofrio et al., 1991), namely arginine (quartet codons), alanine, glycine and proline, rather than those corresponding to codons with only A and/or T in those positions, like lysine, or those corresponding to codons having both G/C and A/T in first and second codon positions, like serine. Finally, the sequences of the H3 family are located in T-bands of metaphase chromosomes (see-next section).

The main reason for proposing the name of *genome core* for the GC-richest isochore family of the human genome is, however, that the strikingly non-uniform gene distribution described here for the human genome and, in particular, the existence of isochores with very high gene concentrations appear to be shared by all warm-blooded vertebrates and, very probably, by all vertebrates (paper in preparation). If the latter point is confirmed, the compositional pattern of warm-blooded vertebrates characterized by GC-rich and very GC-rich isochores, must have been superimposed on a pre-existing gene concentration pattern, which has already present in cold-blooded vertebrates, but was not characterized by any large differences in GC levels. Another genome feature pre-dating the formation of GC-rich isochores is the early and late replication timing of DNA which was already present in cold-blooded vertebrates (Almeida-Toledo et al., 1988; Giles et al., 1988; Yonenaga-Yassuda et al., 1988).

Isochores and chromosomal bands

A number of findings indicate that GC-poor isochores are located in G(iemsa) bands, whereas GC-rich isochores are located in R(everse) bands of human metaphase chromosomes (see Bernardi, 1989, for a review). In the case of R-bands, at least, the correspondence cannot be a direct one, for the simple reason that GC-rich and GC-poor isochores are in a 1:2 ratio, whereas R- and G-bands are in a 1:1 ratio. An

approach developed in order to solve the problem of the correlations between isochores and chromosomal bands is *compositional mapping* (Bernardi, 1989). This consists in hybridizing probes corresponding to landmarks that are localized on a physical map, or on a chromosome band mpa, to compositional DNA fractions. This approach allows to assess the GC level of about 200 Kb around the landmarks, if the fractionated DNA is about 100 Kb in size.

Compositional mapping, as applied to the long arm of human chromosome 21 (Gardiner et al., 1990), showed that practically all probes for loci present in G-bands hybridized to GC-poor isochores, whereas probes located in R-bands hybridized to either GC-poor or GC-rich isochores. In other words, G-bands are GC-poor and at least very largely homogeneous in base composition, whereas R-bands are compositionally heterogeneous. The GC-richest region of the long arm of the human chromosome 21 was shown to correspond to the telomeric band, which is a T-band (Dutrillaux, 1973), namely one of the 20 or so R-bands most resistant to heat denaturation. This observation prompted an analysis of the localization of single-copy sequences from the isochore family H3. As suggested by previous work on the location of genes in T- bands (Ikemura & Wada, 1991; De Sario et al., 1991), this analysis clearly showed that these sequences are precisely located at T-bands (Saccone et al., 1992). This result, a first step towards a compositional map of the human karyotype, is also of interest in that a DNA fraction, isolated on the basis of its nucleotide composition, exactly corresponds to a well-defined cytogenetic compartment of metaphase chromosomes.

Very recent compositional mapping results on the Xq26-Xqter region (Pilia et al., 1992, submitted) confirmed the GC-poorness and compositional homogeneity of G-bands, as well as the heterogeneity of R-bands, and revealed isochores of the H3 family even in chromosomal regions, like the telomeric Xq28 band, which did not correspond to T-bands. This is apparently due to the fact that the GC-richest region of Xq is narrower than T-bands. It is, therefore, possible that a number of "thin" T-bands are still escaping cytogenetic detection. If such is the situation, and if one considers that evidence is available for "intercalary" (non- telomeric) T-bands to be the result of telomeric fusions in evolution (Dutrillaux, 1979), the high concentration of genes in telomeres and "former" telomeres becomes a conspicuous property of the vertebrate genome, whose implications certainly deserve further

investigations. Indeed, human telomeres are tightly associated with the nuclear matrix (De Lange, 1992) *via* their TTAGGG repeats forming the terminal 2-30 Kb of chromosomes, and with the nuclear envelope (Henderson and Larson, 1991).

The correlations between compositional heterogeneity of isochores and chromosomal bands in warm-blooded vertebrates also explain why metaphase chromosomes from cold-blooded vertebrates, whose genomes are characterized by a low degree of compositional heterogeneity, show very poor or no banding (Cuny et al., 1981; Medrano et al., 1988; Schmid & Guttenbach, 1988).

Two modes of genome evolution: the conservative mode

Compositional patterns allow to define two modes of genome evolution: the conservative mode, and the transitional mode. The *conservative mode* is characterized by the absence of compositional changes, as observed by comparing large (*ca.* 100 Kb) DNA fragments, or homologous exons (and their codon positions) and introns from different genomes. In the case of the mouse/rat comparison, the compositional distribution of DNA fragments is practically identical for the two species, and differences in third codon positions GC from homologous genes are less than 1% (Bernardi et al., 1988). Only slightly larger compositional differences were found when comparing third codon positions from human and bovine genes (Figure 6). In the latter case, compositional conservation still holds, in spite of the fact that third codon positions cover a 30-90% GC range, exhibit a nucleotide divergence averaging 20% (without multiple-hit correction) and are from genes separated for 60-70 Myrs.

The simplest explanation for the conservative mode of evolution would be that, for all isochores, GC to AT changes are compensated by an equal number of AT to GC changes and *vice versa*. This would raise no problem for isochores with compositions close to 50% GC. The compositional conservation of isochores (and of the corresponding coding sequences) having extreme GC levels is, however, incompatible with a "random" process of mutation and fixation, which would drive GC levels towards 50%. Maintaining those extreme levels requires a fixation of mutations which are biased in opposite directions for GC-poor and GC-rich isochores. This has been explained by proposing that

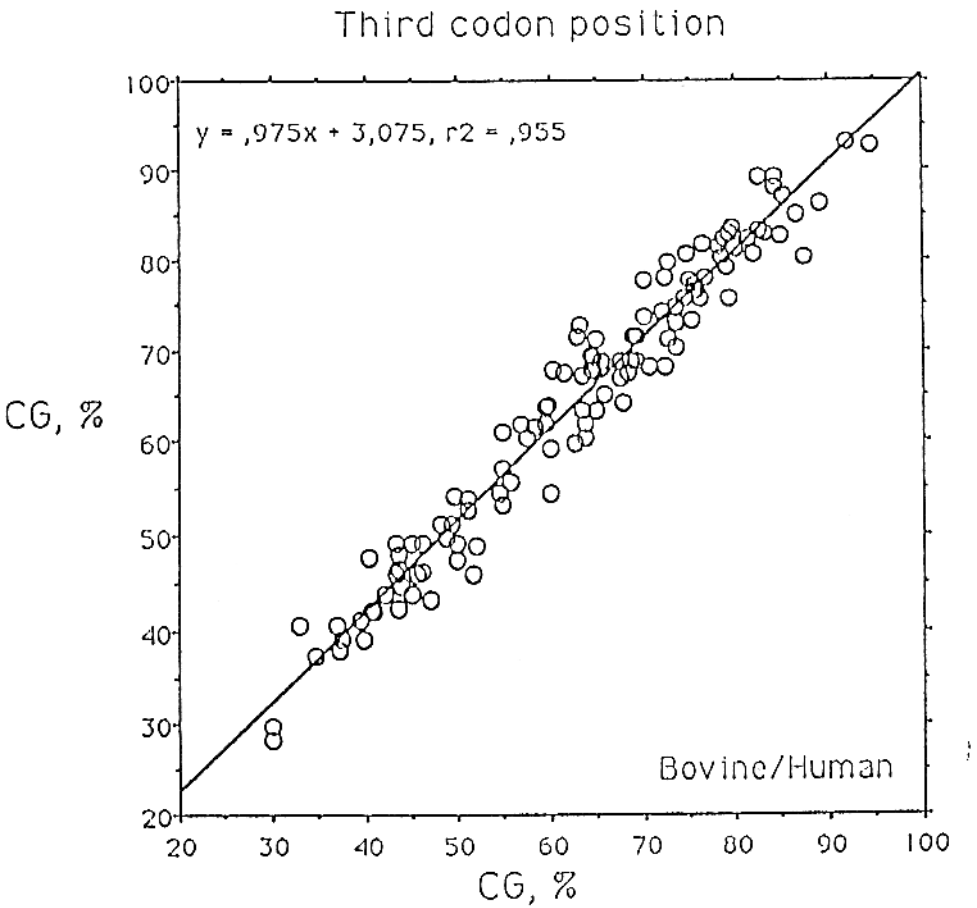


Fig. 6. GC levels of third codon positions of all available pairs of homologous bovine (ordinate) and human (abscissa) genes are plotted against each other (From Mouchiroud et al., 1992).

the bias of the replication/repair machinery is modulated by local chromatin states (Sueoka, 1988), a point discussed later in more detail.

An alternative explanation is that the conservative mode of genome evolution is due to a negative selection against compositional deviations from a narrow GC range (Figure 7A). Negative selection obviously cannot operate at the single nucleotide level, but can do so at a regional (isochore) level (Bernardi et al., 1988). A certain degree of compositional divergence appears to be tolerated, but the upper and lower thresholds seem to be quite close. Indeed, even at size levels as

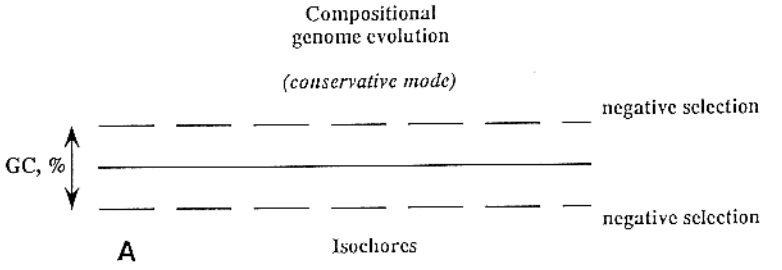
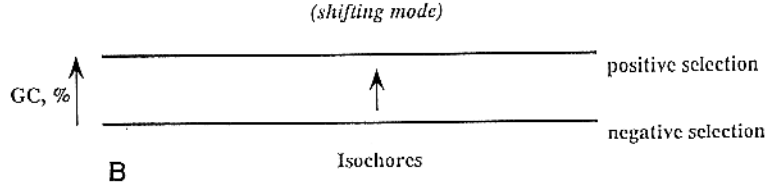


Fig. 7.A. Scheme of negative selection in the conservative mode of evolution. Isochores (solid line) which drift beyond the GC thresholds indicated by the broken lines are counter-selected.



B. Scheme of negative and positive selection in the transitional or shifting mode of genomes evolution. Isochores (solid line) with decreasing GC levels are counter-selected, whereas those with increasing GC levels are selected for.

small as those of genes, compositional divergence remains very low in third codon positions. It may be thought hypothesized, therefore, that cooperative structural changes (which might somehow be compared to phase transitions) take place in isochores beyond the thresholds and that they have deleterious functional consequences (on transcription, for instance), leading to decreased fitness and to negative selection. According to this hypothesis, at least some of the genes present in an isochore which drifted away from its optimal GC level would produce proteins deficient in quantity and/or quality. It should be stressed that this hypothesis does not require negative selection to do more than it admittedly does with regard to classical deleterious mutations in genes.

In this connection, it may be relevant to recall that expressed retroviral sequences integrated in the mammalian genome are located in isochores of matching GC level, whereas non-expressed sequences

from the same viruses are located in isochores characterized by non-matching GC levels (Bernardi, 1989; Rynditch et al., 1991; Zoubak et al., 1992). This may suggest an effect of the chromosomal environment on the expression of integrated viral genomes. A satisfactory demonstration concerning this point can only come, however, from detailed comparisons of integrated viral genomes which are not yet available.

Two modes of genome evolution: the transitional (or shifting) mode

The *transitional (or shifting) mode* of genome evolution is characterized by compositional changes. Major compositional shifts occurred between the genomes of cold- and warm-blooded vertebrates, and minor ones between genomes within each one of these broad classes.

Compositional transitions may be observed at the level of DNA fragments, of exons, of introns, but are best studied by compositional comparisons of different codon positions of homologous genes.

If codon positions of homologous genes from cold- and warm-blooded vertebrates are compared (Bernardi & Bernardi, 1991), GC levels of the coding sequences from warm-blooded vertebrates are either equal or higher (with very few exceptions) than their cold-blooded counterparts (Figure 8), so providing a direct evidence for directional base changes (Perrin & Bernardi, 1987; Bernardi et al., 1988). Such differences are, expectedly, much larger in third than in first + second codon position. Interestingly, the major compositional transitions leading to mammals and birds, although similar, are not identical, the latter attaining slightly higher GC levels in both DNA fragments and third codon positions than the former (Thiery et al., 1976; Bernardi et al., 1988).

In terms of base composition, the genome of warm-blooded vertebrates appears, in fact, to comprise a *paleogenome*, characterized by GC-poor isochores which have not changed in composition relative to the corresponding isochores of cold- blooded vertebrates, and a *neogenome*, characterized by isochores which have become GC-rich (Bernardi, 1989; see Figure 9). Very interestingly, the GC increase in the genome of warm- blooded vertebrates affects only a minority of the genome, about one third of it, which contains, however, the majority of the genes, at least two thirds of them. Indeed, as already mentioned, GC increases parallel gene concentration (see Figure 5).

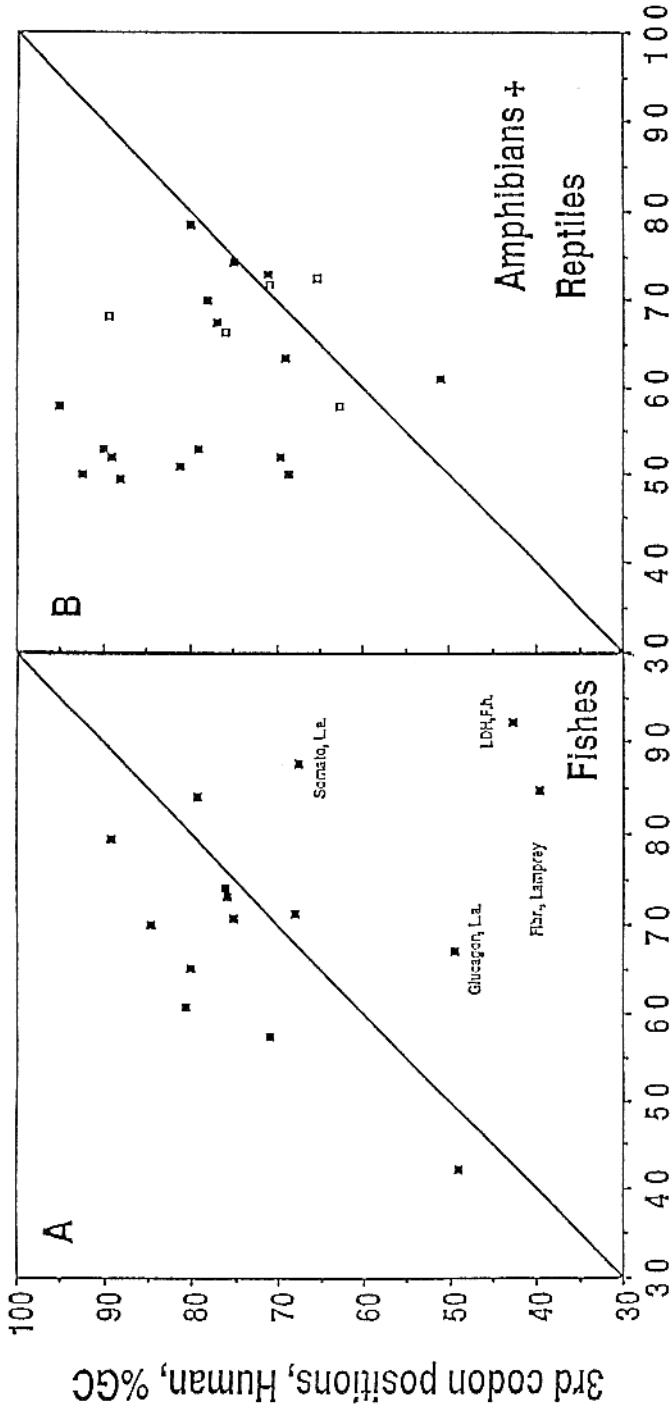


Fig. 8. GC levels of third codon positions of pairs of homologous genes from human (ordinate) and cold-blooded vertebrates (abscissa) are plotted against each other. (From Bernardi & Bernardi, 1991; see this paper for further details).

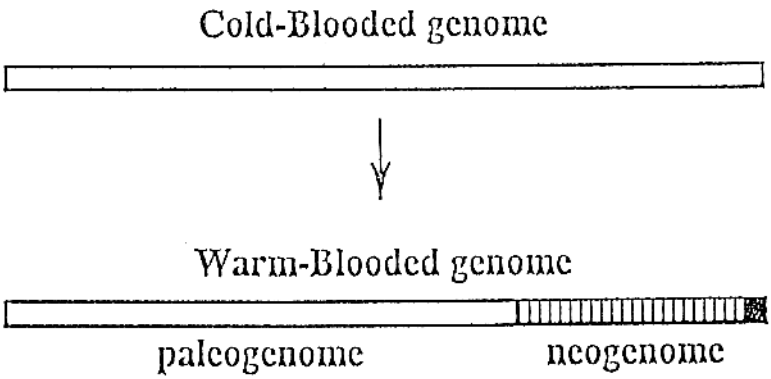


Fig. 9. Scheme of the compositional genome transition accompanying the emergence ofg warm-blooded from cold-blooded vertebrates. The compositionally homogeneous, GC-poor genomes of cold-blooded vertebrates are changed into the compositionally heterogeneous genomes of warm-blooded vertebrates. The latter comprise a paleogenome (corresponding to about two thirds of the genome) which did not undergo any large compositional change and a neogenome (corresponding to the remaining one third of the genome, with the GC-richest part only representing 3% of the genome). In the scheme, the mosaic structure of the warm-blooded vertebrate genome (see Figure 1) is neglected; GC-poor isochores (open bar), GC-rich isochores (hatched bar) and GC-richest isochores (black bar) are represented as three contiguous regions. Gene concentration increases from GC-poor to GC-rich to GC-richest isochores. (From Bernardi, 1992).

Minor compositional transitions occurred among either warm- or cold-blooded vertebrates. In the first case, transitions separate some mammalian orders and families (Thiery et al. 1976; Bernardi et al., 1988; and papers in preparation). A special case is that of murids, cricetids and spalacids which mainly differ from other rodents (as well as from most other mammals investigated) in showing narrower distributions of both DNA fragments and third codon positions (Salinas et al., 1986; Mouchiroud et al., 1987; 1988). Compared to homologous human genes, third codon positions of GC-rich rat and mouse genes are less GC-rich, whereas those of GC-poor genes are less GC-poor, the order of GC levels remaining largely the same in the two species

(Mouchiroud et al., 1987; 1988). Among cold-blooded vertebrates, a number of compositional transitions have been observed (Bernardi & Bernardi, 1990b).

Compositional genome transitions: the mutational bias hypothesis

Two different explanations have been provided for the compositional transitions. The first explanation, originally proposed in order to account for the different composition of prokaryotic genomes, is that compositional patterns (or genome compositions) shift because of directional mutations due to biases in replication/repair enzymes (Freese, 1962; Sueoka, 1962, 1988, 1992).

This explanation is generally considered to have been demonstrated by the existence of an *E coli* mutator strain, *mutT* (Cox and Yanofsky, 1967), which has a mutation rate 1000-fold over the spontaneous mutation rate, which only induces A to C transversions, and which has been reported to cause an increase in GC by 0.3% after 1,200-1,600 generations. The evidence for such a minute change should, however, be viewed with caution, because the buoyant density difference reported is within experimental error.

Moreover, directional mutations found in mutator strains appear to be introduced at a very limited number of hot spots (Yanofsky et al., 1966; Nghiem et al., 1988; Wu et al., 1990) and are unlikely to cause overall compositional genome changes. In any case, for this explanation to be satisfactory, the evidence should be provided that genomes with changed composition are at least not at a disadvantage relative to the unchanged genomes. (Incidentally, this evidence is unlikely ever to be obtained, because of the overwhelming effect of the very high mutation rate).

Obviously, the absence of any direct evidence does not, however, rule out, *per se*, the mutational bias hypothesis. Other problems exist, however. Some of them are of a general nature, other ones are specifically associated with compositional transitions in vertebrates.

Along the first line, the mutational bias hypothesis implies that compositional changes are irrelevant, or neutral, as far as genome organization, function and evolution are concerned, and that they can therefore be left to the vagaries of mutations in the replication/repair systems. This cannot be easily accepted if one considers the present knowledge

concerning the correlations between base composition and DNA structure, the functional importance of the latter, and also the fact that compositional changes in the genome are accompanied by changes in codon usage (reaching even the extreme situation of codon substitutions; Osawa et al., 1987) and by changes in amino acid composition in the encoded proteins. For these reasons, it is difficult to accept the interpretation of differences of base composition in bacterial genomes as simply due to mutational biases as originally proposed (Freese, 1962; Sueoka, 1962).

Another general argument against the mutational bias hypothesis is that the spread of GC levels of genomes from different species decreases from bacteria to protists, to invertebrates, to cold- blooded vertebrates and to warm-blooded vertebrates (Bernardi & Bernardi, 1990b), indicating that the base composition of the genome is certainly not freely drifting in all living organisms. In fact, base composition rather appears to be generally related to the variety of the intra-and extra-cellular environments of the organisms under consideration. In the particular case of vertebrates, the stronger the homeostasis, the narrower the GC spectrum exhibited by genomes from different species.

Additional, very serious problems exist for the mutational bias hypothesis in the case of vertebrates. Indeed, as shown in Figure 10, mutations in the replication/repair machineries leading to mutational biases from AT to GC should have happened only twice, in the two reptilian lineages leading to mammals and birds, respectively (or in the ancestral warm-blooded vertebrates), but never in any other of the extremely numerous families of cold- blooded vertebrates, since the latter never gave rise to genomes compositionally patterned like those of warm-blooded vertebrates. Furthermore, the two series of mutational events led to compositional changes only in the GC-poor isochores that were to form the neogenomes of mammals and birds and not in the more abundant GC-poor isochores which later formed the paleogenomes of warm-blooded vertebrates. Finally, once attained, the GC levels of different GC-rich isochore families, were simply maintained by mutational biases which were different from the original ones. Needless to say, this scenario, which involves processes concerning thousands of physically separated isochores within each genome, is most unlikely to have occurred purely through mutational bias. Indeed, this idea was *de facto* abandoned and the proposal was put forward that

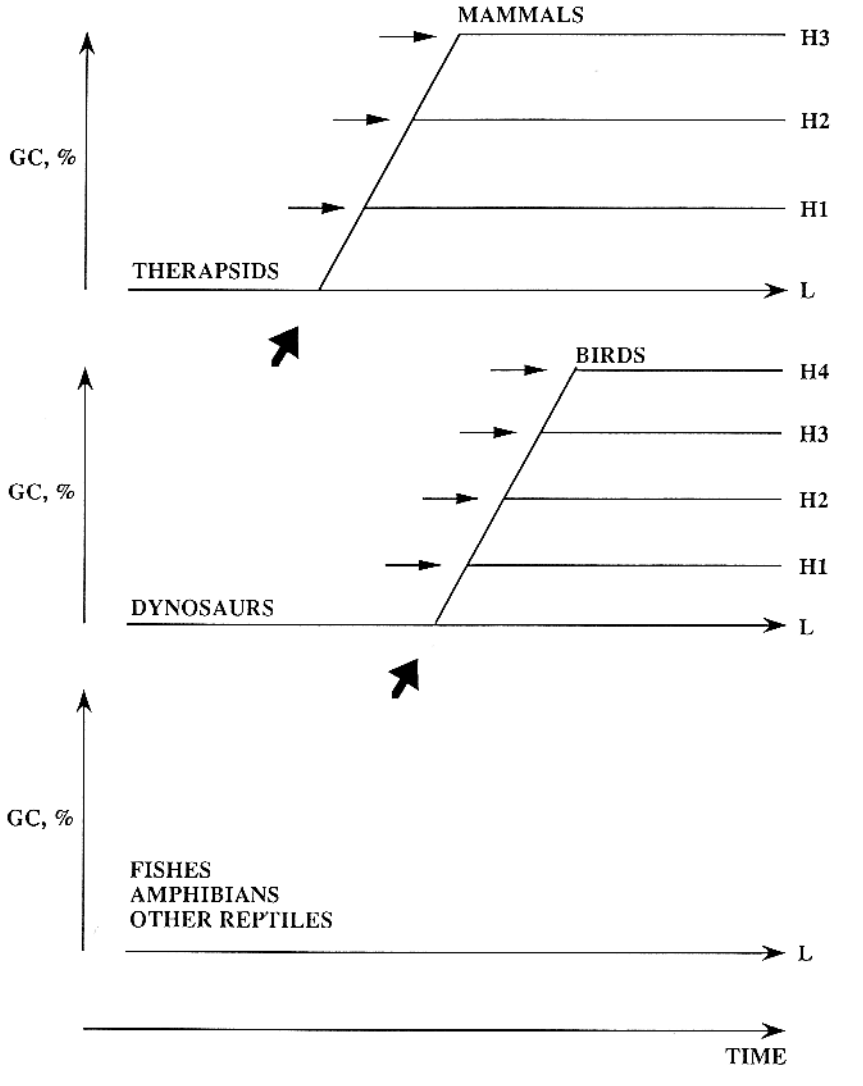


Fig. 10. Scheme of the formation of GC-rich isochores in the genome of warm-blooded vertebrates according to the mutational bias hypothesis. This arrows indicate the times at which mutations in the replication/repair machineries led to GC increases in certain compartments of the genomes of therapsids and dynosaurs (giving rise to GC-rich isochore families H1, H2, H3, H4), but not in other compartments (giving rise to isochore families L). Thin arrows indicate further changes in mutational biases which led to the maintenance of GC-rich isochore families. All changes in mutational biases should be visualized as operating at thousands of physically separated isochores. At the same time, only minor compositional changes took place in the genomes of all other cold-blooded vertebrates (see also Text).

different chromatin structures in different regions of the genome might account for the different directions and effects of mutational biases and that different chromatin transcriptional activities may lead to various extents of repair DNA synthesis (Sueoka, 1988). It is very difficult, however, to see how this *ad hoc* explanation can account for the fact that the original GC increases only occurred in the genomes of the ancestors of mammals and birds. Incidentally, the basic problems just mentioned also made unlikely that changes in nucleotide precursor pools could account for the formation of GC-rich isochores in warm-blooded vertebrates (Wolfe et al., 1989), a hypothesis afflicted by a number of other problems (see Bernardi et al., 1988; Eyre-Walker, 1992).

The minor compositional transitions associated with some cold- or warm-blooded vertebrates also provide arguments against the mutational bias hypothesis. In the case of fish genomes, for example, it was shown that such transitions are unrelated to evolutionary time since the appearance of the order, family or genus, nor are they related to the number of species within a given order (Bernardi & Bernardi, 1990b). Indeed, if compositional changes were only caused by mutations in the replication/repair machineries, compositional divergence should be expected to be larger for older (or larger) groups and smaller for more recent (or smaller) groups, whereas this is not the case. Another observation made on the compositional transitions exhibited by fish genomes is that they show extremely different rates, including very high ones (Bernardi & Bernardi, 1990b). This casts serious doubts about the possibility of constructing correct phylogenetic trees if the genes under consideration (and the isochores harboring them) underwent compositional transitions (Saccone et al., 1990; Bernardi et al., 1992). Finally, if compositional transitions are as frequent as they appear to be from recent work on mammals (paper in preparation), one should pay a special attention to the fact that the molecular clock (Zuckerkandl and Pauling, 1962) may only apply to comparisons between genes that have not been subjected to compositional transitions.

Compositional genome transitions: the selection hypothesis

The second explanation for the compositional genome transitions (like the major transitions that occurred in the genomes of vertebrates) is that directional mutations are fixed, at the isochore level, through both negative and positive selection, the latter only operating, however,

at the initial stage of GC increases. (Incidentally, the need for positive selection would decrease or even disappear if increasing body temperature were to introduce by itself a bias towards high GC). Under this explanation, mutational biases only provide the mechanism for compositional changes, but selection controls the compositional levels of isochores. This explanation deserves several comments.

In general, *selective advantages* associated with compositional patterns of genomes may be elusive, because the patterns are obviously due to a large number of different factors which interplay with each other and are, therefore, impossible to sort out. For instance, it would just not be possible, at present, to provide any explanation for the minor transitions that occurred within either cold- or warm-blooded vertebrates. (Interestingly, however, the narrow compositional patterns exhibited by some rodents, like rat, mouse and hamster, appear to correspond to a partial release, of the compositional constraints operating on the isochores exhibiting extreme GC levels).

This difficult problem may, however, be solved in some particular situations. Indeed, there may be some hope of identifying a selective advantage if the advantage is a predominant one and if it can be evaluated against a relatively similar genomic background, a condition fulfilled by vertebrates. In this case, the major split in genome patterns did not occur at a major step in organismic evolution, like the transitions from anamniotes to amniotes, or from fishes to tetrapods, nor in a gradual way during the evolution from fishes to amphibians to reptiles and to warm-blooded vertebrates, but only and precisely at the transition from cold- to warm-blooded vertebrates. This suggests that the major factor which played a role in the change of compositional patterns of the genome might be related to body temperature. The interest of this suggestion is not only that it fills the empty space left by the old mutational bias hypothesis and by its modern "chromatin" version, but also that it can be tested.

The increase in GC in the genomes of warm-blooded vertebrates makes sense, as far as selective advantages are concerned, because it leads to thermodynamically more stable DNA, RNA and proteins (see Bernardi & Bernardi, 1986, and references quoted therein). Indeed, GC-richness increases the thermal stability of DNA not only in dilute solution but also in chromosomes, as shown by R- and T-bandings, two techniques which show that GC-rich and very GC-rich DNAs are in-

creasingly more stable against thermal denaturation than GC-poor DNAs from G-bands. GC- richness also increases the thermal stability of RNAs, because of the increased secondary structures which make transcripts more stable. Finally, it increases the thermal stability of encoded proteins, because it leads to increased levels of amino acids which confer thermal stability (like arginine, alanine, glycine) and to decreased levels of aminoacids which reduce such stability (like lysine, serine).

The objection that some thermophilic bacteria have AT-rich genomes has no relevance for the body temperature hypothesis in that comparisons of thermophylic and mesophylic bacterial genomes are not warranted when the species under consideration are separated by enormous phylogenetic distances and exhibit extremely large differences in cell physiology. The same warning applies to organelle genomes, in which other selective advantages may be predominant. Another point to be taken into consideration here concerns the fact that thermal stabilization of genomes might be due not to GC but to DNA methylation, or to protein-DNA interactions.

An independent argument favoring the suggestion just presented is the similarity of compositional patterns of mammals and birds, two vertebrate classes characterized by very different genome sizes, which appeared at different geological times (over 200 and about 150 Myrs ago, respectively) and originated from different ancestral reptiles (therapsids and dynosaurs, respectively). Along the same line, and additional argument is provided by the strong compositional heterogeneity of isochores in plants originating from arid climates (like wheat and maize), which stand very high maximal temperatures, and by the weak compositional heterogeneity of isochores in plants from temperate climates (Salinas et al., 1988; Matassi et al., 1989; Montero et al., 1990). The former resemble in their compositional patterns warm-blooded vertebrates, the latter cold-blooded vertebrates. Needless to say that, under the mutational bias hypothesis, all these similarities should be regarded as sheeer coincidences.

ACKNOWLEDGMENT

I wish to thank all my colleagues, and especially Giacomo Bernardi, Giuseppe D'Onofrio and Dominique Mouchiroud, who contributed to the investigations reviewed here. I am particularly indebted to Giacomo

Bernardi, who took an active part in the original formulation of several of the ideas discussed in this review.

The work from the author's laboratory reported here was supported by the Centre National de la Recherche Scientifique (CNRS), the Association Francaise contre les Myopathies (AFM), the Association contre le Cancer (ARC) and the Ligue contre le Cancer.

BIBLIOGRAPHY

AISSANI, B. & G. BERNARDI (1991a). CpG islands: features and distribution in the genome of vertebrates. Gene 106: 173-183.

AISSANI, B. & G. BERNARDI (1991b). CpG islands, genes and isochores in the genome of vertebrates. Gene 106: 185-195.

AISSANI, B., G. D'ONOFRIO, D. MOUCHIROUD., K. GARDINER, C. GAUTIER & G. BERNARDI. (1991). The compositional properties of human genes. J. Mol. Evol. 32: 497-503.

ALMEIDA-TOLEDO, L. F., F. VIEGAS-PEQUIGNOT, F. FORESTI, S.A. TOLEDO-FILHO & B. DUTRILLAUX. (1988). BrdU replication patterns demonstrating chromosome homologies in two species, Genus *Eigenmannia*. Cytogenet. Cell 48: 117-120.

AOTA, S-I. & T. IKEMURA. (1986). Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucl. Acids Res. 14: 6345-6355.

BERNARDI, G. (1985). The organization of the vertebrate genome and the problem of the CpG shortage. in Chemistry, biochemistry and biology of DNA methylation, (Cantoni, G. L. and A. Razin, eds.), pp. 3-10, Alan Liss, New York, N.Y.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL & F. RODIER. (1985). The mosaic genome of warm-blooded vertebrates. Science 228: 953-958.

BERNARDI, G. & G. BERNARDI. (1985). Codon usage and genome composition. J. Mol. Evol. 22: 363-365.

BERNARDI, G. & G. BERNARDI. (1986). Compositional constraints and genome evolution. J. Mol. Evol. 24: 1-11.

BERNARDI, G., D. MOUCHIROUD, C. GAUTIER & G. BERNARDI (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. Ann. Rev. Genet. 23: 637-661.

BERNARDI, G. (1990). Le Génome des Vertébrés: Organisation, Fonction et Evolution. Biofutur 94: 43-46.

BERNARDI, G. & G. BERNARDI. (1990a). Compositional patterns in the nuclear genomes of cold-blooded vertebrates. J. Mol. Evol. 31: 265-281.

BERNARDI, G. & G. BERNARDI. (1990b). Compositional transitions in the nuclear genomes of cold-blooded vertebrates. J. Mol. Evol. 31: 282-293.

BERNARDI, G. & G. BERNARDI. (1991). Compositional properties of nuclear genes from cold-blooded vertebrates. J. Mol. 33: 57-67.

BERNARDI, G., P. SORDINO & D. A. POWERS. (1992). Nucleotide sequence of the 18S ribosomal ribonucleic acid gene from two teleosts and two sharks and their molecular phylogeny. Mol. Mar. Biol. Biotech. 1: 187-194.

BERNARDI, G. (1992). Genome organization and species formation in vertebrates. J. Mol. Evol. in press.

BETTECKEN, T., B. AISSANI, C. R. MULLER & G. BERNARDI. (1992). Compositional mapping of the human dystrophin gene. Gene, in press.

BIRD, A. (1986). CpG-rich islands and the function of DNA methylation. Nature 321: 209-203.

CORTADAS, J., G. MACAYA & G. BERNARDI. (1977). An analysis of the bovine genome by density gradient centrifugation: fractionation in Cs_2SO_4/3,6 bis (acetato-mercurimethyl) dioxane density gradient. Eur. J. Biochem. 76: 13-19.

COX, E. C. & C. YANOFSKY. (1967). Altered base ratios in the DNA of an *Escherichia coli* mutator strain. Proc. Natl. Acad. Sci. USA 88: 1895-1902.

CUNY, G., P. SORIANO, G. MACAYA & G. BERNARDI. (1981). The major components of the mouse and human genomes. Preparation, basic properties and compositional heterogeneity. Eur. J. Biochem. 115: 227-233.

DE LANGE, T. (1992). Human telomeres are attached to the nuclear matrix. EMBO J. 11: 717-724.

DE SARIO, A., B. AISSANI & G. BERNARDI (1991). Compositional properties of telomeric regions from human chromosomes. FEBS Letters 295: 22-26.

D'ONOFRIO, G., D. MOUCHIROUD, B. AISSANI., C. GAUTIER & G. BERNARDI. (1991). Correlations between the compositional properties of human genes, codon usage and aminoacid composition of proteins. J. Mol. Evol. 32: 504-510.

D'ONOFRIO, G. & G. BERNARDI. (1992). A universal compositional correlation among codon positions. Gene 110: 81-88.

DUTRILLAUX, B. (1973). Nouveau système de marquage chromosomique: les bandes T. Chromosoma 41: 395-402.

DUTRILLAUX, B. (1979). Chromosomal evolution in primates: tentative phylogeny from *Microcebus murinus* (Prosimian) to man. Hum. Genet. 48: 251-314.

EYRE-WALKER, A. (1992). Evidence that both G+C rich and G+C poor isochores are replicated early and late in the cell cycle. Nucl. Acids Res. 20: 1497-1501.

FILIPSKI J., J. P. THIERY & G. BERNARDI. (1973). An analysis of the bovine genome by Cs_2SO_4-Ag+ density gradient centrifugation. J. Mol. Biol. 80: 177-197.

FREESE, J. (1962). On the evolution of base composition of DNA. J. Theor. Biol. 3: 82-101.

GARDINER, K., B. AISSANI & G. BERNARDI. (1990). A compositional map of human chromosome 21. EMBO J. 9: 1853-1858.

GARDINER-GARDEN, M. & M. FROMMER. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. 196: 261-282.

GILES, V., G. THODE & M. C. ALVAREZ. (1988). Early replication bands in two scorpion fishes, *Scorpaena porcus* and *S. notata* (order *Scorpaneiformes*). Cytogenet. Cell Genet. 47: 80-83.

HENDERSON, E. R. & D. D. LARSON. (1991). Telomeres - what's new at the end? Curr. Opin. Genet. & Develop. 1: 538-543.

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

IKEMURA, T. & S.-I. AOTA. (1988). Global variation in G+C content along vertebrate genome DNA. J. Mol. Biol. 203: 1-13.

IKEMURA, T., K. WADA & S.-I. AOTA. (1990). Giant G+C mosaic structures of the human genome found by arrangement of Genbank human DNA sequences according to genetic positions. Genomics 8: 207-216.

IKEMURA, T., K. WADA. (1991). Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers, relation between nucleotide sequence data and cytogenetic data. Nucl. Acids Res. 13:1915- 1922.

JUKES, T.H. & V. BUSHAN. (1986). Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J. Mol. Evol. 24: 39-44.

KEREM, B.-S., R. GOITEIN, G. DIAMOND, H. CEDAR & B. MARCUS. (1984). Mapping of DNAase I sensitive regions of mitotic chromosomes. Cell 38: 493-499.

KOPECKA, H., G. MACAYA, J. CORTADAS, J. P. THIERY & G. BERNARDI G. (1978). Restriction enzyme analysis of satellite DNA components from the bovine genome. Eur. J. Biochem. 84: 189-195.

KRANE, D. E., D. L. HARTL & H. OCHMAN. (1991). Rapid determination of nucleotide content and its application to the study of genome structure. Nucl. Acids Res. 19: 5181-5185.

MACAYA, G., J. P. THIERY. & G. BERNARDI G. (1976). An approach to the organization fo eukaryotic genomes at a macromolecular level. J. Mol. Biol. 108: 237-254.

MACAYA, G., J. CORTADAS & G. BERNARDI. (1978). An analysis of the bovine genome by density gradient centrifugation. Eur. J. Biochem. 84: 179-188.

MATASSI, G., L. M. MONTERO, J. SALINAS & G. BERNARDI. (1989). The isochore organization and the compositional distribution of homolo-

gous coding sequences in the nuclear genome of plants. Nucleic Acids Res. 17: 5273-5290.

MEDRANO, L., G. BERNARDI, J. COUTURIER, B. DUTRILLAUX & G. BERNARDI. (1988). Chromosome banding and genome compartmentalization in fishes. Chromosoma 96: 178-183.

MONTERO, L. M., J. SALINAS, G. MATASSI & G. BERNARDI. (1990). Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res. 18: 1859-1867.

MOUCHIROUD, D., G. FICHANT & G. BERNARDI. (1987). Compositional compartmentalization and gene composition in the genome of vertebrates. J. Mol. Evol. 26: 198-204.

MOUCHIROUD, D., C. GAUTIER & G. BERNARDI. (1988). The compositional distribution of coding sequences and DNA molecules in humans and murids. J. Mol. Evol. 27: 311-320.

MOUCHIROUD, D., G. D'ONOFRIO, B. AISSANI, G. MACAYA, C. GAUTIER & G. BERNARDI. (1991). The distribution of genes in the human genome. Gene 100: 181-187.

NGHIEM, Y., M. CABRERA, C. G. CUPPLES & J. H. MILLER. (1988). The *mutY* gene: A mutator locus in *Escherichia coli* that generates G. C ③ T.A transversions. Proc. Natl. Acad. Sci. USA 85: 2709-2713.

OSAWA, S., T. H. JUKES, A. MUTO, F. YAMAO, T. OHAMA & Y. ANDACHI. (1987). Role of directional mutation pressure in the evolution of the eubacterial genetic code. Cold Spring Harbor Symp. Quant. Biol. 52: 777-789.

PERRIN, P. & G. BERNARDI. (1987). Directional fixation of mutations in vertebrate evolution. J. Mol. Evol. 26: 301-310.

PILIA, G., R. D. LITTLE, B. AISSANI, G. BERNARDI & D. SCHLESSINGER. (1992). Isochores and CpG islands in YAC contigs covering the q26.1-qter region of the human X chromosome. Nature Genetics (submitted for publication).

RYNDITCH, A., F. KADI, J. GERYK, S. ZOUBAK, J. SVOBODA & G. BERNARDI. (1991). The isopycnic, compartmentalized integration of Rous sarcoma virus sequences. Gene 106: 165-172.

SACCONE, C., C. LANAVE, G. PESOLE & G. PREPARATA. (1990). Influence of base composition on quantitative estimates of gene evolution. Methods Enzymol. 183: 570-583.

SACCONE, S., A. DE SARIO., G. DELLA VALLE & G. BERNARDI. (1992). The highest gene concentrations in the human genome are inT-bands of metaphase chromosomes. Proc. Natl. Acad. Sci. USA 89: 4913-4917.

SALINAS, J., M. ZERIAL, J. FILIPSKI & G. BERNARDI. (1986). Gene distribution and nucleotide sequence organization in the mouse genome. Eur. J. Biochem. 160: 469-478.

SALINAS, J., G. MATASSI, L. M. MONTERO & G. BERNARDI. (1988). Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucleic Acids Res. 16: 269-4285.

SCHMID, M. & M. GUTTENBACH. (1988). Evolutionary diversity of reverse (R) fluorescent chromosome bands in vertebrates. Chromosoma 97: 101-114.

SUEOKA, N. (1959). A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. Proc. Natl. Acad. Sci. USA 45: 1480-1490.

SUEOKA, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA 48: 582-592.

SUEOKA, N. (1988). Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85: 2653-2567.

SUEOKA, N. (1992). Directional mutation pressure, selective constraints, and genetic equilibria. J. Mol. Evol. 34: 95-114.

TAZI, J. & A. P. BIRD. (1990). Alternative chromatin structure at CpG islands. Cell 60: 909-920.

THIERY, J. P., G. MACAYA & G. BERNARDI. (1976). An analysis of eukaryotic genomes by density gradient centrifugation. J. Mol. Biol. 108: 219-235.

WADA, A., A. SUYAMA & R. HANORI. (1991). Phenomenological theory of GC/AT pressure on DNA base composition. J. Mol. Evol. 32: 374-378.

WADA, A. (1992). Compliance of genetic code with base composition deflecting pressure. Adv. Biophysics (in press).

WOLFE, K. H., P. M. SHARP & W.-H. LI. (1989). Mutation rates differ among regions of the mammalian genome.

WU, T.-H., C. H. CLARKE & M. G. MARINUS. (1990). Specificity of *Escherichia coli mutD mutL* mutator strains. Gene 87: 1-5.

YANOFSKY, C., E. C. COX & V. HORN. (1966). The unusual mutagenic specificity of an *E. coli* mutator gene. Proc. Natl. Acad. Sci. USA 55: 274-281.

YONENAGA-YASSUDA, Y., S. KASAHARA, T. H. CHU & M. T. RODRIGUEZ. (1988). High-resolution RBG-banding pattern in the genus *Tropidurus (Sauria, Iguanidae)*. Cytogenet. Cell Genet. 48: 68-71.

ZOUBAK, S., A. RYNDITCH & G. BERNARDI. (1992). Compositional bimodality and evolution of retroviral genomes. Gene, in press.

ZUCKERKANDL, E. & L. PAULING. (1962). Molecular disease, evolution and genetic heterogeneity. In: M. Kasha, and B. Pullman (eds.), Horizons in biochemistry, Academic Press, New York, pp. 189-225.