J Mol Evol (1988) 28:7–18

Journal of Molecular Evolution

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

Summary. The evolution of vertebrate genomes can be investigated by analyzing their regional compositional patterns, namely the compositional distributions of large DNA fragments (in the 30–100-kb size range), of coding sequences, and of their different codon positions. This approach has shown the existence of two evolutionary modes. In the conservative mode, compositional patterns are maintained over long times (many million years), in spite of the accumulation of enormous numbers of base substitutions. In the transitional, or shifting, mode, compositional patterns change into new ones over much shorter times.

The conservation of compositional patterns, which has been investigated in mammalian genomes, appears to be due in part to some measure of compositional conservation in the base substitution process, and in part to negative selection acting at regional (isochore) levels in the genome and eliminating deviations from a narrow range of values, presumably corresponding to optimal functional properties. On the other hand, shifts of compositional patterns, such as those that occurred between cold-blooded and warm-blooded vertebrates, appear to be due essentially to both negative and positive selection again operating at the isochore level, largely under the influence of changes in environmental conditions, and possibly taking advantage of mutational biases in the replication/repair enzymes and/or in the enzyme make-up of nucleotide precursor pools. Other events (like translocations and changes in chromosomal structure) also play a role in the transitional mode of genome evolution.

The present findings (1) indicate that isochores, which correspond to the DNA segments of individual or contiguous chromatin domains, represent selection units in the vertebrate genome; and (2) shed new light on the selectionist–neutralist controversy.

Key words: Compositional patterns — Compositional shifts — Genome evolution — Isochores — Vertebrates — Selection — Neutral theory

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, that 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

Offprint requests to: G. Bernardi

* This work was presented at the EMBO Workshop on Evolution (Cambridge, UK, 4–6 July 1988) and at the 16th International Congress of Genetics (Toronto, Canada, 20–27 August 1988)

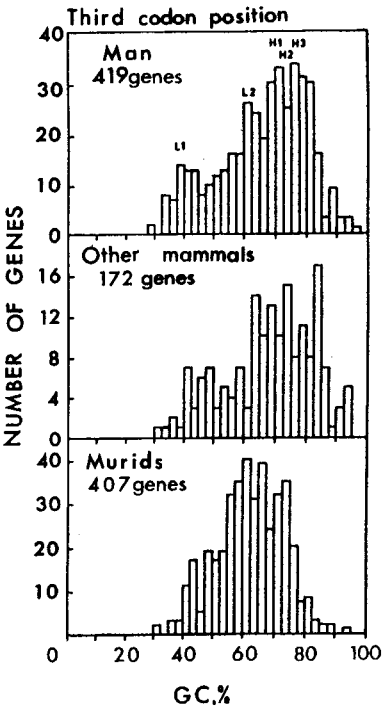

Fig. 1. Compositional distribution of third codon positions for genes (1) from man; (2) from mammals other than man, murids, and hamster; and (3) from murids (rat and mouse). As in previous investigations, all sequences available in GenBank (Bilofsky et al. 1986) were analyzed. Sequences used in the present paper derive from Release 54 (December 1987). The number of genes under consideration is indicated. Genes from "other mammals" comprised 125 genes from artiodactyls (86 genes from calf, 21 genes from pig, 11 genes from sheep, 7 genes from goat), 30 genes from rabbit, 9 genes from dog, 8 genes from horse. Genes from "murids" comprise 181 genes from mouse, 185 genes from rat, and 41 genes belonging to homologous pairs in rat and mouse; these genes were plotted only once, because third codon positions were very similar in the two species (see Figs. 3 and 5). 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, M1, H2, and H3) are indicated following Mouchiroud et al. (1987, 1988). The average GC levels for human and murid third codon positions (two samples of comparable sizes) are 65.4% and 62.2%, respectively.

(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, 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 noncoding 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 underrepresented 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 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 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 then discuss the mechanisms and the causes of these phenomena.

Two Types of Compositional Patterns in Mammals

We have recently identified two types of compositional distributions in the available mammalian

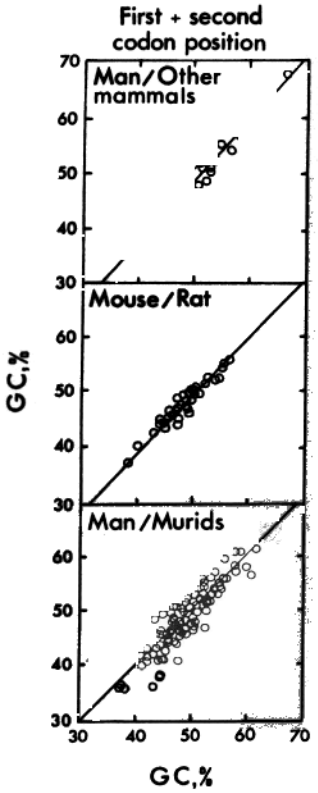

Fig. 2. Relationships between GC levels of first + second codon positions for pairs of homologous genes. For each point, the average GC level for first + second codon positions in one gene is plotted against the average GC level for first + second positions in its homolog. For the upper panel, the ordinate corresponds to the human genes and the abscissa to their homologs in another mammal (rat, mouse, and hamster being excluded). For the middle panel, the ordinate corresponds to mouse genes and the abscissa to their homologs in rat. For the lower panel, the ordinate corresponds to the human genes and the abscissa to their homologs in murids. Lines were drawn using the least-squares method. From the top to the bottom diagram, the numbers of gene pairs investigated were 53, 41, and 141; slopes were 0.99, 0.97, and 1.03; and correlation coefficients were 0.96, 0.97, and 0.90, respectively.

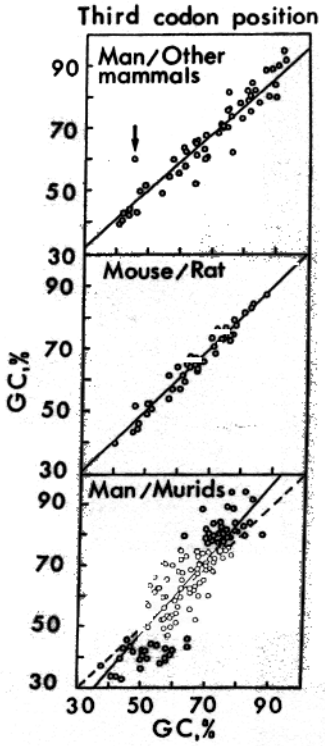

Fig. 3. Relationships between GC levels of third codon position for pairs of homologous genes (1) from man (ordinate) and other mammals (except murids and hamster; abscissa); (2) from mouse (ordinate) and rat (abscissa); and (3) from man (ordinate) and murids (abscissa). The arrow corresponds to the endozepin genes from man and calf. From top to bottom, slopes were 0.95, 0.98, and 1.26, and correlation coefficients were 0.95, 0.98, and 0.86, respectively. In the bottom plot, the slope was significantly different from unity as judged by a statistical test, and the broken line corresponds to a unity slope. For other indications, see legend of Fig. 2.

coding sequences. These types are 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).

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 with unity slopes, in the case of both the "general" and the "murid" distribution (see Figs. 2 and 3 for plots of first + second and of third codon positions, respectively). 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 Figs. 2 and 3). 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, because the order of coding sequences by increasing GC 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

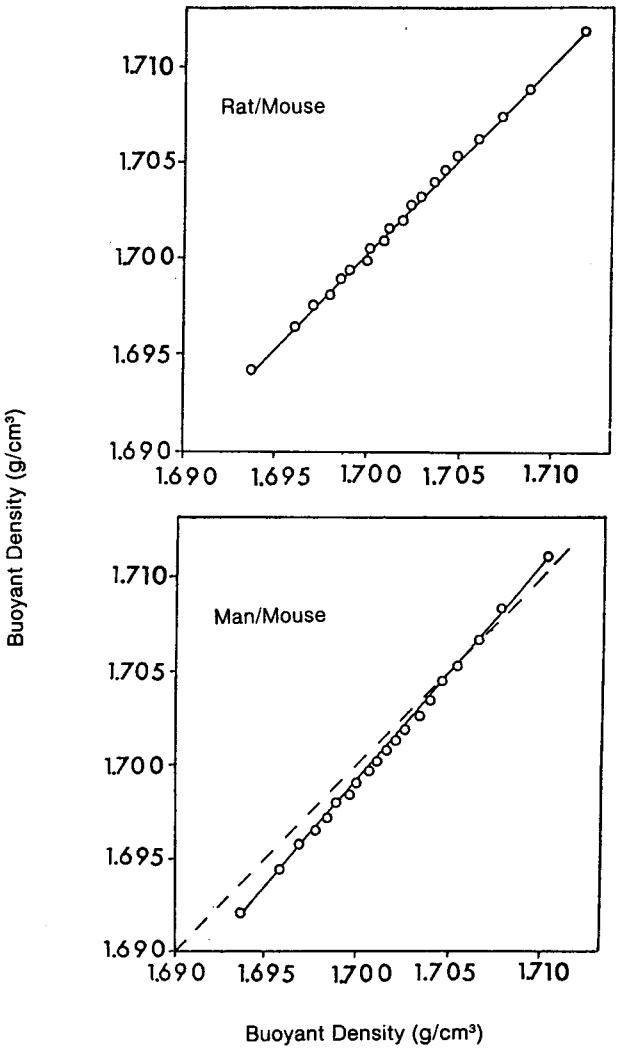

Fig. 4. Comparison of compositional distribution of DNA fragments of rat and mouse and man and mouse. Data for rat or man correspond to the ordinate; data for mouse to the abscissa. Plots were constructed by comparing buoyant densities corresponding to 5% increments in relative amounts of DNAs. Analytical CsCl profiles from Thiery et al. (1976) were used because they concerned DNA preparations of comparable molecular weight. The rat–mouse plot showed a slope of 1.0; the man–mouse plot a slope of 1.13 (the broken line corresponding to the unity slope, passing through the origin). The higher slope of the man–mouse plot found by Mouchiroud et al. (1988) was due to the higher molecular weight of the DNA preparation from mouse compared to that from man. Mouse satellite DNA was eliminated from these comparisons.

(Fig. 4), the results are very similar to those obtained in Fig. 3. 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.

More recent analyses on the genomes of mammals from nine different orders (G. Sabeur, J. Filipski, F. Kadi, G. Bernardi, unpublished observations) have shown that, in fact, the "murid" pattern is present in three families of the suborder Myomorpha, namely murids (rat and mouse), cricetids (hamster), and spalacids (mole rat), but not in the other two rodent suborders, Hystricomorpha (guinea pig) and Sciuromorpha (squirrel), which show the "general" pattern of the other mammals explored. Other slight but significant (mostly "murid-like") deviations from the "general" pattern have been found in some species.

The Conservation of Compositional Patterns in Mammalian Genomes: Compositional Conservation of the Base Substitution Process and Negative Selection at the Isochore Level

The compositional patterns just described have been conserved in mammalian evolution over very extended time intervals, because they have been found in man, artiodactyls, and rabbit at the coding sequence level, and in orders ranging from insectivores to primates at the DNA fragment level. The question then is, how was such a conservation maintained against the enormous numbers of base substitutions that occurred over the many million years separating the mammals under consideration. Before discussing this issue in the following section, several points concerning the conservation of compositional distributions in both coding sequences and isochores should be stressed.

As far as coding sequences are concerned, homologous pairs from either man–other mammals or mouse–rat comparisons were found to be extremely close, not only in overall GC levels and in the GC levels of different codon positions (Mouchiroud et al. 1988; and present work, Figs. 2 and 3), but also at the level of individual nucleotides in third codon positions (Figs. 5 and 6). Now, third codon positions cover a broad GC range in different gene pairs (from less than 40% to over 90%) show large divergences (Britten 1986) (from 20 to 50% for mouse/rat and man/other mammals, respectively; without correction for multiple hits), and belong to sequences that are only about 1 kb in average size. In other words, although third codon positions encompassing a wide compositional range and belonging to short sequences underwent, at the very least, 20–50% base substitutions, their base composition, averaged over the whole coding sequences, changed very little. Moreover, this compositional balance holds separately for transitions (as directly demonstrated by Mouchiroud and Gautier 1988) and for transversions.

In the case of isochores, compositional conservation is indicated (1) by the similarity of the results of Figs. 3 and 4, which show that the compositional conservation of isochores (as detected at the level of DNA fragments) parallels that of third codon positions; (2) by the linear relationship between the GC levels of coding sequences and those of the large

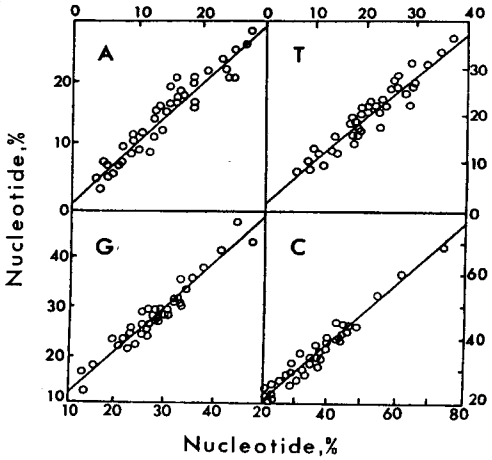

Fig. 5. Relationships between the levels of the different nucleotides in third codon positions for pairs of homologous genes from mouse (ordinate) and rat (abscissa). Slopes were 0.96 (A), 0.92 (T), 1.01 (G), and 0.96 (C). Correlation coefficients were 0.97 (A), 0.96 (T), 0.97 (G), and 0.99 (C). For other indications, see legend of Fig. 2.

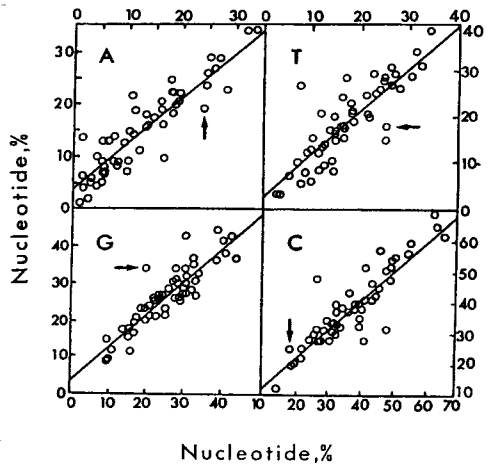

Fig. 6. Relationships between the levels of the different nucleotides in third codon positions for pairs of homologous genes from man (ordinate) and other mammals (except murids and hamster; abscissa). Slopes were 0.90 (A), 0.97 (T), 0.90 (G), and 0.96 (C). Correlation coefficients were 0.94 (A), 0.91 (T), 0.91 (G), and 0.94 (C). Arrows correspond to the endozepin genes from man and calf. For other indications, see legend of Fig. 2.

DNA fragments embedding them (Bernardi et al. 1985), this implies that regions (up to at least 30–100 kb) flanking homologous coding sequences and scattered all over the genome are also close in GC levels; and (3) by the conservation of the levels of individual nucleotides in introns of homologous genes from mouse and rat (Fig. 7).

The Causes of the Conservation of Mammalian Compositional Patterns

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 most unlikely, however, in view of the demonstrated influence of nearest-base composition (see Bernardi and Ninio 1978; Fresco et al. 1980; Bulmer 1986, 1988; Phear et al. 1987), 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 can hardly be, however, the explanation for the findings under consideration, because this would require assigning a significant advantage or disadvantage to any event affecting 1 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, because 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," because these are estimated to comprise 30–300 kb of supercoiled DNA in mammals (Gasser and Laemmli 1987; Luchnik et al. 1988; Goldman 1988). In turn, chromatin loops are supposed to correspond to replicons and to contiguous transcription units (Zehnbauer and Vogelstein 1985).

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 (1) 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 (Thiery et al. 1976; Bernardi et al. 1985); and (2) to changes in the compositional distributions of coding sequences (Bernardi et al. 1985; Mouchiroud et al. 1987; see Fig. 8). These shifts were com-

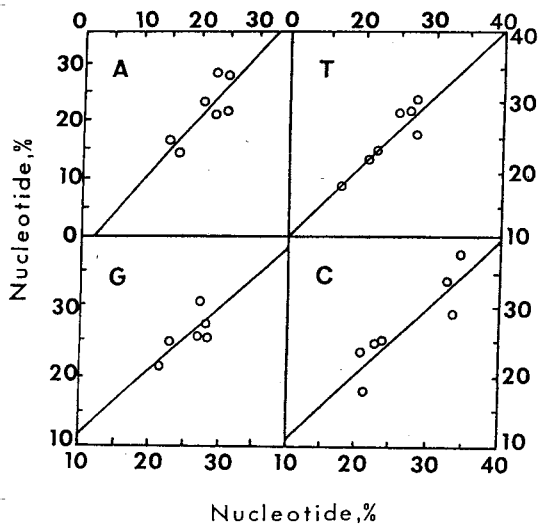

Fig. 7. Relationships between the levels of the different nucleotides in introns of homologous genes of rat (ordinate) and mouse (abscissa). Slopes were 1.25 (A), 1.01 (T), 0.84 (G), and 0.90 (C). Correlation coefficients were 0.87 (A), 0.94 (T), 0.95 (G), and 0.90 (C).

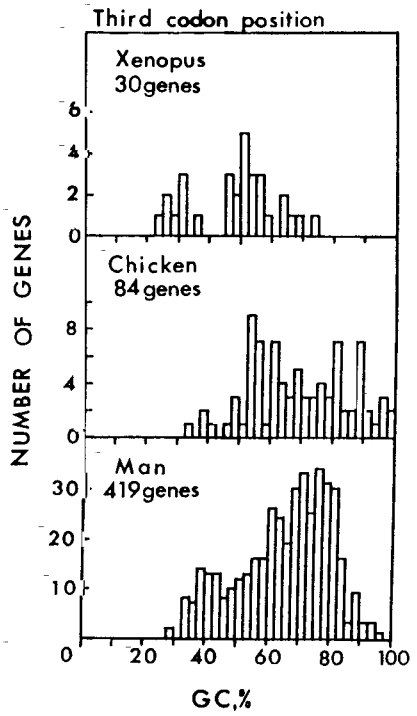

Fig. 8. Compositional distribution of third codon positions for genes from *Xenopus*, chicken, and man. The average GC levels for *Xenopus*, chicken, and human were 48.4%, 69.4%, and 65.4%, respectively. Other indications as in Fig. 1.

pletely independent of each other, because the paleontological record indicates that mammals derived from therapsids over 200 million years (Myr) ago and birds from dinosaurs about 150 Myr ago (Carroll 1987). In agreement with this conclusion, compositional patterns of birds and mammals are different from each other (see Fig. 8 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 cold-blooded and 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 human 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; a similar but expectedly less pronounced situation was found in comparisons for first + second codon positions (Figs. 9 and 10). Expectedly, no significant correlations were found between GC levels of codon positions from homologous genes, contrary to what was found for mammalian genes (Figs. 2 and 3). When mammals and birds were compared, some weak correlation was found in first and second codon positions (Fig. 9) but not in third codon positions (Fig. 10); in the latter case, no trend toward higher or lower GC levels was found. Similar, remarkable differences were also found between the compositional distributions of DNA fragments from chicken (Cortadas et al. 1979) and mammals.

Several additional phenomena accompanied major shifts: (1) 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 α and β globin genes that 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 α gene (in mammals) or in both α and β genes (in birds); (2) large changes in chromosome structure took place, as indicated by the appearance or the sharpening of Giemsa and Reverse bands (see Van Duijn et al. 1985, for the possible involvement of nucleosomes in banding); moreover, the Giemsa light (or Reverse) bands apparently acquired a more complex structure relative to Giemsa dark bands, because their DNA is a mosaic of the different GC-

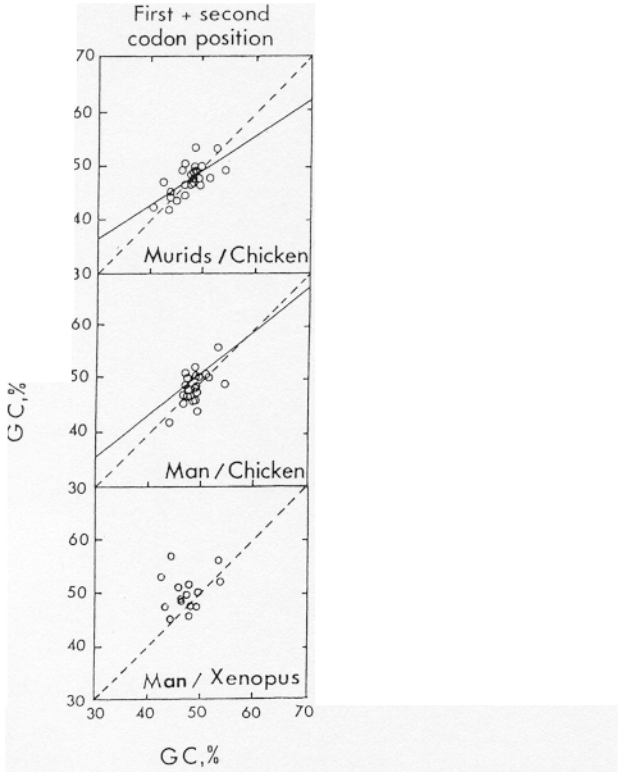

Fig. 9. GC levels of first + second codon positions of pairs of homologous genes are plotted against each other (1) for murids and chicken (27 gene pairs); (2) for man and chicken (25 gene pairs); and (3) for man and *Xenopus* (16 genes). Data for murids or man correspond to the ordinate; data for chicken to the abscissa. From top to bottom, correlation coefficients were 0.71, 0.59, and 0.39, respectively. In the top and middle graphs, slopes were 0.64 and 0.79, respectively; the broken lines correspond to unity slope passing through the origin.

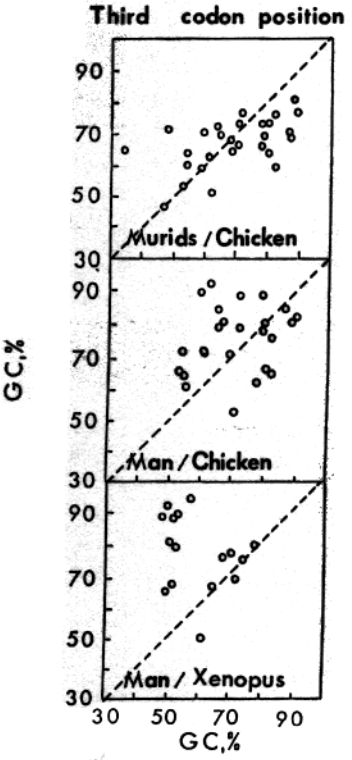

Fig. 10. GC levels of third codon positions of pairs of homologous genes are plotted against each other (1) for murids and chicken; (2) for man and chicken; and (3) for man and *Xenopus*. Data for murids or man correspond to the ordinate; data for chicken or *Xenopus* to the abscissa. Correlation coefficients were 0.56, 0.19, and −0.30, respectively. The broken lines correspond to unity slope passing through the origin.

rich isochores and also seems to have a less-dense packing (G. Bernardi, unpublished observations); (3) nonmethylated 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 (toward 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 contrast with "major" shifts, correlations are still found among GC levels of third codon positions from homologous genes (compare Figs. 2 and 3 with Figs. 9 and 10), and the order of coding sequences by increasing GC is largely preserved.

Finally, even more limited compositional shifts appear to have taken place in individual genes (and their flanking regions) after their translocation from one isochore class to another. One example is that of the endozepin gene that exhibits very different GC levels (see Figs. 3 and 6) in third codon position in calf and man (as compared with other homologous genes also having the same GC levels in first + second positions; see Figs. 2 and 3); this gene is probably located in isochores of very different GC levels in the two species. This case is similar to that of the prolactin genes of man and rat, discussed elsewhere (Mouchiroud et al. 1988).

The Causes of Compositional Shifts in Genome Evolution: Mutational Biases in the Replication and Repair Machineries?

Shortly after large compositional differences were found in genomes of different bacteria (Lee et al. 1956; Belozerski and Spirin 1958), they were explained as being caused by mutational biases in the replication machinery (Freese 1962; Sueoka 1962), namely by differences in the forward and backward mutation rates associated with GC ⇄ AT changes. Mutational biases were presumed to be associated

with mutations in the genes of the replication machinery, because certain mutator strains of *Escherichia coli* do have altered base ratios (Cox and Yanofsky 1967; see also Nghiem et al. 1988). Sueoka's "quantitative theory of directional mutation pressure" still is thought to "explain the wide variation of DNA base composition observed among different bacteria and its small heterogeneity within individual bacterial species," and also to "offer new plausible explanations for the large heterogeneity in G+C content among different parts of the vertebrate genome" (Sueoka 1988).

There are, however, serious problems with the latter point: (1) in the two independent "major" shifts leading from the genomes of cold-blooded to those of warm-blooded vertebrates, such mutational biases occurred only in some isochores that are scattered all over the genome; (2) in the "minor" compositional shift that led to the formation of the "murid" pattern, mutational biases essentially were limited not only to isochores and coding sequences located at both ends of the compositional distribution (yet again physically scattered all over the genome), but also had opposite directions at such ends. These points cannot be explained easily by mutations in the genes of replication/repair enzymes because these enzymes would be required to show a bias, no bias, or opposite biases, depending upon the chromosomal regions in which they work.

These difficulties are so obvious that it has been conceded (Sueoka 1988) that "there might be several different directional mutation pressures (not necessarily mutation rates) in different locations on the genome, and the cause for this difference might reside in the local structural elements in the chromatin. Thus, major mutagenic events (DNA replication and repair) may act differently in different domains." Alternatively (still according to Sueoka 1988), "replication errors and the extent of repair DNA synthesis may vary among various domains of the chromatin because of the different susceptibility of DNA to damage and repair due to differences in chromatin structure." These "plausible explanations" amount, however, to abandoning the biases of the replication and repair machineries as the cause of the compositional shifts undergone by vertebrate genomes in favor of a completely different cause, namely local differences in chromatin structure. Now, either the latter are due to the compositional changes in DNA, and then the argument becomes a circular one (the compositional changes being precisely what is in need of an explanation); or they are not, and then the reason(s) for the appearance of different chromatin structures should be explained.

Three additional considerations weaken even further the point of view according to which mutational biases in the replication/repair enzymes suffice to explain the shifts in base composition. The first one is that the "major" shifts in the compositional patterns of vertebrate genomes are not simply regional GC increases, but are accompanied by a number of phenomena (see preceding section) that have nothing to do with mutational biases. The second consideration concerns the absence of mutational biases over the very extended time of conservative evolution of mammalian genomes, and their appearance in coincidence with the emergence of warm-blooded vertebrates. The third consideration is that the fact that GC-rich isochores have a similar compositional distribution and represent the same fraction (about one-third) in mammalian and avian genomes (which differ by a factor of three in haploid size) would have to be a sheer coincidence.

The Causes of Compositional Shifts in Genome Evolution: Mutational Biases Due to Changes in Precursor Pools?

Another suggestion for the formation of GC-rich isochores is that they are due to changes in the precursor nucleotide pools occurring during the cell cycle. Such changes would have been of such nature as to favor increased GC levels in early-replicating DNA and decreased GC levels in late-replicating DNA (Wolfe, Sharp, and Li, personal communication).

This suggestion is an interesting one in view (1) of the evidence that large changes in the precursor pools do lead to altered base ratios in the newly synthesized DNA and that changes with time in the cell cycle do occur (Leeds et al. 1985); (2) of the early replication of GC-rich isochores and late replication of GC-poor isochores in warm-blooded vertebrates (see Goldman et al. 1984; Bernardi et al. 1985); and (3) of the demonstrated existence of distinct early and late DNA replication in cold-blooded vertebrates (see Giles et al. 1988).

There are, however, two main difficulties with this suggestion: (1) constitutive heterochromatin (which comprises satellite DNAs that are in most cases, but not always, GC-rich), and facultative heterochromatin (such as the inactive X chromosome of mammalian females) replicate at the end of the cell cycle; there is therefore no obvious connection between changes in nucleotide pools, as they are purported to occur, and DNA composition; (2) the problems mentioned at the end of the preceding section also apply to changes in precursor pools.

As a general remark, it should be stressed that our disagreement with the proposals of Sueoka (1988) and Wolfe et al. essentially concerns the idea that mutational biases are the *cause* of the com-

positional shifts, but certainly not the existence of mutational biases nor the role that they may play in the molecular mechanisms leading to the compositional shifts. In other words, we disagree with the idea that DNA composition are just left to the vagaries of mutations in the genes of a few enzymes, because this idea implies that compositional shifts are not important in evolution. This view is not only contradicted by the present work, but also by evidence concerning the dependence of DNA structure and function upon its composition and sequence.

The Causes of Compositional Shifts in Genome Evolution: Negative and Positive Selection at the Isochore Level

An alternative 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 toward 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.

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: (1) GC increases in first and second codon positions lead to amino acid changes that confer thermal stability to proteins (Argos et al. 1979); 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 (Kagawa et al. 1984; Kumai et al. 1986; Nishiyama et al. 1986; Barstow et al. 1987; Kushiro et al. 1987; Kwon et al. 1987; Waldvogel et al. 1987; Wildeman 1988). (2) GC increases in introns, in third codon positions, and in DNA segments corresponding to untranslated regions also conceivably contribute to the thermal stability (Wada and Suyama 1986) of primary transcripts and mRNAs. (3) GC increases in intergenic noncoding sequences can also help in stabilizing DNA structures, possibly through changes in DNA–protein interactions.

A further argument along the same line comes from the very recent observation (Salinas et al. 1988; and paper in preparation) that the genomes of some Gramineae (like those belonging to the *Triticum* and the *Zea* subfamilies) exhibit compositional patterns that are strongly reminiscent of those of warm-blooded vertebrates and that favor thermal stability at the protein, RNA, and DNA levels. In contrast, the genomes of a number of dicotyledons (belonging to several different orders) exhibited compositional patterns similar to those of the genomes of cold-blooded vertebrates. This parallelism can be accounted for by the fact that the former plants originated from arid regions with high maximal temperatures, whereas the latter originated from temperate climates. In agreement with this explanation, the genomes of other Gramineae that originated from humid regions, where water can buffer temperature effects (like those belonging to the *Oryza* subfamily), show compositional patterns that occupy an intermediate position.

In agreement with the selectionist interpretation proposed, the fact that GC increases may concern the totality of the genome, as in thermophilic bacteria, or may be limited to some DNA regions, as in the case of warm-blooded vertebrates, can be understood to be due to different equilibria being reached between genome structure (and function) and environmental parameters. This might explain why approximately the same relative amounts of GC-rich isochores were independently formed in mammals and birds. Once a structural and functional equilibrium is attained with the novel environmental condition, the new compositional pattern appears to be preserved and the conservative mode of evolution to be reinstated.

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.

Obviously, direct evidence for the effect of temperature on compositional patterns would be highly desirable. Along this line, the genomes of some fishes living at high (about 40°C) temperatures were studied, and the presence of GC-rich isochores, not found in congeners living at lower temperature (about 20°C), was detected (Bernardi and Bernardi 1986). Cloned coding sequences from these GC-rich isochores were shown to be rich in GC in third codon positions (71% vs 34% and 33% in first and second positions) and to have undergone amplification events (G. Bernardi and G. Bernardi, unpublished observations). Homologous sequences from the species living at low temperature are currently under investigation.

Two Modes in Genome Evolution

The present work indicates that the evolutionary process as it acts on vertebrate genomes should be visualized as a bimodal one.

In the *conservative mode,* enormous numbers of substitutions accumulated over many million years; however, the composition of coding sequences (and of their different codon positions, including third positions), of the associated introns, and of the intergenic noncoding sequences remained within very narrow limits. Such an invariance of compositional patterns holds for sequences as short as coding sequences (about 1 kb in size) in spite of divergences as large as 50% (with no correction for multiple hits) in third codon positions of homologous coding sequences, and of differences in GC levels that attain 50% in the third codon positions of different pairs of homologous sequences. This invariance appears to be due not only to some compositional conservation in the base substitution process itself, but also to negative selection acting at a regional (isochore) level to eliminate any strong deviation from presumably functionally optimal compositions of isochores.

When the *shifting mode* is operating, large compositional changes occur. Negative selection of isochores with decreasing GC levels and positive selection of isochores with increasing GC levels are apparently at work, under the influence of the functional advantages associated with the compositional shifts. In the case of the two independent transitions between cold-blooded and warm-blooded vertebrates, a great change in body temperature took place, and selective advantages associated with thermal stability of proteins, RNA, and DNA could be identified. Rapid, regional accumulations of mutations biased toward GC increases undoubtedly would facilitate the positive selection process. Such rapid accumulations might be helped (1) by the opening up of AT-rich isochores (which form most of the genomes of cold-blooded vertebrates), perhaps through the action of overproduced (heat-shock?) proteins; (2) by the transcriptional activity of the region (Bohr et al. 1985; Mellon et al. 1986, 1987; Leadon 1986; Okumoto and Bohr 1987); and (3) by mutational biases due to changes in the enzymes of replication/repair and/or in the enzyme make-up of precursor nucleotide pools. Such changes might be due, however, not only to mutations in the corresponding genes, but, conceivably, also to direct or indirect temperature effects on their function.

Very interestingly, the regional negative and positive selections at the isochore level are strongly reminiscent of a process postulated to diminish the genetic load associated with "standard" selection, the "forward creep-back leap" (Zuckerkandl 1975): "In certain parts of the genome, notably in zones of highly repetitive sequences, mutations may be freely accepted as neutral until the sequence adulteration of a larger segment passes a certain threshold. At that time the adulterated sequence may be eliminated by negative selection and may be substituted by a better sequence, one regenerated by amplification from an appropriate master sequence. This process obviously would reduce radically the proportion of events of positive or negative selection necessary to maintain the sequence motifs" (Zuckerkandl 1986).

In conclusion, in both modes of genome evolution, isochores appear to play a role as selection units. Because in neither case selection discriminates between coding and noncoding sequences, the latter must play a functional role, as already suggested, and contribute to the "genome phenotype" (Bernardi and Bernardi 1988). Needless to say, selection at the isochore level is obviously accompanied by both negative and positive selection operating at a small number of genome sites involving changes in critical amino acids and in critical nucleotides in tRNAs, rRNAs, and in signal sequences.

The Selectionist–Neutralist Controversy

At this point, it is of interest to reconsider the selectionist–neutralist controversy that has continued for the past 20 years (Kimura 1968, 1983), in light of the present results.

First of all, it should be stressed that previous investigations (including those on which the neutral theory was built) dealt primarily with the evolution of mammalian proteins and genes and, therefore, concerned only 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 (1) the compositional conservation of homologous coding sequences (and isochores) that characterize the conservative mode of evolution; and (2) 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 considered only at the level of individual base substitutions. On this basis, it was concluded (Kimura 1983) that (1) "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 (2) "only a minute fraction of DNA changes in evolution are adaptive in nature." It is obvious that these conclusions should be revised in the light of the present results.

Indeed, positive selection is not only a 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. This mode, however, appears to be 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; Jukes and Kimura 1984) has proven elusive so far (see Zuckerkandl 1986). A number of individual substitutions occurring in the conservative mode of evolution (which comprises most of the changes that took place 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, although 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.

Acknowledgments. The authors are grateful to Ford Doolittle, Richard Grantham, and Emile Zuckerkandl for very useful comments, and to Cecilia Saccone, Paul Sharp, and Noboru Sueoka for communicating results in advance of publication.

References

Argos P, Rossmann MG, Grau UM, Zuber A, Franck G, Tratschin JD (1979) Thermal stability and protein structure. Biochemistry 18:5698–5703

Barstow DA, Murphy JP, Sharman AF, Clarke AR, Holbrook JJ, Atkinson T (1987) Aminoacid sequence of the L-lactate dehydrogenase of *Bacillus caldotenax* deduced from the nucleotide sequence of the cloned gene. Eur J Biochem 165:581–586

Belozersky AN, Spirin AS (1958) A correlation between the compositions of deoxyribonucleic and ribonucleic acids. Nature 182:111–112

Bernardi F, Ninio J (1978) The accuracy of DNA replication. Biochimie 60:1083–1095

Bernardi G (1985) The organization of the vertebrate genome and the problem of the CpG shortage. In: Cantoni GL, Razin A (eds) Biochemistry and biology of DNA methylation. Alan Liss, New York, pp 3–10

Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 22:363–365

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

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FL, Rindone WP, Swindell CD, Tung CS (1986) The GenBank genetic sequence data bank. Nucleic Acids Res 14:1–4

Bird AP (1987) CpG islands as gene markers in the vertebrate nucleus. Trends Genet 3:342–347

Bohr VA, Smith CA, Okumoto DS, Hanawalt PC (1985) DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. Cell 40:359–369

Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393–1398

Bulmer M (1986) Neighboring base effects on substitution rates in pseudogenes. Mol Biol Evol 3:322–329

Bulmer M (1988) Are codon usage patterns in unicellular organisms determined by selection–mutation balance? J Mol Biol 1:15–26

Carroll RL (1987) Vertebrate paleontology and evolution. WH Freeman, New York

Cortadas J, Olofsson B, Meunier-Rotival M, Macaya G, Bernardi G (1979) The DNA components of the chicken genome. Eur J Biochem 99:179–186

Cox EC, Yanofsky C (1967) Altered base ratios in the DNA of an *Escherichia coli* mutation strain. Proc Natl Acad Sci USA 58:1895–1902

Duijn P Van, Proijev-Knegt AC, Ploeg M Van (1985) The involvement of nucleosomes in Giemsa staining of chromosomes. Histochemistry 82:363–376

Freese E (1962) On the evolution of the base composition of DNA. J Theor Biol 3:82–101

Fresco JR, Broitman S, Lane A-E (1980) Base mispairing and nearest-neighbor effects in transition mutations. In: Mechanistic studies of DNA replication and genetic recombination. Academic Press, New York, pp 753–768

Gasser SM, Laemmli UK (1987) A glimpse at chromosomal order. Trends Genet 3:16–22

Giles V, Thode G, Alvarez MC (1988) Early replication bands in two scorpion fishes, *Scorpaena porcus* and *S. notata* (order Scorpaeniformes). Cytogenet Cell Genet 47:80–83

Goldman MA (1988) The chromatin domain as a unit of gene regulation. BioEssays 9:50–55

Goldman MA, Holmquist GP, Gray MC, Caston A, Naq A (1984) Replication timing of mammalian genes and middle repetitive sequences. Science 224:686–692

Jukes, TH, Kimura M (1984) Evolutionary constraints and the neutral theory. J Mol Evol 21:90–92

Kagawa Y, Nojima H, Nukiwa N, Ishizuka M, Nakajima T, Yasuhara T, Tanaka T, Oshima T (1984) High guanine plus cytosine content in the third letter of codons of an extreme thermophile. J Biol Chem 259:2956–2960

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England

Kumai K, Machida M, Matsuzawa H, Ohta T (1986) Nucleotide sequence and characteristics of the gene for L-lactate dehydrogenase of *Thermus caldophilus* GK24 and the deduced amino acid sequence of the enzyme. Eur J Biochem 160:433–440

Kushiro A, Shimizu M, Tomita K-I (1987) Molecular cloning and sequence determination of the *tuf* gene coding for the elongation factor Tu of *Thermus thermophilus* HB8. Eur J Biochem 170:93–98

Kwon S-T, Terada I, Matsuzawa H, Ohta T (1987) Nucleotide sequence of the gene for aqualysin 1 (a thermophilic alkaline serine protease) of *Thermus aquaticus* YT-1 and characteristics of the deduced primary structure of the enzyme. Eur J Biochem 173:491–497

Leadon SA (1986) Differential repair of DNA damage in specific nucleotide sequences in monkey cells. Nucleic Acids Res 14: 8979–8995

Lee KY, Wahl R, Barbu E (1956) Contenu en bases puriques et pyrimidiques des acides désoxyribonucléiques des bactéries. Ann Inst Pasteur 91:212–224

Leeds JM, Slabourgh MB, Mathews CK (1985) DNA precursor pools and ribonucleotide reductase activity: distribution between the nucleus and cytoplasm of mammalian cells. Mol Cell Biol 5:3443–3450

Luchnik AN, Hisamutdisnov TA, Georgiev GP (1988) Inhibition of transcription in eukaryotic cells by X-irradiation: relation to the loss of topological constraints in closed DNA loops. Nucleic Acids Res 16:5175–5190

Medrano L, Bernardi G, Couturier J, Dutrillaux B, Bernardi G (1988) Chromosome banding and genome compartmentalization in fishes. Chromosoma 96:178–183

Mellon IM, Bohr VA, Smith CA, Hanawalt PC (1986) Preferential DNA repair of an active gene in human cells. Proc Natl Acad Sci USA 83:8878–8882

Mellon IM, Spivak G, Hanawalt P (1987) Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian D4FR gene. Cell 51:241–249

Mouchiroud D, Gautier C (1988) High codon-usage changes in mammalian genes. Mol Biol Evol 5:192–194

Mouchiroud D, Fichant G, Bernardi G (1987) Compositional compartmentalization and gene composition in the genomes of vertebrates. J Mol Evol 26:198–204

Mouchiroud D, Gautier C, Bernardi G (1988) The compositional distribution of coding sequences and DNA molecules in man and murids. J Mol Evol 27:311–320

Nghiem Y, Cabrera M, Cupples CG, Miller JH (1988) The mutY gene: a mutator locus in *Escherichia coli* that generates G.C → T.A transversions. Proc Natl Acad Sci USA 85:2709–2713

Nishiyama M, Matsubara N, Yamamoto K, Iijima S, Uozimi T, Beppu T (1986) Nucleotide sequence of the malate dehydrogenase gene of *Thermus flavus* and its mutation directing an increase in enzyme activity. J Biol Chem 261:14178–14183

Okumoto DS, Bohr VA (1987) DNA repair in the metallothionein gene increases with transcriptional activation. Nature 15:10021–10030

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

Phear G, Nalbantoglu J, Meuth M (1987) Next-nucleotide effects in mutations driven by DNA precursor pool imbalances at the *aprt* locus of Chinese hamster ovary cells. Proc Natl Acad Sci USA 84:4450–4454

Salinas J, Zerial M, Filipski J, Bernardi G (1986) Gene distribution and nucleotide sequence organization of the mouse genome. Eur J Biochem 160:469–478

Salinas J, Matassi G, Montero LM, Bernardi G (1988) Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucleic Acids Res 16:4269–4285

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 85:2653–2657

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

Wada A, Suyama A (1986) Local stability of DNA and RNA secondary structure and its relation to biological function. Prog Biophys Mol Biol 47:113–157

Waldvogel S, Weber H, Zuber H (1987) Structure and function of L-lactate dehydrogenases from thermophilic and mesophilic bacteria VII. Nucleotide sequence of the lactate dehydrogenase gene from the mesophilic bacterium *Bacillus megaterium*. Preparation and properties of a hybrid lactate dehydrogenase comprising moieties of the *B. megaterium* and *B. stearothermophilus* enzymes. Biol Chem Hoppe-Seyler 368: 1391–1399

Wildeman AG (1988) A putative ancestral actin gene present in a thermophilic eukaryote: novel combination of intron positions. Nucleic Acids Res 16:2553–2564

Zehnbauer BA, Vogelstein B (1985) Supercoiled loops and the organization of replication and transcription in eukaryotes. BioEssays 2:52–54

Zerial M, Salinas J, Filipski J, Bernardi G (1986) Gene distribution and nucleotide sequence organization in the human genome. Eur J Biochem 160:479–485

Zuckerkandl E (1975) The appearance of new structures and functions in proteins during evolution. J Mol Evol 7:1–57

Zuckerkandl E (1986) Polite DNA: functional density and functional compatibility in genomes. J Mol Evol 24:12–27

Received May 1, 1988/Revised July 10, 1988