The Human Genome and Its Evolutionary Context

G. BERNARDI AND G. BERNARDI
Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 75005 Paris, France

Our knowledge of the structure, chromosomal location, and transcription of human genes (as well as of eukaryotic genes in general) has made impressive advances in recent years, as witnessed by this Symposium. In contrast, the wider issue of genome organization has largely remained a *terra incognita*. Indeed, the only widely known general picture of the human genome is still based on the reassociation kinetics of short DNA fragments. This information, essentially concerning the relative amounts of sequence classes endowed with different reassociation rates and the interspersion pattern of repetitive and nonrepetitive DNA (Schmid and Deininger 1975), not only is very limited in itself, but also can barely be expanded any further. On the other hand, our advances regarding genes, however important otherwise, concern a minute fraction of the mammalian genome. They are, therefore, unlikely to lead to a breakthrough in our understanding of the overall organization and evolution of the genome. A more hopeful source of information is represented by work on specific families of interspersed repeats (Singer 1982; Singer and Skowronski 1985), but this is a difficult area of research still in a pioneering stage. These points stress the fact that real progress on the issue under consideration here can only come through new approaches, such as those developed in our laboratory over the past years. Here we will review briefly some of our recent results as an introduction, and then present a number of new data on the organization and evolution of the vertebrate genome as well as some general conclusions.

The Compositional Compartmentalization of the Genome of Warm-blooded Vertebrates

Several years ago, we realized that density-gradient centrifugation in the presence of certain DNA ligands could not only provide the expected separation of satellite DNAs, but also a fractionation of the so-called "main-band" DNA from warm-blooded vertebrates into a small number of "light" (GC-poor) and "heavy" (GC-rich) "major components," the modal buoyant densities of which corresponded to maxima of DNA distribution (Fig. 1). The resolution of major components not only was much better in Cs_2SO_4/3,6-bis(acetatomercurimethyl)dioxane (BAMD) or Cs_2SO_4/Ag^+ than in CsCl, but also improved (1) with increasing GC, because GC-rich fractions were less abundant in the genome and less overlapping with other fractions, and (2) with increasing DNA size, because of a lower band-spreading by diffusion. In fact, the families of DNA fragments forming the major components derive by preparative breakage from very long (>200 kb) DNA segments, which are fairly homogeneous in base composition, belong to a small number of major classes distinguished by different GC levels, and probably correspond to the DNA stretches present in Giemsa and reverse chromosomal bands. These DNA segments were called isochores, for "similar regions."

The strong compositional compartmentalization of the genomes of warm-blooded vertebrates contrasts with the weak one shown by those of cold-blooded vertebrates; the latter basically either lack or have scarce GC-rich isochores. This difference is paralleled by chromosomal banding, which is strong in warm-blooded vertebrates and weak or absent in cold-blooded vertebrates.

The separation of major components from the genomes of warm-blooded vertebrates allowed us to study the distribution of any sequence that could be probed, and led to a number of conclusions, which can be summarized as follows (see Bernardi et al. 1985, and references 4–22 therein).

1. Single-copy or clustered genes are usually found in just one of the major components of Figure 1; their distribution appears to be largely conserved within each class of warm-blooded vertebrates (mammals, birds). In contrast, scattered genes and pseudogenes belonging to the same family (like the actin family) may be located on different major components.
2. Genes present in a given major component may be located on different chromosomes; conversely, genes present in different major components may be located on the same chromosome.
3. The distribution of genes (or gene clusters) is highly nonuniform; the majority of those localized so far were found in the heaviest components, which only represent a few percent of total DNA. Families of interspersed repeats are concentrated in given major components. As a consequence, renaturation kinetics of different major components of human DNA are different; these kinetics are also different from those of isopycnic components of mouse DNA. Integrated viral sequences are also mainly located in a given major component.
4. The GC levels of genes, introns, exons, codon third position, specific families of interspersed repeats, and integrated viral sequences, as well as the level of the CpG doublet, are linearly correlated with the

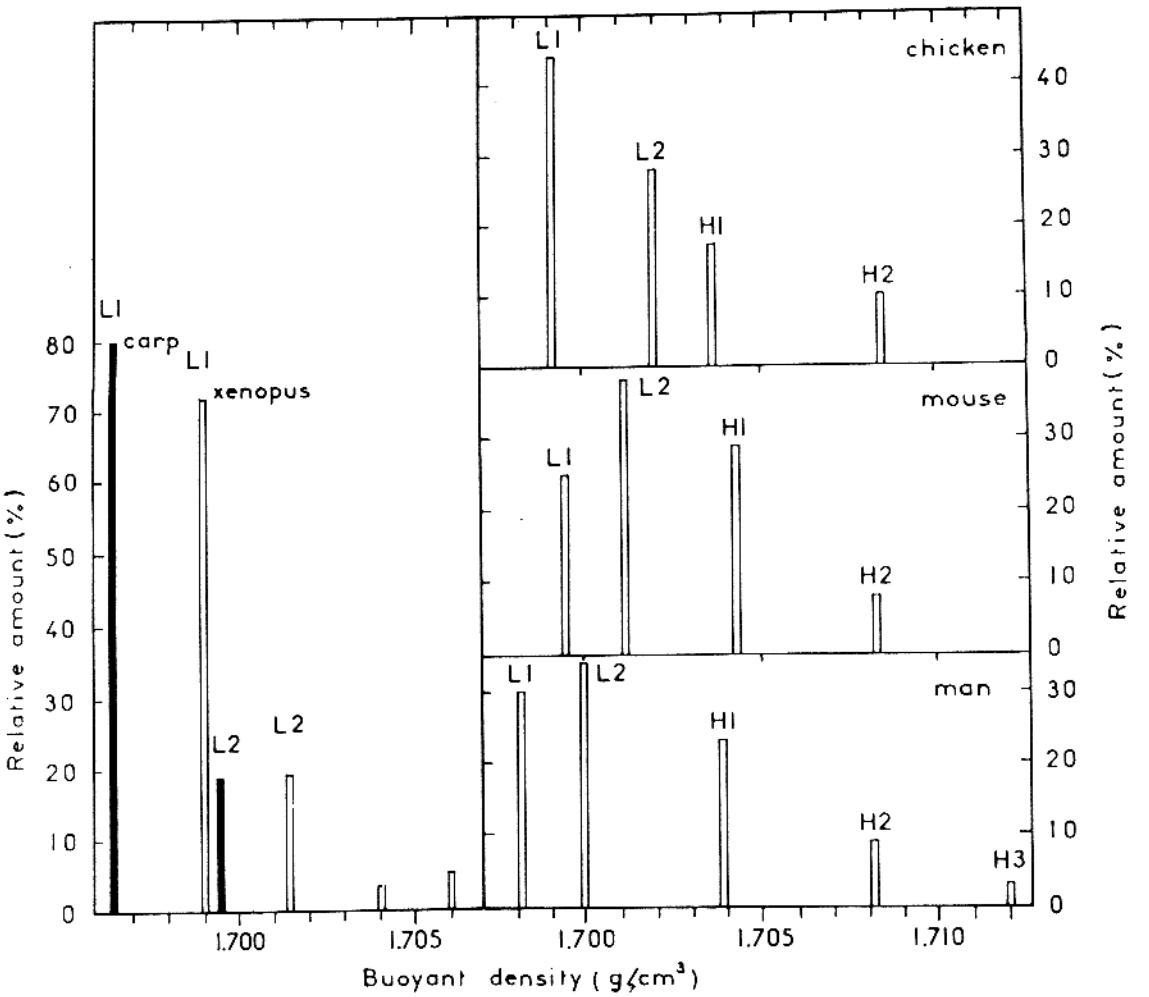

Figure 1. Histograms showing the relative amounts of modal buoyant densities of the major DNA components from *Cyprinus carpio*, *Xenopus laevis* (*left panel*), chicken, mouse, and man (*right panel*), as estimated by analytical CsCl density gradient centrifugation of fractions obtained from preparative Cs_2SO_4 centrifugation in the presence of Ag^+ or BAMD. Satellite and minor components (namely, components representing each less than 3% of DNA) are not shown. Carp and *Xenopus* genomes represent extreme cases of low and high heterogeneity among cold-blooded vertebrates. Notice that even in *Xenopus*, DNA having a density higher than 1.704 represents less than 10% of the genome, as compared with 30–40% for warm-blooded vertebrates. (Modified, from Bernardi et al. 1985.)

GC levels of the major components harboring the sequences under consideration.

5. The evolutionary origin of heavy components is mainly associated with regional GC increases in ancestral sequences from cold-blooded vertebrates, and, occasionally, also with the amplification of preexisting repeated heavy sequences (like the *Alu* sequences). Incidentally, the latter process implies a "targeted" integration of mobile repeats, like that found for viral sequences, and seems, at least in part, to be responsible for the formation of the H3 component in the human genome.
6. The identification of isochores with the DNA segments present in G and R bands rests not only on the points already mentioned (parallelism of compositional compartmentalization and chromosomal banding in vertebrates, presence of different isochores on the same chromosome), but also on the presence in early-replicating DNA (namely in R bands) of early-replicating genes located in the heavy components and in late-replicating DNA (namely in G bands), of late-replicating genes located in light components as well as on other points (Bernardi et al. 1985).

The compositional compartmentalization of the genome of warm-blooded vertebrates indicates that strong compositional constraints are operative on both coding and noncoding sequences. Here we report on (1) the general features of compositional constraints as studied on over 300 genes derived from more than 50 genomes spanning a very wide phylogenetic range (Table 1); (2) the effects of compositional constraints on the structure and function of DNA, RNA, and proteins; (3) the causes of compositional compartmentalization and, more specifically, of the formation of "heavy" isochores; and (4) some more general issues.

Table 1. List of Genomes Examined and Numbers of Genes Analyzed

Genomes	Number of genes	Genomes	Number of genes
Prokaryotes		Viruses	
1A Lambda (left arm)	23	31 Abelson murine leukemia virus	1
1B Lambda (right arm)	39	32 Adenovirus type 12	5
02 *Agrobacterium tumefaciens*	26	33 AKV murine leukemia virus	2
03 *Anacystis nidulans*	1	34 Avian sarcoma virus Y73	1
04 *Bacillus licheniformis*	2	35 Hepatitis B virus	2
05 *Bacillus megaterium*	1	36 Herpes simplex virus type 1	2
06 *Bacillus pumilus*	1	37 Human adult T-cell leukemia virus	3
07 *Bacillus stearothermophilus*	1	38 Human papilloma virus	4
08 *Bacillus subtilis*	7	39 Human papovavirus BK	7
09 *Erwinia amylovora*	2	40 Mouse hepatitis virus	2
10 *Escherichia coli*	43	41 Polyoma virus	6
11 *Haemophilus haemolyticus*	1	42 Tobacco mosaic virus	6
12 *Klebsiella pneumoniae*	4	51* Adenovirus 2 associated virus	
13 *Pseudomonas aeruginosa*	1	52* Adenovirus type 3	
14 *Pseudomonas putida*	1	53* Adenovirus type 25	
15 *Rhizobium* sp.	3	54* Adenovirus type 28	
16 *Salmonella typhimurium*	6	55* ERP	
17 *Shigella dysenteriae*	3	56* Herpes simplex virus type 2	
18 *Streptomyces fradiae*	1	57* Pseudorabies	
19 *Thermus thermophilus*	1	58* Simian adenovirus SA7	
20 *Vibrio cholerae*	4	59* Shope	
		60* Frog virus type 3	
Vertebrates		Vertebrates	
81 *Cyprinus carpio*	1	91 Man	20
82 *Lophius americanus*	1	92 Chicken (L2)	4
83 *Xenopus laevis*	2	93 (H2)	5
84 Chicken	2	94 Mouse (L1)	1
85 Mouse	7	95 (L2)	5
86 Hamster	1	96 (H2)	3
87 Rat	2	97 Rabbit (L2)	1
88 Dog	1	98 (H2)	1
89 Calf	3	99 Man (L2)	5
90 Ape	2	100 (H1)	2
		101 (H3)	6

The nucleotide sequence data for the ~300 genes analyzed in the present study were obtained from the GenBank Genetic Sequence Data Bank (Bilofsky et al. 1986), Release 38.0 (November, 1985). In selecting protein coding sequences we relied mainly on the "Features" tables of the GenBank Data base, and only complete genes, starting with an initiation codon and ending with one of the stop codons, were used in the present work. Numbers with asterisks refer to viral genomes for which only the CpG/GpC ratio was available (Russell 1974). In the case of warm-blooded vertebrates, the table lists the isochore classes or genome compartments (L1, L2, H1, H2, H3) in which several genes were localized (Bernardi et al. 1985). These genes were different from those considered under items 84, 85, and 91. A list of all the genes used in this analysis is available upon request.

Compositional Constraints Affect both Coding and Noncoding Sequences

If the GC levels found at each codon position of genes from different genomes are plotted against the GC levels of the coding sequences from the same genomes, linear relationships characterized by roughly the same slopes and the same intercepts are found for prokaryotes, viruses, and vertebrates (Fig. 2; the only exception concerns second positions of viruses). The slopes generally increase from second-, to first-, and to third-position plots.

In the case of prokaryotic and viral genomes, which are almost exclusively formed by coding sequences, the relationships described for coding sequence GC obviously coincide with those for gene GC and genome GC. In contrast, in the case of vertebrate genomes, where noncoding sequences are abundant, the relationships generally show slight, but significant, changes when plots are made against gene GC or genome GC. The slopes of "gene plots" (not shown) are slightly lower than those of exon plots, due to the influence of intron GC, whereas the linear relationships of genome plots are shifted to the left; this indicates a lower GC level in intergenic sequences compared with exons or genes (Fig. 2).

A special compositional constraint present in both coding and noncoding sequences of vertebrate genomes leads to a shortage of the doublet CpG (Swartz et al. 1962). On the basis of the different CpG levels exhibited by the genomes of small and large vertebrate viruses, it was suggested (1) that CpG shortage is associated with constraints from the translation machinery (Subak-Sharpe et al. 1967; Russell 1974) and (2) that the bulk of vertebrate DNA derives from, and maintains the CpG shortage of, polypeptide-specifying DNA (Russell 1974; Russell et al. 1976). Both suggestions (for a more detailed discussion, see Bernardi 1985)

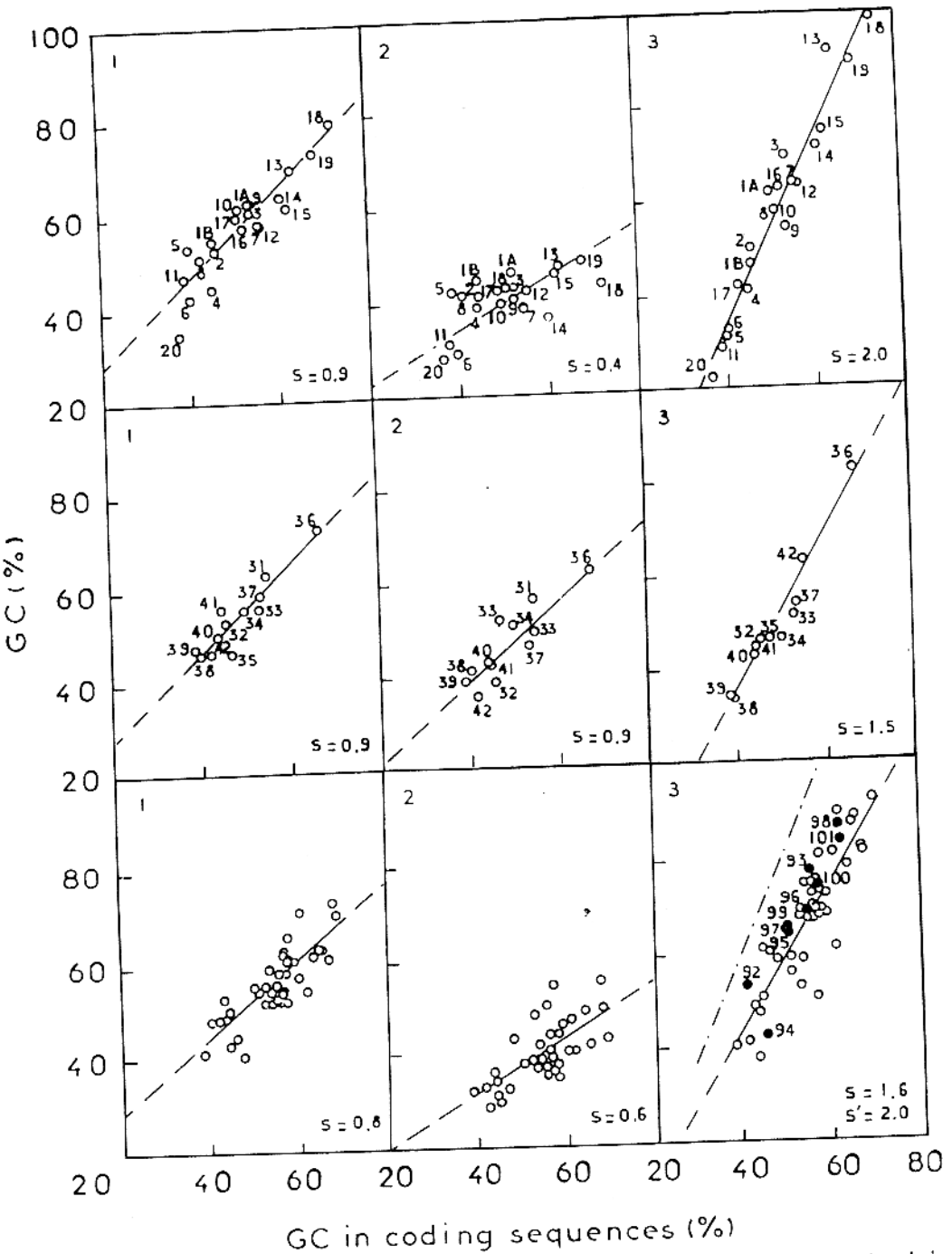

Figure 2. (*Top and middle frames*) GC levels of the three codon positions (1,2,3) of prokaryotic (*top frames*) and viral (*middle frames*) genes are plotted against the GC levels of the corresponding genomes. The scatter of points belonging to different genes from the same genome was small and average values, weighted for gene size, were therefore used. Numbers refer to the genomes listed in Table 1. Lines were drawn using the least-squares method; the slopes, *s*, are indicated; correlation coefficients were 0.91, 0.58, and 0.97 for prokaryotic genomes, and 0.91, 0.88 and 0.95 for viral genomes, respectively. (*Bottom frames*) (○) GC levels of codon positions of individual genes from vertebrates are plotted against the GC levels of the corresponding exons; (●) average values for third position GC of genes belonging to the same compartment of a given genome (see Table 1) are plotted against the GC levels of coding sequences of the genome compartment; correlation coefficients were 0.81, 0.67, and 0.88, respectively; the dash-and-point line is the third position plot against GC of genome compartments; its slope is given by *s'*.

are contradicted, however, by the findings that the CpG level of the genomes of vertebrate viruses is not related to the small or large size of the viral genome (and its supposed dependence or independence upon the translation machinery of the host cell), but simply upon the GC levels of the viral genomes (the shortage disappearing at high GC levels; Fig. 3), and that the relationship for viral genomes is very close to that previously seen

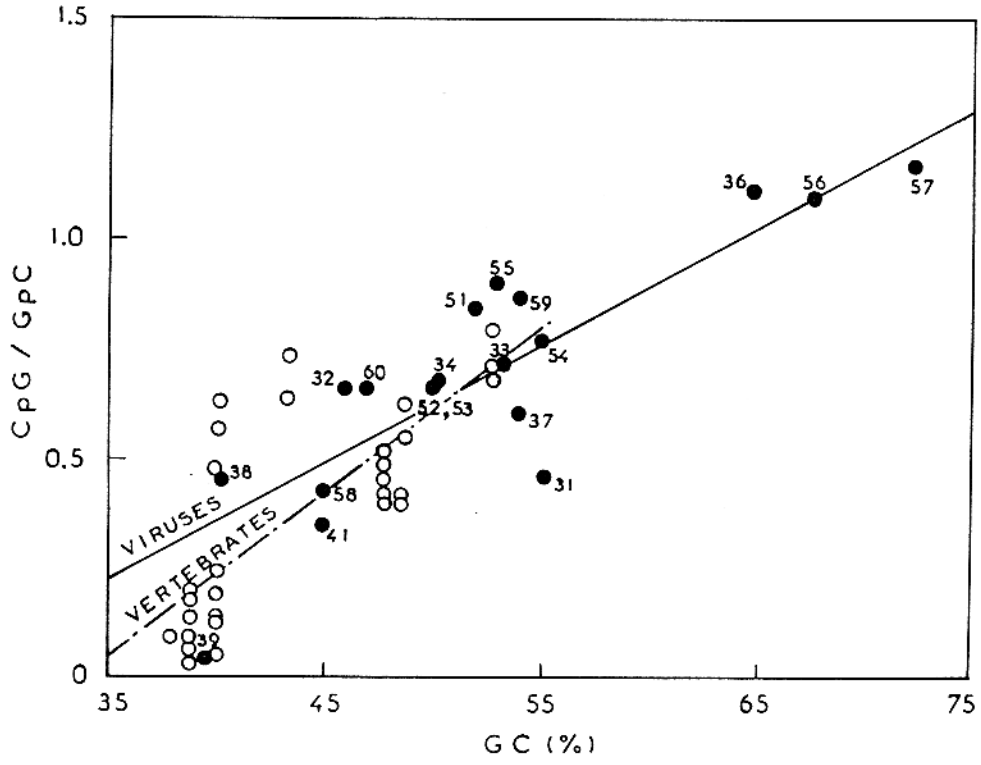

Figure 3. Plot of the CpG/GpC ratio for viral genomes against genome GC (●). Numbers refer to Table 1. Except for frog virus 3 (point 60), all viruses were from warm-blooded vertebrates. The dash-and-point line and the open symbols correspond to a similar plot, as obtained for vertebrate genes or exons (Bernardi et al. 1985).

(Bernardi et al. 1985) for vertebrate genes or exons (Fig. 3).

The general conclusion to be drawn from the above results is that essentially the same compositional constraints affect coding and noncoding sequences (which represent more than 90% of DNA in mammalian genomes). In coding sequences, these constraints concern not only GC levels, but also the levels of individual bases in different codon positions (not shown), and lead to specific relationships for different bases; obviously, such relationships are not the result of a random increase in G and C in different codon positions. Constraints on introns and intergenic sequences are indicated by the linear relationships found between the GC levels of codon positions and those of genes or genomes (Fig. 2; see also Bernardi et al. 1985); the slopes and intercepts of these relationships are close to those exhibited by coding sequences.

GC Increases in Coding Sequences Affect mRNA and Protein Stability

All GC changes in second-codon positions entail changes in the amino acid composition of proteins; so also do most first-position changes and a few third-position changes. An analysis of the amino acid replacements that accompany the GC increases in codon positions has revealed that they comprise those that lead to thermodynamically more stable proteins (Argos et al. 1979; Zuber 1981) (Table 2). Indeed, the amino acids (alanine and arginine) that are most frequently acquired in thermophiles and that most contribute to an increased stability increase, whereas those (serine and lysine) that are correspondingly lost and that diminish stability decrease, with increasing exon GC (Fig. 4). In the case of compartmentalized genomes, these changes may take place within the same genome; for instance, in the human genome the (Ala + Arg)/(Ser + Lys) molar ratio varies by a factor of four between the AT-richest and GC-richest genes (Fig. 4).

In conclusion, the compositional changes that make

Table 2. Amino Acid Exchanges Observed in Thermophiles and the Accompanying GC Changes in Their Codons

Exchanges			Codon GC
mesophiles		thermophiles	
*Gly	→	Ala	0
*Ser	→	Ala	+
Ser	→	Thr	0
Lys	→	Arg	+
Asp	→	Glu	±
Ser	→	Gly	+
*Lys	→	Ala	+

Exchanges are given in order of decreasing frequency. Asterisks indicate the largest expected increase in stability (Argos et al. 1979). Only the Lys → Ala exchange requires more than one base change per codon. The right column indicates increases (+), decreases (−), and no changes (0) in codon GC levels.

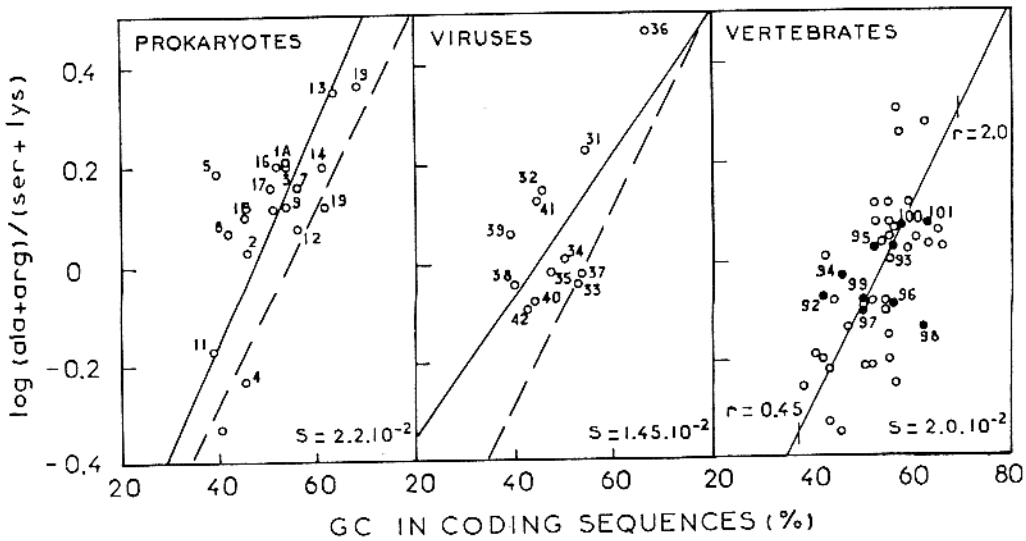

Figure 4. Plot of (alanine + arginine)/(serine + lysine) molar ratio against GC levels of coding sequences from prokaryotes, viruses, and vertebrates (see Table 1). In the vertebrate plot: (○) individual genes (for man, chimpanzee, mouse, hamster, rat, calf, and chicken); (●) average values of genes belonging to the same compartment of a given genome (see Table 1); the two *r* values correspond to the lowest and highest ratios found for human genes. Slopes are indicated by *s*; the vertebrate line is also shown in the other two diagrams (*broken lines*), for the sake of comparison.

DNA thermodynamically more stable also increase the thermodynamic stability of the encoded proteins. The same changes obviously also lead to higher GC levels in mRNAs, a factor known (Hasegawa et al. 1979) to increase their base-pairing and stability. The limited data available so far suggest that similar changes may occur in rRNAs and tRNAs.

These points suggest that the formation of GC-rich isochores may have begun in cold-blooded vertebrates exposed to higher temperatures, by a preferential fixation of A/T→G/C changes in the genes of proteins that could be most affected in their function by temperature. This process was apparently accompanied by a GC enrichment in the neighboring noncoding sequences and led to a limited formation of heavy isochores (experimentally found in a number of cold-blooded vertebrates; see Bernardi et al. 1985). The same process went much farther in warm-blooded vertebrates, where it led to an extensive formation of GC-rich isochores. The suggestion that genes were the initial nuclei for the regional increases in GC would account, at least in part, for the high gene concentration in GC-rich isochores (Bernardi et al. 1985). It also raises the possibility that a particular set of genes was preferentially located in heavy components. This set might correspond to housekeeping genes because these genes appear to be only present in R bands (Goldman et al. 1984), namely in the GC-rich components of the genome, whereas tissue-specific genes are preferentially located in GC-poor G bands. It is clear, however, that a number of the latter genes were later translocated into GC-rich components, where they subsequently underwent a GC increase insuring a better protection against DNA breathing and mutability (L. Orgel, pers. comm.).

Codon Usage Is Largely Determined by Compositional Constraints

Since nonrandomness of codon usage was first discovered, several not mutually exclusive explanations have been provided for this phenomenon. These comprise: (1) the optimization of codon–anticodon interaction energy (Grosjean et al. 1978) and the consequent optimization of translation efficiency in highly expressed genes (Grantham et al. 1981; Bennetzen and Hall 1982); (2) the fulfillment of requirements for mRNA secondary structure and stability (Hasegawa et al. 1979); and (3) an adaptation of codons to the actual populations of isoaccepting tRNAs (Ikemura 1985). These explanations essentially rest on intraspecific differences in the usage of all synonymous codons.

In contrast, our results concern interspecific and intercompartmental differences in the usage of synonymous codons characterized by different GC levels in third positions; this subset of codons corresponds to two-thirds of all synonymous codons. Our results lead to the following conclusions.

1. Interspecific and intercompartmental differences in codon usage largely depend upon the compositional constraints affecting the genome, or the genome compartments. This provides, to a large extent, a rationale for the "genome strategy of codon usage" (Grantham et al. 1980), which comprises several "compartmental strategies" in compartmentalized genomes.
2. The proposals that mRNA structure (Hasegawa et al. 1979) or the abundance of synonymous tRNAs (Ikemura 1985) are the causes and not the effects of codon usage should be reversed. Since third posi-

tions are under essentially the same compositional constraints as noncoding sequences, the primary phenomenon is at the DNA level and the effects are at the mRNA or tRNA levels; this point was already well demonstrated by the changes in tRNA distributions that occur in the silk gland of *Bombyx mori* in connection with the expression of the fibroin and sericin genes (Chevallier and Garel 1979). As already mentioned, our results do not bear on the intraspecific differences in codon usage that have been shown in unicellular organisms, like *Escherichia coli* and *Saccharomyces cerevisiae* (Ikemura 1985). These differences should probably be visualized as due to the selection of a subset of codons for highly expressed genes on the basis of optimization of codon–anticodon interaction energies (Grosjean et al. 1978).

3. The results shown in Figure 2 account for the so-called "contextual constraints" previously seen in different codon positions (Nussinov 1980, 1981; Lipman and Wilbur 1983), for the frequency of pyrimidine-purine doublets in third- and first-codon positions, and for the 3-bp periodicity detected in coding sequences (Trifonov and Sussman 1980; Shepherd 1981).

Mutations Are Fixed Mainly Through Positive Darwinian Selection

The evolution of living organisms is primarily caused by mutations that may subsequently be eliminated or become fixed in the genome. It is generally agreed that elimination affects deleterious mutations and occurs by negative selection. In contrast, fixation has been visualized as due either (1) to positive Darwinian selection acting on advantageous mutations or (2) to random genetic drift acting on selectively neutral (i.e., selectively equivalent) mutations. Our findings have a direct bearing on the mechanism by which mutations become fixed.

Intragenomic changes. (1) Changes in any codon position appear to have been fixed, on the average, under the influence of compositional constraints, in conformity with the base composition of the genome, or the genome compartment, in which genes are located. For instance, the base changes that occurred over evolutionary time in the genomes of warm-blooded vertebrates preferentially kept low and high GC levels in different codon positions of GC-poor and GC-rich genes, respectively. Moreover, the scatter in the base compositions of individual codon positions of different genes belonging to the same genome, or the same genome compartment, is small (a fact justifying our averaging values in Fig. 2). These findings indicate that, on the average, changes are conservative. By analogy with the "genome strategy of codon usage" (Grantham et al. 1980), one should, therefore, take into consideration a more general "compositional strategy of coding sequences," which also concerns nonsilent changes and may comprise several compartmental strategies in compartmentalized genomes. (2) Changes in noncoding sequences of eukaryotes conform with the same general rules as changes in coding sequences. In eukaryotes, the compositional strategy of coding sequences is therefore part of a general compositional strategy which also affects noncoding sequences. (3) The CpG level (which corresponds to the level of potential methylation sites) in both coding and noncoding sequences of vertebrates also appears to be subject to the same compositional constraints as the base changes just discussed; indeed, the CpG shortage is different in different genome compartments and is related to their GC levels (Fig. 3).

To sum up, intragenomic GC changes clearly indicate that most mutations are fixed, in both coding and noncoding sequences, not at random but under the influence of compositional constraints, in compliance with a general compositional strategy. Random fixation of neutral mutations (Kimura 1968, 1983, 1986; King and Jukes 1969) certainly also occurs, but only to such an extent that the general compositional strategy is not blurred.

Intergenomic changes. (1) The decreasing extents of GC changes from the third to the first and to the second positions appear to be correlated with the correponding increasing impacts on amino acid composition of proteins. In other words, the slopes of Figure 2 are correlated with the different fixation rates that have been detected in different codon positions of a number of genes (Zuckerkandl 1976; Kimura 1983, 1986; Li et al. 1985). While the latter results concern differences at specific positions of homologous genes from different organisms, the data of Figure 2 compare average values for codon positions of genes from given organisms with those from other organisms. (2) A clear directionality is shown by the amino acid substitutions, the silent base changes, the changes in noncoding seqeunces, and the CpG changes that accompanied the transition from cold-blooded to warm-blooded vertebrates, to lead to the formation of GC-rich genes and GC-rich isochores in the genomes of the latter.

In conclusion, the compositional constraints within a genome (or within genome compartments) and the directional changes just mentioned can only be explained by a positive Darwinian selection acting on mutations that confer selective advantages in relationship with environmental pressures. Some of these advantages have been identified. For instance, in the transition from cold-blooded to warm-blooded vertebrates, silent changes appear to lead to an optimization of structure and function at the level of both DNA and RNA; nonsilent changes lead, in addition, to an optimization of structure and function at the protein level.

Obviously, our conclusions reverse the proposals of the "neutral mutation–random drift hypothesis" (1) "that 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 and nearly neutral mutants" and (2) "that only a minute fraction of DNA changes are adaptive in nature" (Kimura 1986). Both proposals rest, in fact, on the premise (see next section) that the phenotype of an organism only corresponds to its "gene products"; as a logical consequence, silent mutations and changes in noncoding sequences were visualized as drifting at random and having no evolutionary impact.

CONCLUSIONS

The main finding reported here concerns the demonstration of compositional constraints that affect the genomes of living organisms and are more general and stronger than other genome constraints, namely (1) fixation rate constraints (previously called functional constraints [Kimura 1983]; this terminology should be abandoned since all constraints have functional consequences) that increasingly limit the fixation of mutations taking place in third-, first-, and second-codon positions; (2) constraints associated with the CpG (and potential methylation) levels of vertebrate genomes; and (3) constraints associated with the codon usage of highly expressed genes from unicellular organisms.

Indeed, compositional constraints affect both coding and noncoding sequences from all organisms explored, whereas other constraints are more limited in the range of sequences and/or organisms concerned, are evident in all codon positions independently of differences in fixation rates (Fig. 2), lead to the disappearance of the CpG shortage in GC-rich genomes (Fig. 3), and are not hidden by the preferential use of a subset of codons of highly expressed genes in unicellular organisms.

Compositional constraints indicate that most mutations are not fixed at random, but are fixed in relationship to a general compositional strategy of the genome. This appears to be the result of positive Darwinian selection of mutations that is advantageous as far as environmental pressures are concerned. Neutral mutations obviously also exist, but their fixation is small enough so as not to distort the general compositional strategy. These conclusions lead to two general ideas: (1) that the environment can shape up the genome through selection; and (2) that genome evolution depends more on natural selection than on random events.

Compositional constraints largely affect the structure and stability of the genome (at its different DNA, chromatin, and chromosome levels), of the transcripts, and even of proteins (as exemplified by the higher stability accompanying GC increases), as well as codon usage. At the same time, they also conceivably touch on a number of basic functions, such as replication, recombination, transcription, and translation, that are sensitive to the compositional/structural features just mentioned.

In eukaryotes, both coding and noncoding sequences appear to be under essentially the same compositional constraints, and therefore under the same selection pressures. This finding leads to a number of conclusions. First of all, it stresses the fundamental unity of the genome, already suggested by the genome strategy of codon usage (Grantham et al. 1980), and contradicts what has been called the "bean bag" view of the genes within the genome (Mayr 1976). Second, it points to the genome as the unit upon which natural selection acts. Third, it does not support the view that noncoding sequences can be equated with functionless "junk DNA" (Ohno 1972). In contrast, it suggests that noncoding sequences largely play a physiological role, which may have to do with the modulation of basic genome functions.

As far as the latter suggestion is concerned, it was already pointed out that the fixation of advantageous mutations concerns the majority, but not the totality, of mutations. Likewise, the functional role of noncoding sequences is not an all-or-none issue, and should be visualized as associated with the majority, and not the totality, of noncoding sequences. Moreover, this suggestion, although not a new one (Britten and Davidson 1969; Davidson and Britten 1979), does not rest anymore on "adaptive stories" (Gould and Lewontin 1979), which can be criticized (Doolittle and Sapienza 1980; Orgel and Crick 1980), but on the newly demonstrated existence of compositional constraints. Interestingly, similar conclusions have been reached on the basis of independent evidence for the noncoding sequences of mitochondrial genome of yeast (Bernardi 1982, 1983; de Zamaroczy and Bernardi 1985, 1986, in prep.).

Finally, compositional constraints identify a new component in the organismal phenotype, which may be called the "genome phenotype." This component adds to the other classical component of the phenotype, which is formed by "gene products," and is defined by nonsilent mutations in the genes and by mutations in regulatory signals.

ACKNOWLEDGMENTS

The senior author thanks the Fogarty International Center for Advanced Study in the Health Sciences, National Institutes of Health, Bethesda, Maryland, for a scholarship. Sequence data treatments were performed using computer facilities at CIT12 in Paris on a PDP8 computer with the help of the French Ministere de la Recherche et la Technologie (Programme Mobilisateur "Essor des Biotechnologies").

REFERENCES

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

Bennetzen, J.L. and B.D. Hall. 1982. Codon selection in yeast. *J. Biol. Chem.* **257:** 3026.

Bernardi, G. 1982. The evolutionary origin and the biological role of non-coding sequences in the mitochondrial genome of yeast. In *Mitochondrial genes* (ed. G. Attardi et

al.), p. 269. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
———. 1983. Genome instability and the selfish DNA issue. *Folia Biol.* **29:** 82.
———. 1985. The organization of the vertebrate genome and the problem of the CpG shortage. In *Biochemistry and biology of DNA methylation* (ed. G.L. Cantoni and A. Razin), p. 3. A.R. Liss, New York.
Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228:** 953.
Bilofsky, H.S., C. Burks, J.W. Fickett, W.B. Gould, F.L. Lewitter, W.P. Rindone, C.D. Swindell, and C.S. Tung. 1986. The GenBank genetic sequence data bank. *Nucleic Acids Res.* **14:** 1.
Britten, R.J. and E.H. Davidson. 1969. Gene regulation for higher cells: A theory. *Science* **165:** 349.
Chevallier, A. and J.R. Garel. 1979. Studies on tRNA adaptation, tRNA turnover, precursor tRNA and tRNA gene distribution in *Bombyx mori* by using two-dimensional polyacrylamide gel electrophoresis. *Biochimie* **61:** 245.
Davidson, E.H. and R.J. Britten. 1979. Regulation of gene expression: Possible role of repetitive sequences. *Science* **204:** 1052.
Doolittle, W.F. and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284:** 601.
Goldman, M.A., G.P. Holmquist, M.C. Gray, L.A. Caston, and A. Nag. 1984. Replication timing of genes and middle repetitive sequences. *Science* **224:** 686.
Gould, S.J. and R.C. Lewontin. 1979. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptionist program. *Proc. Roy. Soc. Lond.* **B205:** 581.
Grantham, R., C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9:** r43.
Grantham, R., C. Gautier, M. Gouy, R. Mercier, and A. Paré. 1980. Catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8:** r49.
Grosjean, H., D. Sankoff, W. Min Jou, W. Fiers, and R.J. Cedergren. 1978. Bacteriophage MS 2RNA: A correlation between the stability of the codon:anticodon interaction and the choice of code words. *J. Mol. Evol.* **12:** 113.
Hasegawa, H., T. Yasunaga, and T. Miyata. 1979. Secondary structure of MS2 phage RNA and bias in code word usage. *Nucleic Acids Res.* **7:** 2073.
Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2:** 13.
Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217:** 624.
———. 1983. *The neutral theory of molecular evolution.* Cambridge University Press, England.
———. 1986. DNA and the neutral theory. *Phil. Trans. R. Soc. Lond.* **B312:** 343.
King, J.L. and T.H. Jukes. 1969. Non-darwinian evolution. *Science* **164:** 788.
Li, W.-H., C.-C. Luo, and C.-I. Wu. 1985. Evolution of DNA sequences. In *Molecular evolutionary genetics* (ed. R.J. Mac Intyre), p. 1. Plenum Press, New York.
Lipman, D.J. and W.J. Wilbur. 1983. Contextual constraints on synonymous codon choice. *J. Mol. Biol.* **163:** 363.
Mayr, E. 1976. *Evolution and the diversity of life.* Harvard University Press, Cambridge, Massachusetts.
Nussinov, R. 1980. Some rules in the ordering of nucleotides in the DNA. *Nucleic Acids Res.* **8:** 4545.
———. 1981. The universal dinucleotide asymmetry rules and the amino acid codon choice. *J. Mol. Evol.* **17:** 237.
Ohno, S. 1972. An argument for the genetic simplicity of man and other mammals. *J. Hum. Evol.* **1:** 651.
Orgel, L.E. and F.H.C. Crick. 1980. Selfish DNA: The ultimate parasite. *Nature* **284:** 604.
Russell, G.J. 1974. "Characterization of deoxyribonucleic acids by doublet frequency analysis." Ph.D. Thesis, University of Glasgow, Scotland.
Russell, G.J., P.M.B. Walker, R.A. Elton, and J.H. Subak-Sharpe. 1976. Doublet frequency analysis of fractionated vertebrate nuclear DNA. *J. Mol. Biol.* **108:** 1.
Schmid, C.W. and P.L. Deininger. 1975. Sequence organization of the human genome. *Cell* **6:** 345.
Shepherd, J. 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl. Acad. Sci.* **78:** 1596.
Singer, M.F. 1982. SINES and LINES: Highly repeated short and long interspersed sequences in mammalian genomes. *Cell* **28:** 433.
Singer, M.F. and J. Skowronski. 1985. Making sense out of LINES: Long interspersed repeats in mammalian genomes. *Trends Biochem. Sci.* **10:** 119.
Subak-Sharpe, H., R.R. Burk, L.V. Crawford, J.M. Morrison, J. Hay, and H.M. Keir. 1967. An approach to evolutionary relationships of mammalian DNA viruses through analysis of the pattern of nearest neighbor base sequences. *Cold Spring Harbor Symp. Quant. Biol.* **31:**737.
Swartz, M.N., T.A. Trautner, and A. Kornberg. 1962. Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. *J. Biol. Chem.* **237:** 1961.
Trifonov, E.N. and J.L. Sussman. 1980. The pitch of chromatin is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci.* **77:** 3816.
de Zamaroczy, M. and G. Bernardi. 1985. Sequence organization of the mitochondrial genome of yeast—A review. *Gene* **37:** 1.
———. 1986. The GC clusters of the mitochondrial genome of yeast and their evolutionary origin. *Gene* **41:** 1.
Zuber, H. 1981. Structure and function of thermophilic enzymes. In *Structural and functional aspects of enzyme catalysis* (ed. H. Eggerer and R. Huber), p. 114. Springer Verlag, Berlin.
Zuckerkandl, E. 1976. Evolutionary processes and evolutionary noise at the molecular level. II. A selectionist model for random fixation in proteins. *J. Mol. Evol.* **7:** 269.