Plenary Lecture, Frontiers in Biology: The Challenges of Biodiversity,
Biotechnology and Sustainable Agriculture,
(C. H. Chou and K. T. Shao, eds.), Academia Sinica, Taipei (1998)

THE ORGANIZATION OF THE HUMAN GENOME

Giuseppe D'Onofrio, Oliver Clay, Kamel Jabbari, Fernando Alvarez-Valin,
Salvatore Saccone, and Giorgio Bernardi
Laboratoire de Génétique Moléculaire
Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France
E-mail:Bernardi@citi2.fr

ABSTRACT

Isochores. The human genome is a mosaic of isochores, long (>300 Kb) DNA segments, which are compositionally homogeneous and can be subdivided into a small number of families characterized by different GC levels covering a 30-60% range. Isochore families L1+L2 and H1+H2+H3 represent the GC-poor 2/3 and the GC-rich 1/3 of the genome, respectively. The isochore organization of the human genome is typical of warm-blooded vertebrates, whereas cold-blooded vertebrate genomes are characterized by a much narrower GC range which never reaches the high levels attained by warm-blooded vertebrates. The different isochore patterns just mentioned are paralleled by different compositional patterns of coding sequences. They represent different genome phenotypes.

Isochores and chromosomes. In *situ* suppression hybridization of human DNA fractions characterized by increasing GC levels on human metaphase chromosomes has clarified the correlations between isochores and chromosomal bands: (i) T (elomeric) - bands contain the GC-richest isochores of the H3 family; (ii) R-bands, namely R (everse) - bands exclusive of T-bands, can be divided into two subclasses, T' bands comprising clusters of H3 isochores, and R" bands not containing H3 isochores; (iii)G (iemsa) - bands essentially consist of GC-poor isochores from the L1+L2 families, with a minor contribution of H1 isochores.

Gene distribution. The distribution of genes in the human genome is strikingly non-uniform. Indeed, a low gene concentration is present in the GC-poor isochore families L1 and L2; gene concentration then increases in increasingly GC-richer isochores (isochore families H1 and H2) to attain the highest value (20x higher than in L1+L2) in the GC-richest isochore family H3, that only represents about 4% of the genome. Because isochore distribution in chromosomes is known, these results also provide information on the distribution of genes in chromosomes.

Evolutionary aspects. The reasons for the correlation between gene concentration of isochores and their GC levels is now understood. Indeed, the gene concentration pattern of the human genome is basically present in all vertebrates. A strong GC increase took place, however, in gene-rich regions at the transition between cold-blooded vertebrates on the one hand and mammals and birds (two separate events in time) on the other. A comparison of aligned homologous coding sequences unequivocally demonstrated that GC changes were caused by directional fixation of mutations, which were then maintained at least since the appearance of present-day mammals and birds.

Key Words: Chromosomes, genome, isochores.

The term *genome* was coined over three quarters of a century ago by a German botanist, Hans Winkler (1920) to designate the haploid chromosome set. While current textbooks of Molecular Biology do not yet go beyond this purely operational definition of the eukaryotic genome (updated as the sum total of genes and of intergenic sequences), a number of molecular biologists have been thinking for some time that the genome is more than the sum of its parts. This implies the existence of structural and functional interactions between the small minority of coding sequences and the vast majority of non-coding sequences. This general and rather vague concept has been changed into a precise one by the analysis of the compositional properties of genomes. These properties, which have mainly been investigated in the nuclear genomes of vertebrates (see Bernardi *et al.*, 1985; Bernardi, 1989, 1993, 1995) will be briefly outlined here. They comprise the isochore organization, the compositional patterns of DNA fragments and of coding sequences, the compositional correlations between coding and non-coding sequences and, above all, the gene distribution and its associated functional properties.

The mammalian genomes are mosaics of *isochores* (see Fig. 1a), namely of long (>300 Kb) DNA segments that are homogeneous in base composition and range from 30 to 60% GC in the case of the human genome (Thiery, Macaya, and Bernardi, 1976; Macaya, Thiery, and Bernardi, 1976). This extremely wide range is stable over geological time, since its is shared by all warm-blooded vertebrates (Bernardi and Bernardi, 1990a, b; Sabeur *et al.*, 1993; Kadi *et al.*, 1993). This range is comparable to that of all bacterial DNAs, 25-75% GC (Belozersky and Spirin, 1958), each one of which shows a very low level of intragenomic variability (Rolfe and Meselson, 1959; Sueoka, 1959). In the human genome, isochores can be assigned to two GC-poor families (L1 and L2) representing 2/3 of the genome, and to three GC-rich families (H1, H2 and H3) forming the remaining 1/3 (Fig. 1b).

The *compositional distributions* of large (>100 Kb) genome fragments, such as those forming routine DNA preparations, of exons (and particularly of their third codon positions) and of introns represent *compositional patterns* (Bernardi *et al.*, 1985). These correspond to *genome phenotypes* (Bernardi and

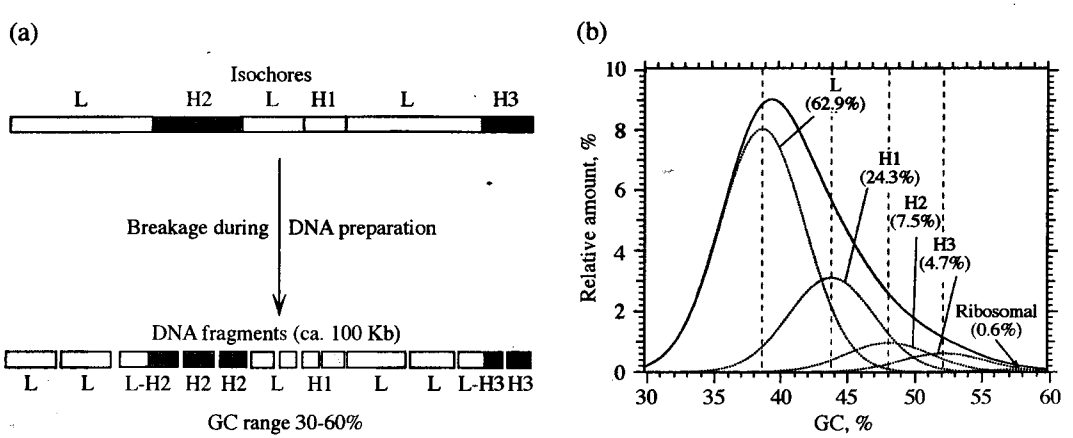

Fig. 1. (a) Scheme of the isochore organization of the human genome. This genome, which is a typical mammalian genome (Sabeur *et al.*, 1993), is a mosaic of large (>300 Kb) DNA segments, the isochores. These are compositionally homogeneous (above a size of 3 Kb) and can be partitioned into a small number of families, GC-poor (L1 and L2), GC-rich (H1) and (H2), and very GC-rich (H3). The GC-range of the isochores from the human genome is 30-60% (modified from Bernardi, 1993). (b) The isochore families of the human genome. The relative amounts of DNA fragments derived from isochore families L (*i.e.*, L1 + L2), H1, H2, H3 are superimposed on the CsCl profile of human DNA (modified from Zoubak *et al.* 1996).

Bernardi, 1986), in that they differ characteristically not only between cold- and warm-blooded vertebrates, but also between mammals and birds and even between murids and most other mammals (see Fig. 2).

Compositional correlations (Bernardi *et al.*, 1985) exist (Fig. 3a, b, c) between exons (and their codon positions) and isochores, as well as between exons and introns (Aïssani *et al.*, 1991; Clay *et al.*, 1996). These correlations concern coding and non-coding sequences and are not trivial since coding sequences only make up about 3% of the genome, whereas non-coding sequences correspond to 97% of the genome. The compositional correlations represent a *genomic code* (Bernardi, 1993). It should be noted that a *universal correlation* (D'Onofrio and Bernardi, 1992) holds among GC levels of codon positions of human genes (Fig. 3d) as it also holds among genes from both prokaryotic and eukaryotic genomes. Thus, the correlation among codon positions is at same time an intra-genomic correlation that can be seen in genomes showing a broad compositional distribution, and an inter-genomic correlation among compositionally homogenous genomes. Both the genomic code and the universal correlation are apparently due to compositional constraints working in the same direction (towards GC or AT), although to different extents, on coding and non-coding sequences, as well as on different codon positions.

The compositional correlation between GC_3 (the GC level of third codon positions) and isochore GC have a practical interest in that allow the positioning of the coding sequence histogram of Fig. 2 relative to the CsCl profile of Fig. 1 and the assessment of the *gene distribution* in the human genome (Mouchiroud *et al.*, 1991; Bernardi, 1995; Zoubak, Clay, and Bernardi, 1996). In fact, if one divides the relative number of genes per histogram bar by the corresponding relative amount of DNA, one can see that the ratio, namely the gene concentration, is low in GC-poor isochores, increases with increasing GC in isochore families H1 and H2, and reaches a maximum in isochore family H3, which exhibits at least a

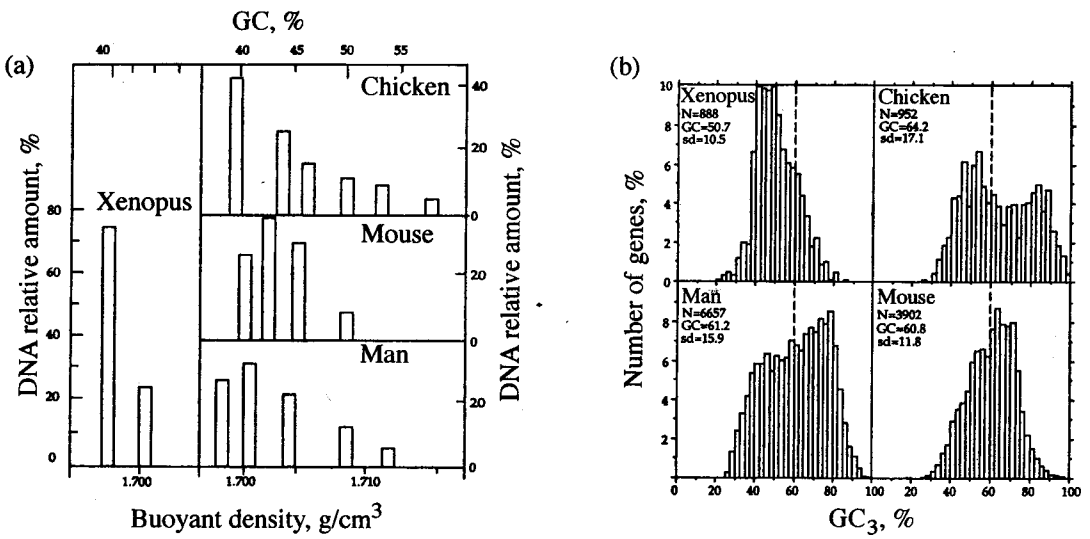

Fig. 2. (a) Compositional patterns of vertebrate genomes. Histograms showing the relative amounts, modal buoyant densities and GC levels of the major DNA components from *Xenopus*, chicken, mouse and man, as estimated after fractionation of DNA by preparative density gradient in the presence of a sequence-specific DNA ligand (Ag^+ or BAMD; BAMD is bis (acetato-mercurimethyl) dioxane). The major DNA components are the families of large DNA fragments derived from different isochore families (see Fig. 1). Satellite and minor DNA components (such as rDNA) are not shown in these histograms (modified from Bernardi, 1993). (b) Compositional distribution of third codon positions from vertebrate genes. The number of genes taken into account is indicated. A 2.5% GC_3 window was used. (Updated from Bernardi, 1993).

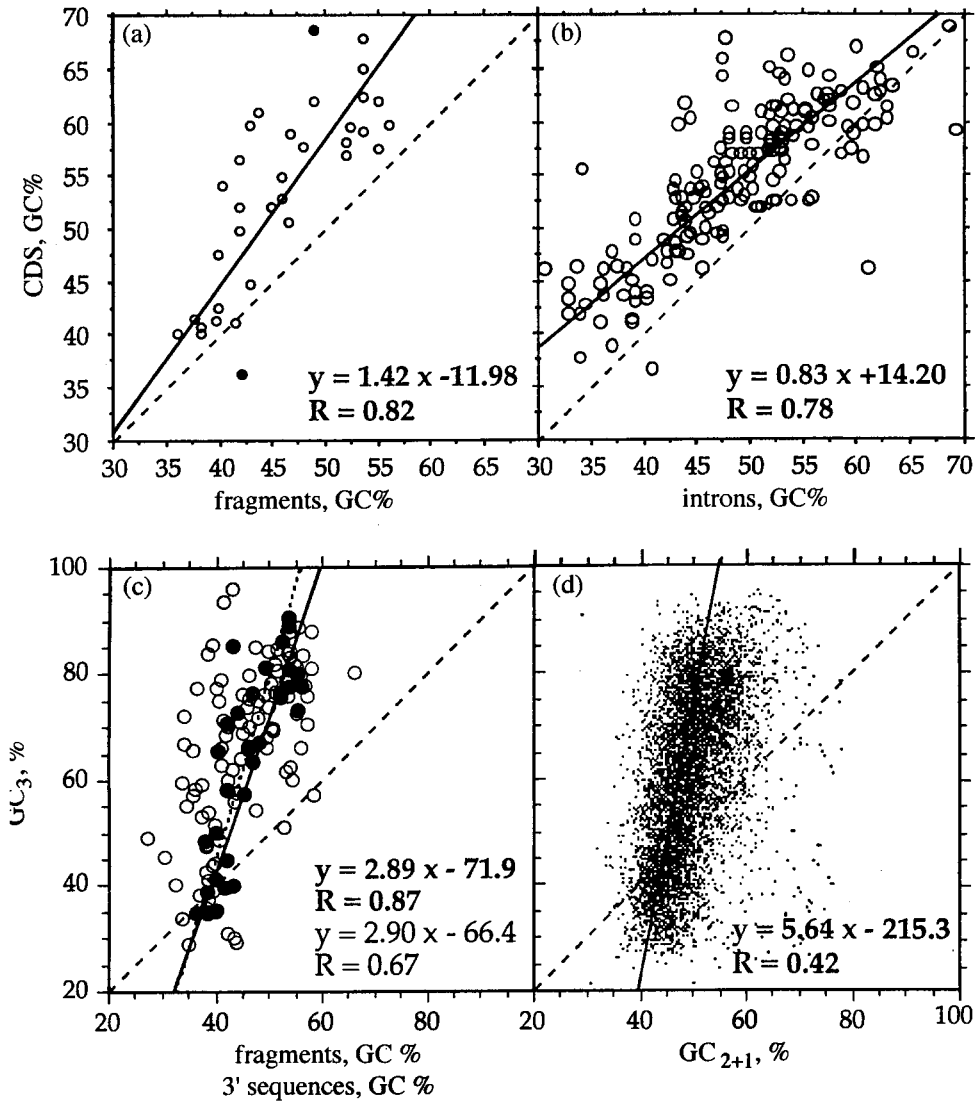

Fig. 3. (GC levels of human coding sequences (CDS) are plotted (a) against GC_3 levels of the DNA fragments in which they were experimentally localized; (b) against the GC levels of the corresponding introns. GC_3 of human coding sequences is plotted (c) against the GC levels of the DNA fragments containing the corresponding genes on the 3' flanking sequences of the latter; (d) against GC_{1+2}. In all plots, orthogonal relationships are shown along with the diagonal (slope = 1), the equations, and the correlation coefficients (modified from Clay *et al.*, 1996).

17-fold higher gene concentration compared to GC-poor isochores (Fig. 4).

The H3 isochore family has been called the *human genome core* (Bernardi, 1993), because it corresponds to the functionally most significant part of the human genome. Indeed, the H3 isochore family is not only endowed with the highest gene (and CpG island) concentration, but also with an open chromatin structure, as witnessed by the accessibility to DNases, by the scarcity of histone H1, the acetylation of histones H3 and H4 and a wider nucleosome spacing (Tazi and Bird, 1991), as well as with the highest transcription and recombination levels and with the earliest replication timing (Federico, Saccone and

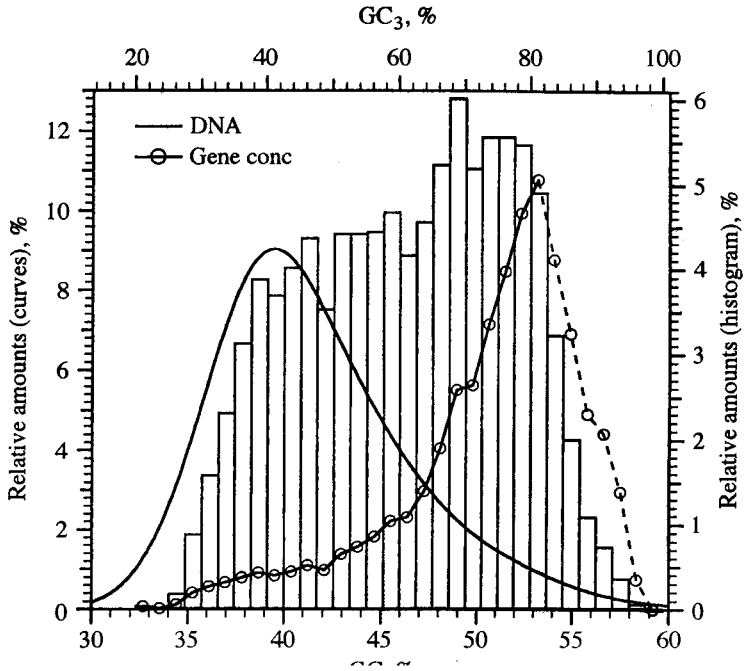

Fig. 4. Profile of gene concentration in the human genome as obtained by dividing the relative amounts of genes in each 2.5% GC_3 interval of the histogram by the corresponding relative amounts of DNA deduced from the CsCl profile. The apparent decrease in gene concentration for very high GC values (broken line) is due to the presence of rDNA in that region. The last concentration values are uncertain because they correspond to very low amounts of DNA (modified from Zoubak *et al.*, 1996).

Bernardi, in press). The genes of the genome core have the highest GC_3 levels relative to their flanking sequences (see Fig. 2a), have the shortest exons and introns (Duret, Mouchiroud, and Gautier, 1995), exhibit an extreme codon usage and encode proteins characterized by amino acid frequencies differing from those of proteins encoded by GC-poor isochores (D'Onofrio *et al.*, 1991).

The human genome core is located in about 30 $H3^+$ or T(elomeric)-bands (Saccone *et al.*, 1992; 1993; 1996), which are largely formed by GC-rich isochores of the H3 family, and about 30 T', bands which contain small amounts of H3 isochores. The remaining 140 R(everse) bands (at a 400 band resolution) comprise both GC-rich isochores (of the H1 family) and GC-poor isochores, but do not contain H3 isochores. G(iemsa) bands are formed almost exclusively by GC-poor isochores (Saccone *et al.*, 1993; see Fig. 5).

It should be stressed that the gene distribution reported for the human genome seems to have been conserved in evolution, genes showing their highest concentration in the GC-richest isochores of all vertebrates (Bernardi, 1995).

As already mentioned, the compositional pattern of the human genome, which is typical of the genomes of most mammals and similar to the genomes of birds, is strikingly different from the compositional patterns of cold-blooded vertebrates, which exhibit a much lower degree of heterogeneity and are characterized by metaphase chromosomes that do not exhibit an R banding. These different genome phenotypes of warm- *versus* cold-blooded vertebrates are due to directional compositional changes. While the gene-poor, GC-poor isochores of cold-blooded vertebrates have undergone little or no com-

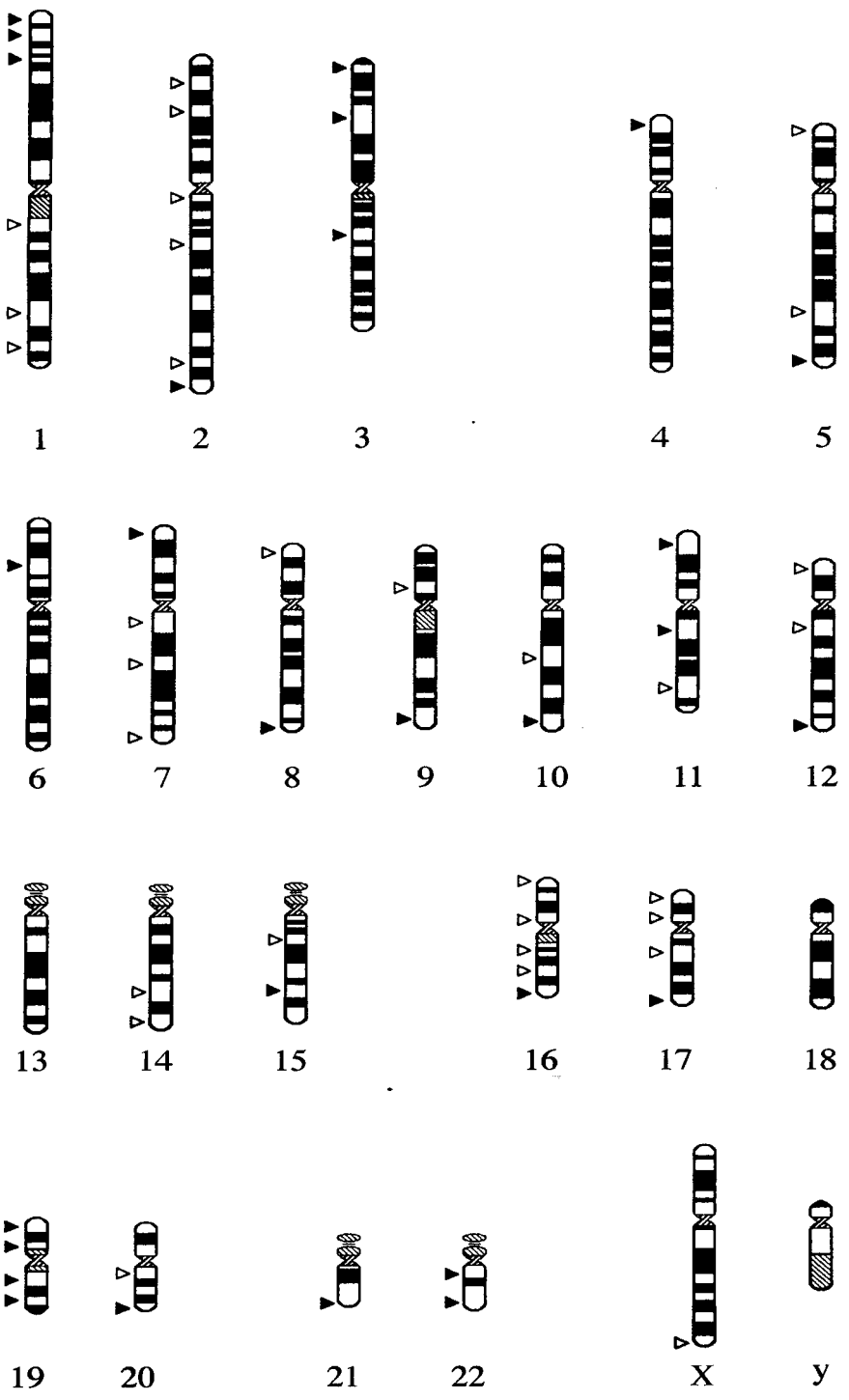

Fig. 5. Distribution of sequences hybridizing DNA from H3 isochores on human chromosomes. $H3^+$ or T bands (solid arrows) correspond to strong signals, H3* or T' bands (open arrows) to medium signals. The remaining $H3^-$ R bands correspond to undetectable signals (modified from Saccone *et al.*, 1996).

positional change in the transition to warm-blooded in vertebrates, the gene-rich, GC-rich isochores underwent dramatic compositional changes at that transition.

In the case of homologous mammalian genes, it has been possible to show that synonymous substitutions in third codon position exhibit frequencies and compositions that strongly suggest natural selection. More specifically, the level of conservation of third codon position is correlated with the degree of amino acid conservation (Cacciò *et al.*, 1995) and the synonymous substitution frequencies are not simply the result of a stochastic process in which nucleotide substitutions accumulate at random over time (Zoubak *et al.*, 1995). Moreover, the rates of synonymous substitutions in third codon position are gene specific, or, in other words, "fast" and "slow" genes in one mammal are fast and slow, respectively, in any other one (Mouchiroud *et al.*, 1995). The synonymous substitution rates are correlated to those of nonsynonymous not only for entire genes (Wolfe and Sharp, 1993; Mouchiroud *et al.*, 1995; Ohta and Ina, 1995), but also at the intragenic level, in other words, all along the coding sequences (Alvarez *et al.*, 1998).

This last point strongly supports the idea that synonymous and nonsynonymous substitution rates are under common selective constraints, since the conservation of the spatial pattern of the amino acids reflects the effect of negative selection for maintaining functionally important amino acids (Kimura, 1991).

Under these circumstances, the compositional changes in non-coding sequences, which are correlated with those occurring in third codon positions, suggest that non-coding sequences are not junk DNA, but must fulfill some functional role.

REFERENCES

Aissani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C., and Bernardi, G. 1991. The compositional properties of human genes. *J. Mol. Evol.* **32**:497-503.

Alvarez-Vallin, F., Jabbari, K., and Bernardi, G. 1998. Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. *J. Mol. Evol.* **46**:37-44.

Belozersky, A. N., and Spirin, A. S. 1958. A correlation between the composition of deoxyribonucleic and ribonucleic acids. *Nature* **182**:111-112.

Bernardi, G. 1989. The isochore organization of the human genome. *Ann. Rev. Genet.*, **23**:0637-661.

Bernardi, G. 1993. The human genome organization and its evolutionary history: a review. *Gene* **135**:57-66.

Bernardi, G. 1995. The human genome: organization and evolutionary history. *Ann. Rev. Genet.* **29**:445-476.

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

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

Bernardi, G., and Bernardi, G. 1990a. Compositional patterns in the nuclear genome of cold-blooded vertebrates. *J. Mol. Evol.* **31**:265-281.

Bernardi, G., and Bernardi, G. 1990b. Compositional transitions in the nuclear genome of cold-blooded vertebrates. *J. Mol. Evol.* **31**:282-293.

Cacciò, S., Zoubak, S., D'Onofrio, G., and Bernardi, G. 1995. Nonrandom frequency patterns of synonymous substitutions in homologous mammalian genes. *J. Mol. Evol.* **40**:280-292.

Clay, O., Cacciò, S., Zoubak, S., Mouchiroud, D., and Bernardi, G. 1996. Human coding and non-coding DNA: compositional correlations. *Mol. Phylogenet. Evol.* **5**:2-12.

D'Onofrio, G., and Bernardi, G. 1992. A universal compositional correlation among codon positions. *Gene* **110**:81-88.

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

Duret, L., Mouchiroud, D., and Gautier, C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* **40**:308-317.

Federico, C., Saccone, S., and Bernardi, G. (in press). The gene-richest bands of human chromosomes replicate at the onset of the S-phase. Cytogenet. & Cell Genet.

Macaya, G., Thiery, J. P., and Bernardi, G. 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.* **108**:237-254.

Kadi, F., Mouchiroud, D., Sabeur, G., and Bernardi, G. 1993. The compositional patterns of the avian genomes and their evolutionary implication. *I. Mol. Evol.* **37**:544-51.

Kimura, M. 1991. Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. *Proc. Natl. Acad. Sci., USA* **88**:5969-5973.

Mouchiroud, D., D'Onofrio G., Aïssani, B., Macaya, G., Gautier, C., and Bernardi, G. 1991. The distribution of genes in the human genome. *Gene* **100**:181-187.

Mouchiroud, D., Gautier, C., and Bernardi, G. 1995. Frequencies of synonymous substitutions in mammals are gene specific and correalted with the frequencies of non-synonymous substitutions. *J. Mol. Evol.* **40**:103-107.

Otha, T., and Ina, Y. 1995. Variation in synonymous substitution rate among mammalian genes and correlation among synonymous and nonsynonymous divergences. *J. Mol. Evol.* **41**:717-720.

Rolfe, R., and Meselson, M. 1959. The relative homogeneity of microbial DNA. *Proc. Natl. Acad. Sci., USA* **45**:1039-1043

Sabeur, G., Macaya, G., Kadi, F., and Bernardi, G. 1993. The isochore patterns of mammalian genomes and their phylogenetic implications. *J. Mol. Evol.* **37**:93-108.

Saccone, S., De Sario, A., Della Valle, G., and Bernardi, G. 1992. The highest gene concentrations in the human genome are in T-bands of metaphase chromosomes. *Proc. Natl. Acad. Sci., USA* **89**:4913-4917.

Saccone, S., De Sario, A., Wiegant, J., Rap, A. K., Della Valle, G., and Bernardi, G. 1993. Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci., USA* **90**:11929-11933.

Saccone, S., Cacciò, S., Kusuda, J., Andreozzi, L., and Bernardi, G. 1996. Identification of the gene-richest bands in human chromosomes. *Gene* **174**:85-94.

Sueoka, N., Marmur, J., and Doty, P. 1959. Heterogeneity in deoxyribonucleic acids. II. Dependence of the density of deoxyribonucleic acids on guanine-cytosine. *Nature* **183**:1427-1431

Tazi, J., and Bird, A. 1991. Alternative chromatin structure at CpG islands. *Cell* **60**:909-920.

Thiery, J. P., Macaya, G., and Bernardi, G. 1976. An Analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.* **108**:219-235.

Winkler, H. 1920. Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreich, Fischer, Jena.

Wolfe, K. H., and Sharp, P. M. 1993. Mammalian gene evolution: nucleotide sequence diveregence between mouse and rat. *J. Mol. Evol.* **37**:441-456.

Zoubak, S., D'Onofrio, G., Cacciò, S., Bernardi, G., and Bernardi, G. 1995. Specific compositional patterns of synonymous positions in homologous mammalian genes. *J. Mol. Evol.* **40**:293-307.

Zoubak, S., Clay, O., and Bernardi. 1996. The gene distribution of the human genome. *Gene* **174**:95-102.