

Isochores and the Regulation of Gene Expression in the Human Genome

Stilianos Arhondakis¹, Fabio Auletta², and Giorgio Bernardi^{*,2,3}

¹Bioinformatics and Medical Informatics Team, Biomedical Research Foundation of the Academy of Athens, Athens, Greece

²Stazione Zoologica Anton Dohrn, Naples, Italy

³Present address: Laboratory of Theoretical Biology and Bioinformatics, Department of Biology, Rome 3 University, Rome, Italy

*Corresponding author: E-mail: gbernardi@uniroma3.it.

Accepted: 28 February 2011

Abstract

It is well established that changes in the phenotype depend much more on changes in gene expression than on changes in protein-coding genes, and that *cis*-regulatory sequences and chromatin structure are two major factors influencing gene expression. Here, we investigated these factors at the genome-wide level by focusing on the trinucleotide patterns in the 0.1- to 25-kb regions flanking the human genes that are present in the GC-poorest L1 and GC-richest H3 isochore families, the other families exhibiting intermediate patterns. We could show 1) that the trinucleotide patterns of the 25-kb gene-flanking regions are representative of the very different patterns already reported for the whole isochores from the L1 and H3 families and, expectedly, identical in upstream and downstream locations; 2) that the patterns of the 0.1- to 0.5-kb regions in the L1 and H3 isochores are remarkably more divergent and more specific when compared with those of the 25-kb regions, as well as different in the upstream and downstream locations; and 3) that these patterns fade into the 25-kb patterns around 5kb in both upstream and downstream locations. The 25-kb findings indicate differences in nucleosome positioning and density in different isochore families, those of the 0.1- to 0.5-kb sequences indicate differences in the transcription factors that bind upstream and downstream of genes. These results indicate differences in the regulation of genes located in different isochore families, a point of functional and evolutionary relevance.

Key words: dinucleotides, nucleosome positioning, promoters, trinucleotides.

Introduction

The concept that the evolution of the phenotype depends much more on changes in regulatory sequences than on changes in protein-coding genes was first proposed a long time ago (Zuckerlandl and Pauling 1965; Britten and Davidson 1971; King and Wilson 1975) and is now well established (see Kasowski et al. 2010; and Felsenfeld and Groudine 2003; Carroll 2008 for reviews). We also know that *cis*-regulatory sequences and chromatin structure are two major factors that influence gene expression. Both act via mechanisms of protein/DNA interaction. On the protein side, transcription factors are concerned in the first case, histones in the second. On the DNA side, short nucleotide sequences are involved in both cases. Understandably, a number of investigations have recently focused on this general problem, increasingly so on a genome-wide level (for a review, see Rando and Chang 2009). Here, we

approached this issue by investigating the short-sequence patterns of 0.1- to 25-kb gene-flanking regions as present in different isochore families of the human genome.

It is well known that the human genome, a typical mammalian genome, is made up of long regions (0.2 Mb up to several Mb), the isochores, that are fairly homogeneous in base composition and belong in five families characterized by widely different GC levels (see Bernardi et al. 1985; Bernardi 2004, 2007; Costantini et al. 2006; see also the note on nomenclature in Materials and Methods). An analysis of di- and tri-nucleotide frequencies has shown large differences among the five isochore families (Costantini and Bernardi 2008). These different short-sequence frequencies or designs: 1) deviate very significantly from those expected for a random nucleotide distribution; 2) are, for a given isochore family, similar in the whole isochores and in the corresponding intergenic sequences and introns that they

contain; 3) are reflected in different frequencies of codons and encoded amino acids; 4) account for the fractionation of vertebrate DNA, as obtained via density gradient ultracentrifugation in the presence of sequence-specific ligands (Corneo et al. 1968; Filipinski et al. 1973); 5) account for the different distribution of DNA methylation, CpG doublets, and CpG islands (Varriale and Bernardi 2009); and 6) correspond to different chromatin structures as judged both at the short-sequence level (Costantini and Bernardi 2008) and at a higher scale level (Saccone et al. 2002; Di Filippo and Bernardi 2008, 2009). In brief, the short-sequence patterns are responsible for many basic properties of the human genome and show large differences in different isochore families.

The aim of the work presented here was to demonstrate the existence of differences in trinucleotide patterns of sequences that influence chromatin structure and/or are directly endowed with regulatory roles. The strategy used involved investigating the trinucleotide frequencies of 0.1- to 25-kb sequences that are immediately upstream and downstream of coding sequences located in different isochore families. Because gradual changes in the short-sequence designs were found in isochores from increasingly GC-richer families, this article will focus on the di- and tri-nucleotide patterns as found in sequences from the GC-poorest L1 and the GC-richest H3 isochore families, the other families showing intermediate patterns that will be presented in [supplementary tables](#) ([Supplementary Material](#) online).

Materials and Methods

Human genes were retrieved from GenBank, and the partial, putative, synthetic construct, predicted, not experimental, hypothetical protein, r-RNA, t-RNA, and mitochondrial genes were eliminated. Afterward, the CLEANUP program (Grillo et al. 1996) was applied in order to eliminate redundancies from the remaining nucleotide sequences. Finally, a script (see Costantini and Bernardi 2008) allowed us to identify and retain only those genes having complete coding sequences that begin with a start codon, end with a stop codon, and contain no internal stop codons, so as to calculate reliable GC values.

The entire chromosomal sequences of the finished human genome assembly were downloaded from UCSC (<http://genome.ucsc.edu>, release hg17, to allow compatibility with the data of Costantini et al. [2006] and Costantini and Bernardi [2008]) together with the positions of the human-coding sequences on those sequences. The isochore localization of genes was obtained using the human isochore map of Costantini et al. (2006). More specifically, 899 and 5,230 coding sequences were localized in L1 and L2 isochore families, respectively, whereas 7871, 7783, and 2410 were identified in the H1, H2, and H3 iso-

chores, respectively. The percentage of trinucleotides was calculated for different size segments flanking the genes, on both 5' and 3' sides using a script written by us.

The nomenclature used in this paper is as follows. GC is the molar ratio of guanine + cytosine in DNA. All dinucleotides are indicated, for brevity, as AA, AT, etc. instead of ApA, ApT, etc. This is even done for GC because this dinucleotide cannot be confused with the molar ratio because of the context. "CG-containing" trinucleotides are those that comprise CpG doublets. A/T, G/C indicate A or T, G or C, respectively.

Results

Trinucleotide frequencies in 25-kb sequences on the 5' side flanking the coding regions of human genes were expectedly found to be different in L1 and H3 isochores (see [fig. 1](#), in which trinucleotide distributions are presented in order of increasing GC) and practically identical to those already reported for the whole isochores L1 and H3 (Costantini and Bernardi 2008) as well as to those found on the 3' side (see [supplementary fig. S1](#), [Supplementary Material](#) online). As expected, complementary trinucleotides showed the same frequencies. The major differences concerned the GC-poor and GC-rich trinucleotides, the former ones being more frequent in L1, the latter ones in H3 isochores. In particular, the "A/T-only" and "G/C-only" classes, which only consist of A and/or T, or G and/or C, were the most divergent ones between L1 and H3 isochores together with the CG-containing trinucleotides that comprise the CpG doublets. Indeed, all CG-containing trinucleotides were poorly represented, much more so, however, in L1 than in H3 isochores. The transition between the two trends was observed to occur precisely when moving from two A's and/or T's to two G's and/or C's in the trinucleotides, as expected.

[Figure 2](#) compares the frequencies of all trinucleotides from the 0.5-kb sequences as present in L1 and H3 isochores on the 5' sides. A histogram for the 0.5-kb sequences on the 3' sides shows differences when compared with those for the 5' sides (see [supplementary fig. S2](#), [Supplementary Material](#) online). On both sides, these histograms are different from those exhibited by the 25-kb sequences, in being much more divergent between L1 and H3. It should be pointed out that we call 0.5-kb sequences the sequences comprised between 0.25 and 0.5kb, a choice justified by the fact that the trinucleotide frequencies are practically stable in this range (see [fig. 5](#)).

Trinucleotide frequencies in the 0- to 5-kb flanking sequences also showed major differences in the GC-poor and GC-rich trinucleotides between L1 and H3 isochores. The patterns were different from those of 25-kb sequences in being more striking, more specific (see below) and also different up to 1kb on the 5' and 3' sides (see [fig. 3](#); for the 3' results, see [supplementary fig. S3](#), [Supplementary Material](#) online, respectively). These patterns completely

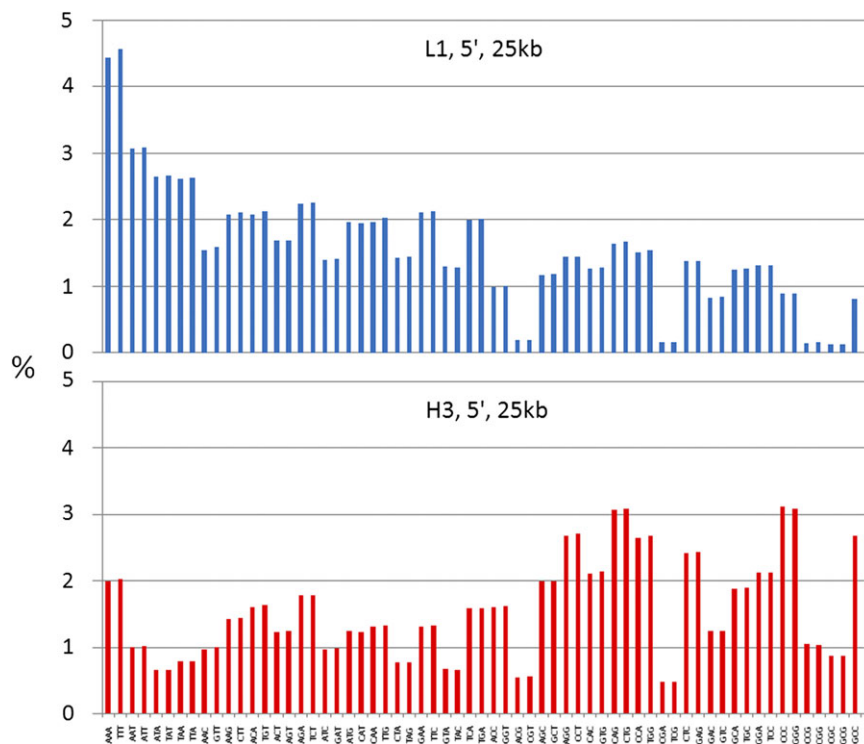


FIG. 1.—Frequencies (%) of trinucleotides in the 25-kb 5' flanking regions of genes located in L1 (blue bars) or H3 (red bars) isochore families.

faded into the 25-kb patterns around 5kb, as also indicated by the analysis of 6-, 8-, and 10-kb sequences (not shown). It should be noted that the patterns of figure 3 are not

“cumulative” patterns (0–0.1, 0–0.25, 0–0.5kb, etc.) but “individual” patterns observed in nonoverlapping segments (0–0.1, 0.1–0.25, 0.25–0.50kb, etc.).

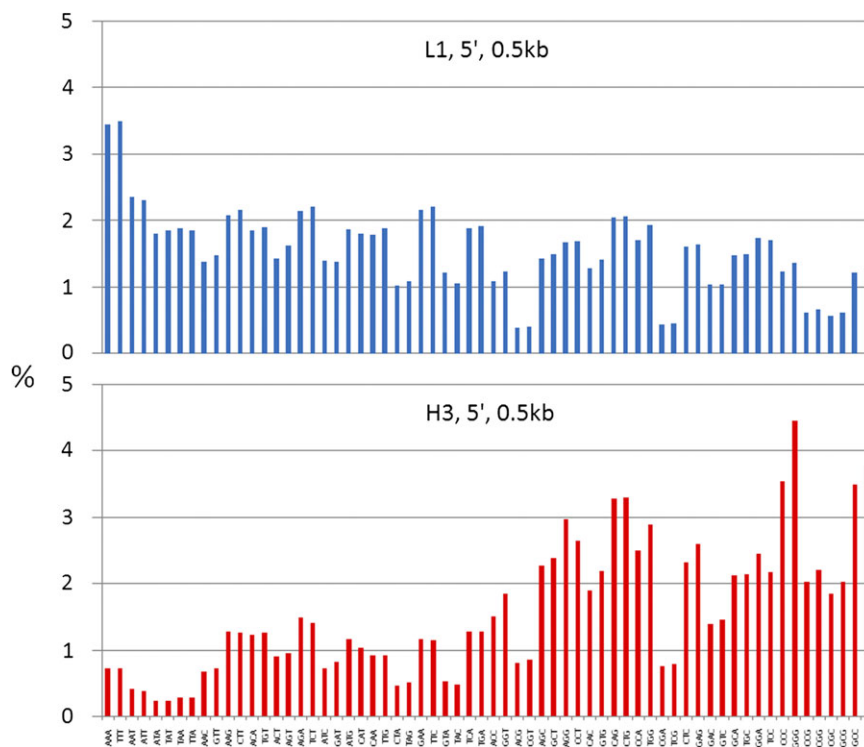


FIG. 2.—Frequencies (%) of trinucleotides in the 0.5-kb 5' flanking regions of genes located in L1 (blue bars) or H3 (red bars) isochore families.

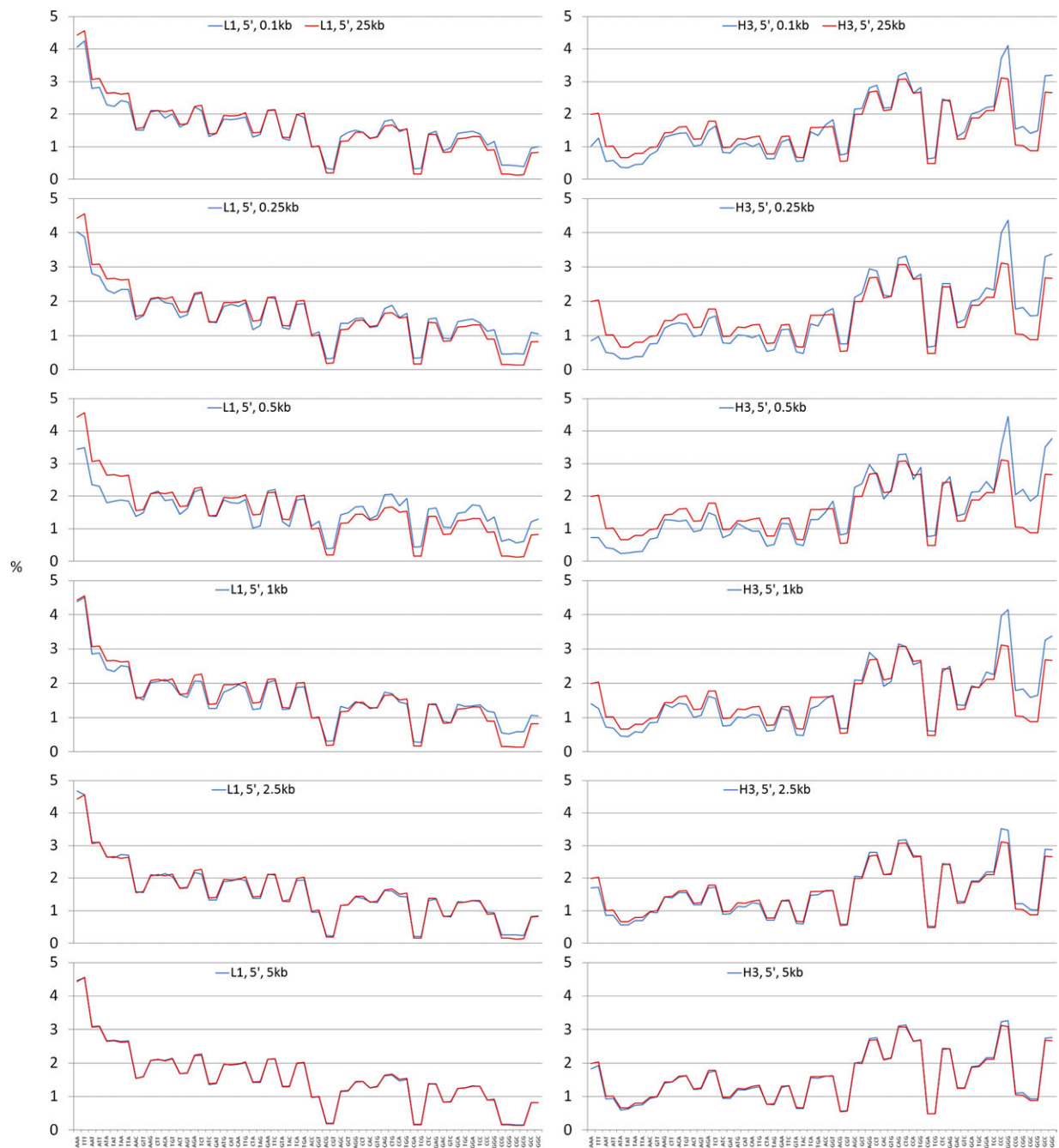


FIG. 3.—Frequencies (%) of trinucleotides of 0.1-kb to 5-kb 5' sequences (blue line) and 25-kb sequences (red line) flanking genes located in L1 (left panel) or H3 (right panel) isochore families.

Because the most striking differences in trinucleotide frequencies between L1 and H3 isochores concerned the A/T-only, the G/C-only, and the CG-containing trinucleotides, figure 4 compares these frequencies from 25-kb sequences of the L1 and H3 isochore families with those from the 0.5-kb sequences from the 5' side (for the 3' side, see [supplementary fig. S4, Supplementary Material](#) online). It should be noted that the CG-containing class of trinucleotides was split into two groups in figure 4, the first one only com-

prising C and G, the second one also A or T. This split was due to the different behavior of the two groups as seen when comparing L1 with H3 results. In the case of the 25-kb sequences from L1 isochores, the A/T-only class represents one quarter of all trinucleotides, a value which drops to less than one tenth in H3 isochores. In contrast, the G/C-only class increases from about 3.5 to 11.4%, and the extremely low CG-containing trinucleotides of L1 isochores show a remarkable increase in H3 isochores. In the case

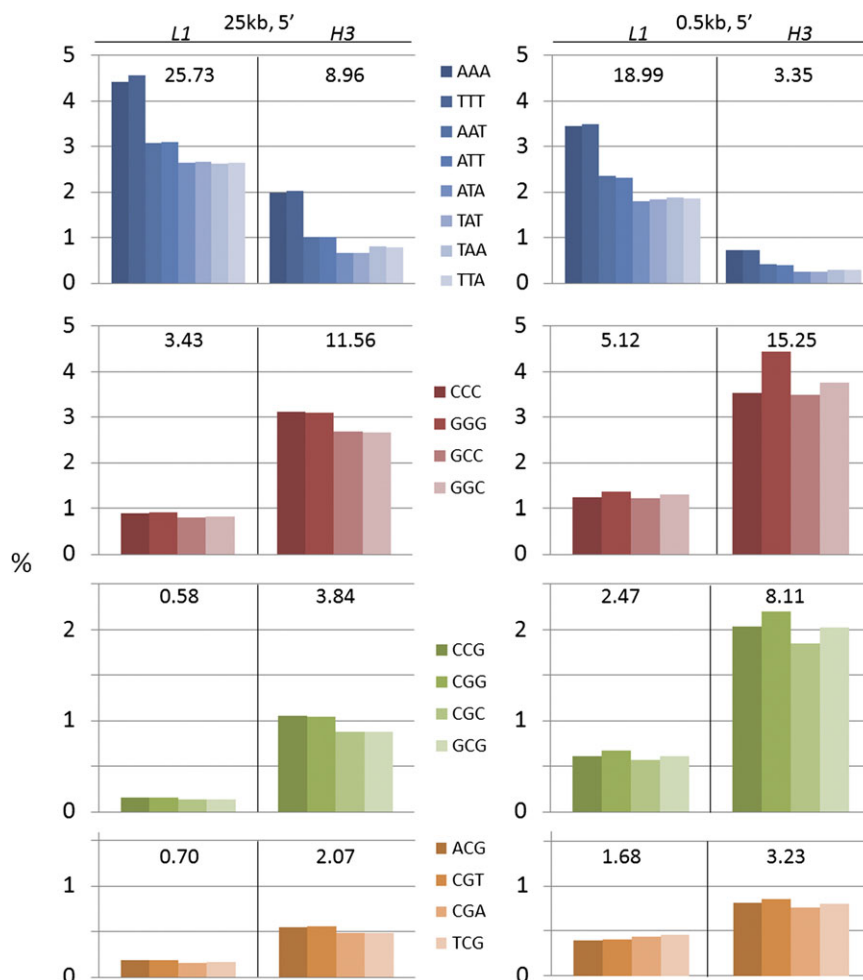


Fig. 4.—Frequencies (%) of A/T-only (blue histograms), G/C-only (red histograms), and CG-containing trinucleotides (green and brown histograms; see text) in the 5' flanking regions of genes; 25kb and 0.5kb results for L1 and H3 isochores families are shown.

of the 0.5-kb sequences (as we call the 0.25- to 0.5-kb sequences; see above), the A/T-only class is less represented in L1 isochores and even less so in H3 isochores, compared with the 25-kb sequences, whereas the opposite is true for the G/C-only class and for the CG-containing class.

Figure 5 shows the detailed results of the 0.1- to 0.5-kb sequences for the same trinucleotides of figure 4 as visualized on both 5' and 3' sides. In L1 isochores, the trinucleotide frequencies are very similar in the 5' and 3' sides, except for the CG-containing trinucleotides which exhibit on the 5' side slighter lower values for CCG, CGG, CGC, GCG and slightly higher values for ACG, CGT, CGA, TCG; in both cases, slightly decreasing trends are observed from 0.1 to 0.5kb. In H3 isochores, the trinucleotide frequencies are definitely different on the 5' and 3' sides. Moreover, an upward trend from 0.1 to 0.5kb is noticeable for the A/T-only trinucleotides, whereas a downward trend is seen in all other trinucleotide classes. Other differences concern the different values for the frequencies of individual tri-

nucleotides on the two sides and the lower values of CCC, GGG, GCC, and GGC for the 0.1-kb sequences.

To sum up the results, 1) the 25-kb sequences mimic the whole isochores results of Costantini and Bernardi (2008); 2) the 0.5-kb sequences show even more striking differences between L1 and H3 results compared with those obtained with the 25-kb sequences, such differences disappearing at about 5kb from the coding sequences; and 3) fine differences were observed when scanning the 0.1- to 0.5-kb flanking sequences.

Supplementary tables S5 and S6 (Supplementary Material online) present the trinucleotide frequencies for all sizes investigated and for all isochores families. The corresponding dinucleotide frequencies are shown in supplementary tables S7 and S8 (Supplementary Material online).

Discussion

Here, we will first take into consideration the trinucleotide patterns of the 25-kb sequences and their implications for

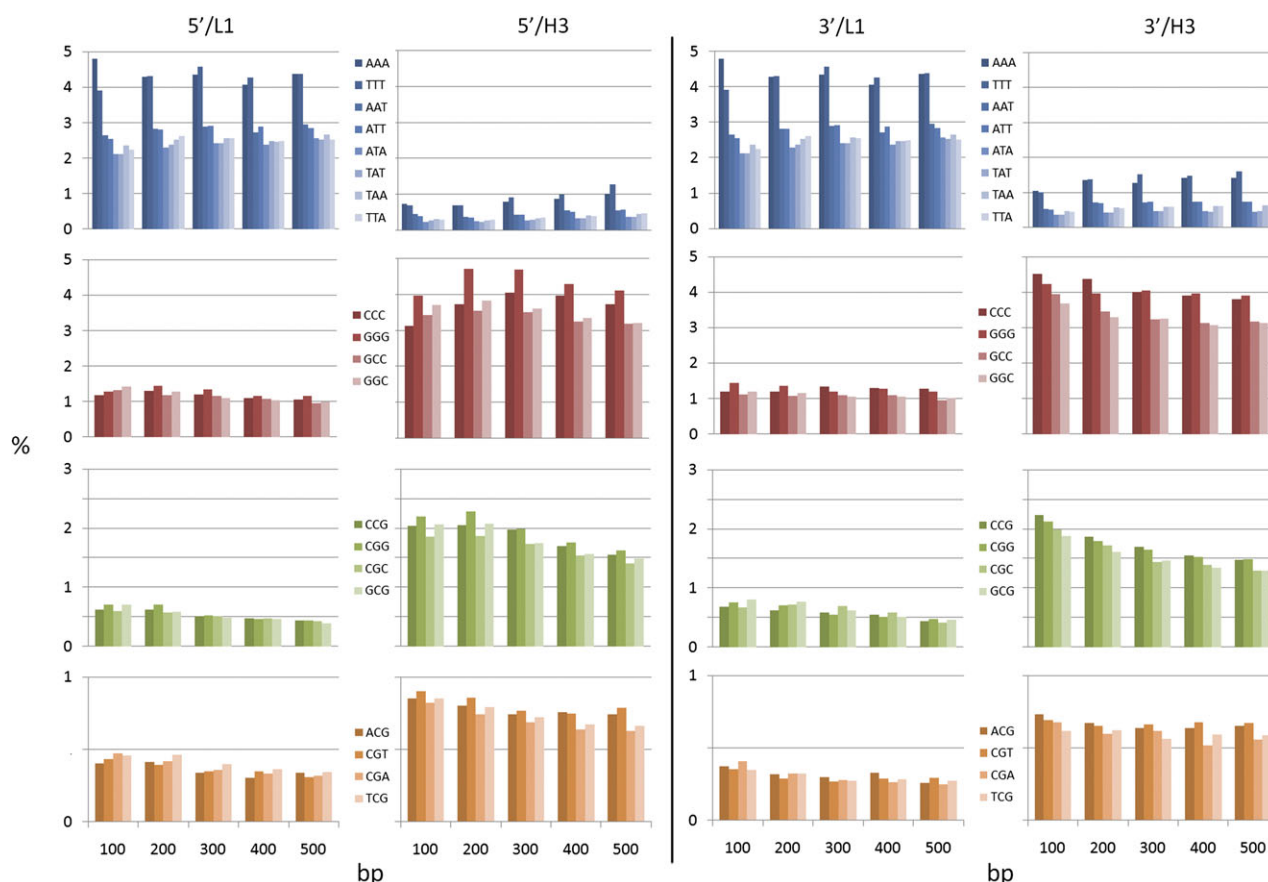


FIG. 5.—Frequencies (%) of A/T-only (blue histograms), G/C-only (red histograms), and CG-containing trinucleotides (green and brown histograms; see text) in the 5' flanking regions of genes; 0.1 to 0.5kb results for L1 and H3 isochores families are shown.

nucleosome positioning, then those of the sequences that comprise *cis*-regulatory regions and, finally, the correlation of the latter with gene function.

Chromatin Structure

Because the GC level of 25-kb sequences is 36% in L1 and 56% in H3 isochores (on both upstream and downstream sides), one would expect that the GC-rich di- and tri-nucleotides should also be higher in H3 compared with L1 isochores families. Indeed, this expectation is fulfilled. However, only a determination of the frequencies of such short sequences can give us a precise picture because the distribution of trinucleotides is nonrandom. This is stressed by the fact that some trinucleotides (CGC and GCG) may differ up to six-fold in frequency between the two isochores families versus a mere 1.5-fold for GC level. An implication of this particular point is that the frequencies of trinucleotides taken over sufficiently large regions are better predictors than GC of the isochores family to which a given sequence belongs. This consideration also applies to some extent to the short "regulatory" sequences upstream and downstream of genes,

which not only deviate from random sequences but also from the 25-kb sequences.

The results obtained for the 25-kb gene-flanking sequences are practically identical to those already reported (Costantini and Bernardi 2008) for the whole isochores of the L1 and H3 families. Those results led us to conclude that the different frequencies of trinucleotides indicate different chromatin structures in different families of isochores (see the introductory remarks). This conclusion, also based on the relation between the periodicities of AA, TT/TA, and the position and stability of nucleosomes (Segal et al. 2006), is now further strengthened by the recent demonstration (Kaplan et al. 2009; Tillo and Hughes 2009; see also Dekker 2008; Hughes and Rando 2009; Segal and Widom 2009, for reviews) that GC level has a widespread and direct influence on chromatin structure, alone explaining ~50% of the variation in nucleosome occupancy *in vitro* (Tillo and Hughes 2009). GC level may dominate because it both reduces frequencies of poly-A-like stretches and correlates with other structural characteristics of DNA.

Some additional considerations are also relevant. The great abundance of the A/T-only trinucleotides (and particularly

that of AAA and TTT) in L1 isochores strongly suggests a particular arrangement of such trinucleotides in the sequences of this isochore family. A similar consideration applies to H3 isochores for the G/C-only and CG-containing trinucleotides. Moreover, several investigations (Englander et al. 1993; Englander and Howard 1994; Bettecken and Trifonov 2009; Tanaka et al. 2010) have stressed two points, namely that Alu sequences play a special role and that CG is a key element in nucleosome positioning. Because both Alu and CG sequences are much more abundant in the GC-richest, H3, isochores, than in the GC-poorest, L1, isochores (Soriano et al. 1983; Pavliček et al. 2001; Bernardi 2004; Varriale and Bernardi 2009), they represent another relevant reason for a different nucleosome positioning in L1 and H3 isochore families and an increased density of nucleosomes in the latter. Needless to say, this increase, possibly clustered on Alu dimers, trimers, etc., may contribute to the thermodynamic stabilization of DNA in H3 isochores.

At a higher size scale, we already showed that the GC-richest and GC-poorest chromosomal regions had a very different compaction, the former corresponding to “open chromatin” and being spread over the center of the interphase nucleus, the latter to “closed chromatin” and being packed at the nuclear periphery (Saccone et al. 2002). Along the same line, we also knew that the density of DNase hypersensitive sites (and thus open chromatin) increases with the increasing GC of isochores (Di Filippo and Bernardi 2008, 2009) and that the latter are preferential regions for insertions and deletions in the human genome (Costantini and Bernardi 2009).

Regulatory Regions

As far as regulatory regions are concerned, compositional approaches have been attempted before the present work. The initiation codon AUG was found to have a different predominant context in the GC-rich isochores, gccAUGg, compared with the GC-poor isochores, aaaAUGg (Pesole et al. 1999). The base composition around transcriptional start sites (TSS's) was investigated in human and shown to be significantly different from those of coding sequences and of intergenic sequences (Aerts et al. 2004). This point is confirmed by the present work because the average GC levels of these sequences range from 41% to 63% on the 5' side and from 39% to 61% on the 3' side, whereas the GC level of 25-kb sequences is 36% in L1 and 56% in H3 isochores (on both upstream and downstream sides). The differences observed were considered to be the consequence of differences in CpG frequencies (Aerts et al. 2004), another point which is also supported by our results. In another investigation (Bajic et al. 2006), four types of compositional landscapes were distinguished in the -100 to +100bp range around TSS's: GC-rich, or GC-poor, both upstream and downstream, and GC-rich upstream/GC-poor downstream (or vice versa), the first type being

predominant and also supported by the high values in the 0.5-kb sequences that we found in H3 isochores. It is also of interest to recall that ~140-bp regions found at the beginnings and ends of genes are nucleosome free (Jiang and Pugh 2009).

At variance with most previous approaches, we took advantage of the isochore organization of the human genome by comparing short regions located upstream and downstream of genes as present in L1 and H3 isochores. We could show that the differences in trinucleotide frequencies of 0.5-kb sequences as present in L1 and H3 isochore families (on the 5' sides) are more pronounced and more specific than in the case of the 25-kb sequences. On a first approximation, this indicates 1) that regulatory sequences predominantly belong in a “TATA-box” model in L1 and in a “GC-rich” model in H3; 2) that the transcription factors bound by the GC-poorest and GC-richest isochores are different; and 3) that the genes located in those isochores may be functionally different. These points will be discussed in the following two paragraphs and in the last section.

As far as the first point is concerned, a genome-wide analysis (Carninci et al. 2006) showed that mammalian promoters can be separated into two classes, conserved TATA-box-enriched promoters (which initiate at a well-defined site) and more plastic, broad and evolvable CG-rich promoters, these different types of promoters being differentially used in different tissues and different gene families. Very high levels of “AT-only” trinucleotides were found in L1 isochores (where they represented almost one fourth of all trinucleotides), and a symmetrical situation was found in H3 isochores for the “GC-only” trinucleotides that represented (together with those containing CG doublets) almost one fourth of all trinucleotides. Both results imply particular patterns of the trinucleotides under consideration into larger sequences (6–9nt). Very interestingly, sequences of the former and the latter type were identified in promoters (Shi and Zhou 2006; Putta and Mitra 2010).

The second point is simply that the differences between the L1 and H3 motifs just outlined are so large that the genes' promoters are most probably interacting with different transcription factors. Incidentally, the compositional profiles of the short gene-flanking regions provide a new approach to investigate regulatory regions. In fact, the profiles of figure 5 suggest that the short untranslated regions likely to be represented in the 0.1-kb sequences (see Mignone et al. 2002) can be distinguished from the following regulatory sequences.

Regulatory Sequences and Gene Function

Before mentioning data that concern this last point, it should be recalled that the distribution of genes in the human genome comprises two “gene spaces” (Mouchiroud et al. 1991; Zoubak et al. 1996; see Bernardi 2004 for a review): 1) gene-rich regions, the “genome core,”

corresponding to the GC-richer isochores of the H2-H3 families that represent 15% of the genome and 2) gene-poor regions, the “genome desert,” corresponding to the isochores of the L1, L2, and H1 families, which represent 85% of the genome. Incidentally, this bimodal distribution of genes was confirmed by the initial sequencing of the human genome (Lander et al. 2001; Venter et al. 2001), by the finding of “RIDGES,” regions of increased gene expression (Caron et al. 2001; Versteeg et al. 2003), by the existence of “transcriptional forests and transcriptional deserts” (Carninci et al. 2005), and by a number of reports on gene deserts (Nobrega et al. 2003; Itoh et al. 2005; Ovcharenko and Nobrega 2005; Ovcharenko et al. 2008; see also below).

As far as the genes that are located in the GC-rich regions are concerned, it was originally proposed, already some 20 years ago (Mouchiroud et al. 1987, 1991; Bernardi 1993), that the GC-rich isochore families were richer in housekeeping genes and poorer in tissue-specific genes compared with GC-poor isochores. This point, originally based on the high frequencies of CpG islands in both GC-rich genes and housekeeping genes, was confirmed by the finding that housekeeping (ubiquitously expressed) genes were on average GC-richer than tissue-specific genes (Vinogradov 2003). Recent work (Bajic et al. 2006) defined the functional specificity (binding and protein transport activities) of GC-rich promoters as opposed to genes that are related to defense responses to the environment and that have AT-rich promoters. Finally, it was shown (D’Onofrio et al. 2007) that different functional classes of genes are characterized by different compositional properties. For example, GC-rich isochores comprise more genes involved in cellular metabolism than genes involved in information storage and processing.

On the other hand, an increasing number of results indicate that the GC-poor regions of the genome preferentially harbor genes that are active during development and are shut off by the compaction of the chromatin structure at the end of the development. More specifically, some studies (Kikuta et al. 2007; Navratilova and Becker 2009), provide support to the idea that the genes that are active during development are located in this GC-poor compartment. The evidence comes from the maintenance of long-range conserved systems across vertebrate genomes characterized by highly conserved noncoding elements and their developmental regulatory target genes. The chromosomal segments involved, the “genomic regulatory blocks,” could be identified with gene deserts (see above), and their presence was also demonstrated in insects (Engström et al. 2007). Interestingly, it was shown that during mouse brain development, most of the expression changes occur in the genes that are localized in GC-poor, LINE-rich, regions (Hiratani et al. 2004). In a similar direction, although through a different approach, Ren et al. (2007) found that the genes expressed in early developmental stages have a preference toward AT-ending codons compared with the genes expressed in later developmental stages.

Because GC3 (the GC level of third codon positions) is correlated with the GC level of the isochores in which the genes are located (Bernardi et al. 1985; Mouchiroud et al. 1991; Zoubak et al. 1996; Clay and Bernardi 2011), those genes are typically located in GC-poor isochores. Interestingly, it was proposed that immune cell-specific promoters are characterized by low GC levels (Kel et al. 1999) in keeping with those genes being localized in GC-poor isochores. Along another line, a study of the domain organization of the human chromosomes as revealed by mapping of nuclear–lamina interaction (Guelen et al. 2008) showed that genome–lamina interactions occur through more than 1,300 sharply defined large domains 0.1–10 Mb in size. These lamina-associated domains (LADs) are typified by low gene-expression levels, suggesting that they represent a repressive chromatin environment. In fact, LADs correspond in their properties to GC-poor isochores, which were previously localized at the periphery of the interphase nucleus (Saccone et al. 2002).

In conclusion, the present results indicate that the expression of genes located in different isochore families are subject to different regulations. A survey of the literature strongly suggests that different functional classes of genes are differentially distributed in isochore families. Further investigations should shed light on the possible association of isochore families (or subsets of them) with the coregulation of genes and with systems biology in general. In addition, these results open the way to an exploration of the changes that occurred in both noncoding sequences and in regulatory sequences as a result of the compositional genome transition which took place at the emergence of warm-blooded vertebrates. Finally, the link between trinucleotide frequencies (which are different in different isochore families) and nucleosome positioning and density, as well as transcription factor binding, indicates that the two major factors influencing gene expression are encoded in the genome sequence. This link adds a new important correlation to those that hold between coding sequences, on the one hand, and their extended noncoding sequences, as well as protein structure, hydrophobicity, and thermodynamic stability on the other and that constitute the genomic code (Bernardi 1990, 2004).

Supplementary Material

Supplementary figure S1–supplementary figure S4 and tables S5–S8 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

Acknowledgments

We thank Fernando Alvarez-Valin, Maria Costantini, Gary Felsenfeld, Graziano Pesole, Edward Trifonov and, especially, Oliver Clay for very relevant comments. This paper was written at Accademia Nazionale delle Scienze, Villa Torlonia, Roma 00161, Italy. Thanks are due to its

President, Gian Tommaso Scarascia-Mugnozza for his hospitality to Giorgio Bernardi. The authors declare no conflict of interest.

Literature Cited

- Aerts S, Thijs G, Dabrowski M, Moreau Y, De Moor B. 2004. Comprehensive analysis of the base composition around the transcription start site in Metazoa. *BMC Genomics*. 5:34.
- Bajic VB, et al. 2006. Mice and men: their promoter properties. *PLoS Genetics*. 2:e54.
- Bernardi G, et al. 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228:953–957.
- Bernardi G. 1990. Le génome des vertébrés: organisation, fonction et évolution. *Biofutur* 94:43–46.
- Bernardi G. 1993. The isochore organization of the human genome and its evolutionary history—a review. *Gene* 135:57–66.
- Bernardi G. 2004. Structural and evolutionary genomics. Natural selection in genome evolution. Amsterdam: Elsevier.
- Bernardi G. 2007. The neoselectionist theory of genome evolution. *Proc Natl Acad Sci U S A*. 104:8385–8390.
- Bettecken T, Trifonov EN. 2009. Repertoires of the nucleosome-positioning dinucleotides. *PLoS One*. 4:e7654.
- Britten RJ, Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Quart Rev Biol*. 46:111–138.
- Carninci P, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* 309:1559–1563.
- Carninci P, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*. 38:626–636.
- Caron H, et al. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291:1289–1292.
- Carroll B. 2008. Evo-Devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134:25–36.
- Clay OK, Bernardi G. 2011. GC3 of genes can be used as a proxy for isochore base composition: a reply to Elhaik et al. *Mol Biol Evol*. 28:21–23.
- Corneo G, Ginelli E, Soave C, Bernardi G. 1968. Isolation and characterization of mouse and guinea pig satellite deoxyribonucleic acids. *Biochemistry* 7:4373–4379.
- Costantini M, Bernardi G. 2008. The short-sequence designs of isochores from the human genome. *Proc Natl Acad Sci U S A*. 105:13971–13976.
- Costantini M, Bernardi G. 2009. Mapping insertions, deletions and SNPs in Venter's chromosomes. *PLoS One*. 4:1–11.
- Costantini M, Clay O, Auletta F, Bernardi G. 2006. An isochore map of human chromosomes. *Genome Res*. 16:536–541.
- Dekker J. 2008. Mapping in vivo chromatin interactions in yeast suggests an extended chromatin fiber with regional variation in compaction. *J Biol Chem*. 283:34532–34540.
- Di Filippo M, Bernardi G. 2008. Mapping DNase-I hypersensitive sites on human isochores. *Gene* 419:62–65.
- Di Filippo M, Bernardi G. 2009. The early apoptotic DNA fragmentation targets a small number of specific open chromatin regions. *PLoS One*. 4:e5010.
- D'Onofrio G, Ghosh TC, Saccone S. 2007. Different functional classes of genes are characterized by different compositional properties. *FEBS Lett*. 581:5819–5824.
- Englander EW, Howard BH. 1994. Nucleosome positioning by human elements in chromatin. *J Biol Chem*. 269:10091–10096.
- Englander EW, Wolffet AP, Howard BH. 1993. Nucleosome interactions with a human Alu element. *J Biol Chem*. 268:19565–19573.
- Engström PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res*. 17:1898–1908.
- Felsenfeld G, Groudine M. 2003. Controlling the double helix. *Nature* 421:448–453.
- Filipski J, Thiery JP, Bernardi G. 1973. An analysis of the bovine genome by Cs₂SO₄-Ag⁺ density gradient centrifugation. *J Mol Biol*. 80:177–197.
- Grillo G, Attimonelli M, Liuni S, Pesole G. 1996. CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. *Comp Appl Biosci*. 12:1–8.
- Guelen L, et al. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453:948–951.
- Hiratani I, Leskovaar A, Gilbert DM. 2004. Differentiation-induced replication-timing changes are restricted to AT-rich/long interspersed nuclear element (LINE)-rich isochores. *Proc Natl Acad Sci U S A*. 101:16861–16866.
- Hughes A, Rando OJ. 2009. Chromatin “programming” by sequence—is there more to the nucleosome code than %GC? *J Biol*. 8:96.
- Itoh T, Toyoda A, Taylor TD, Sakaki Y, Hattori M. 2005. Identification of large ancient duplications associated with human gene deserts. *Nat Genet*. 37:1041–1043.
- Jiang C, Pugh F. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev*. 10:161–172.
- Kaplan N, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458:362–366.
- Kasowski M, et al. 2010. Variation in transcription factor binding among humans. *Science* 328:232–235.
- Kel A, Kel-Margoulis O, Babenko V, Wingender E. 1999. Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J Mol Biol*. 288:353–376.
- Kikuta H, et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res*. 17:545–555.
- King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–116.
- Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
- Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. *Genome Biol*. 3:reviews 0004.1–10.
- Mouchiroud D, et al. 1991. The distribution of genes in the human genome. *Gene* 100:181–187.
- Mouchiroud D, Fichant G, Bernardi G. 1987. Compositional compartmentalization and gene composition in the genome of vertebrates. *J Mol Evol*. 26:198–204.
- Navratilova P, Becker TS. 2009. Genomic regulatory blocks in vertebrates and implications in human disease. *Brief Funct Genomic Proteomic*. 8:333–342.
- Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* 302:413.
- Ovcharenko I, et al. 2008. Evolution and functional classification of vertebrate gene deserts. *Genome Res*. 15:137–145.
- Ovcharenko I, Nobrega MA. 2005. Identifying synonymous regulatory elements in vertebrate genomes. *Nucleic Acids Res*. 33 (Web Server issue): W403–W407.
- Pavliček A, et al. 2001. Similar integration but different stability of Alus and LINEs in the human genome. *Gene* 276:39–45.

- Pesole G, Bernardi G, Saccone C. 1999. Isochore specificity of AUG initiator context of human genes. *FEBS Lett.* 464:60–62.
- Putta P, Mitra CK. 2010. Conserved short sequences in promoter regions of human genome. *J Biomol Struct Dynam.* 5:599–610.
- Rando OJ, Chang HY. 2009. Genome-wide views of chromatin structure. *Annu Rev Biochem.* 78:245–271.
- Ren L, et al. 2007. Developmental stage related patterns of codon usage and genomic GC content: searching for evolutionary fingerprints with models of stem cell differentiation. *Genome Biol.* 8:R35.
- Saccone S, Federico C, Bernardi G. 2002. Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene* 300:169–178.
- Segal E, et al. 2006. A genomic code for nucleosome positioning. *Nature* 442:772–778.
- Segal E, Widom J. 2009. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet.* 10:443–456.
- Shi W, Zhou W. 2006. Frequency distribution of TATA Box and extension sequences on human promoters. *BMC Bioinformatics.* 4:S2.
- Soriano P, Meunier-Rotival M, Bernardi G. 1983. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci U S A.* 80:1816–1820.
- Tanaka Y, Yamashita R, Suzuki Y, Nakai K. 2010. Effects of Alu elements on global nucleosome positioning in the human genome. *BMC Genomics.* 11:309–318.
- Tillo D, Hughes TR. 2009. G + C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics.* 10:442.
- Varriale A, Bernardi G. 2009. Distribution of DNA methylation, CpGs, and CpG islands in human isochores. *Genomics* 95:25–28.
- Venter JC, et al. 2001. The sequence of the human genome. *Science* 291:1304–1351.
- Versteeg R, et al. 2003. The Human Transcriptome Map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* 13:1998–2004.
- Vinogradov AE. 2003. Isochores and tissue-specificity. *Nucleic Acids Res.* 31:5212–5220.
- Zoubak S, Clay O, Bernardi G. 1996. The gene distribution of the human genome. *Gene* 174:95–102.
- Zuckerandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. *Evolving genes and proteins.* New York: Academic Press. p. 97–166.

Associate editor: Bill Martin