ELSEVIER

Gene 276 (2001) 25–31

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Compositional heterogeneity within and among isochores in mammalian genomes II. Some general comments

Oliver Clay, Giorgio Bernardi*

Laboratory of Molecular Evolution, Stazione Zoologica "Anton Dohrn", Villa Comunale, 80121 Naples, Italy

Received 12 May 2001; received in revised form 23 June 2001; accepted 10 August 2001
Received by C.W. Schmid

Abstract

The presence of long-range correlations and/or mosaicism in DNA sequences results in GC fluctuations, even within individual isochores that are much larger than expected correlation-free 'random' sequences. Neglecting the presence of such fluctuations can lead to incorrect conclusions regarding relative homogeneity or isochore borders. In this commentary, we address these and other methodological issues raised by the variations in GC level within human isochores. We also discuss some recent misconceptions. © 2001 Published by Elsevier Science B.V.

Keywords: Analytical ultracentrifugation; Base composition; DNA; Homogeneity; Long-range correlations

1. Introduction

This commentary addresses implications of the observations presented in the preceding paper. We also discuss, in the light of those observations, some recent articles on compositional heterogeneity in DNA sequences, on isochores, and on the prediction or recognition of isochore boundaries.

2. Use of random, correlation-free sequences as null hypotheses in large-scale DNA analyses can lead to serious errors

In the preceding article (Clay et al., 2001), we have reviewed evidence from both recent sequence data and from CsCl analyses dating back over four decades, showing that the compositional heterogeneity in a DNA sequence, as measured by the standard deviation of the GC levels of its fixed-length subsequences, is much higher than that of a random, correlation-free sequence having the same base composition. This observation, which is far from new, raises an inevitable question, namely if, or when, correlation-free sequences can still provide an appropriate null model for the analysis of DNA sequences. It is indeed often of interest, in analyzing or reporting an observed effect, to contrast it with a null hypothesis of what one would expect by chance alone. For such purposes, some software packages for DNA sequence analysis include the option of automatically generating 'random' (usually correlation-free) sequences, or the option of 'randomizing' an existing sequence by permuting its nucleotides (thus destroying any correlations that may be present). For some tasks, such as normalizing frequencies of CpG or other dinucleotides in order to monitor their local elevations using a moving window, such correlation-free null hypotheses can be fully adequate. For other tasks, however, and especially where compositional properties of long expanses of DNA are involved, they are not.

A case in point is the inappropriate use of correlation-free sequences as a standard with which to compare GC fluctuations in DNA sequences, for example in order to determine whether a sequence is homogeneous enough to be contained within a single isochore. The dramatic inadequacy of the 'random sequence' standard in such contexts has apparently been forgotten or left unnoticed by many research groups, and is only now being rediscovered. Other disciplines, ranging from electroencephalography (Dumermuth and Molinari, 1987) to the monitoring of Ethernet traffic

Abbreviations: bp, base pairs; kb, kilobase pairs; Mb, megabase pairs; GC, molar fraction of guanine and cytosine in DNA; MAR, matrix attachment region; σ, standard deviation; $\approx$, approximately equal to; , |...|, absolute value of

* Corresponding author. Fax: +39-081-245-5807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

0378-1119/01/$ - see front matter © 2001 Published by Elsevier Science B.V.
PII: S0378-1119(01)00668-0

(Paxson and Floyd, 1995; Taqqu et al., 1995; Beran, 1994), as well as the early literature on DNA heterogeneity (see Cuny et al., 1981 and references therein), show an acute awareness of the drastic consequences of neglecting dependence when serial correlations are present, and of examples where variability can be as much as an order of magnitude higher than in the uncorrelated case. By contrast, in 3 months prior to submittal of the present article (December 2000 to February 2001), three articles (Nekrutenko and Li, 2000; Häring and Kypr, 2001; IHGSC, 2001) have arduously rediscovered or stumbled upon this fact, apparently independently of each other and of the previous literature.

Considerable confusion on this point still remains, however. It is particularly evident in an incorrect conclusion of the International Human Genome Sequencing Consortium (IHGSC, 2001, p. 877) that "the hypothesis of homogeneity could be rejected for each 300-kb window in the draft sequence", that their and other "results rule out a strict notion of isochores as compositionally homogeneous", and that, consequently, "isochores do not appear to merit the prefix 'iso'". The homogeneity test(s) used by the IHGSC (p. 877 and electronic supplement) are irrelevant, since they simply compare standard deviations or fluctuations of GC observed in DNA with those expected for a sequence of independent, identically distributed nucleotides. As we have seen, an assumption that is implicit in this test, namely that correlations are absent or negligible, is far from correct. For example, a segment flanked by GC-rich segments will be more likely to be GC-rich, and such a dependence will create much higher variation in GC levels within an isochore than one would find in a sequence of independent nucleotides (for quantitative details, see Clay, 2001). The tests reject, therefore, not homogeneity within isochores, but only the independence of nucleotides assumed by the authors. In fact, these tests amount to a redefinition of homogeneity that rejects as heterogeneous any sequence showing more variation than a random sequence. Such tests will therefore reject not only all long DNA sequences, but also any binary sequences in which long-range correlations are present and, in particular, any natural DNA (except for satellite DNAs).

3. Phase plots, inter-segment differences, standard deviations and isochores

We now discuss another recent article in more detail (Nekrutenko and Li, 2000), since it again exemplifies some common misunderstandings, but, in addition, proposes new criteria and measures of heterogeneity that are relevant to the results presented here. In the article in question, the authors propose a new heterogeneity parameter (apparently as an alternative to the standard deviation), namely the average magnitude of the difference between GC levels of successive segments of length l in an n-segment partitioning (or covering by partially overlapping windows) of a DNA sequence of interest,

$$\overline{\Delta_{\mathrm{GC}}} := \frac{1}{n-1}\sum_{i=2}^{n} |\mathrm{GC}_i - \mathrm{GC}_{i-1}| \qquad (1)$$

after appropriate normalization.

The quantity $\overline{\Delta_{\mathrm{GC}}}$ has, in fact, a nice graphical interpretation. If one plots the GC level of each segment i against the GC level of the preceding segment $i-1$, one obtains, in human and even in invertebrate DNA, a good correlation, with the points of the scatterplot lining up near the main diagonal of slope 1. Such plots are called 'time-delay plots' in physics (Ruelle, 1989), or 'phase plots', since they can show the phase of a periodic fluctuation. They can be used to detect and display correlations along DNA sequences, if non-overlapping windows are chosen (see Jabbari and Bernardi, 2000, Fig. 2 for examples). It can be easily seen that $\overline{\Delta_{\mathrm{GC}}}$ is simply the mean vertical (or horizontal) distance of the points from the main diagonal; the greater the (absolute) GC differences between successive segments, the farther the points from the diagonal, and the higher the heterogeneity.

It is, therefore, natural that Nekrutenko and Li (2000) also suggest a criterion, based on the differences $|\mathrm{GC}_i - \mathrm{GC}_{i-1}|$, for the detection of isochore boundaries. Clearly, the points representing pairs of adjacent segments within a single GC-poor isochore will cluster around one part of the diagonal, those representing an adjacent GC-rich isochore will cluster around another part of the diagonal, and the few segment pair(s) spanning a sharp boundary between two adjacent isochores will be found far from the diagonal. A natural question, posed in essence also by the authors, is how far the point(s) need to be from the diagonal, i.e. how big $|\mathrm{GC}_i - \mathrm{GC}_{i-1}|$ needs to be before it is indicative of an isochore boundary. Any such critical distance must depend on the variation in GC level that can be expected within a single isochore (i.e. depend on the fragments' GC distribution), and on some chosen significance level such as 5, 1 or 0.1%.

Unfortunately, the authors do not resolve this problem. Their first normalization of $\overline{\Delta_{\mathrm{GC}}}$ is incorrect: not only do they choose a binomial distribution, i.e. an uncorrelated DNA sequence, as the expected value for normalizing (which, as argued above, and as subsequently realized also by the authors, is inappropriate), but they neglect that the absolute difference $|\mathrm{GC}_i - \mathrm{GC}_{i-1}|$ has a standard deviation that is $2/\sqrt{\pi} \approx 1.128$ times larger than that of GC_i, i.e. than that of a simple binomial distribution or of its normal approximation. A factor $\sqrt{2}$ arises because the variance of a distribution of differences of two random variables is, in the correlation-free case considered here, the sum of the variances of their distributions; an additional factor $\sqrt{2/\pi}$ is introduced by taking absolute values (and using the approximate normality of the binomial distributions of interest; see Shiryayev, 1984, pp. 237 ff.; Bourbaki, 1969, p. 74). As a result, the authors' 400,000 simulations for different fragment lengths and overlaps (Nekrutenko and

Li, 2000, Table 1), using randomly generated, uncorrelated sequences, repeatedly give values close to 1.12–1.13 ≈ $2/\sqrt{\pi}$, instead of 1 as they expect. This apparent 'behavior' or 'baseline value' of their compositional heterogeneity index is, therefore, simply an artifact of incorrect normalization. In addition, since the authors use the binomial distribution, i.e. sequences of independent nucleotides, as their first standard for homogeneity, they find no long homogeneous regions and are obliged to search for a more appropriate standard. (A similar two-step reasoning, amounting to a rediscovery of the results of Cuny et al. (1981) summarized in the preceding paper, can be discerned also in IHGSC (2001) and in Häring and Kypr (2001), a study that has been discussed in more detail elsewhere (Clay and Bernardi, 2001).)

The second attempt by Nekrutenko and Li (2000) to compare $|\mathrm{GC}_i - \mathrm{GC}_{i-1}|$ with a control distribution, namely by assuming that intra-isochore fluctuations in human should not exceed typical intra-chromosomal fluctuations in yeast, is no less arbitrary and extrinsic, but it does lead to somewhat longer 'homogeneous' regions than when the random sequences are used as a yardstick. This partial success leads the authors rather quickly to a bold proposal, namely to a "new definition of an isochore as any genomic fragment longer or equal to 100 kb such that when it is divided into a series of overlapping 10-kb windows, no two windows can differ by >7% GC". In addition to the underestimate of intra-isochore fluctuations imposed by their yeast threshold, there is a second problem inherent in their approach: the probability of witnessing at least one unlikely event (such as four consecutive heads when tossing a coin) will always increase with the number of trials, so that working outwards from an 'isochore seed' until a given fluctuation limit is reached will make it increasingly difficult for long isochores to meet the authors' criterion. It does not come as a surprise, therefore, that their method results in many very short homogeneous fragments, and in no long ones. The short ones are eliminated by the lower bound of 100 kb imposed by the authors, and relegated to regions of DNA assumed to contain no isochores at all, while long 'isochores' remain conspicuously absent. For example, in chromosome 21, where most moving window plots immediately reveal a strikingly homogeneous, long, very GC-poor region covering more than 7 Mb (cf. also De Sario et al., 1997; Hattori et al., 2000; Oliver et al., 2001; Li, 2001), the authors' histogram shows only two 'isochores' longer than 350–370 kb, namely with lengths of ≈ 430 and ≈ 490 kb, respectively. Clearly, such a redefinition of isochores cannot be considered tenable, since it could furnish, at best, only the shortest isochores in the human genome.

4. Phase plot deviations and standard deviations are largely equivalent

Interestingly, the heterogeneity measure $\overline{\Delta_{\mathrm{GC}}}$ used by Nekrutenko and Li (2000) cannot significantly outperform the traditional heterogeneity measure, namely the standard deviation σ, since the two quantities are related by the simple relation

$$\overline{\Delta_{\mathrm{GC}}} = \frac{2}{\sqrt{\pi}}\sqrt{1 - C(1)}.\,\sigma \qquad (2)$$

as can be easily shown by extending the above calculation for the binomial expectation to correlated, but still normally distributed, random variables. Here, $C(1)$ is the correlation between adjacent segments. For the kinds of power-law correlations that characterize DNA sequences (with inter-nucleotide correlation functions of the form $c(d) \approx al^{-2\beta}$, as discussed in the preceding article), we can derive a good approximation to Eq. (2), valid for non-overlapping segments longer than about 500 bp. Indeed, using the expression for $C(1)$ in Clay (2001), we obtain

$$\overline{\Delta_{\mathrm{GC}}} \approx \sqrt{\frac{8}{\pi}(1 - 2^{-2\beta})}.\,\sigma \qquad (3)$$

We first checked that this approximation agrees with Eq. (2), when using the exact expressions for σ and $C(1)$ given in Clay (2001), and different values of a, l and β. We then confirmed its validity directly for database sequences by comparing orthogonal regressions of $\overline{\Delta_{\mathrm{GC}}}$ vs. σ with the expectation in Eq. (3), using estimates of β from standard deviation plots for the genome or isochore family concerned. For segment sizes of 500 bp and higher, the expectations from Eq. (3) were confirmed to a good approximation by both the numerical examples and the DNA sequences analyzed (bacteriophage T4, 169 kb; 9882 human contigs above 175 kb available in GenBank on 2/9/01, with $R \approx 0.7$–0.8 for the different isochore families). Thus, for the scales of interest in Nekrutenko and Li (2000) the two measures of heterogeneity are largely equivalent. The standard deviation σ has the advantage that it has been used and studied for over a century (Pearson, 1893).

5. Intra-isochore heterogeneity and the isochore boundary problem

The first mammalian isochore boundary to be identified, compositionally mapped and extensively studied was in the human major histocompatibility complex (MHC) locus (Fukagawa et al., 1995; Tenzen et al., 1997; and references therein). In this case, and in many other cases, boundaries between isochores are relatively easy to locate on primary DNA sequences, with an accuracy of about ±20 kb. A frequently used method to detect putative isochore boundaries is to scan sequences with a moving window of fixed length, usually about 100 kb, and to look for a visible jump in the GC level. Several regions of the human genome are, however, not easy to partition into isochores by visual inspection alone, and a rigorous rule or partitioning algorithm is necessary. Indeed, the recognition, from DNA sequence data alone, of all isochore borders is a statistically and computationally challenging, and still partly open, problem.

Two related methods to locate potential isochore boundaries are presented in two articles of this issue (Oliver et al., 2001; Li, 2001). These methods, which both derive from the original segmentation method of Bernaola-Galván et al. (1996), recursively segment a sequence into regions so that a measure of GC difference, the Jensen–Shannon distance, is maximized. For example, in the method of Oliver et al. (2001), the recursive segmentation reports first the most certain isochore boundary, then the next most certain one within the entire sequence, and so on, while in Li (2001) and Bernaola-Galván et al. (1996) the most certain isochore boundary is reported first, and subsequent decisions are based on the individual subsequences that are being segmented. In contrast to other methods that explore outwards from a selected 'isochore seed' (Nekrutenko and Li, 2000; Häring and Kypr, 2001), there is no need to fix an extrinsic or arbitrary fluctuation threshold in advance, beyond which a sequence is declared inhomogeneous. All that is needed is a criterion indicating when to stop accepting boundary reports as indicative of isochore boundaries, i.e. when to start classifying GC discontinuities as internal features of isochores rather than as isochore borders. Interestingly, such recursive segmentation methods are either equivalent or closely related to maximum likelihood methods (Li, 2001), which is attractive in view of the generality of this approach. Recursive segmentation is successful in recognizing isochore borders, even where, for simplicity, independence of nucleotides is assumed (see Li, 2001). It therefore appears that even where isochore borders are predicted under such an independence assumption, they will remain accurate in the presence of long-range correlations.

6. The need for realistic likelihood estimates

For some methods of recognizing and testing the significance/strength of isochore boundaries (see above), as well as for other tasks, it would be helpful if we could calculate or estimate the expected frequency (probability, likelihood) with which a particular DNA sequence occurs in a particular region of DNA. In other words, it would be useful to have more accurate likelihood functions than those assuming independence of nucleotides or only short-range correlations. We would also be interested in having a formula for a related probability, namely that a randomly observed or picked sequence of a given length, in a given genome or isochore family, will have a given GC content.

Among the areas of DNA analysis that might profit from such knowledge, we mention here the calculation of GC distributions of sequence segments and their asymmetries; phylogenetic and related inferences based on maximum likelihood, and assuming independence of nucleotides (Kishino and Hasegawa, 1989); tests of temporal stationarity of GC content during evolution (Preparata and Saccone, 1987; Saccone et al., 1990); significance thresholds for BLAST sequence similarities (Karlin and Altschul, 1990, 1993); and some methods of detecting significant changes in GC levels along chromosomes (see Sections 2, 3 and 5). In some of these disciplines, the absence of appropriate likelihood functions for realistic DNA sequences with long-range correlations has repeatedly prompted authors to resort to the expedient of unrealistically assuming that such correlated DNA sequences are correlation-free. In practice, some methods and tests still yield passable results when this is done. Significance levels that are valid for correlation-free sequences often need to be replaced by much less stringent levels, however, which must sometimes be determined empirically or semi-empirically (see, for examples, Preparata and Saccone, 1987; Saccone et al., 1990; Häring and Kypr, 2001), since the variances of correlated sequences are larger than those of uncorrelated sequences, and often unknown in advance. In other cases, the errors incurred cannot be rectified by such adjustments. (For examples of statistical tests in which knowledge of the variance suffices for a correction, see Beran (1994).)

In an artificial random sequence in which no correlations or strand biases are present and the nucleotides are independent, the probability that a particular subsequence of length l and GC level x (expressed as a fraction of 1) will be encountered or picked is equal to the product of the base pairs' individual probabilities, i.e. $\mu^{xl}(1-\mu)^{(1-x)l}$, where μ is the mean or expected GC level, e.g. as estimated from the entire sequence. A related, but distinct, task is to find the probability that a randomly picked subsequence of length l will have a GC level x (where we do not need to know which one of the many sequences having that GC level is the one actually picked). This latter probability is the sum of the probabilities for all possible sequences of GC level x and length l. For the artificial case of independent nucleotides, it is the binomial term $\binom{l}{xl}\mu^{xl}(1-\mu)^{(1-x)l}$.

When long-range correlations are present in a sequence, the situation is more difficult. To our knowledge, no generally valid expressions are available for the first problem mentioned here, namely of calculating the likelihood. For the second problem, namely the calculation of expected GC distributions, the empirical data presented in the preceding paper may be of use. Indeed, if the correlations along an isochore obey a power-law, the GC distribution of its subsequences will usually approach a Gaussian distribution as the subsequence length is increased. The Gaussian will have the same mean GC level as the isochore, and its standard deviation should be given by the corresponding plot (see Fig. 3 of the preceding article) or by the relation that describes it. In practice, the GC distributions become nearly Gaussian for subsequence lengths $l > 1$ kb. For shorter sequences, no approximate formulae have yet been derived for the (asymmetric) GC distributions, as a function of length l: in this case, empirical histograms, from sequence or CsCl analysis, appear to be the only available characterizations.

7. Analytical ultracentrifugation of DNA in CsCl remains a competitive method for measuring intra- and inter-sequence heterogeneity in the era of genomic sequencing

The power of the density gradient ultracentrifugation methodology is that it allows DNA sequence information to be logically inferred without seeing the DNA sequence. It is, therefore, still in the age of genomic sequencing, a rapid and powerful method for quickly characterizing the genome of an unsequenced species.

One still sees claims, which appear periodically in the literature, asserting that the organization of mammalian genomes into isochores was, or still is, only a 'hypothesis' or 'model' that 'has been proposed', the implicit suggestion being that this 'isochore hypothesis' or 'isochore model' might, until very recently, have been in need of 'statistical confirmation' via sequence analyses and/or genetic/physical maps. Such flawed presentations of isochores reflect a widespread failure to understand the reasoning by which the isochore organization of mammalian genomes was deduced, and of which the main idea is sketched in the preceding article (Figs. 3 and 4).

The article by Nekrutenko and Li (2000) presents a variant of such common misconceptions. The authors state, incorrectly (p. 1986): "Because gradient centrifugation separates DNA fragments on the basis of their overall (mean) buoyant density, this crude method does not reveal the full extent of compositional variation within a fragment", suggesting that it is only thanks to the recently available "abundance of genomic sequences" that we can now, finally, "understand the heterogeneity of nucleotide composition along" DNA sequences shorter than about 100 kb. They add (p. 1992) that "if a window size of 100 kb is used … the compositional heterogeneity [of the human genomic sequences] is grossed over and neglected", which is partly correct, but irrelevant, since fragment sizes down to the diffusion resolution limit of ≈ 2 kb have been used experimentally already for decades (see Macaya et al., 1976; Thiery et al., 1976; Cuny et al., 1981). Later in the same article, the authors again refer to "the gradient centrifugation approach, which did not examine the compositional heterogeneity within a segment" (p. 1993). It does not take much imagination to see that intra-fragment (intra-segment) heterogeneity within long fragments of DNA can be experimentally analyzed, without any need for sequence information, by examining shorter fragments of the same DNA: intra-fragment heterogeneity when the fragments are long becomes inter-fragment heterogeneity when the fragments are shorter. For details of how the original gradient centrifugation approach "examine[d] the compositional heterogeneity within" long segments by analyzing shorter fragments, thus "reveal[ing] the full extent of compositional variation within" the longer fragments in different GC classes, we refer to the articles cited above. It will be left to the reader to judge whether this inverse method of deducing detailed information about a genomic sequence, decades before the sequence became available and the quantitative results could be precisely confirmed via an independent route, deserves the designation 'crude', and whether the plots from 1981 (reproduced in Fig. 1 of the preceding article) best summarize an "inability of the gradient centrifugation to reveal the high heterogeneity of the human genome" (Nekrutenko and Li, 2000, p. 1994), or rather an ability to do so.

The only remaining point that could be contended is that, conceivably, the statistical properties of DNA sequences could somehow differ markedly among isochores of similar GC content within the same genome. This would mean that the isochores in a given isochore family, of which the pooled DNA was analyzed in Macaya et al. (1976) and Cuny et al. (1981) in order to estimate intra-isochore heterogeneity before sequences were available, would systematically differ in some important respect. Were this the case, however, the very problem of relating GC heterogeneity to fragment length and to mean GC would be ill-posed, e.g. due to the importance of other, hidden variables such as chromosomal position. Instead, all evidence to date indicates that this is not the case. The evidence includes the observation that the standard deviation vs. length plots for isochores are similar to those for the entire isochore family to which they belong, as illustrated in the preceding article (Fig. 4). It is, therefore, reasonable to assume that the statistical properties within an isochore are very similar to those within the entire isochore family, and that one can reliably estimate the former from the latter. (Technically, this assumption could be regarded as a kind of ergodic hypothesis (see Clay, 2001).) A conclusive statistical demonstration of this point, especially for fragment sizes above 50–100 kb, must however await more *bona fide* contiguous sequences of entire human chromosomes.

8. Genes contribute to the compositional heterogeneity of GC-rich isochores

It should be emphasized that the correlations discussed in the present study pertain to a very simplified, bird's eye view of genomes that we have adopted in order to quantify the large-scale, average heterogeneity of DNA within and among mammalian isochores. In particular, no mention has been made of the compositional structure of DNA at the scale of individual genes, or of the systematic and well-documented differences in GC levels between exons, introns, different codon positions, $3'/5'$ flanks, and CpG islands, all of which contribute to the heterogeneity observed in DNA. For example, in GC-rich human isochores the expected GC level of a coding region is 5–10% higher than that of its flanking intergenic DNA. The existence of such systematic differences indicates that the working hypothesis of similar compositional properties in different parts of an isochore, used here as a simplifying approximation, remains a limited one that is appropriate only when taking averages over large expanses of DNA that span genic as well as intergenic DNA. The differences listed above, and in particular the differences in GC level between genes and intergenic DNA, will contribute to the heterogeneity of an isochore, as well as to differences in GC heterogeneity between isochore families. Indeed, the GC-rich isochore families contain much higher gene densities than the GC-poor families (e.g. see Zoubak et al., 1996; Bernardi, 2000; and references therein). Similarly, matrix attachment regions (MARs), which are often recognizable by regions extending over at least 700 bp and having a GC level of 35% or less (Saitoh and Laemmli, 1994), can be expected to introduce local dips in the GC level of a GC-rich isochore. Thus, the observation that fluctuations of GC are more pronounced in GC-rich isochores may, at least in part, be a result of local compositional constraints imposed by specific features such as genes/exons, CpG islands, $3'$ flanking regions of genes, or MARs, which may set an intrinsic limit on the compositional homogeneity that can be attained. Such alternating features, some of which are associated with GC-rich and others with GC-poor stretches of DNA, can be expected to follow in close succession in the GC-rich, and gene-rich, isochores. This may help explain why such isochores are more heterogeneous than the GC-poor, gene-poor isochores, where at least some of the above features are known to be much more widely spaced. To quantify the precise extent to which gene-related and other known systematic differences in GC levels contribute to the results reported here would, however, be outside the scope of this study, which is devoted to the analysis of DNA sequences found in unannotated contig databases or in experimentally extracted DNA samples.

A final note concerns the question of whether genes and/or repetitive elements can contribute to differences in mean GC levels among isochores, i.e. to inter-isochore heterogeneity. This is the case for genes, since they are GC-richer than intergenic DNA, and since this difference is especially pronounced in GC-rich isochores, which also have the highest gene densities. It does not appear to be the case, however, for (recognizable) repetitive elements such as *Alu*s or LINEs. Indeed, the compositional contrast between sequences from different isochores is, if anything, slightly reduced by the presence of such elements: the removal of human sequences that are significantly similar to repetitive elements slightly raises the mean GC level of long sequences when it is above 50%, and slightly lowers it when it is below 50% (Pavlíček et al., 2001, Fig. 4). This can also be seen by considering that the presence of interspersed repetitive DNA in a GC distribution or CsCl profile will reduce its standard deviation, if located near the center of the main band (as is the case, for example, for *Alu* or L1 elements).

9. Inter-isochore heterogeneity and the shape of CsCl profiles

The organization of mammalian genomes into isochores leads to a natural description of GC distributions of genomic fragments, namely as partitioned into a function describing intra-isochore heterogeneity and a compositional pattern function describing inter-isochore heterogeneity: the folding of the two functions reconstitutes the full GC distribution, i.e. the CsCl profile of the total genomic DNA. The first of the two functions will be essentially Gaussian, for fragment lengths l above 1 kb, with a width proportional to $l^{-\beta}$. As is shown in the preceding article, this Gaussian will be wider (have a lower β) for GC-rich isochores than for GC-poor isochores. In some practical applications, eliminating a parameter, and replacing the general folding operation by a simple convolution, can be more valuable than retaining the GC-dependence of the width. In such cases we sometimes use, for simplicity, an effective standard deviation that is assumed constant for all GC levels in the genome of interest (Zoubak et al., 1996; Douady et al., 2000).

10. Concluding remarks

The recent availability of the draft human genome sequence and of long contigs has allowed certain results, obtained by ultracentrifugation over three decades, to be confirmed at the DNA sequence level. Apart from some conceptual confusion between homogeneity and the absence of correlations, the recent literature shows that there is now a wide interest in applying the isochore concept to sequence analyses, and in particular to the partitioning of human chromosomes into compositionally, functionally and/or structurally distinct regions above the gene level. While for some methods of assessing heterogeneity, the strong compositional fluctuations within isochores pose serious, although probably surmountable, challenges, other methods appear to effectively bypass such problems by inferring relative homogeneity directly from the chromosomes or chromosomal regions themselves.

As more sequenced regions of the human genome become annotated with high-resolution information on experimentally verified genes and other *bona fide* functional and structural features (cf. e.g. Chen et al., 1996; Tenzen et al., 1997; Hattori et al., 2000), it will become possible to align many such feature maps with the pronounced compositional changes that are characteristic of mammalian chromosomes. Analyses of this kind should make it possible to establish more precisely and with higher certainty the functionally relevant boundaries between isochores, and to understand them in more detail.

Acknowledgements

We would like to thank Gabriel Macaya, Wentian Li, José Oliver and an anonymous referee for valuable comments.

References

Beran, J., 1994. Statistics for Long-Memory Processes. Chapman and Hall/CRC, Boca Raton, FL.

Bernaola-Galván, P., Román-Roldán, R., Oliver, J., 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. Phys. Rev. E 53, 5181–5189.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.

Bourbaki, N., 1969. Intégration. Number 35 in Fascicules/Eléments de Mathématique. Hermann, Paris.

Chen, E., Zollo, M., Mazzarella, R., Ciccodicola, A., Chen, C., Zuo, L., Heiner, C., Burough, F., Repetto, M., Schlessinger, D., D'Urso, M., 1996. Long-range sequence analysis in Xq28: thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD loci. Hum. Mol. Genet. 5, 659–668.

Clay, O., 2001. Standard deviations and correlations of GC levels in DNA sequences. Gene 276, 33–38.

Clay, O., Bernardi, G., 2001. The isochores in human chromosomes 21 and 22. Biochem. Biophys. Res. Commun. 285, 855–856.

Clay, O., Carels, N., Douady, C., Macaya, G., Bernardi, G., 2001. Compositional heterogeneity within and among isochores in mammalian genomes. I. CsCl and sequence analyses. Gene 276, 15–24.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes. I. Preparation, basic properties and compositional heterogeneity. Eur. J. Biochem. 115, 227–233.

De Sario, A., Roizes, G., Allegre, N., Bernardi, G., 1997. A compositional map of the cen-q21 region of human chromosome 21. Gene 194, 107–113.

Douady, C., Carels, N., Clay, O., Catzeflis, F., Bernardi, G., 2000. Diversity and phylogenetic implications of CsCl profiles from rodent DNAs. Mol. Phylogenet. Evol. 17, 219–230.

Dumermuth, G., Molinari, L., 1987. Spectral analysis of EEG background activity. In: Gevins, A.S., Remond, A. (Eds.), Methods of Analysis of Brain Electrical and Magnetic Signals, Handbook of Electroencephalography and Clinical Neurophysiology (revised series), Vol. 1. Elsevier, Amsterdam, pp. 85–130.

Fukagawa, T., Sugaya, K., Matsumoto, K., Okumura, K., Ando, A., Inoko, H., Ikemura, T., 1995. A boundary of long-range G + C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. Genomics 25, 184–191.

Häring, D., Kypr, J., 2001. No isochores in the human chromosomes 21 and 22? Biochem. Biophys. Res. Commun. 280, 567–573.

Hattori, M., Fujiyama, A., Taylor, T., et al., 2000. The chromosome 21 mapping and sequencing consortium. The DNA sequence of human chromosome 21. Nature 405, 311–319.

IHGSC, 2001. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Jabbari, K., Bernardi, G., 2000. The distribution of genes in the *Drosophila* genome. Gene 247, 287–292.

Karlin, S., Altschul, S., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 87, 2264–2268.

Karlin, S., Altschul, S., 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. USA 90, 5873–5877.

Kishino, H., Hasegawa, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29, 170–179.

Li, W., 2001. Delineating relative homogeneous G + C domains in DNA sequences. Gene 276, 57–72.

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

Nekrutenko, A., Li, W.-H., 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. 10, 1986–1995.

Oliver, J., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276, 47–56.

Pavlíček, A., Jabbari, K., Pačes, J., Pačes, V., Hejnar, J., Bernardi, G., 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39–45.

Paxson, V., Floyd, S., 1995. Wide-area traffic: the failure of Poisson modeling. IEEE/ACM Trans. Networking 3, 226–244.

Pearson, K., 1893. Contributions to the mathematical theory of evolution. Proc. R. Soc. 54, 329–333.

Preparata, G., Saccone, C., 1987. A simple quantitative model of the molecular clock. J. Mol. Evol. 26, 7–15.

Ruelle, D., 1989. Chaotic Evolution and Strange Attractors: The Statistical Analysis of Time Series for Deterministic Nonlinear Systems. Lezioni Lincee, Cambridge University Press, Cambridge.

Saccone, C., Lanave, C., Pesole, G., Preparata, G., 1990. Influence of base composition on quantitative estimates of gene evolution. Methods Enzymol. 183, 570–583.

Saitoh, Y., Laemmli, U., 1994. Metaphase chromosome structure: bands arise from a differential folding path of the highly AT-rich scaffold. Cell 76, 609–622.

Shiryayev, A., 1984. Probability. Springer-Verlag, New York.

Taqqu, M., Teverovsky, V., Willinger, W., 1995. Estimators for long-range dependence: an empirical study. Fractals 3, 785–788.

Tenzen, T., Yamagata, T., Fukagawa, T., Sugaya, K., Ando, A., Inoko, H., Gojobori, T., Fujiyama, A., Okumura, K., Ikemura, T., 1997. Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex. Mol. Cell. Biol. 17, 4043–4050.

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

Zoubak, S., Clay, O., Bernardi, G., 1996. The gene distribution of the human genome. Gene 174, 95–102.