ELSEVIER

Gene 276 (2001) 15–24

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Compositional heterogeneity within and among isochores in mammalian genomes
I. CsCl and sequence analyses

Oliver Clay[a], Nicolas Carels[a,1], Christophe Douady[a,b,2], Gabriel Macaya[c], Giorgio Bernardi[a,*]

[a]*Laboratory of Molecular Evolution, Stazione Zoologica "Anton Dohrn", Villa Comunale, 80121 Naples, Italy*
[b]*Medical Biology Centre, Biology and Biochemistry, Queen's University, 97 Lisburn Road, Belfast BT9 7BL, UK*
[c]*CIBCM, Universidad de Costa Rica, Ciudad Universitaria Rodrigo Facio, San Pedro de Montes de Oca, San José, Costa Rica*

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

Abstract

GC level distributions of a species' nuclear genome, or of its compositional fractions, encode key information on structural and functional properties of the genome and on its evolution. They can be calculated either from absorbance profiles of the DNA in CsCl density gradients at sedimentation equilibrium, or by scanning long contigs of largely sequenced genomes. In the present study, we address the quantitative characterization of the compositional heterogeneity of genomes, as measured by the GC distributions of fixed-length fragments. Special attention is given to mammalian genomes, since their compartmentalization into isochores implies two levels of heterogeneity, intra-isochore (local) and inter-isochore (global). This partitioning is a natural one, since large-scale compositional properties vary much more among isochores than within them. Intra-isochore GC distributions become roughly Gaussian for long fragments, and their standard deviations decrease only slowly with increasing fragment length, unlike random sequences. This effect can be explained by 'long-range' correlations, often overlooked, that are present along isochores. © 2001 Published by Elsevier Science B.V.

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

1. Introduction

The importance of CsCl density gradient ultracentrifugation for genomics stems from the empirical observation in bacterial genomes, in 1959, that the proportion of the base pairs in a DNA molecule or fragment that are guanine-cytosine (GC level, GC content, GC %) is, to an excellent approximation, linearly related to the molecule's position in a CsCl density gradient at sedimentation equilibrium (Sueoka et al., 1959; Rolfe and Meselson, 1959; Marmur and Doty, 1959; Schildkraut et al., 1962). (A summary of the calculations for CsCl gradients is given in Appendix A.) Density gradient ultracentrifugation, and the methodologies to which it led, thus allowed detailed statistical studies of GC variation along mammalian chromosomes long before their sequences became available. One of the avenues along which such studies developed was that of GC heterogeneity within and among genomes of different species.

In 1966, we were interested in isolating satellite DNAs, which consist of highly repetitive sequences, from the genomes of mouse and guinea pig. Such satellite DNAs could be detected in CsCl density gradients, for example as a separate peak in the case of mouse, or as a shoulder on the CsCl profile in the case of guinea pig. The resolution of such peaks in CsCl, as distinct from the 'main band' consisting of non-satellite, non-ribosomal DNA, was however still modest. To increase resolution, we therefore tried a new approach, which had previously been used with success to isolate the AT satellite of crab DNA: a Cs_2SO_4 gradient (which alone has a lower resolving power than CsCl) in the presence of a DNA ligand, Ag^+. Although the base composition of satellite DNAs is much more

Abbreviations: bp, base pairs; kb, kilobase pairs; Mb, megabase pairs; GC, molar fraction of guanine and cytosine in DNA; MHC, major histocompatibility complex; σ, standard deviation; $\propto$, proportional to; $\approx$, approximately equal to

* Corresponding author. Fax: +39-081-245-5807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).
[1] Present address: Centro de Astrobiologia, INTA edificio S-18, Ctra de Torrejon a Ajalvir, E-28850 Torrejon de Ardoz, Spain.
[2] Present address: Department of Biochemistry and Molecular Biology, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H7.

0378-1119/01/$ - see front matter © 2001 Published by Elsevier Science B.V.
PII: S0378-1119(01)00667-9

complex in rodents than in crab, the new method again augmented the resolution, and thus allowed separation of the satellites by fractionation. To find possible reasons for the high resolution, we then checked the base compositions of the main band and satellites by three independent methods: nucleotide analysis, melting temperature and buoyant density in CsCl. It turned out that the repetitive satellite DNA, but not the main band DNA, had violated the linear relation between GC and buoyant density shown by bacterial genomes (see Appendix A). As a result, the mouse satellite was found further away from the main band in the gradient, compared to its position expected from its GC level, while the guinea pig satellite was even pulled to the opposite side of the main band. The simplest explanation for the unexpected satellite buoyant densities in CsCl was a differential binding of water on the short repeated sequences present in such DNA, while the further increase of resolving power in the Cs_2SO_4/Ag^+ gradients could be explained by a differential frequency of oligonucleotides able to bind the sequence-specific ligand Ag^+ (Corneo et al., 1968).

The success of the Cs_2S0_4/Ag^+ approach encouraged us to tackle the more complex problem of the numerous satellite DNAs present in the bovine genome. The method led not only to good resolution of four of the eight satellite DNAs in this genome, but also to the discovery of a striking and unexpected heterogeneity of the main band DNA, which was far higher than that of any bacterial DNAs, and which persisted even for high molecular weight samples (Filipski et al., 1973).

These and subsequent studies permitted the deduction that mammalian genomes are organized into isochores, regions of DNA that extend over long distances (initially estimated as $\gg 300$ kb on average), that exhibit a marked relative homogeneity in GC compared to the heterogeneity of the total DNA, and that together comprise the large majority of the genome (Macaya et al., 1976).

Furthermore, comparison of individual main band fractions obtained by the Cs_2SO_4/Ag^+ approach revealed the presence of a small number of well-defined major components, ranging from four to six in warm-blooded vertebrate taxa. Thus, simultaneous Gaussian decomposition of the fractions' CsCl profiles shows the presence of individual Gaussian components that increase from one fraction to the next, pass through a maximum and then decrease again in subsequent fractions (Macaya et al., 1976). Each of the major components corresponds to a family of isochores having similar compositional properties. Via a more laborious procedure, the DNA of each major component can also be physically isolated (Cortadas et al., 1979; Cuny et al., 1981).

The present contribution will focus on some of the results obtained in this initial period of genome analysis, and discuss them in the light of recent results from analyses of long human contigs. Such sequence-based results confirm and complement the earlier experimental results, and expose the concordance between the views of mammalian genomes to which each approach has led.

For simplicity, some of the descriptions presented here will pertain to an idealized picture of isochores. In this picture, large-scale statistical properties of base composition (mean GC, fluctuation levels, correlations) are expected to vary only little from one part of the isochore to another, or even among different isochores of the same isochore family. This condition will allow us to derive formulae and conclusions within a clearly defined theoretical framework, which is sketched in another paper in this issue (Clay, 2001). Already the example of the long, ≈ 7 Mb, GC-poor isochore in human chromosome 21 (which is analyzed in Section 2) shows that such isochores exist; moving window scans of GC levels in other chromosomes show that they are frequent (A. Pavlíček et al., unpublished data). In the shorter (<1 Mb) isochores, however, it becomes inherently difficult to verify the above condition for the properties of long segments. It remains to be seen, therefore, to what extent the formulae and conclusions derived for the idealized isochores, and representing genome-wide averages, will remain valid or relevant in all the individual isochores of the human genome.

The compositional correlations in DNA that will interest us in the present study are correlations between the GC levels of different regions of DNA on the same chromosome, which can be as small as a nucleotide each or span hundreds of kilobases, and which can be adjacent to each other or separated by long distances. In a contiguous DNA sequence that is characterized by such compositional correlations among its regions, the nucleotides are never independent (e.g. Shiryayev, 1984, p. 299). We shall call such a sequence serially correlated, or simply (auto)correlated, and say that correlations exist along the sequence. A sequence along which no correlations exist will be called correlation-free, or simply uncorrelated.

In order to relate the analysis of DNA sequences, viewed as binary sequences (GC vs. AT base pairs), to other relevant statistical scenarios that have been studied, we will make use of a mathematical equivalence between GC/AT base pairs of a DNA sequence, digits (1/0) of a binary sequence, and repeated tosses of a coin (head/tail; Bernoulli trial).

Runs of tosses of a biased coin, or the GC and AT base pairs of an artificial, correlation-free random DNA sequence generated by a computer, will follow a binomial distribution. When the coin's bias is not extreme, or the mean GC level chosen for the artificial sequence is not close to 0 or 100% GC, the frequency distribution of the fraction of heads (GC base pairs) will approach a Gaussian distribution, and will become almost indistinguishable from such a distribution for run (fragment) lengths exceeding about 50 tosses (bp). The standard deviation or heterogeneity of the frequency distribution is then given by

$$\sigma_{\text{uncorrelated}} = \sqrt{\frac{\mu(1-\mu)}{l}} \quad (1)$$

where l is the run or fragment length and μ is the probability

of a head or GC base pair, i.e. the mean GC level of the analyzed DNA divided by 100%.

In contrast to this elementary textbook scenario of correlation-free sequences, DNA sequences show, in many species and in particular in vertebrates, strong serial correlations extending over large regions. These were explicitly recognized at the sequence level soon after the first genic sequences became available (Bernardi et al., 1985; Bernardi and Bernardi, 1986), and could be graphically presented as scatterplots in which each point represented GC levels of two parts of the same gene (e.g. introns vs. exons, or 3rd positions vs. 1st and 2nd positions); hybridization experiments on compositional fractions obtained via ultracentrifugation then allowed such correlations to be extended beyond the confines of a sequenced gene, to much longer regions of DNA containing the gene (Bernardi et al., 1985; Zerial et al., 1986; see also Clay et al., 1996; Bernardi, 2000 and references therein).

A frequent form of the autocorrelations present in isochores, when one averages over local intra-genic differences in order to characterize long expanses, is that of power-law correlation functions, which essentially constitute the class of 'long-range' correlations. This fact was explicitly noticed, and confirmed using sequence data, a decade ago (Li and Kaneko, 1992; Voss, 1992; Peng et al., 1992), after it had been predicted on the basis of the DNA duplication alternating with mutation that is ubiquitous in genomes (Li, 1989, 1991). Earlier avatars of this quantitative discovery existed, however, dating back three decades and furnishing the data for a similar conclusion.

An important practical consequence of the presence of long-range correlations is that fluctuations or variations in the GC level along a natural DNA sequence are much larger than those of correlation-free sequences. In other words, the frequency distribution of the GC levels of its fragments is, for any given fragment length above about 20 bp, much broader than one would predict from Eq. (1). In mammalian genomes, the GC distributions can be as much as an order of magnitude broader, already for fragments of a single isochore.

The fact that natural DNA exhibits large compositional fluctuations, compared to artificial random sequences composed of independent nucleotides, was demonstrated more than 20 years ago by density gradient ultracentrifugation. It was noticed early even in the relatively homogeneous genomes of bacteria (Rolfe and Meselson, 1959; Sueoka et al., 1959; Sueoka, 1962; Yamagishi, 1970, 1971, 1974) and subsequently quantified in the much more heterogeneous genomes of eukaryotes (excluding satellite DNA), in particular in vertebrates and their individual isochore families (Filipski et al., 1973; Thiery et al., 1976; Macaya et al., 1976; Cuny et al., 1981; Olofsson and Bernardi, 1983). The results of these early density gradient centrifugation experiments are summarized in the top panel of Fig. 1, from a paper published by Cuny et al. (1981).

In this panel it can be seen that, for a wide range of fragment lengths, the standard deviations for DNA (dashed curves) are all high above those calculated for correlation-free sequences (solid curve at the bottom), and in some cases (long fragments of human and mouse) an order of magnitude higher. Moreover, the curves on this semi-logarithmic plot correspond roughly to straight lines on a double-logarithmic plot. In other words, the standard deviation (which we will denote by σ) shows a power-law dependence on the fragment length l: $\sigma \propto l^{-\beta}$. For the bottom curve, the power $-\beta$ is -0.5, as expected from Eq. (1), while for the less steeply descending curves characterizing natural DNA, it is usually between -0.3 and -0.15. The slow, approximately power-law decrease of σ observed among DNA fragments can be explained by 'long-range' correlations that have been observed in many genomes (Li and Kaneko, 1992; Voss, 1992; Li, 1997). (Details on the nature and implications of such correlations can be found in Beran (1994); a brief summary is given in Clay (2001) in this issue.)

The bottom panel of Fig. 1 illustrates another important

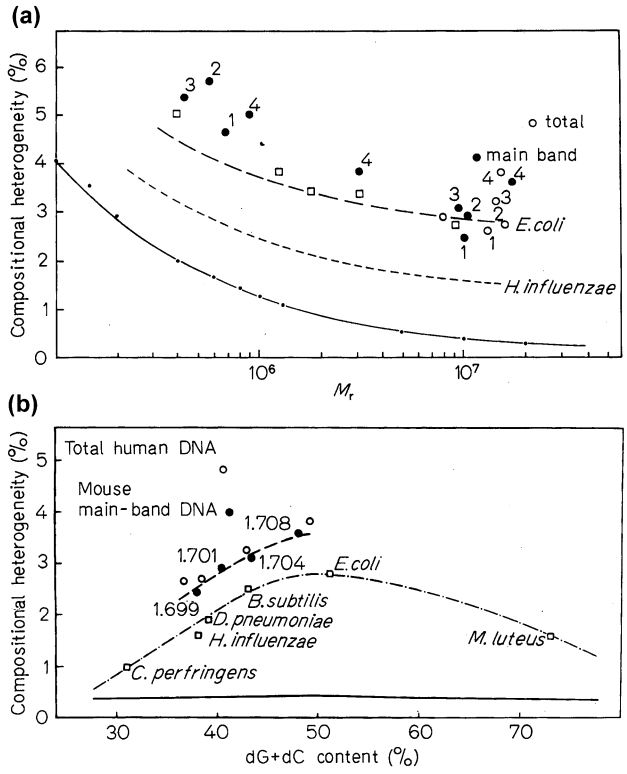

Fig. 1. Compositional heterogeneity measured by the standard deviation among the GC % of fragments in bacterial genomes (squares) and in major components, representing isochore families of human (open circles) and mouse (closed circle) genomes as obtained by CsCl gradient ultracentrifugation. Standard deviations are plotted against mean fragment length (top) and against mean GC content (bottom), and compared with the expectation if no serial correlations were present (independent and identically distributed nucleotides; bottom curve in each panel). A molecular weight M_r of 2×10^6 corresponds to about 3 kb; in the bottom panel, only higher molecular weight data ($l \approx$ 10–20 kb) are shown. From Cuny et al. (1981, Figs. 4 and 5), where further references to data sources are given. The four major components existing in both human and mouse, L1, L2, H1 and H2, are numbered (1–4; top) or denoted by their modal buoyant densities in CsCl in mouse (1.699, 1.701, 1.704 and 1.708 g/cm^3; bottom).

property of GC heterogeneity in DNA. It shows that mammalian GC-rich DNA (GC ≈ 50–55%) exhibits much higher variability than GC-poor DNA (GC ≈ 35–40%) for a given fragment length, a difference that was apparently rediscovered by Nekrutenko and Li (2000) during their recent analysis of human DNA sequence data. In the range 30–70% GC, the dependence of the heterogeneity σ on the mean GC content μ is indeed much stronger than for a random uncorrelated sequence: the solid curve, corresponding to Eq. (1) when l is held constant at ≈10 kb, appears almost flat, and remains far below the points representing actual DNA sequences of comparable lengths. As can be seen from this panel, and as is confirmed by DNA sequences (see below), the heterogeneity among 10–20 kb regions of human DNA is, on average, about 1.6 times higher in isochores with a GC of 50% than in those with a GC of 40%.

In the present paper, we will concentrate on mammalian genomes, since they exhibit, together with avian and Gramineae genomes, some of the strongest compositional (GC) heterogeneities known. In particular, we will focus on compositional heterogeneity within isochores, and on its quantitative description and measurement. Many of the conclusions that will be discussed here in the context of individual isochore families also apply to the relatively homogeneous genomes of prokaryotes or lower eukaryotes or, as a rough approximation, to the genomes of certain cold-blooded vertebrates such as fishes.

2. Results

2.1. Mammalian genomes: intra-isochore heterogeneity

Mammalian nuclear genomes are organized into isochores, long chromosomal regions that were originally estimated as >>300 kb on average and that are characterized by a strong homogeneity of base composition throughout their length, compared to the much larger heterogeneity of the entire genome. In the latter, GC levels of 10–200 kb segments (or fragments) range from around 30% GC to over 60% GC for most mammals, whereas within an isochore the GC levels of such segments rarely vary by more than 5% GC.

Chromosomes are mosaics of isochores. A consequence of this is that two long, adjacent regions of DNA will typically have similar statistical compositional properties if they are in the same isochore, but different properties if they belong to adjacent isochores. Fig. 2 illustrates this point. The two (adjacent) halves of the long, ≈7 Mb GC-poor isochore on chromosome 21 have, at any length scale, practically indistinguishable GC distributions. As a contrast, two adjacent sequences from the major histocompatibility complex (MHC) locus corresponding to different isochores are shown: the two isochores consist of compositionally distinct populations of subsequences.

In a given genome, similar mean GC levels and, as it turns out, similar heterogeneities, are exhibited by many isochores found on the same or on different chromosomes. Together, such isochores are said to belong to a single isochore family (see Section 1).

We have analyzed, using the long contig databases, the compositional heterogeneity of human DNA in each isochore family (Fig. 3) and in long isochores (Fig. 4), as measured by the standard deviation among identical-length fragments. For fragment lengths ranging from less than 100 bp to almost 100 kb, both intra-isochore and intra-isochore family heterogeneities decrease as a power function of the length l, with a power, $-\beta$, that is between about -0.15 and -0.3. Fig. 3 presents double-logarithmic plots showing this behavior, and confirming observations made two decades ago using ultracentrifugation (Cuny et al., 1981). The power $-\beta$ characterizing each of the five plots, which represent the five isochore families, is simply its slope. The slopes are clearly less steep than the slope for correlation-free sequences, which is -0.5 in accordance with Eq. (1). This fact, and the linearity of the slopes, can be explained by the long-range correlations present in the DNA, as was shown in the case of yeast in Li et al. (1998), and as can also be shown in general in the case of power-law correlated sequences (see Clay, 2001; Beran, 1994).

The plots of Fig. 3 also show that the decrease of GC heterogeneity with increasing fragment length is less steep for GC-rich isochores than for GC-poor isochores. For fragments longer than about 10 kb, intra-isochore heterogeneity is therefore much higher when the isochore is GC-rich. In agreement with this latter result, we also found that the heterogeneity in the GC of long contigs increases, over the range 30–50% GC, with increasing GC of the contig (Fig. 5).

Both results, using database sequences, are in agreement with the results from early ultracentrifugation experiments that are shown in Fig. 1. Indeed, the slope of Fig. 5 is essentially the same as that obtained experimentally in Fig. 1 for the same GC range.

2.2. Mammalian genomes: intra-isochore family heterogeneity

The concept of the isochore family, used to describe isochores having similar GC contents, was the key to the statistical characterization of the GC variations along individual isochores by ultracentrifugation, well before their sequences became available.

The GC levels along isochores belonging to an isochore family have, as it turns out, not only similar means, but also similar standard deviations and correlation functions. This is illustrated in Fig. 4 for the long GC-poor isochore of chromosome 21, where the standard deviation vs. length plots of both halves of the isochore can be seen to closely match that of the entire isochore family to which it belongs. In other

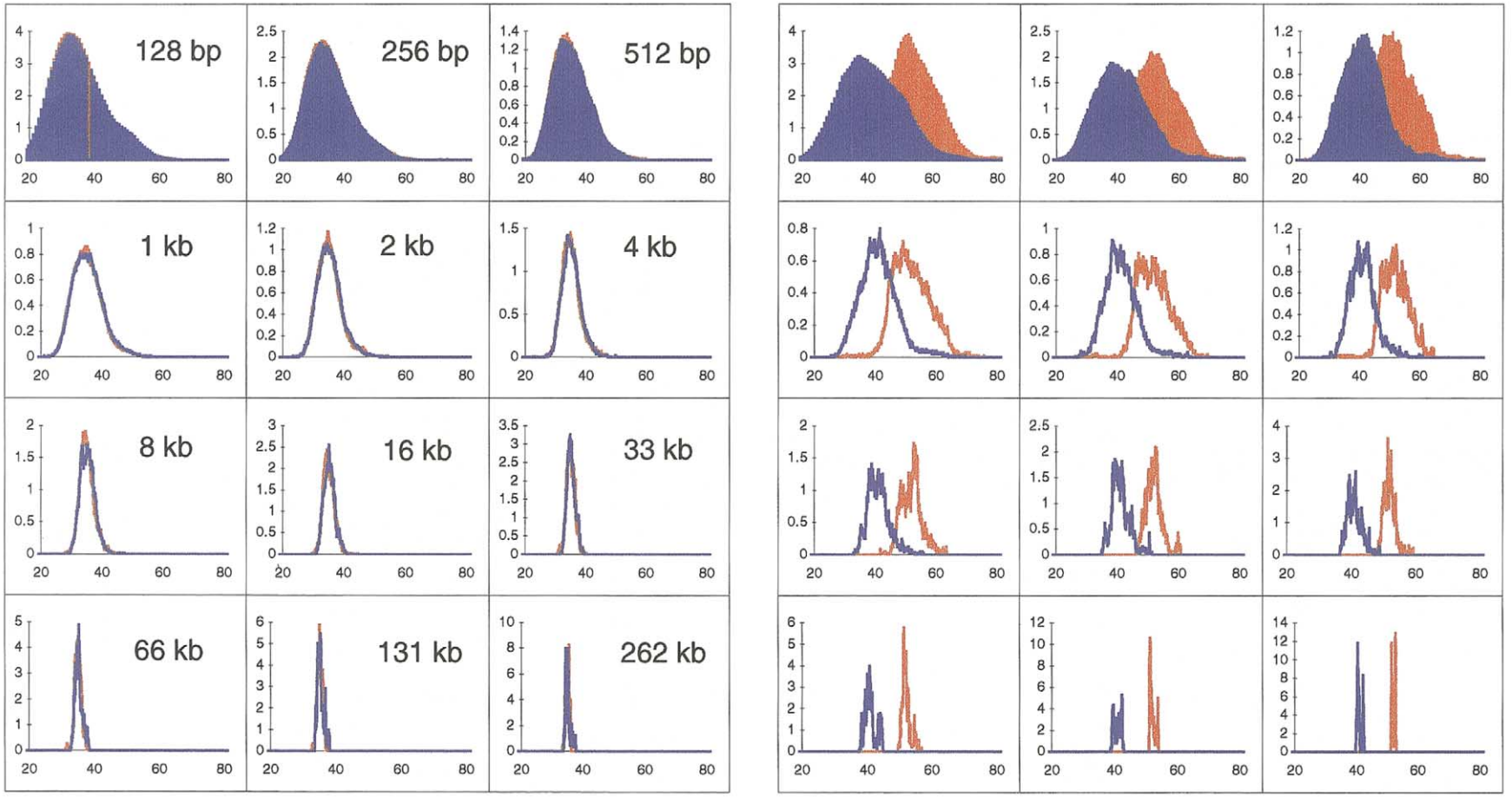

Fig. 2. GC frequency distributions (histograms) of overlapping fragments of human DNA. (Left) The two halves (blue, red) of the long, ≈ 7 Mb GC-poor isochore in chromosome 21. (Right) Two adjacent isochores from the MHC locus (blue, classical class II; red, class III). Fragment sizes are successively doubled, and are indicated (for both panels) on the left. For locations of the isochores shown, see Oliver et al. (2001) and Li (2001). For the choice of overlapping windows, here with a step size 1/128 of the window, see Clay (2001, Section 8). The bin size of the histograms is approximately 0.1% GC (the number of bins is the number of possible GC levels in a sequence of length $2^{10} = 1024$ bp, to allow smooth plots). Horizontal axes show GC % and vertical axes show relative frequencies in %.

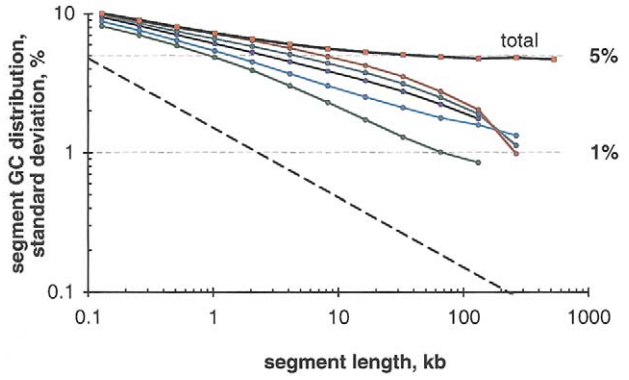

Fig. 3. Double-logarithmic plots of the standard deviation of fragment GC % vs. fragment length in long (>50 kb) contiguous database sequences from the human genome for all segments (top plot) and for five compositional fractions of the long contigs (bottom plots, decreasing GC level from top to bottom), defined by the boundaries <37%, 37–42%, 42–47%, 47–52%, and >52% GC. The fractions represent the five human isochore families L1, L2, H1, H2 and H3. The plots for the individual isochore families form straight lines, over at least three orders of magnitude of the fragment lengths ($\approx$ 100 bp to at least $\approx$ 100 kb), with slopes $-\beta$ from about -0.15 to around -0.3. Expected standard deviations in the absence of correlations (statistical independence of nucleotides, $\beta = 0.5$) are shown by the long dashed line at the bottom, which remains practically unchanged as GC levels vary between 30 and 70%. Similar results, although with no systematic differences between isochore families, were also found when the DNA sequences were represented as purine-pyrimidine (R/Y) or as amino-keto (M/K) sequences instead of as GC/AT sequences.

words, in a given mammalian genome, the mean GC level of an isochore suffices to determine reliable expectation values of its other statistical properties: knowledge of the chromosomal position of the isochore or of other details characterizing the isochore are not needed. As a result, the imaginary DNA sequence consisting of the concatenated isochores of a given family, which in reality are found distributed over all chromosomes of a mammalian genome, would be statistically very similar to a single isochore. This property allows us to often equate intra-isochore heterogeneity with intra-isochore family heterogeneity.

2.3. Mammalian genomes: quantifying inter-isochore heterogeneity

We now turn again to the plots of standard deviation vs. length shown in Fig. 3, in order to compare the heterogeneity of the total genomic DNA (top plot) with the heterogeneities of its compositional fractions (below it). From Fig. 3, it can be seen that the behavior of segments from individual isochores, or from long regions having similar mean GC levels, differs dramatically and qualitatively from the behavior when all these segments are pooled.

In the former case, the compositional classes, or their individual isochores, show approximately straight lines that descend slowly ($\beta \approx 0.15$–0.3), over approximately three orders of magnitude. In the latter case, when the segments represent the total genomic DNA, the plot initially resembles the plots for intra-isochore DNA, but begins to separate from them and level off as segment lengths exceed about 3 kb. This plot reaches a horizontal plateau for segment lengths between 70 kb and at least 500 kb. At such lengths, almost all of the compositional variation is explained by the variation among relatively homogeneous expanses of DNA extending over more than 300 kb, the isochores (Macaya et al., 1976; Cuny et al., 1981; Bernardi et al., 1985); the very small contribution of intra-isochore variation to the total variance is of the order of the variance ratio discernible from the plot ($\approx 1^2/5^2 = 4\%$). The total variance is then mostly due to the variability among isochores from different families: the heterogeneity within isochores or isochore families, although large compared to a random sequence, is very small in comparison to the entire genome.

Relative inter-isochore heterogeneity and intra-isochore homogeneity is an issue that has attracted much attention in recent months, and that will be discussed in the following paper (Clay and Bernardi, 2001). Here, we briefly remark on the non-triviality of the isochore organization of a genome.

For any binary sequence, such as a sequence of base pairs, there are only two ways to build up a substantial heterogeneity that persists at large scales: long-range correlations, and/or a mosaic structure consisting of relatively homogeneous plateaus. The result of Filipski et al. (1973), demonstrating the strong heterogeneity of the bovine genome (see Section 1), therefore implied one and/or the other of these two possibilities. The subsequent demonstration that mammalian genomes had, indeed, a mosaic organization into isochores (Macaya et al., 1976), which were in turn characterized by a much lower heterogeneity resulting from long-range correlations, is summarized in the standard deviation plots of Figs. 3 and 4: the total DNA reaches a plateau, whereas individual isochores and isochore families continue to decrease along essentially straight lines, showing not even an inflexion. Indeed, the curve and subsequent

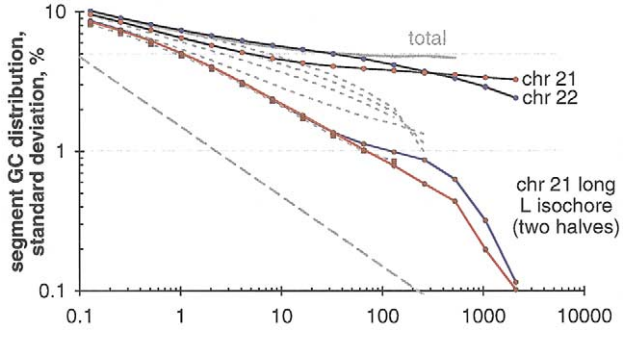

Fig. 4. Double-logarithmic plots of the standard deviation of fragment GC % vs. fragment length in the longest contiguously sequenced parts of chromosomes 21 and 22 (top) and in the two halves of the long, $\approx$ 7 Mb GC-poor isochore in chromosome 21 (bottom; see also Fig. 2). Plot details are as in Fig. 3. For comparison, all plots from Fig. 3 are shown in the background (gray). Both halves of the long isochore of chromosome 21 follow very closely the plot for the entire isochore family to which it belongs.

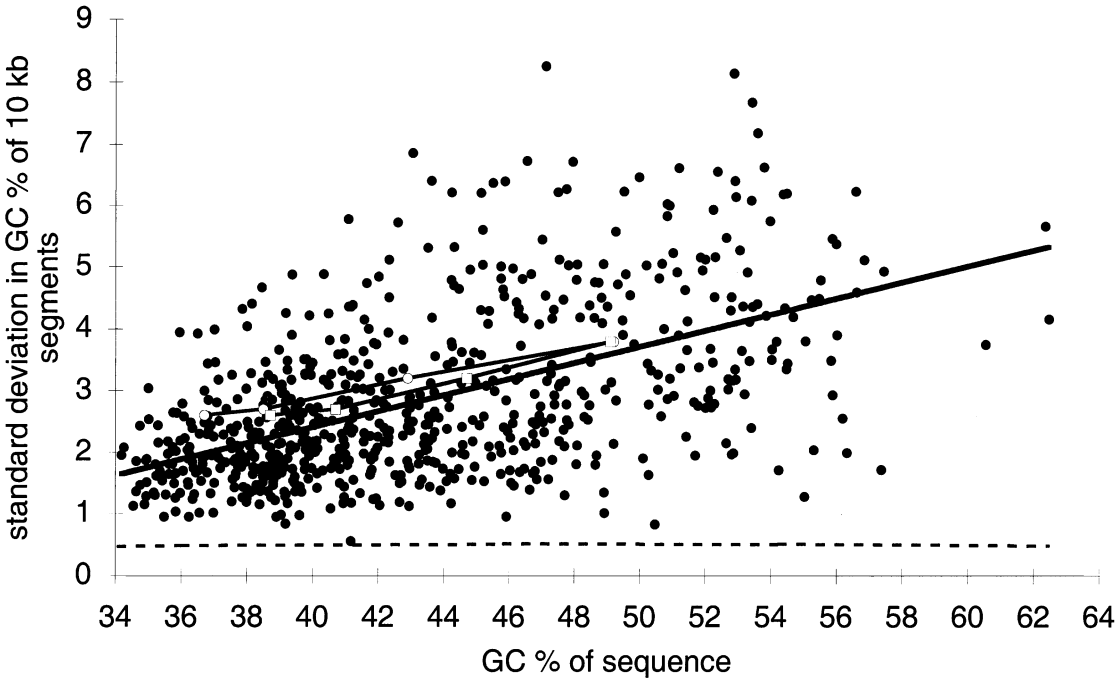

Fig. 5. Standard deviation of fragment GC % vs. mean GC % in long (>50 kb) human genomic sequences from the database. In each sequence, the standard deviation among non-overlapping segments of length 10 kb was calculated and plotted. The orthogonal regression line (long bold line; $R = 0.53$, $y = 0.13x - 2.83$) is seen to be very similar to the relation for human obtained from CsCl analysis, in the range 30–55% GC (upper and lower short bold lines correspond to the human data from Fig. 1, with mean GC levels calculated via nucleotide analysis and buoyant density in CsCl, respectively). A linear relationship is used here to characterize the human sequences, since the scarce data for GC > 50% do not allow a reliable fit to a concave curve (cf. Fig. 1, bottom). For clarity of display, data from an earlier release of GenBank are shown; recent (2/9/01) sequence data sets give denser scatterplots with very similar regression lines and R values. The dashed, barely curved line at the bottom is the expected relationship for sequences of independent nucleotides. Points near this line (lowest heterogeneities) correspond to satellite or other repetitive sequences. Points at the top of the scatterplot (highest heterogeneities) correspond to working draft 'contigs' that were apparently patched from distant regions of a chromosome, with patches separated by short runs of Ns and coinciding with some unrealistically abrupt jumps in GC.

plateau exhibited by the total human DNA sample are very similar to those that were observed by ultracentrifugation for the mouse genome 25 years ago (Macaya et al., 1976, Fig. 8).

As is shown in Fig. 4, similar behavior is exhibited by human chromosome 21, which has a simple isochore structure. In contrast, chromosome 22, which contains regions that are not easy to partition into isochores by visual inspection alone (see the compositional maps of these chromosomes in Oliver et al. (2001) and Li (2001)), does not exhibit a horizontal plateau, but only a brief inflexion after which the plot continues to descend. These comparisons already suggest that the unexpectedly turbulent changes in GC levels found in chromosome 22 are an anomaly, compared to the remaining 97–98% of the human genome, and that most of the other chromosomes will have a clearer isochore structure, as in the case of chromosome 21. This is confirmed by 100 kb moving-window scans of all chromosomes of the draft genome sequence (A. Pavlíček et al., unpublished data). The different plot for chromosome 22 in Fig. 4 also shows that the clear plateau exhibited by the total DNA, and the isochore organization to which it corresponds, is not trivial.

3. Discussion

3.1. GC distributions and CsCl profiles can reveal long-range correlations

'Long-range', or slowly decreasing, correlations in DNA often correspond to (GC-GC) correlation functions $c(d)$, or correlograms, that depend on the distance d between two base pairs along the sequence according to a power-law formula, $c(d) \propto d^{-\gamma}$. 'Short-range' correlations instead resemble an exponential dependence of the form $c(d) \propto e^{-\lambda d}$, leading to a more rapid decrease with distance. Unlike long-range correlations, they exhibit a characteristic length or scale, which is defined by the rate constant λ.

Long-range power-law correlations can be detected, for example, by Fourier transformation of DNA sequences (e.g. coded as 0 = AT, 1 = GC). They correspond to a power-law dependence of the squared Fourier amplitudes on frequency, sometimes over two or more orders of magnitude of the frequency. If the exponent of this power-law dependence is close to 1, the DNA sequence is said to be characterized by '1/*f* noise' or '1/*f*-like noise'.

The results obtained so far, and theoretical arguments

(see Clay, 2001; Beran, 1994), indicate that long-range correlations in DNA within an isochore or isochore family can be detected, and the correlations' exponents estimated, from GC distributions obtained at different fragment lengths. These frequency distributions can be obtained either by CsCl gradient ultracentrifugation of samples having different molecular weights (e.g. Macaya et al., 1976, Fig. 8) or by scanning long contiguous database sequences with moving windows of different sizes. If the sequence is power-law correlated, with slope $-\gamma$, a double-logarithmic plot of the standard deviation $\sigma(l)$ vs. the fragment length l will be a straight line over several orders of magnitude, with slope $-\beta \approx -\gamma/2$ (as for the isochores and the compositional classes of long contigs shown in Figs. 3 and 5). Perfect $1/f$ noise should correspond to a correlation that does not decrease at all when distances are increased, and to a horizontal plateau of the double-logarithmic plot of σ vs. l, while $1/f$-like noise should be recognizable by a slope of this latter plot that is close to 0 (for details of such correspondences, see Clay, 2001). By comparison, an uncorrelated sequence will show a straight line with a steep negative slope of -0.5, in accordance with the binomial relation in Eq. (1). The existence of long-range correlations in many genomes has been confirmed during the past 10 years using a variety of methods, including but not limited to Fourier analysis, so that little doubt remains as to the widespread existence of this phenomenon in DNA (Li, 1997; Li et al., 1998; and references therein).

Short-range correlated sequences differ, as their name suggests, only transiently from uncorrelated sequences. On double-logarithmic plots of the standard deviation as a function of l, short-range correlated sequences, exemplified by an exponential decrease of the correlation function, can at best briefly mimic the plots of power-law correlated sequences, and only for fragment lengths in a narrow interval. When fragment lengths are further increased the plots curve rapidly downwards into the steep slope, -0.5, that characterizes correlation-free sequences. It is, therefore, highly improbable that an exponential correlation could mimic a power-law correlation over almost three orders of magnitude, i.e. over the range that would be necessary in order to model the DNA data from the GC-richer human isochores in Figs. 3 and 4. Such a masquerade would require a low rate constant of the exponential correlation function, lying in a very narrow interval. Thus, a correlation function $c(d) \propto e^{-\lambda d}$ would require $\lambda \approx -0.0005$ to (unsatisfactorily) approximate the human sequence data of Fig. 3: slightly lower rate constants would give a horizontal plateau of the double-logarithmic plot in the region of interest, and slightly higher ones a strong curve throughout this region.

There is a caveat that should be kept in mind when interpreting the wide range over which the standard deviation vs. length plots of isochores, or of isochore families, exhibit power-law behavior. Whereas the existence and exponent of a power-law correlation can be estimated from such plots, its range can be narrower than they suggest, since the major contributions to the large standard deviations of fragment GC levels come from correlations between nucleotides at distances shorter than the fragment length (see Clay, 2001, Eq. (1)). For this reason, failure of the serial correlations to exhibit a power-law for distances above some threshold d may not be noticeable in the standard deviation vs. length plot until lengths well above d are reached. Another reason to suspect a narrower range for the correlation function's power-law behavior is that naturally occurring behaviors of this type rarely span much more than two orders of magnitude.

3.2. Isochores exhibit fluctuations much higher than those of independent, identically distributed sequences, yet can be defined as homogeneous

Subsequent studies may be necessary to clarify some of the quantitative points addressed above, but some very simple conclusions emerge from the results we have presented. In particular, as illustrated in Figs. 1, 3 and 4, the compositional heterogeneity in an isochore, when measured by the standard deviation of the GC levels of its fixed-length subsequences, is far higher than that of a random sequence having the same base composition, but consisting of independent, identically distributed (i.i.d.) nucleotides. This property does not prevent such isochores from qualifying as homogeneous, either in the relative sense (as compared to the heterogeneity among isochores, and as shown in Fig. 4), or in the original sense of homogeneity of populations as used by Pearson and Weldon in the 1890s (Pearson and Kendall, 1970, pp. 336–338). In this latter sense, when one considers populations of fixed-length segments of an isochore, homogeneity can be confirmed by the invariance, throughout the isochore, of the statistical properties that we have considered above (mean value, correlations, standard deviations, and even GC distributions of its segments of any reasonable fixed length, as shown in Fig. 2). To our knowledge, no other, more recent definition of homogeneity has since become established in the statistical literature. Thus, the original description of isochores as 'fairly homogeneous', yet at the same time endowed with fluctuations much larger than those of an i.i.d. sequence (Cuny et al., 1981), remains, apparently, at least as appropriate as when it was first proposed.

Acknowledgements

We thank José Oliver, Wentian Li, Pedro Bernaola-Galván, Pedro Carpena and Carl W. Schmid for important discussions, helpful ideas, and clarifications of issues that were relevant to this paper. We also thank David Haussler for useful critical comments. The work presented here was funded in part by a TMR grant from the European Community (ERBFRMX-CT98-0221, Network on Mammalian Phylogeny).

Appendix A. Obtaining GC distributions from CsCl absorbance profiles

The GC level of a DNA molecule or fragment can be calculated from its buoyant density ρ in a CsCl density gradient (Sueoka et al., 1959; Rolfe and Meselson, 1959; Marmur and Doty, 1959; Schildkraut et al., 1962), using the equation

$$\mathrm{GC} = \frac{\rho - 1.660\ \mathrm{g\ cm^{-3}}}{0.098} \times 100\% \quad (2)$$

The buoyant density ρ is defined as the density of the CsCl solution displaced by, or surrounding, the molecule at sedimentation equilibrium: it, in turn, is a simple function of an easily measurable quantity, namely the distance r of the molecule from the axis of an analytical ultracentrifuge at sedimentation equilibrium (Ifft et al., 1961):

$$\rho = \rho_m + \kappa\omega^2(r^2 - r_m^2) \quad (3)$$

Here, ρ_m and r_m are the buoyant density in CsCl and radial position of a suitable marker (such as bacteriophage 2c), ω is the angular speed and κ is a constant that depends on the details of the ultracentrifuge cell and the rotor ($\kappa \approx 4.2 \times 10^{-10}$ for Beckman models E and XL-A). Since the differences in radial position are very small compared to the distances from the axis, $r^2 - r^2{}_m \approx 2r_m(r - r_m)$, i.e. the non-linearity in Eq. (3) is negligible: differences in radial position are proportional to differences in GC level. In other words, for a high molecular weight DNA sample in which diffusion of the fragments can be neglected, the radial distribution of its fragments in the CsCl gradient is, after a linear calibration, identical to their GC level distribution. Only in samples consisting of very short, strongly diffusing DNA fragments ($\ll 10$ kb) (Rolfe and Meselson, 1959; Fujita, 1962; Schmid and Hearst, 1972), of highly methylated fragments (Kirk, 1967), of certain satellite DNAs (Corneo et al., 1968), or of highly concentrated (Schmid and Hearst, 1969) and/or aggregating DNA (Macaya et al., 1976, Appendix B) does this picture require modification. The absorbance profile of the DNA, or CsCl profile, is usually measured at 260 nm and 44,000 rev./min; sedimentation equilibrium is usually reached after 20–24 h.

References

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

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

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

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

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

Clay, O., Bernardi, G., 2001. Compositional heterogeneity within and among isochores in mammalian genomes. II. Some general comments, Gene, 276, 25–31.

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

Corneo, G., Ginelli, E., Soave, C., Bernardi, G., 1968. Isolation and characterization of mouse and guinea pig satellite DNAs. Biochemistry 7, 4373–4379.

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

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.

Filipski, J., Thiery, J., Bernardi, G., 1973. An analysis of the bovine genome by $Cs_2SO_4^-Ag^+$ density gradient centrifugation. J. Mol. Biol. 80, 177–197.

Fujita, H., 1962. Mathematical Theory of Sedimentation Analysis. Academic Press, New York.

Ifft, J., Voet, D., Vinograd, J., 1961. The determination of density distributions and density gradients in binary solutions at equilibrium in the ultracentrifuge. J. Phys. Chem. 65, 1138–1145.

Kirk, J.T., 1967. Effect of methylation of cytosine residues on the buoyant density of DNA in caesium chloride solution. J. Mol. Biol. 28, 171–172.

Li, W., 1989. Spatial 1/*f* spectra in open dynamical systems. Europhys. Lett. 10, 395–400.

Li, W., 1991. Expansion-modification systems: a model for spatial 1/*f* spectra. Phys. Rev. A 43, 5240–5260.

Li, W., 1997. The study of correlation structures of DNA sequences: a critical review. Comput. Chem. 21, 257–272.

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

Li, W., Kaneko, K., 1992. DNA correlations. Nature 360, 635–636.

Li, W., Stolovitzky, G., Bernaola-Galván, P., Oliver, J., 1998. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. Genome Res. 8, 916–928.

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.

Marmur, J., Doty, P., 1959. Heterogeneity in deoxyribonucleic acids. I. Dependence on composition of the configurational stability of deoxyribonucleic acids. Nature 183, 1427–1429.

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.

Olofsson, B., Bernardi, G., 1983. Organization of nucleotide sequences in the chicken genome. Eur. J. Biochem. 130, 241–245.

Pearson, E., Kendall, M., 1970. Studies in the History of Statistics and Probability, Charles Griffin, London.

Peng, C., Buldyrev, S., Goldberger, A., Havlin, S., Sciortino, F., Simons, M., Stanley, H., 1992. Long-range correlations in nucleotide sequences. Nature 356, 168–170.

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

Schildkraut, C., Marmur, J., Doty, P., 1962. Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl. J. Mol. Biol. 4, 430–443.

Schmid, C.W., Hearst, J.E., 1969. Molecular weights of homogeneous coliphage DNAs from density-gradient sedimentation equilibrium. J. Mol. Biol. 44, 143–160.

Schmid, C.W., Hearst, J.E., 1972. Sedimentation equilibrium of DNA samples heterogeneous in density. Biopolymers 11, 1913–1918.

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

Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA 48, 582–592.

Sueoka, N., Marmur, J., Doty, P., 1959. Heterogeneity in deoxyribonucleic acids. II. Dependence of the density of deoxyribonucleic acids on guanine-cytosine content. Nature 183, 1429–1433.

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

Voss, R., 1992. Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences. Phys. Rev. Lett. 68, 3805–3808.

Yamagishi, H., 1970. Nucleotide distribution in the DNA of *Escherichia coli*. J. Mol. Biol. 49, 603–608.

Yamagishi, H., 1971. Heterogeneity in nucleotide composition of *Bacillus subtilis*. J. Mol. Biol. 57, 369–371.

Yamagishi, H., 1974. Nucleotide distribution in bacterial DNAs differing in G + C content. J. Mol. Evol. 3, 239–242.

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