Published online 22 June 2010

Nucleic Acids Research, 2010, Vol. 38, No. 15 ***e158***
doi:10.1093/nar/gkq532

Identifying compositionally homogeneous and nonhomogeneous domains within the human genome using a novel segmentation algorithm

Eran Elhaik[1,*], Dan Graur[2], Krešimir Josić[3] and Giddy Landan[2]

[1]McKusick - Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, 21205, [2]Department of Biology and Biochemistry, University of Houston, Houston, TX, 77204-5001 and [3]Department of Mathematics, University of Houston, Houston, TX, 77204-3008, USA

Received April 20, 2010; Revised May 23, 2010; Accepted May 26, 2010

ABSTRACT

It has been suggested that the mammalian genome is composed mainly of long compositionally homogeneous domains. Such domains are frequently identified using recursive segmentation algorithms based on the Jensen–Shannon divergence. However, a common difficulty with such methods is deciding when to halt the recursive partitioning and what criteria to use in deciding whether a detected boundary between two segments is real or not. We demonstrate that commonly used halting criteria are intrinsically biased, and propose IsoPlotter, a parameter-free segmentation algorithm that overcomes such biases by using a simple dynamic halting criterion and tests the homogeneity of the inferred domains. IsoPlotter was compared with an alternative segmentation algorithm, D_{JS}, using two sets of simulated genomic sequences. Our results show that IsoPlotter was able to infer both long and short compositionally homogeneous domains with low GC content dispersion, whereas D_{JS} failed to identify short compositionally homogeneous domains and sequences with low compositional dispersion. By segmenting the human genome with IsoPlotter, we found that one-third of the genome is composed of compositionally nonhomogeneous domains and the remaining is a mixture of many short compositionally homogeneous domains and relatively few long ones.

INTRODUCTION

Mammalian guanine–cytosine (GC) content is known to have a complex internal compositional organization (1). For example, the human genome is known to contain long compositionally homogeneous domains whose GC contents range from ~33% to ~60%. Evidence for the non-uniformity and non-randomness of nucleotide composition was first discovered several decades ago when bulk DNA sequences that had been randomly sheared into long fragments were separated by their base composition using thermal melting and gradient centrifugation (2). The fragments were grouped into a small number of classes distinguished by their buoyant densities that correlate with GC content. These findings led Bernardi and co-workers (3–5) to propose the isochore theory for the structure of homeotherm genomes.

The isochore theory posits that mammalian genomes are mosaics of isochores: long (≥300 kb), relatively homogeneous regions, each with a typical GC content (6). The theory further posits that these regions are separated by boundaries of sharp GC content changes (7) and that all isochores can be divided into a handful of compositional domain classes or families (8).

With the advent of mammalian genomics, it became feasible to attempt to detect isochores using a segmentation algorithm with the genomic sequence as the sole input. Indeed, many methods were proposed to detect isochores by partitioning genomic sequences into compositional domains according to predefined criteria (9–14).

In a previous study, we proposed a benchmark for testing the abilities of segmentation algorithms to identify compositionally homogeneous regions and isochores within genomic sequences (15). Surprisingly, the various segmentation methods, such as sliding-window (16,17), recursive segmentation (10,12,18) and least-square segmentation (14,19), yielded inconsistent results. Recursive segmentation algorithms based on the Jensen–Shannon divergence (20), such as D_{JS} (10,11), significantly outperformed all other segmentation algorithms.

*To whom correspondence should be addressed. Tel/Fax: +1 410 502 7544; Email: eelhaik1@jhmi.edu

The authors wish it to be known that, in their opinion, the last three authors should be regarded as joint First Authors.

Recursive segmentation methods find cutting points (known also as segmentation points or partition points) that maximize the difference in base compositions between adjacent subsequences. Because there is at least one position in the sequence that maximizes the difference in base composition of any two subsequences, the recursive partitioning process can (in theory) continue until the number of segments equals the number of nucleotides (11). Therefore, a central part of all segmentation algorithms is the criterion used to halt the segmentation when the differences between adjacent segments become 'insignificant'.

Criteria that are too stringent or relaxed can lead to under- and over-segmentation, respectively (15). For instance, the halting criterion of the D_{JS} algorithm works as follows: first, the threshold is set by partitioning multiple homogeneous sequences of a certain size and composition and obtaining the maximal Jensen–Shannon divergence statistic (D_{JS}). Next, a fixed threshold for the D_{JS} entropy statistic is set *a priori* by establishing a minimum statistical significance level for the Jensen–Shannon entropy below which segmentation cannot take place (11). Finally, the D_{JS} statistic is calculated over all possible cutting points along the candidate sequence, and the maximal D_{JS} value is compared to the threshold value. The candidate sequence is partitioned in the position of the maximal D_{JS} if it is higher than the threshold. Segmentation continues recursively for segments for which D_{JS} exceeds the given threshold. Unfortunately, any choice of an *a priori* threshold affects the lengths of the inferred domains (15). For example, lowering the threshold to improve short domain inference decreases the ability to detect large domains. Although this problem has been reported previously (18,21,22), no solution has been proposed. The problem is caused by the initial choice of sequences used to determine the threshold and, therefore, cannot be solved by changing their properties.

To overcome this difficulty, we introduce IsoPlotter, an improved recursive segmentation algorithm that employs a 'dynamic threshold' that takes into account the composition and size of each segment. IsoPlotter calculates the D_{JS} statistic over all possible cutting points and compares its maximum to a dynamic threshold. In contrast to the standard D_{JS} algorithm, the threshold is not determined *a priori*, but separately for each segment to be partitioned. The length and standard deviation of GC content averaged over small windows along the segment are used to determine the dynamic threshold. If the maximum D_{JS} statistic exceeds this dynamically determined threshold, the segment is partitioned and segmentation continues recursively.

In this study, we tried to avoid some of the semantic confusion regarding isochores by simulating two sets of genomic sequences containing predetermined compositionally homogeneous domains, each separated from adjacent domains by sharp changes in GC content. These simulated domains should be considered isochores and may be used to compare the detection capability of D_{JS} (10,11) and IsoPlotter. The first set consisted of tri-domain sequences with a short central-domain flanked by two long domains. In the second set, multi-domain sequences were generated with varying standard deviations of GC content for each domain. In the last part of our study, we studied the compositional architecture of the human genome and compared the results obtained by using both IsoPlotter and D_{JS}.

METHODS

The D_{JS} and IsoPlotter segmentation algorithms

The D_{JS} is a binary, recursive segmentation algorithm that splits a DNA sequence by finding a point that maximizes the difference in GC content between adjacent subsequences. The resulting subsequences are then recursively segmented until a halting condition is satisfied (11).

Briefly, a sequence of length L, GC content F_{GC} and AT content $F_{AT} = 1 - F_{GC}$, is divided into two continuous subsequences (s = left, right) of length l_s, GC content $f_{s,GC}$, and AT content $f_{s,AT} = 1 - f_{s,GC}$. These subsequences are split at the point i that maximizes the entropic measure D_{JS} defined as the difference between the overall entropy of the sequence H_{tot} and the weighted sum of the entropies of both subsequences H_{left} and H_{right}:

$$D_{JS}(i) = H_{tot} - \left(\frac{l_{left}}{L}H_{left} + \frac{l_{right}}{L}H_{right}\right) \quad (1 < i < L) \quad (1)$$

where the entropy of the right and left subsequences is

$$H_s = -f_{s,GC}\log_2 f_{s,GC} - f_{s,AT}\log_2 f_{s,AT} \quad (0 \le f \le 1) \quad (2)$$

and the entropy of the whole sequence is

$$H_{tot} = -F_{GC}\log_2 F_{GC} - F_{AT}\log_2 F_{AT} \quad (0 \le F \le 1) \quad (3)$$

The maximal D_{JS} value is denoted by $\hat{D}_{JS} = \max D_{JS}(i)$ as the point of maximum difference between the left and right subsequences. The process of segmentation is terminated when $\hat{D}_{JS}$ is smaller than a predetermined threshold. We used a threshold of 5.8×10^{-5} (Dr. Tal Dagan, University of Düsseldorf, personal communication). Instead of comparing $\hat{D}_{JS}$ to a predetermined threshold, the IsoPlotter algorithm compares it to a dynamic threshold computed from the length and composition of the candidate subsequence.

The algorithms were implemented in Matlab 7.5 and are available at http://code.google.com/p/isoplotter/.

Simulated data analyses

To test the capabilities of IsoPlotter and D_{JS} to detect compositionally homogeneous domains, we performed two analyses. First, we simulated 1000 tri-domain sequences, each composed of two 10-Mb-long compositionally homogeneous domains at the 5′ and 3′ ends and a short, compositionally homogeneous, central-domain of length 100 kb. Domain mean GC contents were chosen from a normal distribution with a mean of 50% and a standard deviation of 5%. Negative values of the mean GC contents or those higher than 100% were unlikely

given the low standard deviation. We repeated the simulation with central-domains of 300 kb and 1 Mb in size.

In the second analysis, we simulated 1000 multi-domain sequences composed of 13 equally sized compositionally homogeneous domains, each 10 kb in length. Domain mean GC contents were chosen from a normal distribution with a mean of 50% and a standard deviation of 1%. The domain GC content standard deviation (σ_{GC}) ranged from 10^{-2} to 10^{-1} with the between-domain variation and the within-domain variation increasing accordingly.

We used simulated sequences comprising of compositionally homogeneous domains within predefined borders separated from adjacent domains by sharp changes in GC content. These simulated domains should be recognized as isochores. Domains were composed of 32 bp non-overlapping windows to reduce computation time without compromising accuracy. The GC content GC_{window} and standard deviation $\sigma_{GC\ window}$ of each window i were calculated from a uniform distribution and its full-width standard deviation as

$$GC_{window}(i) = \left(GC_{domain} - \frac{\sum \sigma_{GCwindow}}{n}\right) + \sigma_{GCwindow} \quad (4)$$

$$\sigma_{GCwindow}(i) = \sigma_{GC} \times \sqrt{12} \times R_{uniform} \quad (5)$$

where GC_{domain} and σ_{GC} are the domain GC content and standard deviation, respectively, n is the number of windows in a domain and $R_{uniform}$ is a random variable drawn from a uniform distribution between 0 and 1; for example, a domain of size 160 bp with a mean GC content of 54% and a GC content standard deviation of 5% composed of five 32-bp windows with mean GC contents of 49%, 58%, 45%, 60% and 58%. The segmentation algorithms were applied to the resulting sequences of GC frequencies.

In our analyses, we chose to work on short windows of 32 bp rather than on single nucleotides to save computation time without sacrificing accuracy (10,15). To exclude a possibility of a bias due to window size, we repeated the second analysis without using windows.

After sequences were partitioned using IsoPlotter and D_{JS} we tested the performance of the two algorithms by computing correct domain inferences, defined as the number of domains whose borders were identified within <1000 bp or a distance of <5% of their size, whichever is smaller (15). To evaluate the segmentation results, we used two statistics: sensitivity and precision. Sensitivity is the proportion of correctly inferred domains out of all predetermined domains. Precision quantifies the probability for positive prediction of the algorithm, i.e., the proportion of correctly inferred domains out of all reported domains. For example, if a sequence composed of 100 compositionally homogeneous domains was partitioned by an algorithm that inferred 50 domains but only 25 of them were correctly inferred, then the algorithm sensitivity would be 25% and its precision would be 50%. Sensitivity and precision quantify an algorithm's accuracy and prediction power, respectively. To test whether the differences between the algorithms are significant, we used the one-tailed Wilcoxon rank-sum test with $\alpha = 0.01$ (23) and the false discovery rate (FDR) correction for multiple tests (24). All reported results were obtained with a minimum domain length of 3 kb for both IsoPlotter and D_{JS}, consistent with the literature (6,8,25).

Segmentation of the human genome

The human genome assembly (build 36) was obtained from the NCBI FTP website ftp://ftp.ncbi.nlm.nih.gov/genomes/. Each chromosome was divided into non-overlapping windows of 32 bp in length, and their GC content was calculated.

The genome was partitioned by using both IsoPlotter and D_{JS} with a minimum domain length of 3 kb. Because the actual genome segmentation is unknown, the algorithms could not be evaluated for accuracy. Only their results can be compared.

In order to compare the output of the two algorithms, we divided the detected domains into nine groups. The cutoffs (l_c) that define the boundaries between the groups were chosen as 3 kb, 10 kb, 50 kb, 100 kb, 200 kb, 300 kb, 500 kb, 1 Mb and 10 Mb. For each group, the 'genome coverage' was calculated by dividing the sum of the domain lengths by the genome length. The homogeneity of the domains was assessed using a homogeneity test (see below). The domain lengths and the proportion of compositionally homogeneous domains were compared between the two algorithms using a one-tailed t-test with $\alpha = 0.05$ (23).

Homogeneity test

Inferred domains were classified into two types, compositionally homogeneous and nonhomogeneous, based on their homogeneity relative to the chromosome (Figure 1). We used the F-test to compare the GC content variance of a domain with that of the sequence on which it resides (26). The GC content for each 32-bp non-overlapping window was calculated for the domain in question and for the entire sequence. Because the F-test assumes the data are normally distributed, we followed Cohen *et al.* (10) and applied the arcsine-root

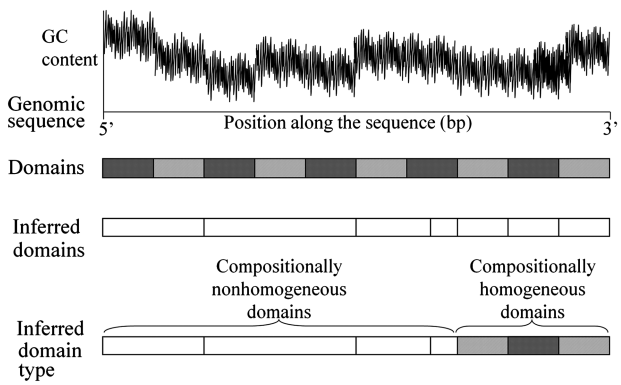

Figure 1. Example of a hypothetical genomic sequence composed of 10 compositionally homogeneous domains used to demonstrate two aspects of sequence analysis: partitioning and testing for homogeneity. Note that the partitioning yielded seven domains, out of which three were found to be homogeneous and the rest were nonhomogeneous.

transformation to the GC content values of the windows within each domain (and sequence) before calculating the variance.

A one-tailed *F*-test with a null hypothesis H_0: $\sigma^2_{\text{domain}} \geq \sigma^2_{\text{sequence}}$, and an alternative hypothesis, H_1: $\sigma^2_{\text{domain}} < \sigma^2_{\text{sequence}}$, were applied with n_1–1 and n_2–1 degrees of freedom, where n_1 and n_2 are the numbers of windows in the domain and in the sequence, respectively. If the variance over a domain turned out to be significantly lower ($P < 0.05$) than that of the corresponding sequence, then the domain was considered homogeneous compared to the sequence. We improved the procedure proposed by Cohen *et al.* (10) by adjusting for multiple comparisons using the FDR correction (24).

RESULTS

Modeling the dynamic threshold, D_t

Ideally, the segmentation halting threshold should be calculated analytically from the distribution of the D_{JS} entropy statistic (11). Because the theoretical distribution of the D_{JS} entropy is unknown, the threshold had to be obtained empirically. This was previously done by simulating uniform (homogeneous) regions of a certain length and standard deviation of the GC content (σ_{GC}), obtaining the maximal D_{JS} entropy statistic ($\hat{D}_{JS}$) for every sequence, calculating the cumulative distribution of $\hat{D}_{JS}$, and choosing a threshold value corresponding to some type I error rate (10,11). For example, Cohen *et al.* (10) obtained the threshold using 100 000 sequences of size 1 Mb and σ_{GC} of 1%. This practice leads to biases in the length and GC variability of inferred domains.

To demonstrate the relationship between $\hat{D}_{JS}$ and the sequence length, we generated a random sequence of length 1 Mb with a GC content of 50% and GC content standard deviation σ_{GC} of 1%. We then calculated $\hat{D}_{JS}$ for the entire sequence and subsequences of lengths 100 kb and 10 kb (Figure 2). The resulting $\hat{D}_{JS}$ values show a 10-fold increase with every 10-fold decrease in sequence length. A similar relationship exists between $\hat{D}_{JS}$ and the standard deviation σ_{GC} of the sequence. No relationship was found between $\hat{D}_{JS}$ and the sequence GC content (data not shown).

Reducing this bias in IsoPlotter required modeling the dependencies between a segment length and its standard deviation σ_{GC} on the one hand, and the $\hat{D}_{JS}$ statistic on the other hand. We generated 10 000 sequences for each of 13 length parameters L ranging from 1 kb to 1 Mb and for each of five GC content standard deviation parameters σ_{GC} ranging from 1% to 10%. For each simulation setting, we calculated $\hat{D}_{JS}$ by allowing IsoPlotter to partition the sequence once and obtained the threshold (D_t) from the top 0.01% percentile of the cumulative $\hat{D}_{JS}$ distribution (Figure 3). A log-scale plot of D_t as a function of sequence length and variability reveals a near perfect linear relationship (Figure 4). A log-scale linear regression on the length (L) and GC content dispersion (σ_{GC}) fits the

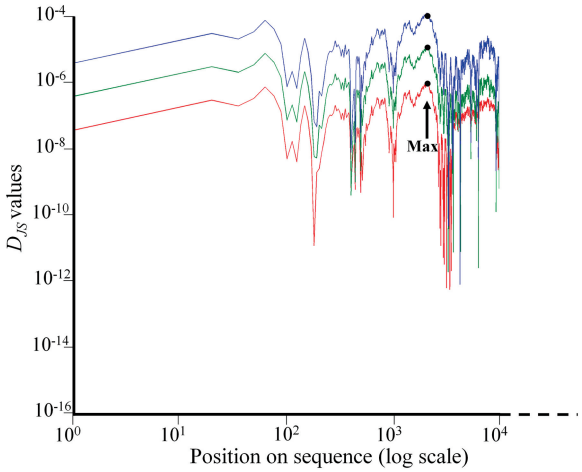

Figure 2. A demonstration of the relationship between D_{JS} entropy statistic, sequence length and variability (σ_{GC}), using 1-Mb simulated sequence (red) with σ_{GC} of 1% and two initial subsequences of sizes: 100 kb (green) and 10 kb (blue). Segmentation is carried out by obtaining the D_{JS} statistic values for all possible bipartition for each of the three sequences and partitioning them at the point of maximal D_{JS} ($\hat{D}_{JS}$). The D_{JS} values are plotted for each position in a different color for each sequence. The $\hat{D}_{JS}$ entropy measures are marked for each plot by a large dot. The plot illustrates the difference in D_{JS} values obtained for sequences with similar initial content but different lengths. A fixed threshold of 5.8×10^{-5} will not partition the shortest sequence although it has the same initial content as the other sequences.

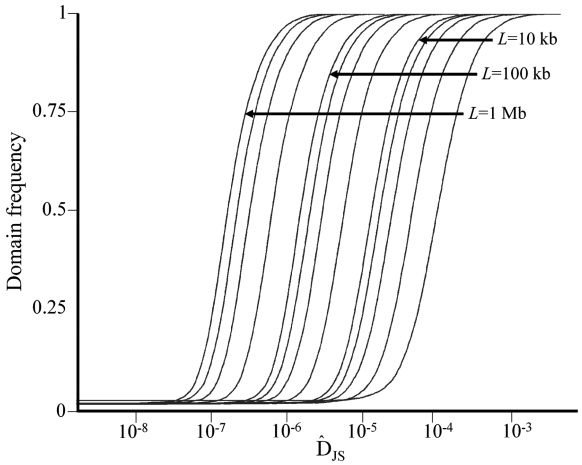

Figure 3. Cumulative $\hat{D}_{JS}$ distributions calculated for 10 000 sequences of σ_{GC} = 1% and different lengths L. The thresholds (D_t) are obtained from the top 0.01% percentile of each cumulative distribution and are marked for three sequence lengths.

empirical data extremely well ($r^2 = 0.995$, $P < 10^{-16}$) with the following coefficients:

$$\ln D_t = -0.97 \ln L + 0.7 \ln \sigma_{GC} + 4.42 \quad (6)$$

The threshold D_t depends on the length and GC content dispersion for each subsequence. We therefore refer to it is a 'dynamic threshold'.

Comparing segmentation results for simulated sequences

To gauge domain homogeneity in the simulated sets, we compared the variance in GC content within each domain

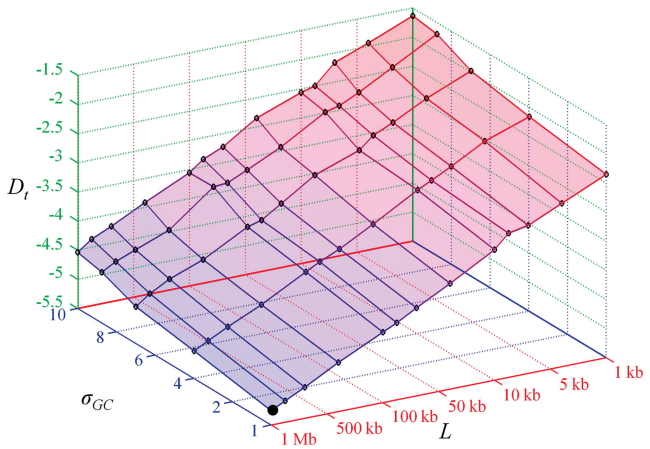

Figure 4. The threshold D_t values obtained from simulated sequences with various lengths and GC standard deviations are plotted against each sequence length L and variability σ_{GC} on a log–log scale. The threshold chosen by Cohen *et al.* (10) is marked by a large dot.

to that of the sequence on which it resides (see the 'Homogeneity test' section). Domains in all the sequences were constructed to be homogeneous, and the two algorithms were expected to detect them.

We first tested the abilities of IsoPlotter and D_{JS} to detect a short compositionally homogeneous central-domain (100 kb, 300 kb and 1 Mb) within two large compositionally homogeneous domains (10 Mb). Inferences of central-domain borders were divided into three types: no detection, partial detection (one border detected) and full detection (Table 1). IsoPlotter identified at least one domain border in all sequences and both domain borders in more than 86% of the sequences. By contrast, D_{JS} missed both domain borders in 15–20% of the sequences, identified one domain border in 46–75% of the sequences, and fully identified the two domain borders in <40% of the sequences. IsoPlotter inferences were unaffected by central-domain size, while D_{JS} inferences and sensitivity were highest for longer domains. Interestingly, both algorithms had high precision (97–100%) despite of D_{JS}'s poor performances. IsoPlotter performances were significantly higher than those obtained by D_{JS} (Wilcoxon rank-sum test, $\alpha < 0.01$).

Next, we tested the abilities of IsoPlotter and D_{JS} to infer domains with differing GC content standard deviations σ_{GC} and fixed lengths. An example of a simulated sequence is shown in Figure 5a. The mean sensitivity for each domain σ_{GC} is shown in Figure 5b. For these data, IsoPlotter sensitivity was 98%, with a precision approaching 100%. The D_{JS} sensitivity was 19% and strongly dependent on σ_{GC}, while the precision was near 100%. IsoPlotter significantly outperformed D_{JS} for every domain variation tested (Wilcoxon rank-sum test, $\alpha < 0.01$). Results were robust to the choice of window size that composed the compositionally homogeneous domains.

Segmentation of the human genome

One of the main premises of isochore theory is that nearly the entire genome of homeotherms (warm-blooded animals) consists of compositionally homogeneous domains that exceed 300-kb in length (4–6,27). Figure 6 presents the genome coverage as a function of domain length. IsoPlotter 'isochoric' domains (≥300 kb) cover <30% of the human genome, while D_{JS}'s 'isochoric' domains cover <50% of the genome.

Table 1. Proportion of central-domain detection inferred by IsoPlotter and D_{JS}

Segmentation algorithm	Central-domain				Sensitivity (%)	Precision (%)
	Size	No detection (%)	Partial detection (%)	Full detection (%)		
IsoPlotter	100 kb	0	13	86	93	97
	300 kb	0	12	87	94	97
	1 Mb	0	12	87	94	97
D_{JS}	100 kb	20	75	5	42	99
	300 kb	20	60	20	50	99
	1 Mb	15	46	39	62	100

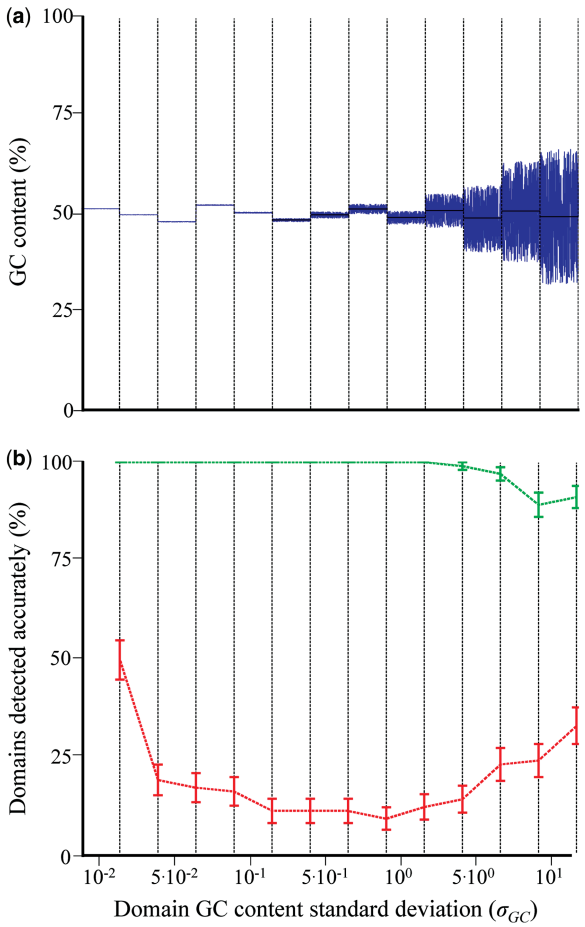

Figure 5. (**a**) An illustration of the GC content spatial distribution along a 130-kb simulated sequence that contains 13 equally sized domains (dotted bars). Domain mean GC content is marked by black lines. (**b**) The mean sensitivities (with error bars) of IsoPlotter (green) and D_{JS} (red) are shown for every domain (dotted bars).

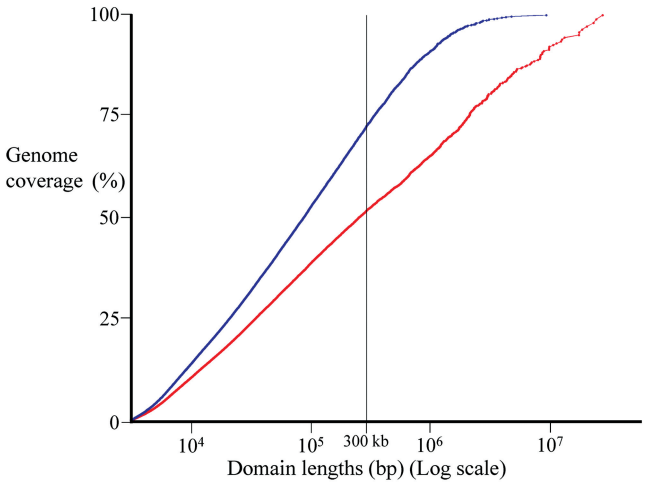

Figure 6. Cumulative spatial coverage for human genome using all domains inferred by IsoPlotter (blue) and D_{JS} (red). Seventy percent of IsoPlotter inferences were homogeneous compared to 67% for D_{JS}.

Classifying domains according to their lengths shows that overall IsoPlotter inferred a higher proportion of short to medium-size domains than D_{JS} (10–200 kb), while D_{JS} inferred more long domains (≥300 kb), including very long domains (≥10 Mb). Overall, domains inferred by D_{JS} are significantly longer than those inferred by IsoPlotter (*t*-test, $\alpha < 0.05$) illustrating the bias of D_{JS} toward long domains.

Classifying domains inferred by both algorithms according to their homogeneity shows that for each length-cutoff group 70% of IsoPlotter inferences were homogeneous compared to only 53–67% of D_{JS}. Moreover, the ratio of homogeneous/nonhomogeneous inferences for IsoPlotter is significantly higher than that of D_{JS} (*t*-test, $\alpha < 0.05$) for all length cutoffs (Table 2) except $l_c = 10$ Mb (see Supplementary Tables S1 and S2). These results suggest that the low proportion of compositionally homogeneous domains detected by D_{JS} is an outcome of the segmentation quality and does not depend on domain lengths.

In terms of coverage, 19% of the human genome is composed of 820 compositionally homogeneous domains longer than 300 kb (Table 2). Furthermore, these domains constitute only 1% of the total number of compositionally homogeneous domains (117 391). Therefore, restricting genome compositional studies to 'isochoric' domains, as is commonly done (e.g., 17,28,29), necessitates ignoring 99% of all domains that cover over 80% of the genome.

Compositional domain ideograms of the human genome

Using IsoPlotter inferences, three compositional domain ideograms of the human genome were drawn (Figures 7a–c). The ideograms illustrate our major findings, and allow us to compare compositional patterns among chromosomes. The first ideogram shows that compositionally homogeneous domains cover between 62% (chromosome 4) and 78% (chromosome 11) of the chromosomes (Figure 7a). Dividing compositionally homogeneous domains to long (≥300 kb) and short domains reveals that 'isochoric domains' are heterogeneously distributed along chromosomes covering between 5% (chromosome 19) and 29% (chromosomes 5) of the chromosomes (Figure 7b). By contrast, short compositionally homogeneous domains cover between 38% (chromosome 4) and 72% (chromosome 22) of the chromosomes. 'Isochoric domains' can further be classified into low GC domains ranging from 20% to 40% (574 domains) and high GC domains ranging from 40% to 60% (245 domains) with a single rich GC domain (61%) in chromosome 16 (Figure 7c). Thus, 70% of all 'isochoric domains' are AT rich. We find no evidence for the five-family division proposed by Bernardi *et al.* (4).

Table 2. Comparison of IsoPlotter and D_{JS} segment results for the human genome as a function of length cutoff (l_c)

	Length cutoff (l_c)	Domains longer than l_c		Compositionally homogeneous domains longer than l_c	
		Number (%)	Genome coverage (%)	Number (%)	Genome coverage (%)
IsoPlotter	3 kb	117 391 (100)	100	82 186 (70)	70
	10 kb	45 466 (39)	86	31 867 (27)	60
	50 kb	9861 (8)	60	6880 (6)	42
	100 kb	4691 (4)	48	3265 (3)	33
	200 kb	2081 (2)	35	1431 (1)	24
	300 kb	1199 (1)	27	820 (1)	19
	500 kb	561 (1)	19	386 (<1)	13
	1 Mb	163 (<1)	9	113 (<1)	7
	10 Mb	0 (0)	0	0 (0)	0
D_{JS}	3 kb	87 794 (100)	100	58 554 (67)	64
	10 kb	33 470 (38)	89	22 365 (26)	57
	50 kb	6986 (8)	70	4600 (5)	44
	100 kb	3337 (4)	61	2180 (3)	38
	200 kb	1598 (2)	53	1019 (1)	33
	300 kb	1042 (1)	48	667 (1)	30
	500 kb	650 (1)	43	412 (1)	26
	1 Mb	327 (<1)	35	208 (<1)	21
	10 Mb	15 (<1)	9	8 (<1)	5

DISCUSSION

The study of genome composition has been hampered for decades by conflicting results and uncertain methodology. Schmidt and Frishman (30) proposed to address this problem by using a consensus method based on the results of several algorithms (such as 12,17,19). This approach is problematic because combining correct inferences with incorrect ones only serves to dilute the truth. Instead, we proposed a benchmark to test the performance of different segmentation algorithms (15). We showed that recursive segmentation algorithms based on the Jensen–Shannon divergence (20) performed significantly better than all other segmentation algorithms.

However, even recursive segmentation algorithms can perform poorly because of their use of fixed thresholds as halting criteria. Here, we show that the D_{JS} entropy measure is correlated with sequence length and the standard deviation of its GC content (σ_{GC}), and that

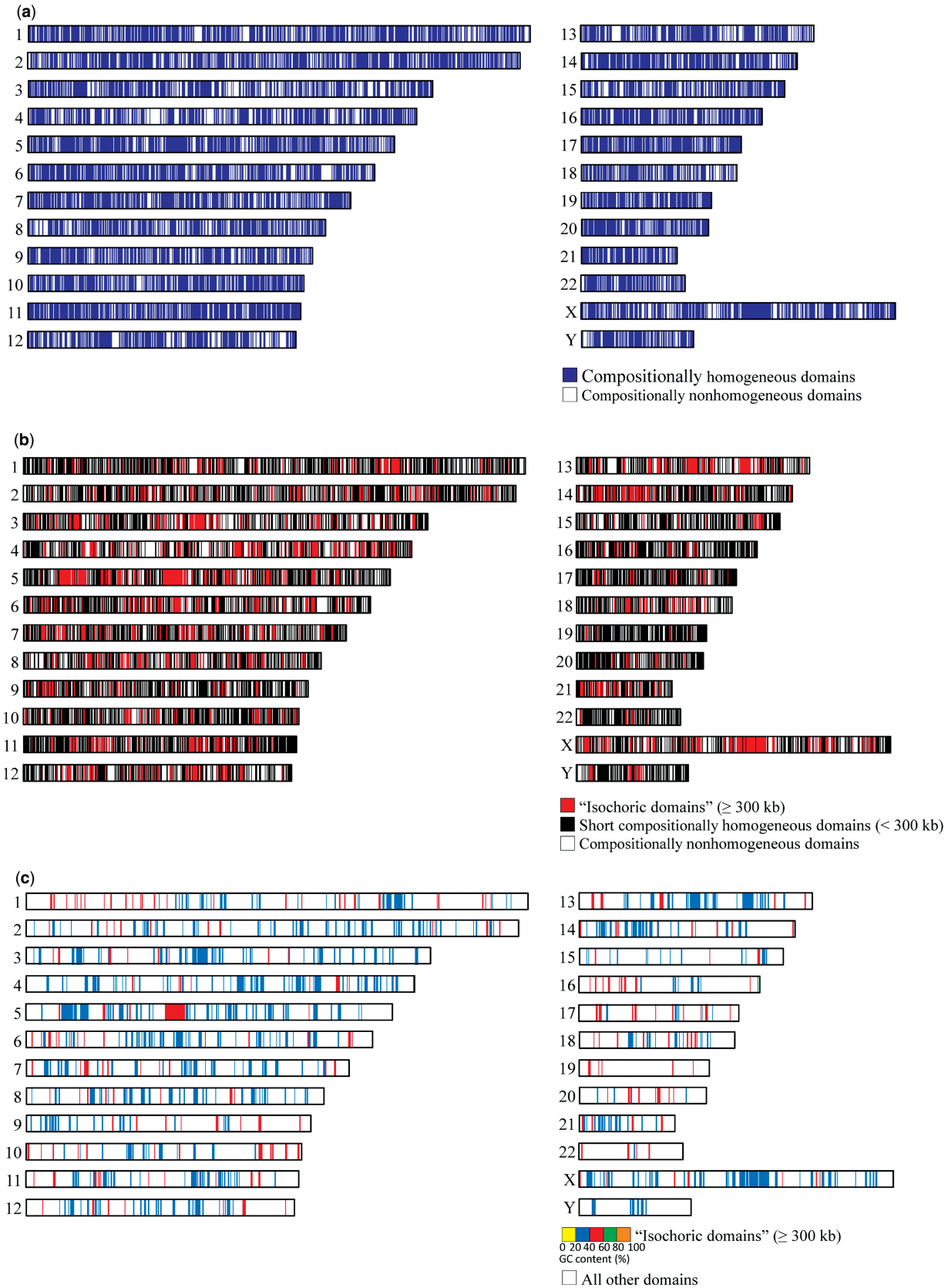

Figure 7. Ideograms of compositional domains inferred by IsoPlotter and mapped to chromosomes. Using data mining approach, the ideograms uncover the compositional patterns of long homogeneous domains ('isochoric') in three layers (from top to bottom): (**a**) compositionally homogeneous domains and nonhomogeneous domains; (**b**) long compositionally homogeneous domains (≥300 kb), short compositionally homogeneous domains (<300 kb) and compositionally nonhomogeneous domains; and (**c**) long compositionally homogeneous domains (≥300 kb) color coded by their mean GC content and all other domains (short compositionally homogeneous and nonhomogeneous domains).

this dependence introduces biases in the segmentation. Consequently, recursive segmentation algorithms employing a fixed threshold, such as D_{JS}, cannot be expected to perform well on sequences containing isochores of different lengths and compositions. To overcome this problem, we modeled these relationships to create a dynamic-threshold algorithm.

The log-linear relation between the threshold value, D_t, and sequence length, L [Equation (6)] is not surprising. Let us consider an idealized model of genomic sequence as a series of Bernoulli trials with P the probability of a G or a C nucleotide. The mean GC content of a sequence of length N is, therefore, a random variable that approximately follows a normal distribution with mean P and standard deviation $\sqrt{P(1-P)/N}$. The standard deviation of the mean GC content decreases with sequence length.

Since the entropy H_{tot} is defined as a function of the mean GC content, H_{tot} is itself a random variable. In the case of a Bernoulli sequence, H_{tot} is approximately normally distributed as it is a function of mean GC content, and the standard deviation of the mean GC content is relatively small. Furthermore, the D_{JS} statistic is a function of H_{tot}. Hence, D_{JS} is also a random variable with variance that decreases with sequence length. Although real genomic sequences cannot be expected to be modeled by perfect Bernoulli trials, the decrease of the standard deviation of the mean GC content with the increase of sequence size is expected due to the central limit theorem (31), if the correlations between nucleotide content along the sequence are not too strong.

The main idea proposed here is expected to hold despite the mild long-range correlations described in genomic sequences (11,32–35), which may increase the standard deviation. These correlations reduce domain homogeneity and produce a higher proportion of false positive inferences (type I errors) because of the fluctuations in nucleotide composition. Moreover, since recursive segmentation algorithms compare the composition of subsegments residing on adjacent subsequence to each other, the resulting domains are not necessarily homogeneous compared to the whole sequence. For these reasons, it is essential to assess the homogeneity of the inferred domains using a homogeneity test.

The high sensitivity and precision obtained using the dynamic threshold were demonstrated in two analyses in which IsoPlotter successfully detected compositionally homogeneous domains of various lengths and GC content standard deviations σ_{GC}. IsoPlotter's high sensitivity and precision are a direct result of its dynamic threshold. Such results are not achievable with other segmentation algorithms. Moreover, these results suggest that segmentation approaches that filter or concatenate 'short' segments to eliminate GC content fluctuations (e.g., 12,28) may be misleading.

A homogeneity test was applied to the domains inferred in the human genome. A classification of these domains revealed, surprisingly, that the majority of the genome (70%) consisted of compositionally homogeneous domains, but only 19% of them can be considered isochores in the traditional sense (8). We note that IsoPlotter was not artificially tuned to detect domains of a particular length (e.g., 28) and, unlike D_{JS}, it is not biased toward short domains (Table 1).

Cohen *et al*. (10) used the D_{JS} algorithm to partition the human genome and then classified the inferred domains according to their lengths using a length cutoff, l_c. A comparison of our results with those obtained by Cohen *et al*. (10) reveals two major differences. First, Cohen *et al*. (10) reported that the proportion of domains found to be 'putatively homogeneous' out of all inferred domains was dependent on the domain length cutoff, l_c. That is, the longer domain was more likely considered the 'putatively homogeneous' domain. In contrast, our results show that the proportion of compositionally homogeneous domains out of all inferred domains slightly decreased with the increase in length cutoff, l_c. Second, they reported a possible bias in their homogeneity test, which qualified almost all long domains (>50 kb) as 'putatively homogeneous'. However, we did not observe this bias using our homogeneity test. The difference in our results can be explained by our improved homogeneity test that corrects for multiple comparisons.

Segmenting the human genome with IsoPlotter revealed a new genomic compositional architecture consisting of a mixture of compositionally nonhomogeneous domains with numerous short compositionally homogeneous domains and relatively few long ones (Figures 7a–c). A preliminary analysis of eight species indicates that this salient description holds for other mammalian genomes (Elhaik E. and Graur D., unpublished data). To understand how such structures emerged in an evolutionary perspective, a comparative analysis using different genomes is currently underway. Using IsoPlotter, we now have the ability to use the same analytical tool on genomes that were heretofore considered too heterogeneous to be partitioned, such as the yeast genome (36).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: The National Science Foundation (grant number DBI-0543342 to D.G.); University of Houston (Small Grant award number I098048 to D.G.); US National Library of Medicine (grant number LM010009-01 to D.G. and G.L.); National Science Foundation (grant numbers DMS-0604429, DMS-0817649); Texas Advanced Research Program and the Advanced Technology Program (Project no. 003652-0024-2007 to K.J.).

Conflict of interest statement. None declared.

REFERENCES

1. Bernardi,G. (1965) Chromatography of nucleic acids on hydroxyapatite. *Nature*, **206**, 779–783.
2. Filipski,J., Thiery,J.P. and Bernardi,G. (1973) An analysis of the bovine genome by Cs2SO4-Ag density gradient centrifugation. *J. Mol. Biol.*, **80**, 177–197.

3. Thiery,J.P., Macaya,G. and Bernardi,G. (1976) An analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.*, **108**, 219–235.
4. Bernardi,G., Olofsson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–958.
5. Macaya,G., Thiery,J.P. and Bernardi,G. (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.*, **108**, 237–254.
6. Bernardi,G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241**, 3–17.
7. Fukagawa,T., Sugaya,K., Matsumoto,K., Okumura,K., Ando,A., Inoko,H. and 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.
8. Bernardi,G. (2001) Misunderstandings about isochores. Part 1. *Gene*, **276**, 3–13.
9. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
10. Cohen,N., Dagan,T., Stone,L. and Graur,D. (2005) GC composition of the human genome: in search of isochores. *Mol. Biol. Evol.*, **22**, 1260–1272.
11. Bernaola-Galván,P., Román-Roldán,R. and Oliver,J.L. (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys. Rev. E.*, **53**, 5181–5189.
12. Oliver,J.L., Carpena,P., Hackenberg,M. and Bernaola-Galván,P. (2004) IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res.*, **32**, W287–W292.
13. Guéguen,L. (2005) Sarment: Python modules for HMM analysis and partitioning of sequences. *Bioinformatics*, **21**, 3427–3428.
14. Haiminen,N., Mannila,H. and Terzi,E. (2007) Comparing segmentations by applying randomization techniques. *BMC Bioinformatics*, **8**, 171.
15. Elhaik,E., Graur,D. and Josić,K. (2010) Comparative testing of DNA segmentation algorithms using benchmark simulations. *Mol. Biol. Evol.*, **27**, 1015–1024.
16. Costantini,M., Auletta,F. and Bernardi,G. (2007) Isochore patterns and gene distributions in fish genomes. *Genomics*, **90**, 364–371.
17. Costantini,M., Clay,O., Auletta,F. and Bernardi,G. (2006) An isochore map of human chromosomes. *Genome Res.*, **16**, 536–541.
18. Li,W., Bernaola-Galván,P., Haghighi,F. and Grosse,I. (2002) Applications of recursive segmentation to the analysis of DNA sequences. *Comput. Chem.*, **26**, 491–510.
19. Haiminen,N. and Mannila,H. (2007) Discovering isochores by least-squares optimal segmentation. *Gene*, **394**, 53–60.
20. Lin,J. (1991) Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory*, **37**, 145–151.
21. Li,W. (2001) New stopping criteria for segmenting DNA sequences. *Phys. Rev. Lett.*, **86**, 5815–5818.
22. Li,W. (2001) Delineating relative homogeneous G+C domains in DNA sequences. *Gene*, **276**, 57–72.
23. Sokal,R.R. and Rohlf,F.J. (1995) *Biometry*, 3rd edn. W.H. Freeman and Company, NY.
24. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.
25. Li,W., Bernaola-Galván,P., Carpena,P. and Oliver,J.L. (2003) Isochores merit the prefix 'iso'. *Comput. Biol. Chem.*, **27**, 5–10.
26. Zar,J.H. (1999) *Biostatistical Analysis*. Prentice-Hall, Upper Saddle River, NJ.
27. Cuny,G., Soriano,P., Macaya,G. and Bernardi,G. (1981) The major components of the mouse and human genomes: preparation, basic properties and compositional heterogeneity. *Eur. J. Biochem.*, **115**, 227–233.
28. Oliver,J.L., Bernaola-Galván,P., Carpena,P. and Román-Roldán,R. (2001) Isochore chromosome maps of eukaryotic genomes. *Gene*, **276**, 47–56.
29. Oliver,J.L., Carpena,P., Román-Roldán,R., Mata-Balaguer,T., Mejías-Romero,A., Hackenberg,M. and Bernaola-Galván,P. (2002) Isochore chromosome maps of the human genome. *Gene*, **300**, 117–127.
30. Schmidt,T. and Frishman,D. (2008) Assignment of isochores for all completely sequenced vertebrate genomes using a consensus. *Genome Biol.*, **9**, R104.
31. Feller,W. (1968) *An Introduction to Probability Theory andIts Applications*, 3rd edn. John Wiley & Sons, Inc., NY.
32. Li,W. and Kaneko,K. (1992) Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys. Lett.*, **17**, 655–660.
33. Peng,C.K., Buldyrev,S.V., Goldberger,A.L., Havlin,S., Sciortino,F., Simons,M. and Stanley,H.E. (1992) Long-range correlations in nucleotide sequences. *Nature*, **356**, 168–170.
34. Peng,C.K., Buldyrev,S.V., Havlin,S., Simons,M., Stanley,H.E. and Goldberger,A.L. (1994) Mosaic organization of DNA nucleotides. *Phys. Rev. E.*, **49**, 1685–1689.
35. Stanley,H.E., Buldyrev,S.V., Goldberger,A.L., Havlin,S., Peng,C.K. and Simons,M. (1999) Scaling features of noncoding DNA. *Physica A*, **273**, 1–18.
36. Li,W., Stolovitzky,G., Bernaola-Galván,P. and Oliver,J.L. (1998) Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. *Genome Res.*, **8**, 916–928.