



ELSEVIER

Gene 300 (2002) 129–139

**GENE**  
AN INTERNATIONAL JOURNAL ON  
GENES AND GENOMES

[www.elsevier.com/locate/gene](http://www.elsevier.com/locate/gene)

# Are isochore sequences homogeneous?

Wentian Li\*

Center for Genomics and Human Genetics, North Shore – LIJ Research Institute, 350 Community Drive, Manhasset, NY 11030, USA

Received 21 December 2001; received in revised form 13 May 2002; accepted 17 July 2002

## Abstract

Three statistical/mathematical analyses are carried out on isochore sequences: spectral analysis, analysis of variance, and segmentation analysis. Spectral analysis shows that there are GC content fluctuations at different length scales in isochore sequences. The analysis of variance shows that the null hypothesis (the mean value of a group of GC contents remains the same along the sequence) may or may not be rejected for an isochore sequence, depending on the subwindow sizes at which GC contents are sampled, and the window size within which group members are defined. The segmentation analysis shows that there are stronger indications of GC content changes at isochore borders than within an isochore. These analyses support the notion of isochore sequences, but reject the assumption that isochore sequences are homogeneous at the base level. An isochore sequence may pass a homogeneity test when GC content fluctuations at smaller length scales are ignored or averaged out. © 2002 Elsevier Science B.V. All rights reserved.

*Keywords:* Isochore sequences; Homogeneous; Statistical/mathematical analyses

## 1. Introduction

In the paper on the human genome draft sequence (Lander et al., 2001), the following comments were made on long-range variation of GC content in DNA sequences and on the topic of ‘isochores’ (Macaya et al., 1976; Cuny et al., 1981; Bernardi, 1995): “We studied the draft genome sequence to see whether strict isochores could be identified. For example, the sequence was divided into 300-kb windows, and each window was subdivided into 20-kb subwindows. We calculated the average GC content for each window and subwindow, and investigated how much of the variance in the GC content of subwindows across the genome can be statistically ‘explained’ by the average GC content in each window. About three-quarters of the genome-wide variance among 20-kb windows can be statistically explained by the average GC content of 300-kb windows that contain them, but the residual variance among subwindows (standard deviation, 2.4%) is still too large to be consistent with a homogeneous distribution. In fact, the hypothesis of homogeneity could be rejected for each 300-kb window in the draft genome sequence. ... These results rule out a strict

notion of isochores as compositionally homogeneous. Instead, there is a substantial variation at many different scales, ... Although isochores do not appear to merit the prefix ‘iso’, the genome clearly does contain large regions of distinctive GC content and it is likely to be worth redefining the concept so that it becomes possible rigorously to partition the genome into regions” (p. 877 of Lander et al., 2001).

Several sentences in the above paragraph need further examination. Besides an inappropriate test used in Lander et al. (2001) (see Section 4.1), the discussion on variances of GC contents in windows and subwindows of certain sizes (300 and 20 kb) raises questions. As discussed in Li (2001c), the concept of homogeneity is relative: not only does it depend on the stringency of the criterion, but it also depends on the length scale at which GC contents are examined. It is natural to ask whether other choices of the window and subwindow sizes may change the conclusion concerning homogeneity. Sometimes, short segmentations of DNA sequences (e.g. 1 kb) with extreme high or low GC content are the reason for the sequence to fail a homogeneity test. Nevertheless, these segments are much shorter than the sequence being examined, and might be ignored. With these short-scale fluctuations of base composition averaged out (or ‘coarse graining’, borrowing from a term in statistical physics which specializes in connecting the microscopic and the macroscopic worlds), can a claimed heterogeneous sequence become homogeneous?

Besides using variances within and between windows to test homogeneity, in a branch of statistics called ‘change-

*Abbreviations:* ANOVA, analysis of variance; BIC, Bayesian information criterion; bp, base pair; GC, nucleotides of either guanine or cytosine; kb, kilo (1000) bases; LLR, log likelihood ratio; Mb, mega (1,000,000) bases; MHC, human major histocompatibility complex.

\* Tel.: +1-516-562-1076; fax: +1-516-562-1683.

*E-mail address:* wli@linkage.rockefeller.edu (W. Li), wli@nshs.edu (W. Li).

point problem', there is another way to test the hypothesis of a sequence being random and homogeneous. In this test, the hypothesis (null) that a sequence is a random sequence is compared to another hypothesis (alternative) that the sequence consists of one random sequence followed by another random sequence with a different bias at some point (change-point). This test can be carried out in the framework of the (log maximum) likelihood ratio test. Given a significance level, any sequence can be tested for homogeneity by the presence or absence of a change-point. Even if the number of domains with different GC contents is not two, but three, four, or more, this test should still be able to determine whether to reject or accept the null hypothesis. This method was not applied to the human genome draft sequence in Lander et al. (2001).

Because the interest in the field of change-point problems is not just to test the null hypothesis, but also to determine the change-point position, it provides a rigorous tool to partition a given sequence into relatively homogeneous domains, as desired in Lander et al. (2001). In fact, a recursive segmentation has been applied to DNA sequence analysis for the last few years (Bernaola-Galván et al., 1996; Román-Roldán et al., 1998; Oliver et al., 1999; Li et al., 2002). This method is an extension of the 1-to-2 segmentation discussed in the change-point literature: rather than applying it once, it was applied repeatedly. When the focus is on homogeneity in GC content, recursive segmentation successfully delineates isochore sequences (Li, 2001c; Oliver et al., 2001, 2002). Again, these research developments were not included or referred to in Lander et al. (2001).

Finally, it was mentioned in passing in the above paragraph that "there is a substantial variation at many different scales". Probably unknown to the authors of Lander et al. (2001), this topic has been actively studied under names such as 'long-range correlation', 'self-similarity', 'hierarchical pattern', and 'fractal landscapes' (Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992; Li et al., 1994; Li, 1997a). For example, it was pointed out explicitly in Li et al. (1994) that the fluctuation of base composition at small length scales can be independent from that at larger length scales. Using moving windows of different sizes, the base composition variation may look different. A similar discussion can be found in Bernaola-Galván et al. (1996). The terms 'complex heterogeneity' and 'domains within domains' used in Li (1997a,b) are attempts to capture the essence of this complicated situation. The multi-scaled fluctuation of base composition was also discussed in the framework of 'complexity measure' (Román-Roldán et al., 1998; Li, 1997b). Interestingly, an old mathematical technique, spectral analysis, remains an effective tool for distinguishing sequences characterized by different multi-scaled fluctuations in base composition (Li, 1997b).

As pointed out in Bernardi (2001), the Lander et al. (2001) paper may have been too ambitious in its endeavor to summarize every aspect of the human genome sequence. Understandably, it may not have provided the state-of-art

discussion on each topic it addressed either due to time limitation or the range of expertise of the authors. Here I will attempt to supply some analyses missing from Lander et al. (2001) that address the topic of base composition homogeneity in isochore sequences: analysis of variance using different group and member definitions, segmentation analysis, and spectral analysis. From these analyses, it is hoped that a clearer picture will emerge concerning the homogeneity of isochore sequences.

## 2. Methods

### 2.1. Analysis of variance

Analysis of variance (ANOVA), first developed by Ronald Fisher (Fisher, 1925), is a test of equal mean values among different sample groups. Since a difference of mean values in two groups can either be due to a true difference or a large within-group variation, it is essential in ANOVA to estimate and compare variances, besides mean values, in different groups. When the number of groups is reduced to two, the ANOVA test is equivalent to the well known *t*-test.

Suppose there are  $n$  members separated in  $g$  groups, with group  $i$  having  $n_i$  samples ( $\sum_{i=1}^g n_i = n$ ). The sample values  $y_{ij}$  in group  $i$ , individual  $j$ , are modeled by the following equation:

$$y_{ij} = \mu + \mu_i + \epsilon_{ij} \quad i = 1, 2, \dots, g \quad j = 1, 2, \dots, n_i \quad (1)$$

where  $\mu$  is the common mean (average) among groups,  $\mu_i$  is the deviation from the common mean for each group  $i$ , and  $\epsilon_{ij}$  is a Gaussian (normally distributed) noise. The null hypothesis to be tested is that all groups have the same mean, i.e.  $\mu_1 = \mu_2 = \dots = \mu_g = 0$ . A test result is summarized in a  $P$  value: it is the probability of obtaining the observed value, and other more extreme ones, of the test statistic under the null hypothesis.

ANOVA models can also be called 'factorial models' in the regression framework (McCullagh and Nelder, 1983). This is because Eq. (1) can be viewed as a regression of  $y_{ij}$  over a categorical variable  $i$ . Since regression over a non-numerical variable is not possible, the effect of the categorical variable (i.e. the group label) is represented by a distinct  $\mu_i$  term, instead of index  $i$  multiplied by a coefficient.

To apply ANOVA to our problem, we partition a given DNA sequence twice. The first time, the sequence is evenly divided into  $g$  windows. The second time, each window is evenly divided in  $n/g$  subwindows. A subwindow is a sample, whose GC content is a sample value. All subwindows in a window form a group.

### 2.2. Recursive segmentation

The recursive segmentation used here was introduced in Bernaola-Galván et al. (1996). The objective of a segmentation is to divide a sequence into subsequences of different

base compositions that are internally homogeneous. To decide whether two neighboring subsequences are different in composition, one can use the likelihood ratio test. For the same purpose, a model selection version for the decision to continue the recursion can also be used (Li, 2001a,b,c).

The decision to segment a DNA sequence into two domains requires a comparison of two statistical models of the sequence: (1) the null model: one random sequence with one set of base frequencies; and (2) the alternative model: two random subsequences with different base frequencies. For the isochore problem, bases G and C are combined into symbol S (strong) and bases A and T into symbol W (weak).

The likelihood function of the null model is

$$L_1(P_S) = \prod_{\alpha=(S,W)} P_\alpha^{N_\alpha} \quad (2)$$

where  $P_S$  is the probability of finding either G or C ( $P_W$  is simply  $1 - P_S$ ), and  $N_S$  and  $N_W$  are the number of Ss and Ws in the sequence ( $N_S + N_W = N$  is the sequence length). Similarly, the likelihood function of the alternative model is

$$L_2(P_S^l, P_S^r, N_1) = \prod_{\alpha=(S,W)} (P_\alpha^l)^{N_{\alpha,l}} \prod_{\alpha=(S,W)} (P_\alpha^r)^{N_{\alpha,r}} \quad (3)$$

where a label l and r for left and right subsequences is used,  $N_1$  and  $N - N_1$  are the length of the left and right subsequence, and  $N_{S,l} + N_{W,l} = N_1$ ,  $N_{S,r} + N_{W,r} = N - N_1$ .

When both likelihood functions are maximized over the parameter values (the maximum likelihood estimation of the parameter  $P_\alpha$  is simply  $\hat{P}_\alpha = N_{\alpha}/N$ ), under certain conditions, the ratio of the maximum likelihoods (after taking its logarithm and multiplying by 2) follows a  $\chi^2$  distribution if the null model is correct. This condition is satisfied when  $L_2$  is maximized only over two GC content parameters  $P_S^l, P_S^r$  but not over the segmentation point (change-point)  $N_1$ .

When  $L_2$  is maximized over all three parameters ( $P_S^l, P_S^r, N_1$ ), however, the situation is more complicated (Horvath, 1989; Csorgo and Horvath, 1997; Grosse et al., 2002). The main reason why the  $\chi^2$  distribution will not be the correct distribution is that the (2)log-(max)likelihood-ratio (LLR),  $2\log\hat{L}_2/\hat{L}_1$ , will increase with the sequence length  $N$ , whereas a  $\chi^2$  distribution does not depend on  $N$ . If we assume that LLR( $N_1$ ) is a random walk in  $N_1$  (a ‘bridged’ random walk because it starts and ends at zero), LLR is the maximum of LLR( $N_i$ ) over  $N_i$ , then LLR may increase with  $N$  as  $\sqrt{N}$ . The discussions in Horvath (1989) and Csorgo and Horvath (1997), however, conclude that LLR increases with  $N$  as  $\log(\log(N))$ . In the model selection procedure discussed below, a  $\log(N)$  term is used, which is a compromise between  $\sqrt{N}$  and  $\log(\log(N))$ .

In a likelihood ratio test, the null model is rejected if the LLR is too large, or equivalently, if the  $P$  value obtained from the null distribution of LLR is too small. Alternatively, one can use the model selection technique which addresses the ‘merit’ of two models directly. The ‘merit’ of a model is a combination of both its data-fitting performance (e.g.

maximum likelihood) and its ‘cost’ (e.g. number of free parameters used in the model). Larger (max)likelihood is not enough to select a model because a perfect data-fitting performance can be accomplished by using an overly sophisticated model. It was proposed in Li (2001a,b) that the Bayesian information criterion (BIC) (Schwarz, 1978) could be used in the segmentation decision. BIC is defined as:

$$\text{BIC} = -2\log\hat{L} + \log(N)K \quad (4)$$

where  $\hat{L}$  is the maximized likelihood,  $N$  is the sample size, and  $K$  is the number of free parameters in the model. The smaller the BIC, the better the model. A straightforward use of BIC leads to the following criterion for continuing the recursion in S-W sequences ( $K_2 = 3$  for the parameters  $P_S^l, P_S^r, N_1$ ;  $K_1 = 1$  for the parameter  $P_S$ ):

$$2\log\frac{\hat{L}_2}{\hat{L}_1} > 2\log(N) \quad (5)$$

By further examination of this problem, we see that the segmentation point  $N_1$  is more an index of models than a regular parameter in the model. It is still an open question whether  $N_1$  should be considered to contribute more than one parameter (Li, 2001c; David Sigmund, pers. commun.).

To mimic something similar to the  $P$  value (and significance level) in the hypothesis testing framework, the percentage increase of the (2 log max) likelihood-ratio over the allowed threshold is used as a ‘segmentation strength’ (Li, 2001a,b):

$$s = \frac{2\log\hat{L}_2/\hat{L}_1 - 2\log(N)}{2\log(N)} = \frac{\log\hat{L}_2/\hat{L}_1}{\log(N)} - 1 \quad (6)$$

A recursive segmentation can be carried out in the following way: the original sequence is segmented at a particular segmentation point to create two subsequences; if the segmentation strength  $s$  is larger than a pre-specified value  $s_0$ , the segmentation continues for each one of the two subsequences. Whether the segmentation should further continue is determined by the  $s$  at each stage of the recursion. Requiring  $s > s_0 = 0$  to continue the recursion is equivalent to the model selection using BIC. Requiring  $s > s_0 > 0$  is to use a more stringent criterion, usually when one is interested in obtaining larger DNA domains.

### 2.3. Spectral analysis

Spectral analysis is a method to decompose the variance of a sequence into contributions from different periodic components. Considering a sequence of  $n$  points,  $x_j$  ( $j = 1, 2, \dots, n$ ), the standard and the oldest spectral analysis, Fourier analysis (Fourier, 1822; Bracewell, 1965), is to use sine and cosine periodic functions to decompose a sequence:

$$A_k = \frac{1}{n} \sum_{j=1}^n x_j \cos(2\pi j \frac{k}{n}), \quad k = 0, 1, 2, \dots, \frac{n}{2}$$

$$B_k = \frac{1}{n} \sum_{j=1}^n x_j \sin(2\pi j \frac{k}{n}), \quad k = 1, 2, \dots, \frac{n}{2} - 1 \quad (7)$$

where  $k$  is the wavelength number. The power spectrum or variance spectrum is:

$$P_k = A_k^2 + B_k^2, \quad k = 0, 1, 2, \dots, \frac{n}{2} \quad (8)$$

(though the  $k = 0$  component is simply the mean, and usually not displayed).

For a DNA sequence with length  $N$ , if one considers the presence or absence of particular bases (e.g. G and C), then our sequence  $x_j$  consists of  $n = N$  binary (0/1) values. For a longer DNA sequence, it is impractical to consider the base at each position in a spectral analysis. One can then partition the sequence evenly into  $n$  subsequences, and our sequence consists of GC contents in these subsequences. All patterns within a subsequence are averaged out and thus destroyed. For example, the periodicity of three bases in coding regions should present itself as a peak at frequency  $f = k/N = 1/3$ . When a window size is larger than 3, however, this periodicity will not be revealed. Generally speaking, the choice of window size does not affect the power spectrum at lower frequencies, despite its impact on the high-frequency spectrum.

### 3. Results

In this paper, only sequences that are widely considered to be isochore sequences are analyzed. Towards this purpose, two well known isochores in the human major histocompatibility complex (MHC) on chromosome 6 (Fukagawa et al., 1995, 1996; Stephens et al., 1999), and one long GC-poor region in human chromosome 21 are used. There are four regions in the human MHC sequence, class I, class III, class II and extended class II (Beck et al., 1999). Class III and class II are more homogeneous than the other two regions, and are considered to be isochores (Stephens et al., 1999). Among the completely sequenced human chromosomes, the chromosome 21 (Hattori et al., 2000) sequence is much more homogeneous, at the isochore scale, than the chromosome 22 sequence (Dunham et al., 1999), as shown in several studies (e.g. Nekrutenko and Li, 2000; Clay et al., 2001; Li, 2001c; Saccone et al., 2001; Pavlíček et al., 2002). One GC-poor, gene-poor, and Alu-poor region on chromosome 21 is a safe bet for an isochore (Saccone et al., 2001). The boundaries of the three isochore sequences are delineated by recursive segmentation (Li, 2001c). These segmentation results are reproduced here in Fig. 1.

#### 3.1. ANOVA test result depends on window sizes

Our ANOVA test requires the partitioning of a sequence twice: the first time into subsequences (windows) of equal sizes to get groups, and the second time into sub-

subsequences (subwindows) of equal sizes to get members. The value of a member is the GC content of a subwindow. Tables 1–3 show the test results of the three isochore sequences for various choices of window and subwindow sizes.

For the MHC class III sequence, when the number of windows and number of subwindows per window are (2,10), (10,10), (2,100), (10,20), (20,10), and (10,40), the ANOVA test is unable to reject the null hypothesis that the mean values of groups are the same (at the significance level of 0.01). In other words, in this test of sequence homogeneity, at these length scales, the MHC class III sequence is considered to be homogeneous. However, when the number of windows and subwindows per window increases to (20,20) or (30,30), the null hypothesis is rejected (see Table 1). Significance levels other than 0.01 could also be used, but the  $P$  values listed in Table 1 remain the same.

Similarly, for the MHC class II sequence, the null hypothesis, that the mean in different groups is the same, can not be rejected when the number of windows and subwindows per window is (2,10), (4,20) and (10,10) (at a significance level of 0.01), but it can be rejected when the number of windows and subwindows per window is (2,100), (20,10), and (15,15) (Table 2). With five windows (groups) and 30 subwindows per window (members per group), the  $P$  value is around 0.01, a borderline for rejection at the 0.01 significance level. For the human chromosome 21 isochore, the null hypothesis is not rejected at the 0.01 significance level for (number of windows, number of subwindows per window) equal to (2,10), (10,10), (3,40), and (2,100), but is rejected for (4,30), (4,50), and (20,10) (Table 3).

The size of the subwindows is also indicated in Tables 1–3. It can be seen that generally speaking, the smaller the subwindow size, the more likely it is that the null hypothesis can be rejected (the smaller the  $P$  values). Nevertheless, from Tables 1 and 3, even with the same subwindow size, the ANOVA test can sometimes lead to either no-rejection or rejection, depending on the number of windows (groups). Tables 1–3 clearly show that the ANOVA test result is not unique for a given sequence: the choice of the number of windows and subwindows, which determines the number of groups and number of members per group, may alter the conclusion. Changing the significance level for rejecting the null hypothesis is another (relatively trivial) way to alter the conclusion.

#### 3.2. Isochore sequences can be further segmented

As can be seen from Fig. 1, the two isochore sequences in the MHC are segmented with very strong segmentation strengths ( $s = 237, 171, \text{ and } 288$ ), and so is the isochore sequence in human chromosome 21 ( $s = 148 \text{ and } 188$ ). Here we would like to examine whether these three sequences themselves can be segmented at a much lower segmentation criterion, such as  $s > s_0 = 0$ .

Figs. 2–4 show the segmentation results of MHC class

III, MHC class II, and the human chromosome 21 isochore. For graphical illustration, the thresholds for segmentation strength,  $s_0$ , are set at 0.333, 1, and 1, respectively, for these three sequences. The numbers of domains resulting from the segmentation are 138, 83, and 92.

For the MHC class III sequence, the strongest segmentation is achieved at  $s = 24.5$  at position 469,430 bp from the left end of the sequence. This segmentation point separates two domains with GC contents of 58.6% and 45.8% (and lengths of 5.8 and 2.8 kb). There are 16 segmentations with  $s$

larger than 5, and only four segmentations with  $s$  larger than 10. The longest domain at  $s_0 = 0.333$  is 116 kb. One can see from Fig. 2 that many smaller domains (a few hundred kb) have much higher (e.g. larger than 70%) or lower (e.g. lower than 50%) GC contents as compared to the bulk GC content (51.9%). The segmentation of MHC class II sequence (Fig. 3) is very similar to that of the class III sequence. The strongest segmentation has  $s = 22$ , and only nine segmentations have  $s$  larger than 10.

Segmentation of the chromosome 21 isochore sequence

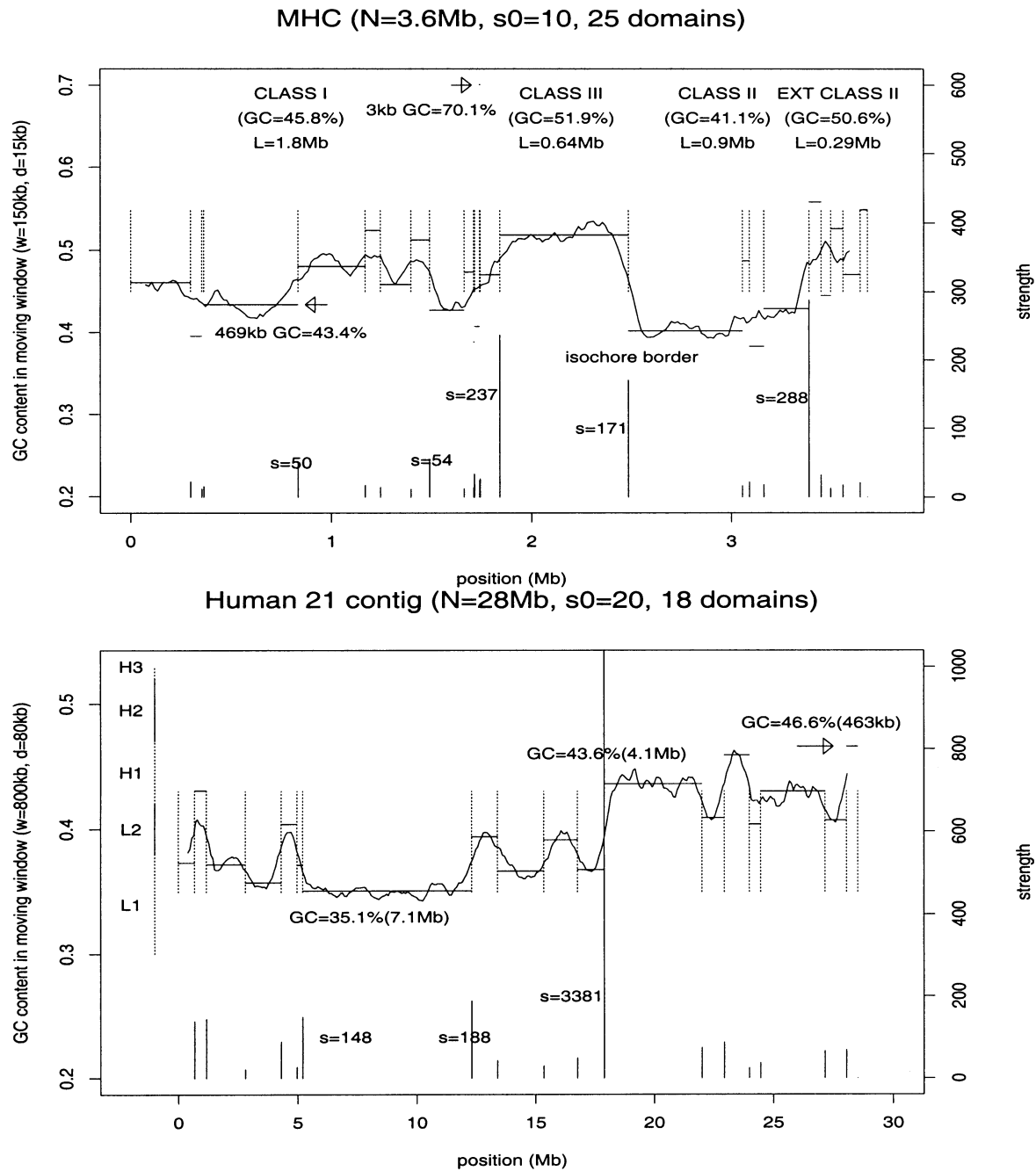


Fig. 1. (Top) Two isochore sequences in human MHC region (class III and class II). (Bottom) One isochore sequence in human chromosome 21. The threshold for segmentation strength ( $s_0$ ) is set at 10 and 20, respectively. The lengths of the three isochores are 642 kb (class III in MHC), 901 kb (class II in MHC), and 7.104 Mb (chromosome 21 isochore). (Reproduced from Li (2001c).)

Table 1  
ANOVA test results for MHC class III isochore sequence ( $N = 642,095$  bp)<sup>a</sup>

# windows	# subwindows	Total # elements	Subwindow size (bp)	<i>P</i> value
2	10	20	32,104	0.199
10	10	100	6420	0.2834
2	100	200	3210	0.10
10	20	200	3210	0.075
20	10	200	3210	0.0424
10	40	400	1605	0.0124
20	20	400	1605	<b>0.0021</b>
30	30	900	713	<b><math>6.4 \times 10^{-8}</math></b>

<sup>a</sup> ‘# windows’ is the number of groups, and ‘# subwindows’ is the number of members for each group. ‘# elements’ is the total number of subwindows (members) in the whole sequence, ‘subwindow size’ is the number of base pairs contained in a subwindow (sequence length divided by the total number of subwindows), and *P* value provides the degree of evidence for rejecting the null hypothesis. Significant results at the 0.01 significance level (i.e.  $P < 0.01$ ) are given in bold font.

(Fig. 4) illustrates clearly that there are smaller domains with GC contents different from that of the overall sequence. For example, for 46 segmented domains in Fig. 4 with GC% larger than 37% (as a comparison, the overall GC% is around 35%), there is only one relatively large domain: 324.8 kb with GC% = 37.5%. The remaining domains with GC% larger than 37% mostly contain a few hundred or a few thousand bases.

### 3.3. Isochore sequences exhibit $1/f^{0.5-0.7}$ power spectra

Spectral analyses were carried out on GC contents in non-overlapping windows along a DNA sequence. To present a spectral analysis result consistently for sequences of different lengths, we set a fixed number of non-overlapping windows ( $2^{16} = 65,536$ ). For the three isochore sequences (MHC class III, MHC class II, and chromosome 21), a GC content value is obtained from a window with 9.80, 13.75, and 108.40 bp, respectively. Although the number of spectral components is equal to the number of GC content measurements, only half of them are not redundant. This leaves  $65,536/2 = 32,768$  spectral components. To smooth a noisy spectrum, 16

Table 2  
ANOVA test results for MHC class II isochore sequence ( $N = 900,941$ )<sup>a</sup>

# windows	# subwindows	Total # elements	Subwindow size (bp)	<i>P</i> value
2	10	20	45,047	0.079
4	20	80	11,262	0.016
10	10	100	9009	0.099
5	30	150	6006	0.0102
2	100	200	4504	<b>0.0069</b>
20	10	200	4504	<b>0.0028</b>
15	15	225	4004	<b>0.000224</b>

<sup>a</sup> See footnote to Table 1.

Table 3  
ANOVA test results for the human chromosome 21 isochore sequence ( $N = 7,104,025$ )<sup>a</sup>

# windows	# subwindows	Total # elements	Subwindow size (bp)	<i>P</i> value
2	10	20	355,201	0.52
10	10	100	71,040	0.053
3	40	120	59,200	0.40
4	30	120	59,200	<b>0.0081</b>
2	100	200	35,520	0.32
4	50	200	35,520	<b>0.0056</b>
20	10	200	35,520	<b>0.00065</b>

<sup>a</sup> See footnote to Table 1.

neighboring spectral components are added to form one ‘smoothed’ spectral component. These 32,768/16 = 2048 spectral components are plotted in Fig. 5 (in log-log scale) for the three isochore sequences. The spectra of MHC class III and MHC class II are very close to each other. To show them more clearly, the spectrum value of the class II sequence is multiplied by 0.5.

The spectra of both MHC isochores exhibit the  $1/f^\alpha$  shape with  $\alpha \approx 0.7$ . No attempt was made to determine the value of exponent  $\alpha$  more accurately. But when a  $1/f^{0.7}$  line is drawn side-by-side with the two spectra, the similar trend is clearly visible. The  $1/f^\alpha$  functional shape spans more than three decades (from 100 kb to 100 bp). Although  $1/f^\alpha$  power spectra in DNA sequences were previously reported (Li, 1992; Li and Kaneko, 1992; Voss, 1992; Li et al., 1998; Vieira, 1999), for the first time, here we show the presence of  $1/f^\alpha$  spectra in individual isochore sequences. Besides the general ‘background’ of  $1/f^\alpha$ , the spectra of two MHC isochore sequences also contain a peak around 160–180 bp. This periodicity may be related to the ‘spacings’ between nucleosomes that range from 170 to 260 bp (Hewish and Burgoyne, 1973; Calladine and Drew, 1992), or less interestingly, related to a tandem repeat. Note that in Audit et al. (2001), an attempt was made to relate the  $1/f^\alpha$  spectrum, not the 170–260 bp periodicity, to the nucleosome structure.

The spectrum of the chromosome 21 isochore is not as ‘good’ a  $1/f^\alpha$  spectrum as those of the two MHC isochore sequences. In particular, the low-frequency spectral components follow a flattened trend (‘white noise’), indicating a loss of statistical correlation at longer length scales. However, for middle-ranged frequencies (from 10 to 1 kb), the shape of the spectrum is approximately  $1/f^{0.5}$ . Because of our particular choice of the window size, and due to the number of spectral components being averaged, the smallest periods (highest frequencies) displayed in Fig. 5 are 19.6 bp (MHC class III), 27.5 bp (MHC class II), and 216.9 bp (chromosome 21 isochore).

Besides the power spectra of GC content fluctuations, Fig. 5 also shows those of the single base composition fluctuations. The dots in Fig. 5 are the sum of four spectral components for the compositional fluctuations of each

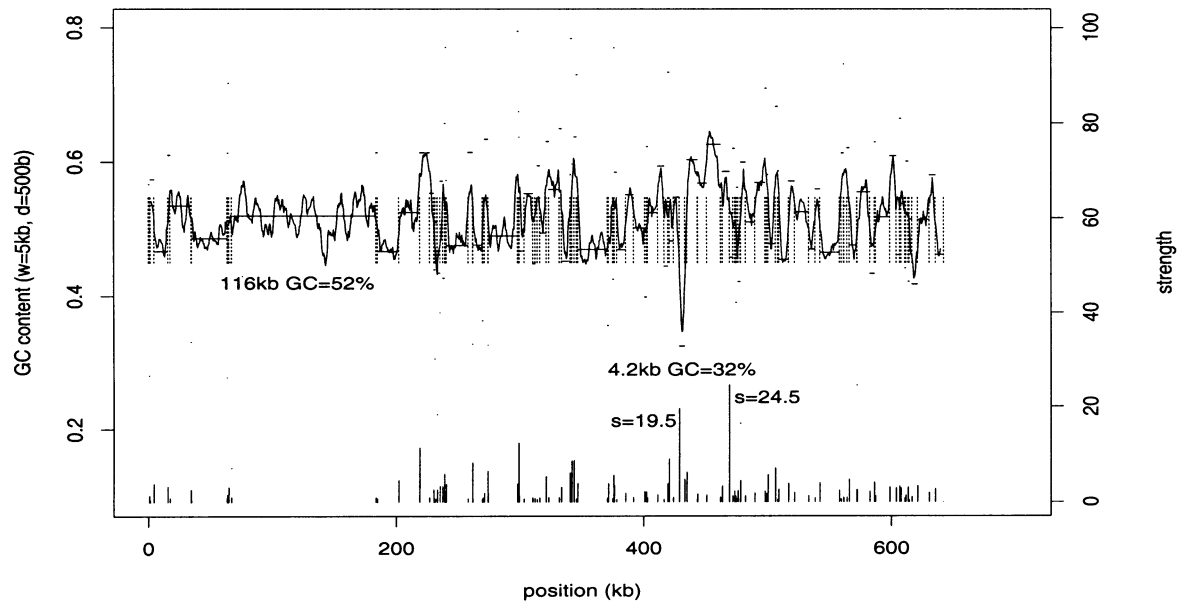
Segmentation of MHC class III ( $s_0=0.333$ , 138 domains)

Fig. 2. Segmentation result of MHC class III sequence (with the threshold for segmentation strength set at  $s_0 = 0.333$ ). The curve shows the GC content within a moving window (the window length is 5 kb, the moving distance is 0.5 kb). Vertical dash lines indicate the positions of segmentation points, whose segmentation strengths are shown below. Horizontal lines indicate the GC content of segmented domains. The total number of domains is 138.

nucleotide. A comparison between the GC content power spectra and single base content power spectra shows that the  $1/f^\alpha$  shape is more flat for the latter than the former. It might be explained by the extra long-range statistical correlation between G and C, or between A and T. These correlations contribute to the GC content power spectra, but not to the single base content power spectra.

#### 4. Discussions and conclusions

##### 4.1. Comparison to the conclusion in [Lander et al. \(2001\)](#)

In this paper, we do not argue the semantic meaning of the word 'isochore'. We only take what are considered to be isochore sequences, i.e. MHC class III, MHC class II, and a

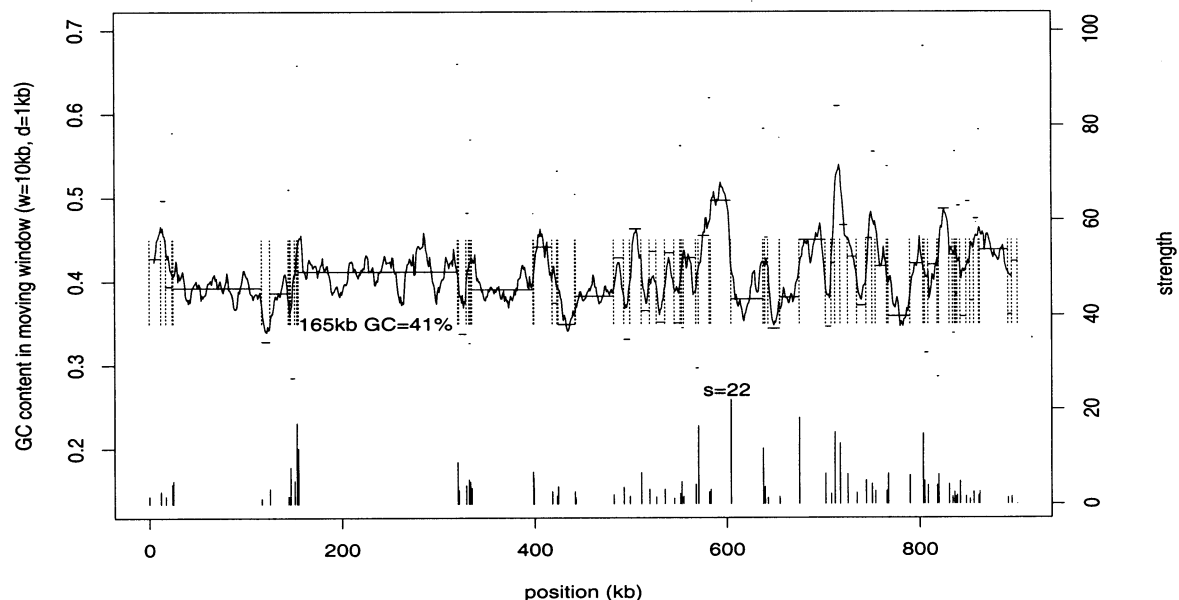
Segmentation of MHC class II ( $s_0=1$ , 83 domains)

Fig. 3. Segmentation result of MHC class II sequence with the threshold for segmentation strength  $s_0 = 1$ . See the caption for Fig. 2 for explanations of the plot.

### Segmentation of human ch21 ( $s_0=1$ , 92 domains)

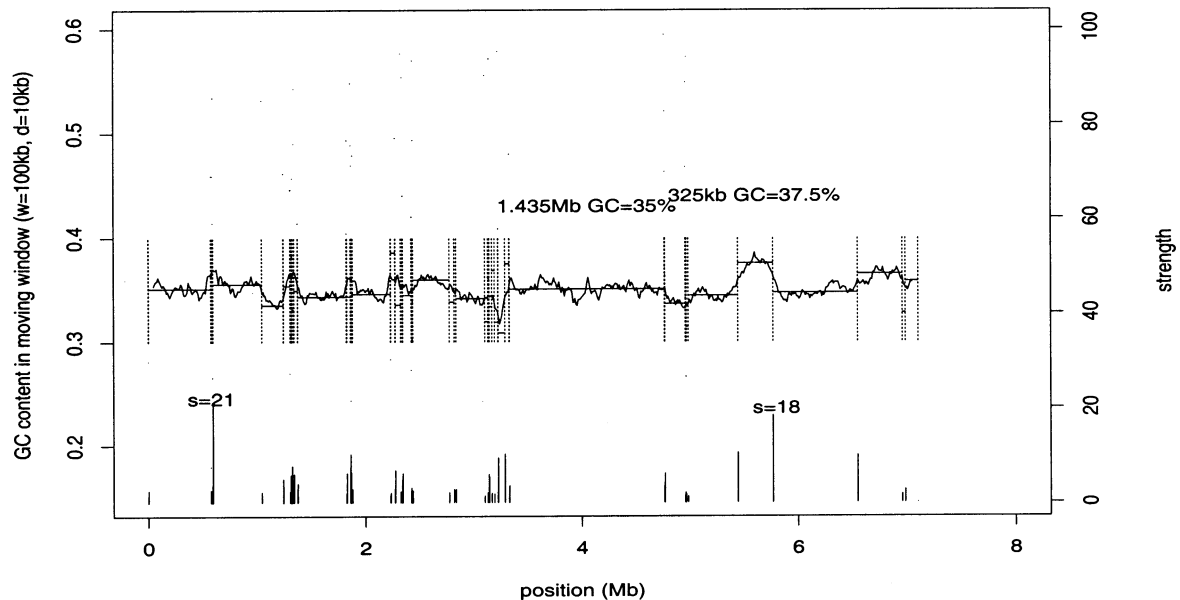


Fig. 4. Segmentation result of the human chromosome 21 isochores sequence with the threshold for segmentation strength  $s_0 = 1$ . See the caption for Fig. 2 for explanations of the plot.

GC-poor isochores on human chromosome 21, and apply various statistical analyses to these sequences. The three analyses, ANOVA, segmentation, and power spectra, reveal a consistent picture of isochores sequences: they are relatively more homogeneous compared to the rest of the genome, but very likely they will fail various homogeneity tests if the GC content fluctuation at the short length scales is considered.

The ANOVA test result shows that the three isochores sequences can be considered to be homogeneous when GC content fluctuations are measured at subwindow sizes larger than a certain value. These subwindow sizes are around 2–3, 5–6, and 30–60 kb for the MHC class III, class II, and chromosome 21 isochores, respectively. When GC contents are calculated from smaller subwindow sizes, ANOVA tests tend to, but do not always, reject this null hypothesis. Note that even with a fixed subwindow size, the choice of windows, or the allocation of members in groups, may still affect the test result, as seen in Tables 1–3.

Our conclusion, that the null hypothesis (homogeneity) is not rejected when subwindows are large enough, is in contrast with the conclusion in the paper by Lander et al. (2001), i.e. that “the hypothesis of homogeneity could be rejected for each 300-kb window in the [human] draft genome sequence”. The reason for this difference is because Lander et al. (2001) did not select a correct distribution for GC content as well as a correct statistical test. In Lander et al. (2001), the GC content of a window is assumed to follow a binomial distribution. It is well known that, unlike normal distribution, the mean and the variance of a binomial distribution are related. In other words, the variance of the

distribution is fixed once the mean value is determined. For a binomial distribution to be true, one has to assume DNA sequences to be uncorrelated random sequences.

Since DNA sequences are not random sequences, it is incorrect to use the binomial distribution to estimate the variance for GC content in a window. The correct method is to use more than one window to estimate the variance. This multi-window design naturally leads to the ANOVA test. Although in Lander et al. (2001), variances of GC content at 20 and 300 kb windows were calculated, these were never used in a statistical test, such as the ANOVA test. Also, even without a test, the conclusion that “three-quarters of the genome-wide variance among 20-kb windows can be statistically explained by the average GC content of 300-kb windows that contain them” is based on a specific length scale, i.e. 20 and 300 kb. It is expected that the percentage value (3/4) could be changed if the window sizes are not 20 and 300 kb. More discussions on the analysis in Lander et al. (2001) will be presented elsewhere (Li et al., in preparation).

#### 4.2. Some details about the ANOVA test

The windows and subwindows are all chosen to have equal sizes in this paper. It is possible that some GC content fluctuations are not captured because they are averaged out by being in the same subwindow. One window partitioning scheme to avoid this problem is to segment the DNA sequence at two different stringency conditions: first a more stringent criterion so large windows are generated, then a more relaxed criterion so smaller subwindows are created. The resulting windows or subwindows may not have the

same sizes. This two-level segmentation test is discussed in Oliver and Li (1998). It is also possible to segment DNA sequences once at a given stringency level to obtain windows, then partition windows into subwindows of equal sizes.

The  $P$  value for an ANOVA test is determined by a known distribution ( $F$  distribution). For this distribution to be true, several conditions need to be satisfied, including the normal distribution of samples in a group, equal variance in different windows, independent samplings of members in a group, and independent groups. Due to the long-range correlation in DNA sequences, GC contents obtained from neighboring subwindows are not independent, and those from neighboring windows are also correlated (Clay et al., 2001; Clay, 2001). It is not clear how these correlations might affect the  $P$  value. If these correlations introduce an error in the  $P$  value, it is also unclear whether the error is always biased in one direction or randomly in both directions.

#### 4.3. Different levels of segmentation strength and isochores

The segmentation results in Figs. 2–4 show that at short length scales, such as individual bases or a few kb, perhaps no long (e.g. 300 kb) DNA sequences can be considered to be homogeneous. A DNA sequence passes the change-point test (by the BIC criterion) if the strongest segmentation has strength  $s > 0$ . The largest segmentation strengths of the three isochore sequences, however, are 24.5, 22, and 21. Not only are these strong segmentation signals, but also there are up to hundreds of other segmentations that are, though weaker, significant (i.e.  $s > 0$ ).

Interestingly, as the segmentation strengths at isochore borders ( $s = 237, 171, 288, 148$  and  $188$ , see Fig. 1) are compared to the maximum segmentation strengths within isochores ( $s = 24.5, 22$  and  $21$ , see Figs. 2–4), the concept of an isochore actually makes sense. It is because there is an order of magnitude difference between the two sets of

### Power spectra of 3 isochores (line: W/S, dot: A/C/G/T)

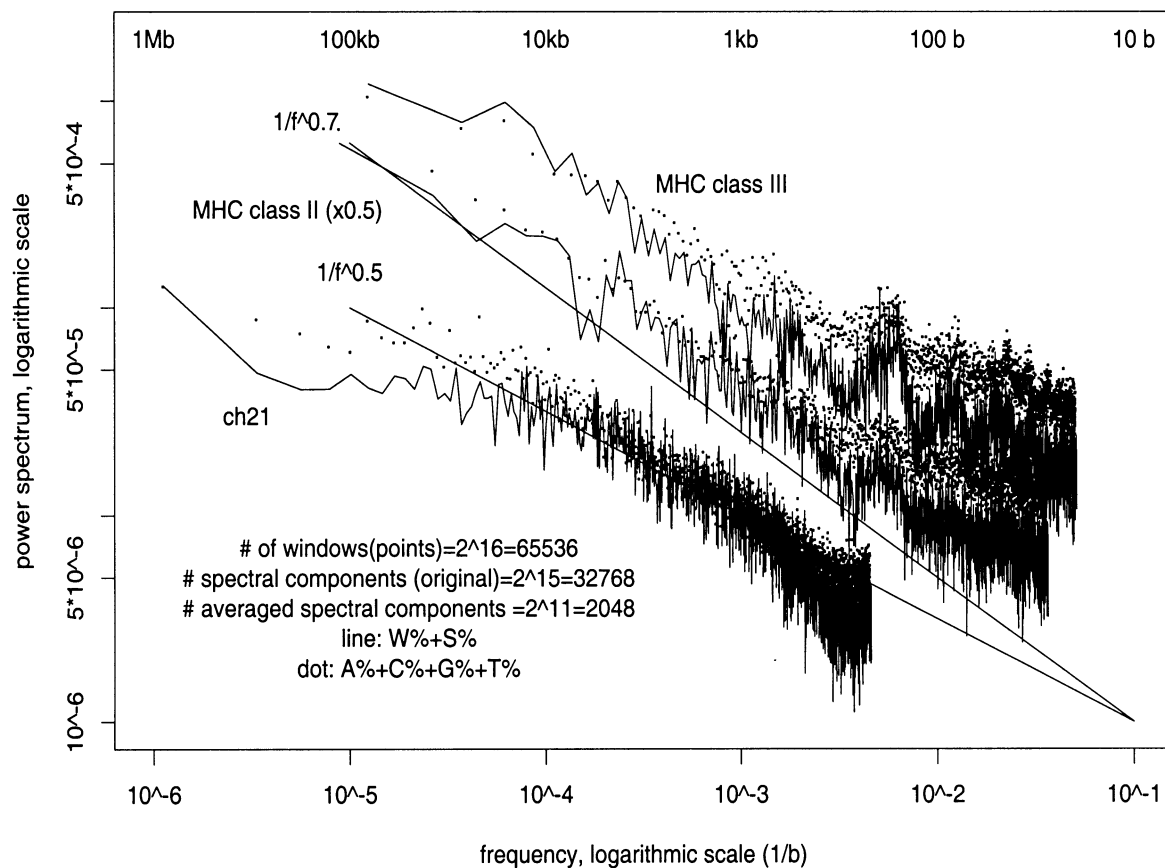


Fig. 5. Power spectra of base composition fluctuation of three isochore sequences. The spectrum of the MHC class II sequence is multiplied by a factor of 0.5 so that it is separated from the MHC class III spectrum. Each isochore sequence is partitioned into 65,536 windows, from which 65,526 base compositions are obtained (e.g. A%, C%, ..., GC%, AT%). Only half of these 65,536 spectral components (32,768) are non-redundant. Sixteen neighboring spectral components are added, and their central location is used as the new frequency value (2048 of them). Solid lines represent the sum of power spectra of GC% and AT% fluctuations; dots represent the sum of power spectra of four nucleotide base fluctuations. Since power spectra are drawn in log-log scale, a straight line is equivalent to a power-law function  $1/f^\alpha$  ( $f$  is the frequency). A  $1/f^{0.5}$  and a  $1/f^{0.7}$  function is plotted as a baseline functional shape. Length scales (one over the frequencies) are indicated at the top of the plot.

segmentation strengths. Of course, we are using the best examples here because the three isochore sequences are well acknowledged and their existence is firmly established. In some other genomic regions such as human chromosome 22, well established isochores may be more difficult to find (Nekrutenko and Li, 2000; Li, 2001c).

#### 4.4. Small-scale base composition fluctuation and relative nature of homogeneity tests

Figs. 2–4 also show that one factor that contributes to the GC content heterogeneity within an isochore sequence is the existence of smaller domains with higher or lower GC% (a few hundred or a few thousand bp). These short length scale deviations may not be visible in the drawing of GC% within a large moving window, but they can be detected by the recursive segmentation procedure. We may want to remove these small domains first so that other factors that contribute to the heterogeneity can be studied separately.

The ANOVA results in Tables 1–3 are also a good illustration of the relative nature of the homogeneity test. In an extreme, when GC contents are more or less averaged out in large windows, any DNA sequence can be homogeneous. In another extreme, when GC content fluctuations at very small length scales (including individual bases) are ‘magnified’, almost any DNA sequence can be heterogeneous. Without specifying the length scales, the debate on whether a DNA sequence is homogeneous or not will not end.

#### 4.5. The shape and the area of a power spectrum

The spectra of isochore sequences in Fig. 5 represents a different characterization of the GC content fluctuation. The functional shape of a power spectrum is related to the distribution of length scales at which GC% fluctuation occurs, whereas the area underneath the function, which is equal to the total variance (power) of the sequence, is related to the amount of GC% fluctuation. Because of the separated roles of shape and area, one has to be careful in relating the functional form of a power spectrum to heterogeneity. For example, a non-white power spectrum may not be used as an evidence for heterogeneity, as in Li et al. (1998).

It is well known that a random sequence exhibits a flat power spectrum (white noise). Any introduction of non-randomness or correlations will change the shape of the power spectrum. But does the change of the shape of the spectrum increase the total variance, which is the area underneath the spectrum? This question can be discussed from two different perspectives. In the first argument, the shape of a function and the area underneath the function are totally independent. If this argument is true, we can, at least in principle, introduce non-randomness into a sequence without increasing the overall heterogeneity (as measured by the sequence variance).

In a second argument, if we restrict our spectrum to a specific functional form, then indeed change in the shape of

the spectrum may change the total variance. As we know by now (Li, 1992; Li and Kaneko, 1992; Voss, 1992; Li et al., 1998; Vieira, 1999), low-frequency  $1/f^\alpha$  power spectra are rather common for long DNA sequences, even though the spectral shape at high frequencies may vary from one sequence to another. Because of the increase of the spectrum at low frequencies, the area underneath the spectrum will also increase. In fact, if  $1/f^\alpha$  is the true functional form of the power spectrum, the total variance diverges to infinity – an extreme example that supports the existence of a relationship between shapes and areas.

The exponent  $\alpha$  in a  $1/f^\alpha$  spectrum may also affect the total variance. The larger the  $\alpha$  value, the more severe the low-frequency divergence. Interestingly, the range for  $\alpha$ , somewhere around 0.5–0.7 as seen in Fig. 5, is consistent with the original finding on how the standard deviation of GC% in DNA fragments changes with the fragment sizes (Cuny et al., 1981). As discussed in detail in Clay et al. (2001) and Clay (2001), if the standard deviation of GC% decreases with the fragment length  $l$  as  $1/l^\beta$ , then the exponent  $\alpha$  in  $1/f^\alpha$  power spectra should be  $\alpha = 1 - 2\beta$  (random sequences have  $\beta = 1/2$  and  $\alpha = 0$ ). The range of values for  $\beta$  was determined experimentally to be 0.15–0.3 (Cuny et al., 1981; cf. Clay et al., 2001). This translates to the range for  $\alpha$  to be 0.4–0.7, which is consistent with the results in Fig. 5.

#### 4.6. Conclusions

In conclusion, we present a more complete picture of the isochore sequence than that provided in Lander et al. (2001). The spectral analysis shows that there are fluctuations of GC content at various length scales for isochore sequences. These multi-scaled fluctuations make a test of homogeneity, such as ANOVA, sensitive to the length scale at which GC contents are sampled. Despite the relative nature of homogeneity or heterogeneity, we recognize that the concept of isochore is reasonable as shown by the segmentation analysis. In this analysis, the signal for a change-point is much stronger at the isochore borders than within the sequence, justifying the use of the prefix ‘iso’, in a relative sense, for these three sequences.

#### Acknowledgements

I would like to thank Giorgio Bernardi for inviting me to the 5th Anton Dohrn Workshop at Ischia, which motivated the analyses discussed in this paper. The financial support from the Anton Dohrn Workshop is acknowledged. I also thank Oliver Clay for continuous discussions on this topic, Yaning Yang for assistance on the ANOVA analysis, and Sabyasachi Guharay for his participation in the initial stage of the spectral analysis. Partial support from NIH contract N01-AR12256 is acknowledged.

## References

- Audit, B., Thermes, C., Vaillant, C., d'Aubenton-Carafa, Y., Muzy, J.F., Arneodo, A., 2001. Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Phys. Rev. Lett.* 86, 2471–2474.
- The MHC Sequencing Consortium, Beck, S., Geraghty, D., Inoko, H., Rowen, L., et al., 1999. Complete sequence and gene map of a human major histocompatibility complex. *Nature* 401, 921–923.
- Bernaola-Galván, P., Román-Roldán, R., Oliver, J.L., 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys. Rev. E* 53, 5181–5189.
- Bernardi, G., 1995. The human genome: organization and evolutionary history. *Annu. Rev. Genet.* 23, 637–661.
- Bernardi, G., 2001. Misunderstandings about isochores. Part I. *Gene* 276, 3–13.
- Bracewell, R.N., 1965. *The Fourier Transform and Its Applications*, McGraw-Hill, New York.
- Calladine, C.R., Drew, H.R., 1992. *Understanding DNA*, Academic Press, New York.
- Clay, O., 2001. Standard deviations and correlations of GC levels in DNA sequences. *Gene* 276, 33–38.
- 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.
- Csorgo, M., Horvath, L., 1997. *Limit Theorems in Change-Point Analysis*, Wiley, New York.
- 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.
- Dunham, I., Shimizu, N., Roe, B.A., Chisoe, S. et al., 1999. The DNA sequence of human chromosome 22. *Nature* 402, 489–495.
- Fisher, R.A., 1925. *Statistical Methods for Research Workers*, Oliver and Boyd, London.
- Fourier, J., 1822. *Théorie Analytique de la Chaleur* (Paris); translated in 1878 by Alexander Freeman, *The Analytical Theory of Heat*. Cambridge University Press, Cambridge.
- 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.
- Fukagawa, T., Nakamura, Y., Okumura, K., Nogami, M., Ando, A., Inoko, H., Saito, N., Ikemura, T., 1996. Human pseudoautosomal boundary-like sequences: expression and involvement in evolutionary formation of the present-day pseudoautosomal boundary of human sex chromosomes. *Hum. Mol. Genet.* 5, 23–32.
- Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J.L., Stanley, H.E., 2002. Analysis of symbolic sequences using the Jensen-Shannon divergence measure. *Phys. Rev. E* 65, 041905.
- The Chromosome 21 Mapping and Sequencing Consortium, Hattori, M., Taudien, S., Kudoh, J., Nordsiek, G., Ramser, J., et al., 2000. The DNA sequence of human chromosome 21. *Nature* 405, 311–319.
- Hewish, D., Burgoyne, L., 1973. Chromatin sub-structure: the digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochem. Biophys. Res. Commun.* 52, 504–510.
- Horvath, A.L., 1989. The limit distributions of likelihood ratio and cumulative sum tests for a change in a binomial probability. *J. Multivar. Anal.* 31, 148–159.
- International Human Genome Sequencing Consortium, Lander, E.S., Waterston, R.H., Sulston, J., Collins, F.S., et al., 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Li, W., 1992. Generating non-trivial long-range correlations and  $1/f$  spectra by replication and mutation. *Int. J. Bifurcation Chaos* 2, 137–154.
- Li, W., 1997a. The study of correlation structures of DNA sequences – a critical review. *Comput. Chem.* 21, 257–271. special issue on open problems of computational molecular biology.
- Li, W., 1997b. The complexity of DNA: the measure of compositional heterogeneity in DNA sequences and measures of complexity. *Complexity* 3, 33–37.
- Li, W., 2001a. New stopping criteria for segmenting DNA sequences. *Phys. Rev. Lett.* 86, 5815–5818.
- Li, W., 2001b. DNA segmentation as a model selection process. *Proceedings of the Fifth Annual International Conference of Computational Molecular Biology (RECOMB 2001)*, pp. 210–216.
- Li, W., 2001c. Delineating relative homogeneous G + C domains in DNA sequences. *Gene* 276, 57–72.
- Li, W., Kaneko, K., 1992. Long-range correlation and partial  $1/f$  spectrum in a noncoding DNA sequence. *Europhys. Lett.* 17, 655–660.
- Li, W., Marr, T., Kaneko, K., 1994. Understanding long-range correlations in DNA sequences. *Phys. D* 75, 392–416.
- Li, W., Stolovitzky, G., Bernaola-Galván, P., Oliver, J.L., 1998. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. *Genome Res.* 8, 916–928.
- Li, W., Bernaola-Galván, P., Haghghi, F., Grosse, I., 2002. Applications of recursive segmentation to the analysis of DNA sequences. *Comput. Chem.* 26, 491–510.
- Macaya, G., Thiery, J.P., Bernardi, G., 1976. An approach to the organization of eukaryotic genomes at the macromolecular level. *J. Mol. Biol.* 108, 237–254.
- McCullagh, P., Nelder, J.A., 1983. *Generalized Linear Models*, Chapman and Hall, London.
- Nekrutenko, A., Li, W.H., 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.* 10, 1986–1995.
- Oliver, J.L., Li, W., 1998. Quantitative analysis of compositional heterogeneity in long DNA sequences: the two-level segmentation test (abstract). *Genome Mapping, Sequencing & Biology*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, p. 163.
- Oliver, J.L., Román-Roldán, R., Perez, J., Bernaola-Galván, P., 1999. SEGMENT: identifying compositional domains in DNA sequences. *Bioinformatics* 15, 974–979.
- Oliver, J.L., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. *Gene* 276, 57–72.
- Oliver, J.L., Carpena, P., Román-Roldán, R., Mata-Balaguer, T., Mejias-Romero, A., Hackenberg, M., Bernaola-Galván, P., 2002. Isochore chromosome maps of long human contigs. *Gene* this issue.
- Pavliček, A., Pačes, J., Clay, O., Bernardi, G., 2002. A compact view of isochores in the draft human genome sequence. *FEBS Lett.* 511, 165–169.
- Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simon, M., Stanley, H.E., 1992. Long-range correlations in nucleotide sequences. *Nature* 356, 168–170.
- Román-Roldán, R., Bernaola-Galván, P., Oliver, J.L., 1998. Sequence compositional complexity of DNA through an entropic segmentation method. *Phys. Rev. Lett.* 80, 1344–1347.
- Saccone, S., Pavliček, A., Federico, C., Pačes, J., Bernardi, G., 2001. Gene, isochores and bands in human chromosomes 21 and 22. *Chromosome Res.* 9, 533–539.
- Schwarz, G., 1978. Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.
- Stephens, R., Horton, R., Humphray, S., Rowen, L., Trwosdale, J., Beck, S., 1999. Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *J. Mol. Biol.* 291, 789–799.
- Vieira, M., 1999. Statistics of DNA sequences: a low frequency analysis. *Phys. Rev. E* 60, 5932–5937.
- Voss, R., 1992. Evolution of long-range fractal correlations and  $1/f$  noise in DNA base sequences. *Phys. Rev. Lett.* 68, 3805–3808.