Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 311 (2003) 215–222
BBRC
www.elsevier.com/locate/ybbrc

Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis☆

Sheng-Yun Wen and Chun-Ting Zhang*

Department of Physics, Tianjin University, Tianjin 300072, PR China

Received 16 September 2003

Abstract

Incorporated with the Z curve method, the technique of wavelet multiresolution (also known as multiscale) analysis has been proposed to identify the boundaries of isochores in the human genome. The human MHC sequence and the longest contigs of human chromosomes 21 and 22 are used as examples. The boundary between the isochores of Class III and Class II in the MHC sequence has been detected and found to be situated at the position 2,490,368 bp. This result is in good agreement with the experimental evidence. An isochore with a length of about 7 Mb in chromosome 21 has been identified and found to be gene- and Alu-poor. We have also found that the G + C content of chromosome 21 is more homogeneous than that of chromosome 22. Compared with the window-based methods, the present method has the highest resolution for identifying the boundaries of isochores, even at a scale of single base. Compared with the entropic segmentation method, the present method has the merits of more intuitiveness and less calculations. The important conclusion drawn in this study is that the segmentation points, at which the G + C content undergoes relatively dramatic changes, do exist in the human genome. These 'singularity' points may be considered to be candidates of isochore boundaries in the human genome. The method presented is a general one and can be used to analyze any other genomes.

Keywords: Human genome; Isochores; Z curve; Wavelet multiresolution analysis; Segmentation points

The mosaic organization of mammalian genomes composed of many regions of rather homogeneous G + C content was revealed by the ultracentrifugation experiments of bulk DNA in the mid-1970s [1–4]. The long DNA segments (≫300 Kb, on average) of fairly homogeneous G + C content lately were given the name 'isochores' [5]. According to Bernardi's analysis, there are five isochore families. Two are G + C-poor isochore families L1 (G + C < 38%) and L2 (38% ⩽ G + C < 44%). The other three are G + C-rich isochore families H1 (44% ⩽ G + C < 48%), H2 (48% ⩽ G + C < 52%), and H3 (G + C ⩾ 52%) [3,6]. Nowadays the availability of the human genome draft sequences offers an unprecedented opportunity to explore and understand the genomic organization at the sequence level.

The traditional methods for analyzing the G + C content use the technique of overlapping or non-overlapping moving-window. For example, the window-based methods were used to study the sequence heterogeneity [7,8]. The main disadvantage of such methods is that they cannot detect boundaries of isochores precisely. Especially, the window-based methods failed to precisely detect the unique boundary of isochores that has been experimentally characterized so far [9–12]. In addition, some statistical analyses based on such methods resulted in some debates and misunderstandings on isochores [8,13,14]. The essence of the debates is how to define the homogeneity of G + C content within a window or a fragment of genome sequences.

Recently a windowless technique (cumulative GC profile) has been proposed to calculate the G + C content and detect the isochore boundaries in the human genome [15,16]. As pointed out in [13,16], the homogeneity of G + C content of the human genome should be considered as relative. Boundaries of isochores seem to be more important than the homogeneity of G + C content within isochores in some sense. In this paper, a new algorithm is proposed, which combines the wavelet

☆ *Abbreviation*: MHC, major histocompatibility complex.

* Corresponding author. Fax: +86-22-2740-2697.
E-mail address: ctzhang@tju.edu.cn (C.-T. Zhang).

0006-291X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2003.09.198

multiresolution analysis and the cumulative GC profile, to precisely detect boundaries of isochores in the human genome. The wavelet multiresolution analysis is a new technique for studying non-stationary signals. The technique of wavelet transform is widely applied in many fields including image processing, signal analysis, and geophysics as well as bioinformatics [17–20]. Compared with the window-based approaches, the present algorithm detects isochore boundaries more accurately and easily.

Materials and methods

The draft sequence of the human genome and the complete sequence of the human major histocompatibility complex (MHC) were downloaded from the websites http://genome.ucsc.edu/ and http://www.sanger.ac.uk/HGP/Chr6/, respectively.

The cumulative GC profile and its derivative. The G + C content is a statistical quantity of biological importance. Usually it is calculated within a window of sufficient size. However, as pointed out previously [16], the window-based method is not applicable in the study of isochores. To solve the problem, a windowless technique to calculate the G + C content was proposed [15], which is derived from the Z curve method [21,22]. Based on the Z curve, any DNA sequence can be uniquely described by three independent distributions, i.e., x_n, y_n, and z_n. In particular, z_n displays the distribution of bases of GC/AT types along the sequence, which is calculated as follows [21,22]

$$z_n = (A_n + T_n) - (C_n + G_n), \quad n = 0, 1, 2, \ldots, N, \;\; z_n \in [-N, N], \tag{1}$$

where A_n, C_n, G_n, and T_n are the *cumulative* numbers of the bases A, C, G, and T, respectively, occurring in the subsequence from the first base to the nth base in the DNA sequence inspected. $A_0 = C_0 = G_0 = T_0 = 0, z_0 = 0$.

For almost all genome or chromosome sequences, the curves of $z_n \sim n$ are roughly straight lines (data not shown here). To amplify the variations of the curve, first of all, the curve of $z_n \sim n$ is fitted by a straight line using the least square technique,

$$z = kn, \tag{2}$$

where (z, n) is the coordinate of a point on the fitted straight line and k is its slope. Instead of using the curve of $z_n \sim n$, we will use the $z'_n \sim n$ curve, or simply z' curve hereafter, where

$$z'_n = z_n - kn. \tag{3}$$

Therefore, the variations of $z_n \sim n$ curve deviated from the straight line, which corresponds to a constant G + C content (see Eq. (4) below), are protruded by the $z'_n \sim n$ curve. The z' curve or the cumulative GC profile [23] is used interchangeably in this paper. Let $\overline{\mathrm{G+C}}$ denote the average G + C content within a region Δn in a sequence. It was shown that [15]

$$\overline{\mathrm{G+C}} = \frac{1}{2}\left(1 - k - \frac{\Delta z'_n}{\Delta n}\right) = \frac{1}{2}(1 - k - k'_n), \tag{4}$$

where $k'_n = \Delta z'_n/\Delta n$ is the average slope of the z' curve within the region Δn. When $\Delta n \to 0, k'_n \to \mathrm{d}z'_n/\mathrm{d}n$, it is difficult to calculate, because z'_n has definition only at $n = 1, 2, \ldots, N$, where N is the length of the sequence being studied. Consequently, the following difference formula is used to approach $\mathrm{d}z'_n/\mathrm{d}n$

$$k'_n = \frac{(z'_{n+1} - z'_n) + (z'_n - z'_{n-1})}{2} = \frac{(z'_{n+1} - z'_{n-1})}{2}. \tag{5}$$

We define $z'_{-1} = 2z'_0 - z'_1 = -z'_1$ and $z'_{N+1} = 2z'_N - z'_{N-1}$. The difference k'_n calculated in Eq. (5) will be used to approximate the derivative of the cumulative GC profile in the study below. Note that the cumulative GC profile (or the z' curve) is not the G + C content itself. Rather, the derivative of z' with respect to the base position n is negatively proportional to the G + C content at the given position, i.e., $\mathrm{G+C} \propto -\mathrm{d}z'/\mathrm{d}n$ or $\mathrm{G+C} \propto -k'_n$. Therefore, $k'_n > 0$ indicates a decrease of the G + C content, whereas $k'_n < 0$ indicates an increase of the G + C content. For convenience, k'_n is normalized such that $k'_n \in [0, 1]$ hereafter. Based on the properties of the Z curve [22], it is deduced that k'_n takes three values only, i.e., $k'_n \in \{0, 1/2, 1\}$. It can be derived from Eq. (4) that $(\mathrm{G+C})_n = 1 - k'_n$, where $(\mathrm{G+C})_n$ is the G + C content at the region between the positions of n and $n + \Delta n$, and k'_n is normalized. A wavelet multiresolution analysis is performed to $(\mathrm{G+C})_n = 1 - k'_n$ below.

The wavelet multiresolution analysis. A discrete dyadic wavelet transform can be calculated using a fast filter bank algorithm proposed by Mallat [17]. Suppose that $\{h_n, n \in Z\}$ is a low-pass filter and $\{g_n, n \in Z\}$ is a high-pass filter, where Z is the set of integers. Let $\phi(t)$ and $\psi(t)$ be the scaling function and wavelet function, respectively. The two-scale relations read

$$\phi(t) = \sum_n h_n \sqrt{2}\phi(2t - n), \tag{6}$$

$$\psi(t) = \sum_n g_n \sqrt{2}\phi(2t - n). \tag{7}$$

Let $f(t) \in L^2(R)$ be a function of t. Let $a_j[n], j \geqslant 0$ and $d_j[n], j > 0$ denote the coefficients of scaling and wavelet transforms of $f(t)$ at scale j, respectively, then

$$a_j[n] = \int_{-\infty}^{\infty} f(t) 2^{-j/2} \phi(2^{-j}t - n)\,\mathrm{d}t, \quad j \geqslant 0, \tag{8}$$

$$d_j[n] = \int_{-\infty}^{\infty} f(t) 2^{-j/2} \psi(2^{-j}t - n)\,\mathrm{d}t, \quad j > 0. \tag{9}$$

Using the orthonormality of the scaling and wavelet functions, we have

$$a_{j+1}[n] = \sum_k h_{k-2n} a_j[k], \quad j \geqslant 0, \tag{10}$$

$$d_{j+1}[n] = \sum_k g_{k-2n} a_j[k], \quad j \geqslant 0. \tag{11}$$

Repeatedly using Eqs. (10) and (11), one can decompose $a_0[n]$ into $a_1[n]$ and $d_1[n], a_1[n]$ into $a_2[n]$ and $d_2[n], \ldots$, and so forth until $a_{J-1}[n]$ into $a_J[n]$ and $d_J[n]$, where J is the terminal scale of the decomposition. Therefore, $a_0[n]$ can be expressed by $a_J[n]$ and $d_1[n], \ldots, d_J[n]$, or $a_0[n] \Rightarrow \{a_J, d_j[n]|_{1 \leqslant j \leqslant J}\}$. Note that $a_j[n], j \geqslant 0$ and $d_j[n], j > 0$ represent the approximation and the detail components of the signal, respectively.

Conversely, starting from $a_J[n]$ and $d_J[n], a_0[n]$ can be reconstructed by repeatedly using the following reconstruction formula:

$$a_j[n] = \sum_k \{h_{n-2k} a_{j+1}[k] - g_{n-2k} d_{j+1}[k]\}, \quad j \geqslant 0, \tag{12}$$

or $\{a_J, d_j[n]|_{1 \leqslant j \leqslant J}\} \Rightarrow a_0[n]$. What specialty presented in this paper is that at each scale $j, 1 \leqslant j \leqslant J$, a threshold c_j is selected

$$c_j = \sqrt{2 \log_2 L_j}, \tag{13}$$

where L_j is the length of $d_j[n]$. At each scale $j, 1 \leqslant j \leqslant J$ those $d_j[n]$ with their magnitudes greater than c_j are retained, whereas those smaller than c_j are forced to be 0. This is a denoising procedure, after which the reconstructed signal becomes $\bar{a}_0[n]$, rather than $a_0[n]$. The latter is the original signal, while the former is the reconstructed signal with noises reduced. Note that the function $f(t)$ in Eqs. (8) and (9) is generally unknown. The sample of $f(t)$ at $t = n$ is $f[n]$. We set $f[n] \equiv (\mathrm{G+C})_n = 1 - k'_n = a_0[n]$ hereafter. Accordingly, the multiresolution analysis is performed to $(\mathrm{G+C})_n = 1 - k'_n$.

Determination of segmentation points and mergence of regions. Given a sequence, the above multiresolution analysis is applied to $(\mathrm{G+C})_n = 1 - k'_n$ using the Haar wavelet. Starting from the original signal $a_0[n] = 1 - k'_n$, the reconstructed signal $\bar{a}_0[n]$ with noises reduced

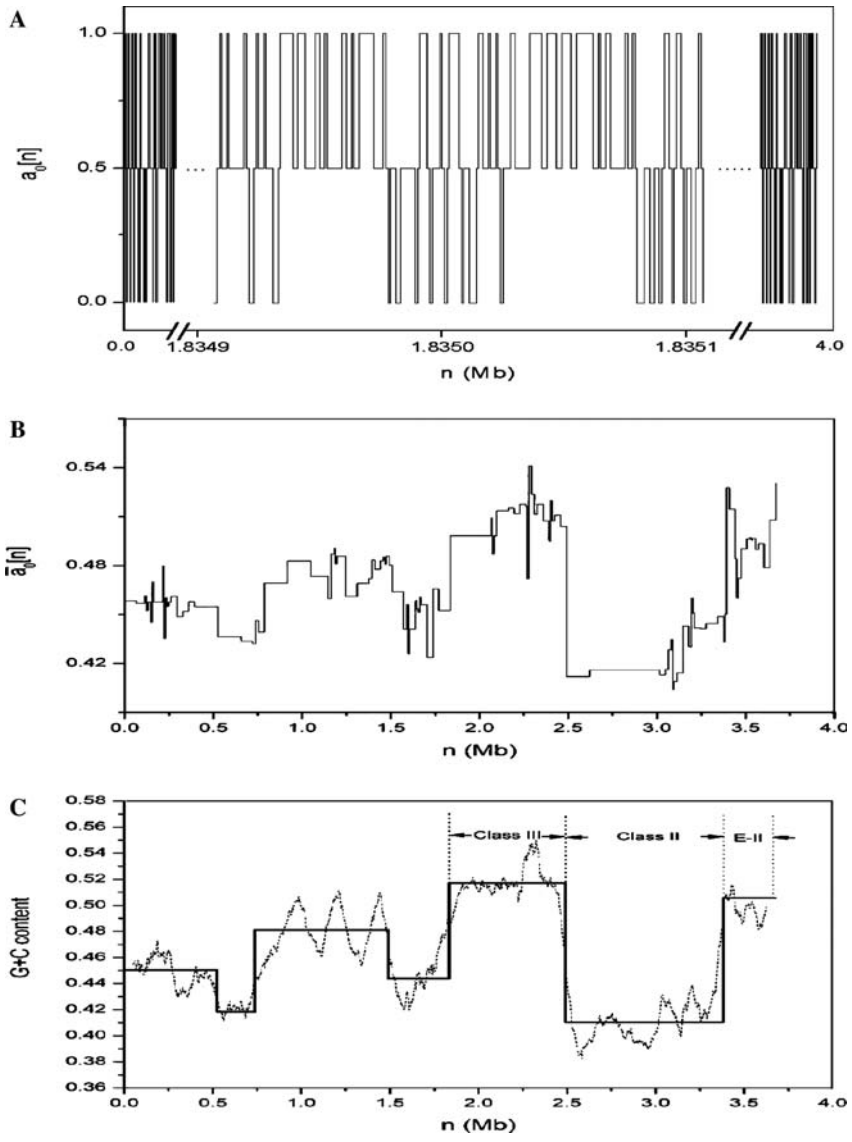

Fig. 1. (A) The original signal for the MHC sequence, $(G+C)_n = 1 - k'_n$, is denoted by $a_0[n]$. Note that the normalized k'_n takes only three values, i.e., $k'_n \in \{0, 0.5, 1\}$, and so is the G + C content, $(G+C)_n$. For clarity, only a small part of the signal is shown. The omitted part is similar to this local signal, but much denser than this. (B) The reconstructed signal $\bar{a}_0[n]$ based on the Haar wavelet multiresolution analysis at the terminal level $J = 34$. The positions at which jumps appear are considered to be the segmentation points of the MHC sequence. Since $(G+C)_n = a_0[n]$, in fact $\bar{a}_0[n]$ shows the distribution of the G + C content along the MHC sequence, but with reduced noises. (C) The distribution of the G + C content along the human MHC sequence (3,673,777 bp) based on the segmentation points shown in (B), but after the merging procedures. The segmentation points situated at the positions 1,835,008, 2,490,368, and 3,383,296 bp correspond to the boundaries between Class I and Class III, Class III and Class II as well as Class II and the Extended Class II regions, respectively. For comparison, the distribution of the G + C content using a moving-window method (window size, 100 Kb, step, 1 Kb) is also shown (dash line). Note that the window-based method cannot precisely detect the boundaries of isochores.

Table 1
The coordinates of six segmentation points detected by the method of wavelet multiresolution analysis for the human MHC sequence[a]

Start (bp)	End (bp)	Length (bp)	(G + C)%	Family	Notes
1	524,287	524,287	45.03	H1	
524,288	737,279	212,992	41.85	L2	
737,280	1,490,943	753,664	48.11	H2	
1,490,944	1,835,007	344,064	44.40	H1	
1,835,008	2,490,367	655,360	51.71	H2	Class III
2,490,368	3,383,295	892,928	41.04	L2	Class II
3,383,296	3,673,777	290,462	50.56	H2	Extended class II

[a] The sequence length is 3.6 Mb. The terminal scale is at $J = 34$ and the minimum isochore length is taken as 200 Kb. Starting from the first base, the segmentation points are arranged in an ascending order of their coordinates. Refer to Fig. 1C for the distribution of the seven domains.

Table 2
The coordinates of 25 segmentation points detected by the method of wavelet multiresolution analysis for the longest contig NT_011512 of human chromosome 21[a]

Start (bp)	End (bp)	Length (bp)	(G + C)%	Family	Notes
1	671,743	671,743	37.34	L1	
671,744	1138,687	466,944	43.08	L2	
1,138,688	2,097,151	958,464	36.65	L1	
2,097,152	2,621,439	524,288	38.17	L2	
2,621,440	4,325,375	1,703,936	36.00	L1	
4,325,376	4,980,735	655,360	40.38	L2	
4,980,736	12,393,507	7,412,772	35.16	L1	Gene-poor, Alu-poor
12,393,508	13,442,083	1,048,576	39.36	L2	
13,442,084	15,611,975	2,169,892	36.79	L1	
15,611,976	16,922,695	1,310,720	39.24	L2	
16,922,696	17,840,199	917,504	36.55	L1	
17,840,200	18,888,775	1,048,576	42.40	L2	
18,888,776	19,109,959	221,184	45.82	H1	
19,109,960	19,331,143	221,184	42.55	L2	
19,331,144	19,658,823	327,680	46.11	H1	
19,658,824	20,052,039	393,216	41.09	L2	
20,052,040	20,461,639	409,600	45.60	H1	
20,461,640	21,362,759	901,120	42.62	L2	
21,362,760	21,927,019	564,260	44.67	H1	
21,927,020	23,024,747	1,097,728	41.19	L2	
23,024,748	24,106,091	1,081,344	45.77	H1	
24,106,092	25,744,491	1,638,400	41.91	L2	
25,744,492	26,170,475	425,984	45.16	H1	
26,170,476	28,136,555	1,966,080	41.78	L2	
28,136,556	28,390,507	253,952	45.21	H1	
28,390,508	28,602,511	212,004	48.04	H2	

[a] The sequence length is 28 Mb. The terminal scale is at $J = 34$ and the minimum isochore length is taken as 200 Kb. Starting from the first base, the segmentation points are arranged in an ascending order of their coordinates. Refer to Fig. 2 for the distribution of the 26 domains.

can be obtained. We notice that there exist a number of jumps in the reconstructed signal $\bar{a}_0[n]$ (see the next section). The locations at which the jumps are situated are deemed as the segmentation points of the sequence studied. The G + C contents at both sides of any segmentation point have relatively dramatic variations. Given a sequence, the number of segmentation points obtained depends on the choice of the threshold c_j. In the case of c_j calculated in Eq. (13), there are relatively a large number of segmentation points found. A sequence region is formed with any two adjoining segmentation points as its boundaries. To study the isochore structure, two merging procedures are performed for adjoining regions. (i) If the G + C contents of two adjoining regions belong to the same isochore family defined by Bernardi [3], the two regions are merged into a new and larger one. The merging procedure is in an iteration mode. Repeatedly perform the above merging procedure until no further merging task is needed. (ii) If the length of a region after the above merging procedure is less than 200 Kb, merge it with its adjoining region, to form a new and larger one. Repeatedly perform the above merging procedure until the minimum length of regions obtained is greater than 200 Kb.

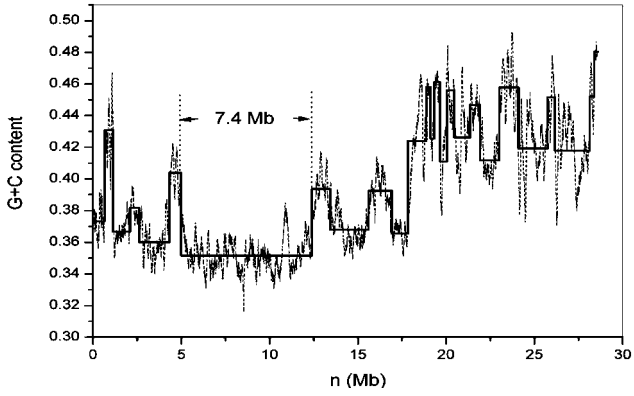

Fig. 2. The distribution of the G + C content along the longest contig (NT_011512) of human chromosome 21 (sequence length = 28,602,511 bp), based on the segmentation points obtained by the technique of wavelet multiresolution analysis, after the merging procedures. The isochore with gene- and Alu-poor has been identified by the present method and found to be at the region between the coordinates of 4,980,736 and 12,393,507 bp. The G + C content of this isochore (with length = 7.4 Mb) is 35.2%, belonging to the L1 isochore family. For comparison, the distribution of the G + C content using a moving-window method (window size, 100 Kb, step, 1 Kb) is also shown (dash line).

Results and discussions

The isochores in the sequence of the human major histocompatibility complex

The human major histocompatibility complex (MHC) sequence situated at human chromosome 6p21 region has been completely sequenced [24]. The MHC plays a key role in some human diseases, most of them being of autoimmune or infectious features. This 3.6 Mb

long sequence codes for 224 genes, some of which have functions related to immune response and participate in diverse pathways, e.g., antigen processing, antigen presentation, and T-cell interaction. Several classes of proteins such as MHC Class I and II proteins are encoded in this region. In the human genome, these proteins are also known as human leukocyte antigens (HLAs). Owing to its importance, the MHC sequence has been subjected to intensive studies [24–28]. It was found that the sequence can be divided into four regions roughly. They are distributed on chromosome 6 in the order of telemere $\Rightarrow$ Class I $\Rightarrow$ Class III $\Rightarrow$ Class II $\Rightarrow$ extended Class II $\Rightarrow$ centromere [24]. The G + C contents of Class II and III are fairly homogeneous and the two regions are thought to be isochores [27,28]. It has been found that the switch point of DNA replication timing is precisely located at the boundary between Class II and Class III isochores [12].

To test the present method, the MHC sequence is analyzed by the multiresolution technique. The original signal $a_0[n] = 1 - k'_n$ is decomposed at different scales until at $J = 34$, using a Haar wavelet. We have tested all the Daubechics wavelets [18,19], db1 $\sim$ db10. Consequently, only the db1 (i.e., the Haar) wavelet results in the best result in the present study. Fig. 1A shows the original signal $a_0[n]$, whereas the reconstructed signal $\bar{a}_0[n]$ is shown in Fig. 1B with the terminal scale $J = 34$. We have found that the segmentation points obtained are not sensitive as $J = 17$–34. When the threshold c_j in Eq. (13) is adopted, 107 segmentation points are found, corresponding to 108 regions in the human MHC sequence. Among them, the shortest region found is

Table 3
The coordinates of 37 segmentation points detected by the method of wavelet multiresolution analysis for the longest contig NT_011520 of human chromosome 22[a]

Start (bp)	End (bp)	Length (bp)	(G + C)%	Family	Notes
1	1,097,727	1,097,727	48.62	H2	
1,097,728	1,507,327	409,600	52.89	H3	
1,507,328	1,736,703	229,376	47.91	H1	
1,736,704	2,031,615	294,912	43.62	L2	
2,031,616	2,785,279	753,664	45.80	H1	
2,785,280	3,055,615	270,336	53.82	H3	
3,055,616	4,063,231	1,007,616	50.45	H2	
4,063,232	4,276,223	212,992	44.56	H1	
4,276,224	4,980,735	704,512	48.22	H2	
4,980,736	5,794,552	813,817	45.39	H1	
5,794,553	6,056,696	262,144	43.40	L2	
6,056,697	7,154,424	1,097,728	45.81	H1	
7,154,425	7,637,752	483,328	49.55	H2	
7,637,753	8,465,144	827,392	39.76	L2	
8,465,145	8,989,432	524,288	45.64	H1	
8,989,433	9,595,640	606,208	49.83	H2	
9,595,641	9,988,856	393,216	41.47	L2	
9,988,857	10,447,608	458,752	51.88	H2	
10,447,609	11,998,705	1,551,097	46.76	H1	
11,998,706	12,285,425	286,720	43.35	L2	Gene-poor, Alu-poor
12,285,426	13,710,833	1,425,408	44.60	H1	Gene-poor, Alu-poor
13,710,834	14,734,833	1,024,000	41.16	L2	Gene-poor, Alu-poor
14,734,834	15,095,281	360,448	45.13	H1	
15,095,282	15,390,193	294,912	51.05	H2	
15,390,194	15,750,641	360,448	40.28	L2	
15,750,642	16,012,785	262,144	46.73	H1	
16,012,786	16,717,297	704,512	49.36	H2	
16,717,298	18,186,475	1,469,178	53.28	H3	
18,186,476	18,432,235	245,760	45.21	H1	
18,432,236	18,825,451	393,216	50.88	H2	
18,825,452	19,439,851	614,400	53.60	H3	
19,439,852	20,660,459	1,220,608	43.40	L2	
20,660,460	20,889,835	229,376	45.34	H1	
20,889,836	21,315,819	425,984	52.01	H3	
21,315,820	21,962,987	647,168	48.01	H2	
21,962,988	22,495,467	532,480	53.00	H3	
22,495,468	22,757,611	262,144	48.53	H2	
22,757,612	23,178,213	420,602	53.27	H3	

[a] The sequence length is 23 Mb. The terminal scale is at $J = 34$ and the minimum isochore length is taken as 200 Kb. Starting from the first base, the segmentation points are arranged in an ascending order of their coordinates. Refer to Fig. 3 for the distribution of the 38 domains.

8192 bp in length. The G + C contents of these regions vary from 45% to 54%. These small regions with large fluctuations of G + C content may be caused by repeats, insertions, etc. After the merging procedures described above, seven regions are retained in the MHC sequence. The coordinates of the boundaries of the seven regions obtained by the present method are listed in Table 1. The distribution of G + C content along the MHC sequence calculated by the present method is shown in Fig. 1C. As can be seen from Fig. 1C and Table 1, the position at 2,490,368 bp corresponds to the boundary of the Class III and Class II regions. The positions at 1,835,008 and 3,383,296 bp correspond to the boundary of Class I and Class II regions and that of Class II and the extended Class II regions, respectively. The G + C contents of Class III, Class II, and the extent Class II regions are 51.7%, 41.0%, and 50.6%, respectively. All of them have relatively sharp boundaries, as reflected by the fact that they correspond to the relatively dramatic jumps in Fig. 1B. The location of the boundary between the Class III and Class II isochores, i.e., 2,490,368 bp, is well consistent with the experimental evidence [9,12].

It should be pointed out that the boundaries found by the present method are in good agreement with those of the recursive entropic segmentation method used by Li [29] and Oliver et al. [27,28]. See the text below for a more detailed comparison between the two methods. The boundaries obtained by the present method have the highest resolution, in a scale of only one base-pair. It is hard to imagine that a moving-window approach can achieve similar resolution. For comparison, the distribution of G + C content of the MHC sequence based on a moving-window method (window size 100 Kb, step 1 Kb) is also shown in Fig. 1C. As we can see from Fig. 1C, the window method cannot find accurate boundaries of isochores in the MHC sequence.

The isochores of the longest contigs in human chromosomes 21 and 22

The human chromosome 21 is particularly of interest because it is involved in Down's syndrome and other diseases. Chromosome 21 is about 33.8 Mb in length and contains about 225 protein coding genes. The multiresolution analysis performed to the longest contig NT_011512 of chromosome 21 shows that there are 315 segmentation points with the threshold adopted in Eq. (13). After the merging procedures, 25 points are retained, corresponding to 26 regions. The coordinates of the boundaries of these regions are listed in Table 2. There is a region of gene- and Alu poor, which is about a 7 Mb stretch in chromosome 21 [30]. This region of gene- and Alu poor is considered to be a typical isochore [13], which can be easily identified by the method presented in this paper. The gene- and Alu-poor isochore is found to be situated at the region between the positions 4,980,736 and 12,393,507 bp in the longest contig NT_011512 of chromosome 21. The G + C content of this region is 35.2%, while the G + C contents of flanking regions (up and down stream) are 40.4% and 39.4%, respectively. The variations of G + C content at both boundaries are relatively dramatic. This isochore, with a length of about 7 Mb, belongs to the L1 isochore family. The distribution of G + C content of the contig NT_011512 of chromosome 21 is graphically displayed in Fig. 2. It can be clearly seen from Fig. 2 that the distribution of G + C content undergoes relatively dramatic variations along the sequence. Fig. 2 also shows that a region with relatively high G + C content is flanked by two adjoining regions with relatively low G + C content and vice versa. For comparison, the distribution of G + C content of the contig NT_011512 of chromosome 21 based on a moving window method (window size 100 Kb, step 1 Kb) is also shown in Fig. 2. As we can see the window method cannot find accurate boundaries of isochores in the contig NT_011512 of chromosome 21.

The human chromosome 22 is the second smallest autosome [31]. There is an area between 16,000 and 18,800 Kb from the centromeric end of the sequence, in which the G + C content is less than 45%. This region contains only three genes and it is almost depleted in Alu repeats and CpG islands [31]. When the wavelet multiresolution analysis is applied to the longest contig NT_011520 of chromosome 22, 597 segmentation points are found. After the merging procedures, 37 are retained, corresponding to 38 regions. The coordinates of the boundaries of these regions are listed in Table 3. Three isochores identified should be mentioned, in which the

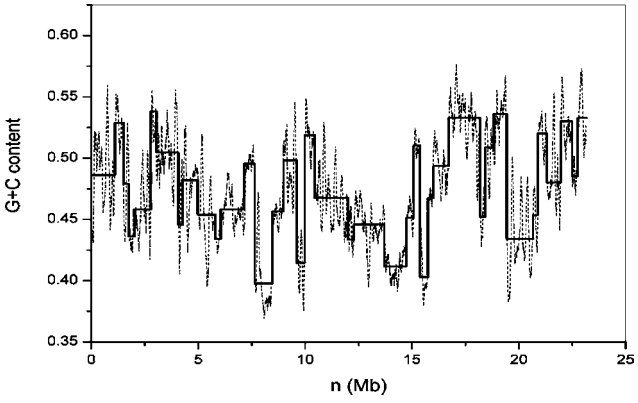

Fig. 3. The distribution of the G + C content along the longest contig (NT_011520) of human chromosome 22 (sequence length = 23,178,213 bp), based on the segmentation points obtained by the technique of wavelet multiresolution analysis, after the merging procedures. Note that a region with gene- and Alu-poor composed of three isochores has been identified and found to be situated at the positions from 11,998,706 to 14,734,833 bp. This region is relatively GC-poor with the G + C content less than 45%. For comparison, the distribution of the G + C content using a moving-window method (window size, 100 Kb, step, 1 Kb) is also shown (dash line).

Table 4
Comparison of the coordinates of segmentation points obtained by the entropic and wavelet transform methods, respectively[a]

Sequence	Entropic (bp)	WT (bp)	Difference (bp)	Percentage error (%)
MHC	1,841,872	1,835,008	6864	0.37
	2,483,967	2,490,368	6401	0.26
	3,384,908	3,383,296	1612	0.05
NT_011512 of chromosome 21	4,970,507	4,980,736	**10,229**	0.21
	12,310,289	12,393,507	**83,218**	0.67

[a] The entropic segmentation and wavelet transform methods are abbreviated as 'Entropic' and 'WT', respectively. The absolute value of the difference between the figures at the second and third columns is denoted by 'Difference', whereas the percentage of the absolute difference over the average of the figures at the second and third columns is denoted by 'Percentage error'. Note that the length of the longest contig NT_011512 of chromosome 21 used by Oliver et al. [27,28] is 28,515,771 bp, whereas ours is 28,602,511 bp, due to different sequence versions used in the study. The difference of lengths between the two sequences is 86,740 bp. The bold figures in this table indicate that the length difference caused by different versions adopted has not been taken into account.

G + C content is relatively low, all less than 45%. The first one is located between the positions 11,998,706 and 12,285,425 bp with the G + C content equal to 43.4%, belonging to the L2 family. The second one is from 12,285,426 to 13,710,833 bp, whose G + C content is 44.6%, belonging to the H1 isochore family. The third one is from 13,710,834 to 14,734,833 bp with the G + C content equal to 41.2%, belonging to the L2 isochore family. The total length of the three isochores is about 2.7 Mb. Compared Fig. 2 with Fig. 3, it can be seen that the distribution of G + C content of chromosome 21 is more homogeneous than that of chromosome 22.

Comparison of the segmentation points obtained with those of other methods

Of the methods in detecting isochore boundaries, the recursive entropic segmentation algorithm [26–28] should be compared with the present method. The method developed by Li, Oliver, and co-workers is a highly precise one with the highest resolution at the scale of only one base-pair. Based on this algorithm, a program, IsoFinder, has been developed by Oliver and co-workers [27,28]. The boundaries of isochores in many eukaryotic genomes are calculated and shown. For more details, refer to the website: http://bioinfo2.ugr.es/isochores/. To compare the segmentation points obtained by both methods, some important segmentation points in the human MHC sequence and the chromosome 21 sequence are listed and compared in Table 4. For the MHC sequence, the boundaries of regions of Class III, Class II, and Extended-Class II are considered. For the chromosome 21 sequence, the boundaries of the isochore with about 7 Mb in length are considered. As we can see from Table 4 that the five segmentation points obtained by the two quite different methods are highly consistent. The relative error of coordinates between the two kinds of points is less than 0.67%. Considering the fact that the two methods are based on quite different principles, the high agreement of the results derived from both methods indicates that the segmentation points in the sequences being studied exist objectively. The boundaries of isochores can be precisely detected regardless of using any methods of sequence analysis. Compared with the entropic segmentation method, the present method has the merits of more intuitiveness and less calculations.

Conclusion

A new algorithm based on the wavelet multiresolution analysis has been proposed in detecting the boundaries of isochores in the human genome. As an example, the boundaries of isochores in the human MHC sequence and chromosomes 21, 22 are determined. The advantages of the current algorithm include: (i) Isochore boundaries can be detected with the highest resolution, even at a scale of a single base. (ii) The mosaic distribution of the G + C content along a genome or chromosome is simultaneously displayed. (iii) The method has the flexibility that a user can choose appropriate minimum isochore length and the terminal scale decomposed in the wavelet transform to meet the user's interest and need. (iv) The method is a general one and can be applied to any genome. Compared with the entropic segmentation method, the present method has the advantages of more intuitiveness and less calculations.

Acknowledgments

Suggestions, discussions, and helps from Yonghong Wang, Hong-Yu Ou, Qiang Li, Ren Zhang, Ling-Ling Chen, and Feng-Biao Guo are gratefully acknowledged. This work was supported in part by the 973 Project of China (Grant 1999075606).

References

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

[2] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, F. Rodier, The mosaic genome of warm-blooded vertebrates, Science 228 (1985) 953–958.
[3] G. Bernardi, The human genome: organization and evolutionary history, Annu. Rev. Genet. 29 (1995) 445–476.
[4] G. Bernardi, Isochores and the evolutionary genomics of vertebrates, Gene 241 (2000) 3–17.
[5] S. Saccone, A. De Sario, J. Wiegant, A.K. Rap, G. Della Valle, G. Bernardi, Correlations between isochores and chromosomal bands in the human genome, Proc. Natl. Acad. Sci. USA 90 (1993) 11929–11933.
[6] A. Pavlicek, K. Jabbari, J. Paces, V. Paces, J. Hejnar, G. Bernardi, Similar integration but different stability of Alus and LINEs in the human genome, Gene 276 (2001) 39–45.
[7] J.C. Venter et al., The sequence of human genome, Science 291 (2001) 1304–1351.
[8] E.S. Lander et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[9] T. Fukagawa, K. Sugaya, K. Matsumoto, K. Okumura, A. Ando, H. Inoko, T. Ikemura, A boundary of long-range G + C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary, Genomics 25 (1995) 184–191.
[10] T. Fukagawa, Y. Nakamura, K. Okumura, M. Nogami, A. Ando, H. Inoko, N. Saiton, T. Ikemura, Human pseudoautosomal boundarylike sequences: expression and involvement in evolutionary formation of the present-day pseudoautosomal boundary of human sex chromosomes, Hum. Mol. Genet. 5 (1996) 123–132.
[11] R. Stephens, R. Horton, S. Humphray, L. Rowen, J. Trowsdale, S. Beck, Gene organization, sequence variation and isochore structure at the centromeric boundary of the human MHC, J. Mol. Biol. 291 (1999) 789–799.
[12] T. Tenzen, T. Yamagata, T. Fukagawa, K. Sugaya, A. Ando, H. Inoko, T. Gojobori, A. Fujiyama, K. Okumura, T. Ikemura, Precise switching of DNA replication timing in the GC content transition area in the human MHC, Mol. Cell. Biol. 17 (1997) 4043–4050.
[13] G. Bernardi, Misunderstanding about isochores. Part 1, Gene 276 (2001) 3–13.
[14] D. Haring, J. Kypr, No isochores in human chromosome 21 and 22?, Biochem. Biophys. Res. Commun. 280 (2001) 567–573.
[15] C.-T. Zhang, J. Wang, R. Zhang, A novel method to calculate the G + C content of genomic DNA sequences, J. Biomol. Struct. Dyn. 19 (2001) 333–341.
[16] C.-T. Zhang, R. Zhang, An isochore map of the human genome based on the Z curve method. Gene, 2003, in press.
[17] S. Mallat, Multiresolution approximation and wavelets, Trans. Am. Math. Soc. 315 (1989) 69–88.
[18] I. Daubechics, Ten lectures on wavelets, SIAM, Philadephia, 1992, CBMS Lecture Series.
[19] I. Daubechics, The wavelet transform, time-frequency localization and signal analysis, IEEE Trans. Inform. Theory 36 (1990) 961–1006.
[20] P. Lio, Wavelet in bioinformatics and compositional biology: state of art and perspectives, Bioinformatics 19 (2003) 2–9.
[21] C.-T. Zhang, R. Zhang, Analysis of distribution of bases in the coding sequences by a diagrammatic technique, Nucleic Acids Res. 19 (1991) 6313–6317.
[22] R. Zhang, C.-T. Zhang, Z curves, an intuitive tool for visualizing the DNA sequences, J. Biomol. Struct. Dyn. 11 (1994) 767–782.
[23] C.-T. Zhang, R. Zhang, Isochore structures in the mouse genome, Genomics, 2003, in press.
[24] S. Beck et al., The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex, Nature 401 (1999) 921–923.
[25] A. Eyre-Walker, L.D. Hurst, The evolution of isochores, Nat. Rev. Genet. 2 (2001) 549–555.
[26] W. Li, P. Bernaola-Galvan, F. Haghighi, I. Grosse, Applications of recursive segmentation to the analysis of DNA sequences, Comput. Chem. 26 (2002) 491–510.
[27] J.L. Oliver, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, Isochore chromosome maps of eukaryotic genomes, Gene 276 (2001) 47–56.
[28] J.L. Oliver, P. Carpena, R. Roman-Roldan, T. Mata-Balaguer, A. Mejias-Romero, M. Hackenberg, P. Bernaola-Galvan, Isochore chromosome maps of the human genome, Gene 300 (2002) 117–127.
[29] W. Li, Delineating relative homogeneous G + C domains in DNA sequences, Gene 276 (2001) 57–72.
[30] M. Hattori et al., The DNA sequence of human chromosome 21, Nature 405 (2000) 311–319.
[31] I. Dunham et al., The DNA sequence of human chromosome 22, Nature 402 (1999) 489–495.