Segmentation algorithm for DNA sequences

Chun-Ting Zhang,[1,*] Feng Gao,[1] and Ren Zhang[2]
[1]*Department of Physics, Tianjin University, Tianjin 300072, China*
[2]*Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060, China*
(Received 7 March 2005; published 17 October 2005)

A new measure, to quantify the difference between two probability distributions, called the quadratic divergence, has been proposed. Based on the quadratic divergence, a new segmentation algorithm to partition a given genome or DNA sequence into compositionally distinct domains is put forward. The new algorithm has been applied to segment the 24 human chromosome sequences, and the boundaries of isochores for each chromosome were obtained. Compared with the results obtained by using the entropic segmentation algorithm based on the Jensen-Shannon divergence, both algorithms resulted in all identical coordinates of segmentation points. An explanation of the equivalence of the two segmentation algorithms is presented. The new algorithm has a number of advantages. Particularly, it is much simpler and faster than the entropy-based method. Therefore, the new algorithm is more suitable for analyzing long genome sequences, such as human and other newly sequenced eukaryotic genome sequences.

DOI: 10.1103/PhysRevE.72.041917 PACS number(s): 87.15.Cc

I. INTRODUCTION

The completion of genome sequencing projects for humans and many other organisms has produced a huge amount of DNA sequence information. Mining useful biological knowledge from these DNA sequences currently represents a challenge to the biological (if not the whole scientific) community. Accumulating evidence shows that there are a number of turning points in most genome sequences, through which the nucleotide composition undergoes sudden changes. Usually, clear biological implications are associated with turning points. For example, in bacterial [1] and archaeal [2] genomes, turning points generally correspond to replication origins. Turning points of $G+C$ (guanine +cytosine) content distributions of bacterial and archaeal genomes may correspond to integration sites of horizontally transferred genes or genomic islands [3]. In human and many eukaryotic genomes, turning points of $G+C$ content distributions frequently associate with boundaries of isochores [4]. Turning points may also be called segmentation points. Therefore, given the availability of an increasing number of genome sequences, algorithms to identify genome segmentation points will play a more and more important role to gain an understanding of the genome organization. Historically, many segmentation algorithms were proposed [5–12], such as those based on the walking Markov model [5], hidden Markov model [6], change-point problem [7], recursive entropy [8–10], the cumulative GC profile [11], and the wavelet multiple scale analysis [12]. Recently, a computer program (ISOFINDER), based on a modified version of the entropic compositional segmentation algorithm, has been available online and can be used to identify isochores [13]. Here, a segmentation algorithm is proposed. This algorithm has a number of differences. Particularly, it is simple and fast. Therefore, this algorithm is suitable for analyzing long genome sequences, such as human and other newly sequenced eukaryotic genome sequences.

II. THE NEW SEGMENTATION ALGORITHM

Let $P=(p_1,p_2,\ldots,p_k)$ and $Q=(q_1,q_2,\ldots,q_k)$ be two probability distributions, where $0\leqslant p_i,q_i\leqslant 1$, for $i=1,2,\ldots,k$, and $\Sigma_{i=1}^{k}p_i=1$, $\Sigma_{i=1}^{k}q_i=1$. Define

$$S(P)\equiv\sum_{i=1}^{k}p_i^2, \tag{1}$$

which was called the genome order index in the case of $k=4$ [14]. Let w_1 and w_2 be two weights, where $0<w_1$, $w_2<1$, and $w_1+w_2=1$. Define the quadratic divergence between the two distributions P and Q by

$$\Delta S(P,Q)=w_1S(P)+w_2S(Q)-S(w_1P+w_2Q). \tag{2}$$

The quadratic divergence $\Delta S(P,Q)$ quantifies the difference between the distributions P and Q. Simple derivation shows that

$$\Delta S(P,Q)=w_1w_2S(P-Q)=w_1w_2\sum_{i=1}^{k}(p_i-q_i)^2. \tag{3}$$

Based on Eq. (3), some mathematical properties of the quadratic divergence $\Delta S(P,Q)$ may be derived

(i)

$$\Delta S(P,Q)\geqslant 0, \tag{4}$$

with $\Delta S(P,Q)=0$, if and only if $P=Q$.

(ii)

$$\Delta S(P,Q)=\Delta S(Q,P). \tag{5}$$

(iii) For three probability distributions P, Q, and M

*Email: ctzhang@tju.edu.cn

$$\Delta S^{1/2}(P,M) + \Delta S^{1/2}(M,Q) \geqslant \Delta S^{1/2}(P,Q). \quad (6)$$

The above three properties are obvious, because $\Delta S(P,Q)$ calculated in Eq. (3) is the Euclidean distance between the two vectors P and Q.

(iv) The quadratic divergence $\Delta S(P,Q)$ reaches its maximum when the difference between the distributions P and Q is the maximum among other pairs of P and Q.

The mathematical property (iv) constitutes the basis of the present segmentation algorithm.

When $k=4$, as in the case of DNA sequence, Eq. (1) may be rewritten as

$$S \equiv S(P) = a^2 + c^2 + g^2 + t^2, \quad (7)$$

where a, c, g, and t denote the occurrence frequencies of A, C, G, and T, respectively, in a genome or a DNA sequence. The genome order index S defined in Eq. (7) is a useful statistical quantity to reflect the compositional characteristics of a genome [14]. The phenomenon $S<1/3$ was observed previously for each of the 90 species, such as human and *E. coli* [15], and confirmed by subsequent work [14]. Obviously, $1/4 \leqslant S \leqslant 1$, and particularly, for almost all (more than 1000) genomes currently available, it was found that $1/4 \leqslant S < 1/3$ [14]. Furthermore, it was also found that S has different values in coding and noncoding regions. This fact was used to recognize protein-coding genes in the budding yeast genome [16]. The genome order index S may serve as an appropriate divergence measure to quantify the compositional difference between two DNA sequences. The new segmentation algorithm proposed here is based on the quadratic divergence. Consider a genome with N bases. Let n be an integer, $2 \leqslant n \leqslant N-1$. For a given n, the genome sequence is partitioned into two subsequences, one left and the other right. Let $w_1=n/N$ and $w_2=(N-n)/N$. Let $P_l=(a_l,c_l,g_l,t_l)$ and $P_r=(a_r,c_r,g_r,t_r)$, where a_l,c_l,g_l,t_l and a_r,c_r,g_r,t_r are the occurrence frequencies of bases A, C, G, and T in the left and right subsequences, respectively. Define

$$\Delta S(P_l,P_r) = (n/N)S(P_l) + [(N-n)/N]S(P_r) - S\{(n/N)P_l + [(N-n)/N]P_r\}. \quad (8)$$

Suppose that n^* is a position at which $\Delta S(P_l,P_r)$ reaches maximum; then, n^* is a compositional segmentation point of the genome first found. The new algorithm is also recursive as in Refs. [8–10], i.e., after n^* is determined, the same procedure is applied to both the left- and right subsequences, respectively. Recursively apply the procedure until $\Delta S(P_l,P_r)$ is less than a given threshold. Note that for a given genome or DNA sequence to be segmented, the third term of Eq. (8) is a constant. Therefore, it may be ignored. Define

$$\Sigma(n) = (n/N)S(P_l) + [(N-n)/N]S(P_r), \quad 2 \leqslant n \leqslant N-1, \quad (9)$$

then $\Sigma(n^*)=$maximum. Of course, to search for segmentation points, one may adopt Eq. (3) directly. That is, let

$$\Delta S'(P_l,P_r) = \frac{n(N-n)}{N^2} S(P_l - P_r); \quad (10)$$

the maximum of $\Delta S'(P_l,P_r)$ leads to the same segmentation points. All Eqs. (8)–(10) result in identical segmentation points for a given genome or DNA sequence.

Note that S may be rewritten as

$$S(P) = \bar{P} = \sum_{i=1}^{k} p_i p_i = \sum_{i=1}^{k} p_i^2, \quad (11)$$

indicating that S in fact represents the average probability. In the case of DNA sequence ($k=4$), S is the average probability of occurrence of A, C, G, and T in a DNA sequence. Given a DNA sequence to be segmented, the segmentation point is at the position where the difference between the average probabilities of occurrence of four bases in the two resulting subsequences reaches the maximum. In general, $S(P)=\Sigma_{i=1}^{k} p_i^{\beta} (\beta>2)$ can also be used to quantify the difference between two probability distributions. Both theoretical proof and computational experiments show that positions of segmentation points are independent of the choice of β values. However, $\beta=2$ is chosen because of obvious physical and biological implications and computation efficiency. In other words, the choice of $\beta=2$ is reasonable to partition a given genome or DNA sequence into compositionally distinct domains.

A question needed to be answered is the halting condition of the segmentation algorithm. We define a halting parameter t

$$t = N\Delta S(P_l,P_r), \quad (12)$$

where N is the length of the sequence or subsequence to be segmented. If $t<t_0$, the segmentation procedure halts; otherwise, the procedure continues until $t<t_0$. Since we are only interested in segmenting concrete genomes, the choice of t_0 is based on heuristic considerations. Larger threshold t_0 leads to fewer segmentation points and longer segmented subsequences, whereas smaller threshold t_0 leads to more segmentation points and shorter segmented subsequences. It should be noted that in some cases the segmentation procedure also halts when the resulting subsequence is shorter than a given minimum length. For prokaryotic genomes, we choose 1000 bp as the minimum length, roughly the size of a prokaryotic gene. For eukaryotic genomes, we choose 3000 bp as the minimum length according to a requirement imposed by the experimental characterization of isochores through DNA centrifugation [4]. In general, the choice of t_0 and the minimum length is heuristic and depends on each case.

III. COMPARISON WITH THE ENTROPIC SEGMENTATION ALGORITHM

It is interesting to compare the new segmentation algorithm with the entropic segmentation algorithm [8–10]. Note that the Shannon entropy H for a DNA sequence is defined by

$$H = -a \log_2 a - c \log_2 c - g \log_2 g - t \log_2 t, \quad H \in [0, 2]. \tag{13}$$

The Jensen-Shannon divergence is defined by [8–10]

$$D(n) = H - \left(\frac{n}{N} H_{\text{left}} + \frac{N-n}{N} H_{\text{right}} \right), \quad n = 2, \ldots, N-1, \tag{14}$$

where H_{left} and H_{right} are the Shannon entropy for the left and right subsequences, respectively. Suppose that n^* is calculated by $D(n^*) = \max D(n)$; if $D(n^*)$ is above a given threshold, then n^* is deemed a segmentation point [8–10].

Both the new and entropic segmentation algorithms were used to locate segmentation points for a given genome or DNA sequence. It is interesting to see if the two algorithms result in the same or different results. Here, we used some well-established isochore examples to test both algorithms. The human major histocompatibility complex (MHC) sequence is situated at the human chromosome 6*p*21 region. This sequence carries 224 genes, many of which are involved in some important human genetic diseases such as arthritis and diabetes. The MHC sequence is 3 673 778 bp in length, which was sequenced before the completion of the whole human genome project. The isochore structure of the MHC sequence has undergone extensive studies during the past few years, because the best known isochore boundary within this sequence was confirmed experimentally [17]. Therefore, the sequence becomes a touchstone for testing any segmentation algorithm. In order to explore the isochore structure, it is necessary to convert the DNA sequence into a binary sequence of *S* (strong H-bond) and *W* (weak H-bond) bases. Table I shows the segmentation points obtained by the present method based on this binary sequence. The most remarkable result is that the coordinates of segmentation points derived from the genome order index and entropy-based segmentation algorithms are all identical for each figure at each digit! The fact that the two sets of numbers are all identical indicates that both algorithms are accurately consistent with each other. We have also applied our algorithm to other human chromosomes (data not shown here). The coordinates of segmentation points obtained are all identical with those derived from the entropic segmentation algorithm.

One may wonder why the two algorithms yield identical results. Our explanation is as follows. According to the information theory, the Shannon information entropy H is defined in Eq. (13) in the case of DNA sequence. The difference between $H_{\max}$ and H, denoted by D

$$D = H_{\max} - H = 2 - H, \tag{15}$$

is called the negative entropy. Using the values of S and H for 627 virus genomes, the correlation coefficient between S and H is calculated and found to be equal to -1 [14]. To further study the relation between the Shannon entropy H and the genome order index S, the values of H and S for more than 1000 genome sequences were calculated, including the human, mouse, rat, *C. elegans*, yeast, *Arabidopsis thaliana*, more than 100 bacterial and archaeal genomes, and more than 600 virus genomes. Figure 1 shows the relation of $H \sim S$. It shows that the correlation coefficient between H and S is almost -1, indicating that they are negatively correlated. The above result also implies that the genome order index S defined in Eq. (7) plays a role of some kind of negative entropy. As pointed out above, in the case of DNA sequence ($k=4$), S is the average probability of occurrence of A, C, G, and T in a DNA sequence. Given a DNA sequence to be segmented, the segmentation point is at the position where the difference between the average probabilities of occurrence of four bases in the two resulting subsequences reaches the maximum. In contrast, H represents an appropriate measure of average randomness or uncertainty for a given probability distribution. The segmentation of a DNA sequence using $H(P)$ is based on the consideration of maximum difference between the average uncertainty of nucleotide distribution in the two resulting subsequences. Obviously, our segmentation algorithm is more straight to partition a DNA sequence into compositionally distinct domains, and needs less calculation than $H(P)$. The above facts may give an explanation of the equivalence of the two segmentation algorithms, but it cannot serve as a mathematical

TABLE I. The coordinates of segmentation points obtained by the present segmentation method for the human MHC sequence. Note that the coordinates of segmentation points derived from the genome order index and the entropy-based segmentation methods ($s_0=10$) are all identical for each figure at each digit! Refer to Ref. [10] for detail of the entropic segmentation algorithm and the definition of the strength parameter s_0.

No.	Segmentation points
1	299270
2	354226
3	364029
4	833239
5	1168415
6	1244738
7	1396980
8	1490846
9	1662756
10	1709646
11	1712970
12	1715527
13	1739420
14	1742437
15	1841871
16	2483966
17	3054365
18	3088089
19	3159420
20	3384907
21	3444780
22	3491519
23	3552176
24	3638110

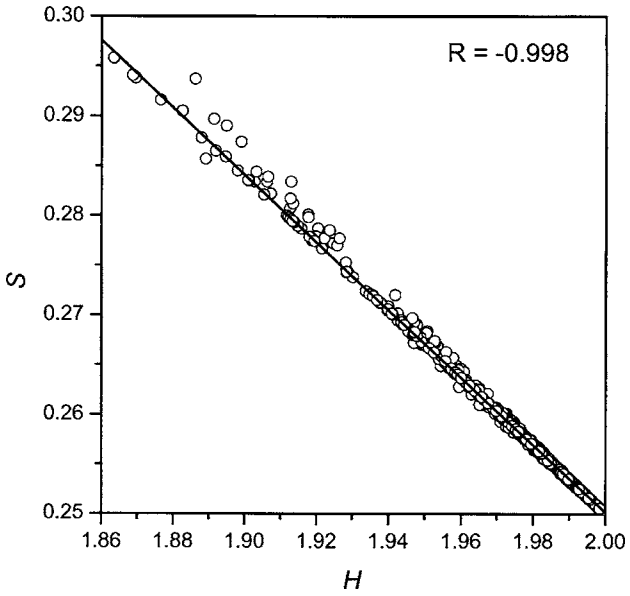

FIG. 1. To study the relation between the Shannon entropy H and the genome order index S, the values of H [defined in Eq. (13)] and S [defined in Eq. (7)] for more than 1000 genome sequences were calculated, including the human, mouse, rat, *C. elegans*, yeast, *Arabidopsis thaliana*, more than 100 bacterial and archaeal genomes, and more than 600 virus genomes. This figure shows that the correlation coefficient R between H and S is almost -1, indicating that H and S are negatively correlated, also implying that the genome order index S plays a role of some kind of negative entropy.

proof. We will leave the mathematical proof to mathematicians who are interested in this issue.

Our algorithm has a series of merits. First, the genome order index S possesses simpler mathematical form than that of the Shannon entropy H. Second, the algorithm needs less calculation than that of the entropy-based algorithm, and is fast. This feature is particularly useful for segmenting long genome sequences, such as the human genome and other eukaryotic genomes. Third, the genome order index S has a clear geometrical meaning, i.e., it is a square of a Euclidean distance [14]. It represents a deviation of the given probability distribution in a genome from the equal distribution of nucleotides ($a=c=g=t=1/4$). Fourth, S possesses clear biological implications. Note that S may be rewritten as [14]

$$S = \frac{1}{2} + \frac{1}{2}(a-t)^2 + \frac{1}{2}(g-c)^2 - (a+t)(g+c). \qquad (16)$$

The second and third terms are directly related to the deviations from the Chargaff parity rule 2 (PR2). If the PR2 is strictly correct, the two terms should be equal to 0. In fact, both $a-t$ and $g-c$ are small quantities in real genomes. Therefore, the genome order index S is mainly relevant to the $G+C$ or $A+T$ content of the genome. It appears that S contains more information than the $G+C$ content contains. Therefore, in addition to the widely used $G+C$ content, S would be a new biological statistical quantity useful to describe the compositional features of genomes. Usually, S has different values in coding and noncoding regions. This fact was used to recognize protein-coding genes in the budding yeast genome [16]. Finally, the segmentation algorithm is different from the entropic one in that the former is able to provide an intuitive picture by incorporating with the Z-curve representation of DNA sequences [18]. For example, in the case of the $G+C$ content, Eq. (10) reduces to

$$\Delta S' \propto ((G+C)_l - (G+C)_r)^2, \qquad (17)$$

where $(G+C)_l$ and $(G+C)_r$ are the average $G+C$ content of the subsequences at the left- and right hand sides of the segmentation point concerned, respectively. Therefore, such segmentation point is exactly a turning point of the $G+C$ content, which corresponds to an extreme point in the cumulative GC profile [11]. Consequently, we may use the segmentation coordinate to annotate the related cumulative GC profile, giving researchers an intuitive picture. In what follows, we will show some concrete examples.

IV. APPLICATION EXAMPLES OF THE SEGMENTATION ALGORITHM

The segmentation algorithm may find many applications in genome sequence analysis. For example, in Ref. [9] the entropic segmentation algorithm has been applied to study a number of genome problems, including the isochore mapping, *CpG* island detection, identification of the replication origin and terminus in bacterial genomes, identification of complex repeats in telomere sequences, and delineating coding and noncoding regions. Additionally, the possibility to detect horizontally transferred genes in a genome using a segmentation method has been proposed [3]. The segmentation algorithm proposed here may find applications in all of the above areas, but a simple form and fast calculations. As mentioned above, collaborating with the technique of cumulative GC profile [11], a more intuitive picture to display the distribution of segmentation points will be presented. In what follows, as an example, we will study the isochore structure of human chromosome 21 and chimpanzee chromosome 22. The human genome sequences, release *hg*17, the chimpanzee genome sequences, release panTro1, and the annotation files were downloaded from http://genome.ucsc.edu/.

Figure 2 shows the negative cumulative GC profiles for human chromosome 21 and chimpanzee chromosome 22. Note that in each chromosome there are a number of larger or smaller gaps. Here, only gaps more than 1% of the chromosome size were retained; gaps less than 1% of the chromosome size were simply deleted. Consequently, each chromosome was split into two contigs. The first contig was not taken into consideration due to small size. For the larger contig, the constituting subcontigs are simply merged, as if there were no gaps. Applying the segmentation algorithm to the resulting contig of each chromosome, the segmentation points were obtained at $t_0=1000$. In addition, the cumulative GC profile is also called the z_n' curve, which is a discrete function of the nucleotide position n in a genome or chromosome. It was shown that the average $G+C$ content of a genome or chromosome at position $n \to n+\Delta n$ is calculated by [11]

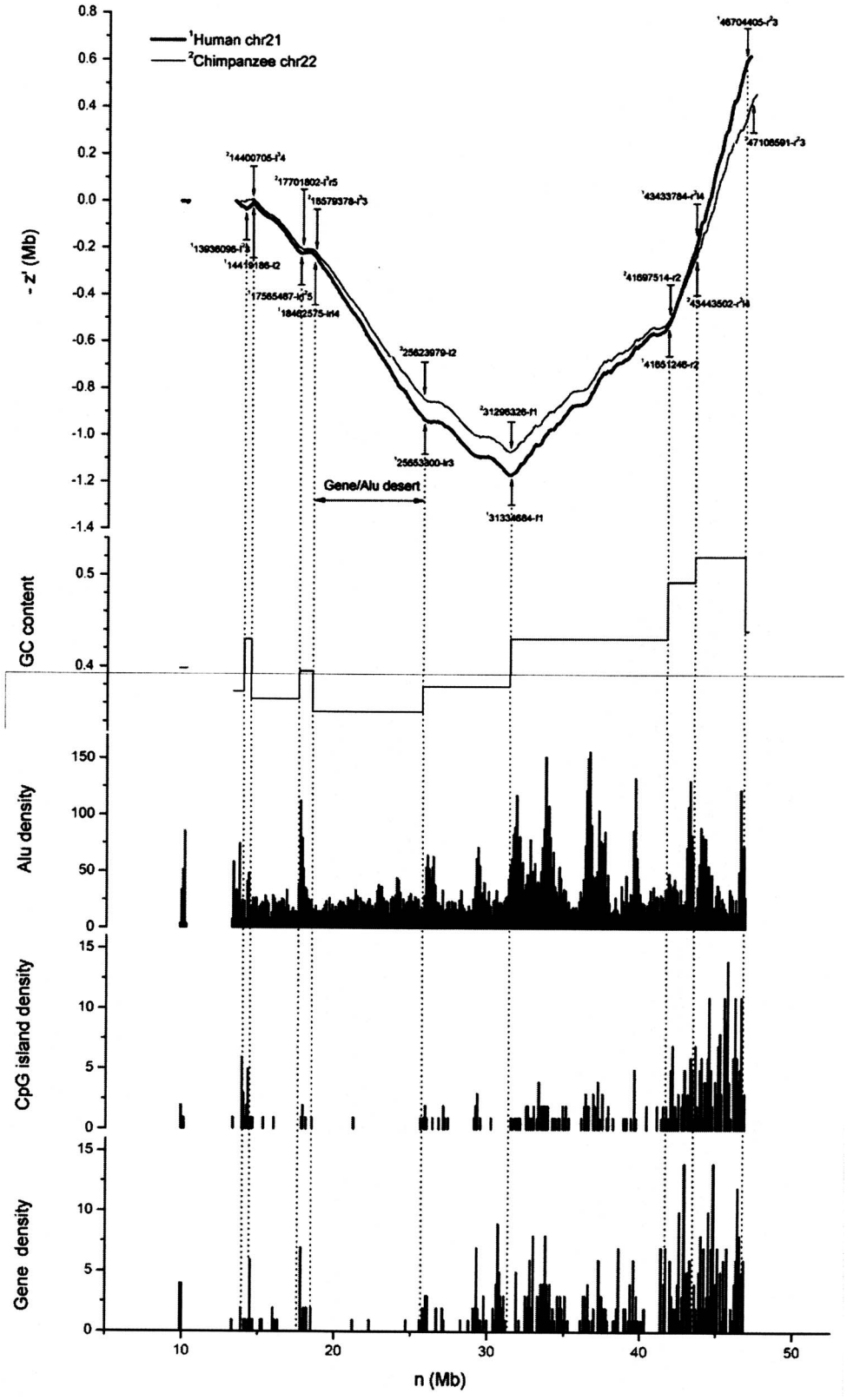

FIG. 2. The negative cumulative *GC* profiles for human chromosome 21 and chimpanzee chromosome 22 marked with the segmentation points obtained. The notation used here is described as follows. Besides the position coordinates, the order of occurrence for each point in the segmentation process is also labeled in the figure. We use "*f*," "*l*," "*r*," and an integer to label the order of occurrence, where *f* denotes the first point occurring during the course of segmentation, whereas *l* and *r* denote the point to occur in the left- and right subsequence, respectively. The integer denotes the times of segmentation. Take the point $43\,443\,502-r^2l4$ as an example. The first part 43 443 502 is the position coordinates. The second part "r^2l4" denotes the order of occurrence. The last integer "4" in the second part means that this point occurs after four times of segmentation. In the symbol "r^2l," *r* appears two times, so we used "r^2" instead of "*rr*" for convenience. The bottom four plots show the distributions of the *G+C* content, *Alu*, *CpG* island, and gene along human chromosome 21, respectively. Note that all of these distributions are closely correlated with the segmented regions with distinct *G+C* content. Also note that the coordinate value of each segmentation point has been corrected by taking the gap length into account. For instance, there is a gap occurring at $n_0 \rightarrow n_0+\Delta$, where Δ is the gap length. If a segmentation point obtained is situated at n, and $n > n_0$, then the actual coordinate of n adopted in this plot is $n+\Delta$. Meanwhile, the gap region $n_0 \rightarrow n_0+\Delta$ is represented by a blank interval in this plot. Here, n_0 and n are the relative coordinates with respect to the contig without gaps. Other gaps are treated with a similar procedure.

$$\overline{\mathrm{G+C}} \propto \Delta(-z_n')/\Delta n. \tag{18}$$

Therefore, an up jump in the $-z_n'$ curve indicates an increase of the *G+C* content, whereas a drop in the $-z_n'$ curve indicates a decrease of the *G+C* content. An approximate straight region in the $-z_n'$ curve implies that the *G+C* content in this region is roughly constant. Bearing the above points in mind, let us study the negative cumulative *GC* profiles shown in Fig. 2. It is seen that when the threshold t_0 was set to 1000, nine segmentation points were obtained in the hu-

man chromosome 21 (the larger contig), resulting in ten regions with distinct $G+C$ contents. The region from 18 462 576 to 25 653 300 bp was deemed as an isochore [10,12]. The $G+C$ content of this isochore (with length =7.2 Mb) is 35.1%, the lowest $G+C$ content among the resulting ten regions. It is clearly shown that this region is the desert region of gene/*Alu*/*CpG* island distributions, which were calculated in 100 kb long, nonoverlapping windows. The positive correlation between the $G+C$ content and the density of *Alu*, *CpG* island, and gene is a well-known fact; however, it is noteworthy that the segmentation points obtained here are exactly the boundaries of the related regions. For example, there is an abrupt increase (decrease) of the densities of *Alu*/*CpG* island (gene) at the first (second) boundary of the short $G+C$-rich region between 17 565 487 and 18 462 575 bp. Similar phenomena were observed in other $G+C$ distinct regions. The precise boundary coordinates obtained by the segmentation algorithm and the associated cumulative *GC* profile provide a useful platform to analyze a genome or chromosome. For instance, gene-finding algorithms would benefit from these boundary coordinates. To gain better gene-finding results, different parameters would be adopted in a gene-finding algorithm by considering different regions of distinct $G+C$ content with precise boundary coordinates.

It was reported [19] that the difference between human chromosome 21 and chimpanzee chromosome 22 caused by single base substitution is 1.44%. In addition, there are about 68 000 indels of the two chromosomes, where most indels are about 30 bp in length, with a few reaching 54 000 bp. Consequently, human chromosome 21 is about 400 kb longer than chimpanzee chromosome 22. The above two variations lead to a difference of about 5% between the two chromosomes. Although there are relatively large sequence variations between the human chromosome 21 and chimpanzee chromosome 22, comparison based on their cumulative *GC* profiles showed that these two chromosomes have similar isochore structures, including the numbers and positions of segmentation points. This fact suggests that the same evolutionary forces might have shaped genome organization in both organisms, and the isochore structure is highly conserved during the evolution of these two organisms. It is reasonable to deduce that such genome organization (the isochore structure) plays an important role in keeping the survival for both organisms.

In summary, the cumulative *GC* profile marked with the coordinates of resulting segmentation points is a useful tool for genome analysis. This leads to a neat graphical representation of $G+C$ content variation along a genome or chromosome, and a clear-cut definition of isochores. This technique allowed us to show and confirm that $G+C$-rich isochores in a human or chimpanzee chromosome have higher gene, *CpG* island, and *Alu* densities than $A+T$-rich isochores. Although these are well-known characteristics of isochores of the vertebrate organisms, the advantage of the technique is that an investigator is able to study all of these in a perceivable and precise manner. We believe that plots similar to Fig. 2 would become a common tool for analyzing the $G+C$ content variations for genome or chromosome sequences. For higher eukaryotic genomes, the cumulative *GC* profile equipped with the segmentation algorithm would be an appropriate starting point for analyzing isochore structures.

ACKNOWLEDGMENT

The present work was supported in part by the NNSF of China Grant No. 90408028.

[1] J. R. Lobry, Biochimie **78**, 323 (1996).
[2] R. Zhang and C. T. Zhang, Archaea **1**, 335 (2004).
[3] R. Zhang and C. T. Zhang, Bioinformatics **20**, 612 (2004).
[4] G. Bernardi, Gene **241**, 3 (2000).
[5] J. W. Fickett, D. C. Torney, and D. R. Wolf, Genomics **13**, 1056 (1992).
[6] G. A. Churchill, Comput. Chem. (Oxford) **16**, 107 (1992); L. Peshkin and M. S. Gelfand, Bioinformatics **15**, 980 (1999).
[7] *Change-point Problems*, Lecture Notes and Monograph Series, Vol. 23, edited by E. Carlstein, H. G. Muller, and D. Siegmund, (Institute of Mathematical Statistics, Hayward, CA, 1994).
[8] P. Bernaola-Galvan, R. Roman-Roldan, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996); J. L. Oliver, P. Bernaola-Galvan, P. Carpena, and R. Roman-Roldan, Gene **276**, 47 (2001); J. L. Oliver, P. Carpena, R. Roman-Roldan, T. Mata-Balaguer, A. Mejias-Romero, M. Hackenberg, and P. Bernaola-Galvan, *ibid.* **300**, 117 (2002).
[9] W. Li, P. Bernaola-Galvan, F. Haghighi, and I. Grosse, Comput. Chem. (Oxford) **26**, 491 (2002).
[10] W. Li, Gene **276**, 57 (2001).
[11] C. T. Zhang and R. Zhang, Genomics **83**, 384 (2004).
[12] S. Y. Wen and C. T. Zhang, Biochem. Biophys. Res. Commun. **311**, 215 (2003).
[13] J. L. Oliver, P. Carpena, M. Hackenberg, and P. Bernaola-Galvan, Nucleic Acids Res. **32**, W287 (2004).
[14] C. T. Zhang and R. Zhang, Comput. Biol. Chem. **28**, 149 (2004).
[15] C. T. Zhang and R. Zhang, Nucleic Acids Res. **19**, 6313 (1991).
[16] C. T. Zhang and J. Wang, Nucleic Acids Res. **28**, 2804 (2000).
[17] T. Tenzen, T. Yamagata, T. Fukagawa, K. Sugaya, A. Ando, H. Inoko, T. Gojobori, A. Fujiyama, K. Okumura, and T. Ikemura, Mol. Cell. Biol. **17**, 4043 (1997).
[18] C. T. Zhang, R. Zhang, and H. Y. Ou, Bioinformatics **19**, 593 (2003).
[19] H. Watanabe *et al.*, Nature (London) **429**, 382 (2004).