ELSEVIER

Computers and Chemistry 26 (2002) 491–510

Computers & Chemistry

www.elsevier.com/locate/compchem

Applications of recursive segmentation to the analysis of DNA sequences

Wentian Li [a,b,*], Pedro Bernaola-Galván [c], Fatameh Haghighi [d], Ivo Grosse [e]

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

[b] *Laboratory of Statistical Genetics, The Rockefeller University, New York, NY 10021, USA*

[c] *Departmento de Física Aplicada II, Universidad de Málaga, E-29071 Málaga, Spain*

[d] *Columbia Genome Center, Columbia University, New York, NY 10032, USA*

[e] *Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA*

Received 22 May 2001; received in revised form 25 October 2001; accepted 12 November 2001

Abstract

Recursive segmentation is a procedure that partitions a DNA sequence into domains with a homogeneous composition of the four nucleotides A, C, G and T. This procedure can also be applied to any sequence converted from a DNA sequence, such as to a binary strong(G + C)/weak(A + T) sequence, to a binary sequence indicating the presence or absence of the dinucleotide CpG, or to a sequence indicating both the base and the codon position information. We apply various conversion schemes in order to address the following five DNA sequence analysis problems: isochore mapping, CpG island detection, locating the origin and terminus of replication in bacterial genomes, finding complex repeats in telomere sequences, and delineating coding and noncoding regions. We find that the recursive segmentation procedure can successfully detect isochore borders, CpG islands, and the origin and terminus of replication, but it needs improvement for detecting complex repeats as well as borders between coding and noncoding regions. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Recursive segmentation; DNA sequence; Dinucleotide

1. Introduction

One generic feature of DNA sequences is that their statistical properties are not homogeneously distributed along the sequence (Sueoka, 1962). For practically any feature of interest, such as the G + C content, the CpG dinucleotide content, the periodicity-of-three, the A–T strand asymmetry, the presence of protein-binding motifs, or the origin or terminus of replication, its density along almost any DNA sequence fluctuates from position to position. For many situations, these fluctuations can be better explained by alternating homogeneous domains (called segments in Elton (1974)) than by random fluctuations in a homogeneous sequence.

If we accept a domain picture of DNA sequences, it is natural to design computational approaches that segment a DNA sequence into homogeneous domains, and computer algorithms that accomplish such a segmentation are commonly called segmentation algorithms. Two well-known examples of segmentation algorithms are the one based on hidden Markov model by Churchill (1989, 1992) and the walking Markov model algorithm by Fickett et al. (1992). Parallel studies on similar problems called ‘change-point problems’ (Carlstein et al., 1994) have been recently performed and applied to segment DNA sequences (Braun and Müller, 1998; Braun et al., 2000). In the biology community, however, most people still use the old-fashioned ‘moving window’ approach, as in the case of the

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

PII: S0097-8485(02)00010-4

two recent papers on human genome sequence (Venter et al., 2001; Lander et al., 2001).

One advantage of the widely-used sliding-window methods is that their implementation is straightforward: one calculates the density of a sequence feature of interest within a window, moves the window along the sequence, and recalculates the density again. However, the choice of the window size and the moving distance are, in general, arbitrary. If the window size is too large, local fluctuations that contain significant biological information may be averaged out. If the moving distance is too long, one domain can be split between two windows and its distinctive feature may not be revealed. One example of the window size effect on the $G+C$ content fluctuation is illustrated in Li et al. (1994).

Another drawback for the moving window approach as well as any other sequential (left-to-right) approaches is that they are not appropriate for sequences that exhibit hierarchical patterns. Consider the $G+C$ content, for example. If a seemingly homogeneous $G+C$ domain becomes heterogeneous under a more relaxed criterion for being homogeneous, subdomains may emerge. This 'domains-within-domains' phenomenon can continue to emerge on smaller and smaller length scales. There are many consequences of the domains-within-domains phenomenon, such as the co-existence of small and large domains, the slow decay of the variance with the window size (Cuny et al., 1981; Li et al., 1998; Clay et al., 2001), or the slow decay of base–base correlations with length. Studies over the last 10 years have shown that this hierarchical description of DNA sequences is quite prevalent (Li, 1992; Li et al., 1994; Bernaola-Galván et al., 1996; Li, 1997a).

If the domains are organized in a hierarchical manner, it is natural to segment DNA sequences recursively (top-to-bottom). One such recursive segmentation for DNA sequences was proposed in Bernaola-Galván et al. (1996). Although this recursive segmentation algorithm was originally developed for finding domains that are homogeneous in base composition or $G+C$ content (Bernaola-Galván et al., 1996; Oliver et al., 1999, 2001; Li, 2001c), it is the purpose of this paper to show that there are many other applications of the recursive segmentation algorithm to the analysis of DNA sequences.

The original recursive segmentation algorithm segments an input sequence into domains with a homogeneous composition of the four nucleotides. For many other potential applications, we apply a filter that converts the original four-base DNA sequence into a k-symbol sequence. The number of symbols, k, can either be smaller or larger than four. For example, for detecting $G+C$ domains, a DNA sequence is converted to a binary sequence with the two symbols S(strong) = {C, G} and W(weak) = {A, T}; for detecting domains with a homogeneous dinucleotide composition, we construct a 16-symbol sequence, in which each symbol represents one dinucleotide. It is clear that applications of the recursive segmentation algorithm are not restricted to the five examples discussed in this paper. For any new application, one only needs to design a new filter for the sequence conversion.

2. Recursive segmentation

We follow the divide-and-conquer approach (see, Cormen et al., 1990) proposed in Bernaola-Galván et al. (1996). For a k-symbol sequence of length N, we calculate at each position i ($0<i<N$) the entropy H of the whole sequence, the entropy H_l of the subsequence on the left side of the partition point, and the entropy H_r of the subsequence on the right side of the partition point defined by Shannon (1948)

$$\hat{H} = -\sum_{j=1}^{k} \frac{N_j}{N}\log\frac{N_j}{N}, \quad \hat{H}_l = -\sum_{j=1}^{k}\frac{N_{j,l}}{i}\log\frac{N_{j,l}}{i},$$
$$\hat{H}_r = -\sum_{j=1}^{k}\frac{N_{j,r}}{N-i}\log\frac{N_{j,r}}{N-i}, \tag{1}$$

where N_j, $N_{j,l}$ and $N_{j,r}$ are the counts of symbol j in the whole, the left, and the right sequence. As a measure of the heterogeneity of the sequence we choose the maximized Jensen–Shannon divergence:

$$\hat{D}_{JS} = \max_i \hat{D}_{JS}(i) = \max_i\left[\hat{H} - \frac{i}{N}\hat{H}_l - \frac{N-i}{N}\hat{H}_r\right]. \tag{2}$$

If $\hat{D}_{JS}$ is large enough, we say that the sequence is heterogeneous and should be segmented. We recursively apply the same procedure to both the left and the right subsequence, as long as $\hat{D}_{JS}$ stays above a given threshold. If $\hat{D}_{JS}$ falls below that given threshold, the recursion along the current path is stopped. This recursive segmentation procedure is very similar to the procedure of growing a binary tree. When the segmentation is continued, two branches of the tree are generated; if it is stopped, that branch becomes a leaf.

The stopping criterion can be handled either in the hypothesis testing or in the model selection framework. In the hypothesis testing framework, we compute the probability that the observed value of $\hat{D}_{JS}$ or a greater value can be obtained by chance by the null hypothesis that the sequence is homogeneous. The exact form of the null distribution is hard to obtain (Pettitt, 1980). Although an asymptotic approximation (large sample size limit) is available in Horvath (1989), Csorgo and Horvath (1997), it was questioned on whether this approximation is good for finite sample sizes (Grosse et al., 2002). An empirical functional form of the null distribution is also suggested by numerical simulation in Grosse et al. (2002).

An alternative approach to determine the stopping criterion was recently proposed in the model selection framework (Li, 2001a,b), where a model is judged by a combination of how good the model fits the data and how complex the model is. The goal is to find a model at the border between underfitting models (those that do not fit the data well) and overfitting models (those that fit the data too well by using too many parameters). In order to balance the goodness-of-fit of the model to the data with the number of parameters of the model we use the Bayesian Information Criterion (BIC) (Schwarz, 1978; Akaike, 1978; Raftery, 1995) defined by

$$\mathrm{BIC} = -2\log(\hat{L}) + \log(N)\,K, \qquad (3)$$

where $\hat{L}$ is the maximum likelihood of the model, K is the number of free parameters, and N is the sample size (in our example, it is the sequence length). Among all considered models we choose that model which minimizes the value of BIC.

Deciding whether to proceed with a segmentation can be considered as a comparison between two models: modeling the sequence as one single random sequence, and modeling it as two random subsequences with different base compositions. If the first model is better by a model selection criterion, the segmentation is stopped; otherwise, it is continued. It is easy to show that with BIC as the model selection criterion, in order for the recursive segmentation to continue, the value of $2N\hat{D}_{\mathrm{JS}}$ has to exceed the threshold $\log(N)[2(k-1)+1-(k-1)]$, or,

$$2N\hat{D}_{\mathrm{JS}} > \log(N)\,k, \qquad (4)$$

where k is the number of different symbols in the sequence.

We propose to use the relative increase of $2N\hat{D}_{\mathrm{JS}}$ from the BIC threshold, $\log(N)k$, as a measure of stringency for the segmentation, and we call this relative increase the 'segmentation strength' (Li, 2001a,b)

$$s = \frac{2N\hat{D}_{\mathrm{JS}} - \log(N)\,k}{\log(N)\,k}. \qquad (5)$$

We continue the recursive segmentation process as long as $s > s_0$, where s_0 is a threshold predefined by the user. In the following sections, we will present five applications of the recursive segmentation algorithm to the analysis of DNA sequences.

3. Detection of isochores

Isochores are large homogeneous G + C domains that have been studied for over 20 years (Cuny et al., 1981; Bernardi, 1989, 1995; Clay et al., 2001). Isochores, whose definition was originally experimentally based, were noticed as distinct components in density gradient ultra-centrifugation (Thiery et al., 1976; Macaya et al., 1976; Cuny et al., 1981). With the recent completion of full genome sequencing projects, such as the yeast, worm, fruit fly, mouse, or human genome project, there is a renewed debate about the characteristic properties of isochores (Nekrutenko and Li, 2000; Häring and Kypr, 2001; Lander et al., 2001). One reason for this debate is the lack of a sequence-based definition of isochores. We use the recursive segmentation procedure to address this issue directly.

In order to apply the recursive segmentation algorithm to the isochore problem, we convert a DNA sequence to a binary sequence with the two symbols S (C or G) and W (A or T). Given a threshold s_0 for segmentation strength, the algorithm segments the input sequence into domains that are homogeneous by that criterion. Since the degree of homogeneity is a relative quantity, changing s_0 leads to different segmented domains. A large value of s_0 results in fewer large G + C domains, and a small value of s_0 leads to more small domains. There is no a priori reason to select a certain values of s_0, except for $s_0 = 0$, which corresponds to the default criterion based on BIC. We determine an optimal value of s_0 by segmenting sequences with a known isochore structure.

Some well studied isochores are the isochores in the major histocompatibility complex (MHC) sequence on human chromosome 6p21 (Beck et al., 1999). This 3.67 Mb long region contains 224 genes and has been linked to many human genetic diseases, including common diseases like rheumatoid arthritis and diabetes. There are four domains in this region, with two of them homogeneous enough to be called isochores. They are positioned on the chromosome as follows: (telomere)–(class I)–(class III)–(class II)–(extended class II)–(centromere) (Beck et al., 1999). Fig. 1 shows a recursive segmentation result for the MHC sequence, with s_0 set to 20, 10, and 5.

Fig. 1 shows clearly that the segmentations at the three known domain borders have the highest segmentation strength ($s = 236.79$, 170.84 and 288.49) (Li, 2001b,c). Class-III is the most homogeneous domain among the four, class-II is the domain with the second highest degree of homogeneity, and class-I is the least homogeneous domain. The 642 kb class-III domain is clearly a 'good' example of an isochore: it is larger than 300 kb, and it is homogeneous (at $s_0 = 10$) except for a few-kb region with a higher (G + C)% that can be detected at $s_0 = 5$. The border between the class-III and -II domains is a 'good' example of an isochore border because the (G + C)% change at this point is large. It is not surprising that this is by far the best studied isochore border (Fukagawa et al., 1995).

There are two technical remarks that requires further explanation. The first concerns the choice of s_0. Besides the general requirement that $s_0 > 0$, a selection of s_0

usually reflects the choice of the length scale. Fig. 2 shows, in a log–log representation, the distribution of segmentation strength of all recursive segmentations at $s_0 = 0$ when these segmentations are ranked. This plot is roughly the same as the similar segmentation strength profile at a $s_0 > 0$ value, with all low-ranked segments being removed. Since the number of segmentations is the number of domains minus 1, it is easy to know the required s_0 value for obtaining certain number of domains (e.g. $s_0 = 100$ for four domains).

A second technical remark concerns the question whether the domain borders determined by a recursive segmentation algorithm are accurate. This problem will be discussed more in Section 5 with the example of detecting the origin and terminus of replication. Generally speaking, a border position is more accurately determined when the sequence being segmented contains only two domains. For the determination of the isochore border in Fig. 1, for example, it is ideal to include only class-III and -II domains with only one segmentation. In most situations, however, the determined borders are robust and not very sensitive to an addition or deletion of sequences far away from the border.

We also apply the recursive segmentation algorithm to detect isochores in the longest contig on chromosome 1 of *C. elegans* (Ainscough et al., 1998), with 8.57 Mb. Although it is possible to patch other contigs (this contig–100 B gap–230 kb contig–1 kb gap–4.2 Mb contig) to form a longer sequence, we adhere our analysis to the originally given contig sequence in order to avoid potential problems caused by the treatment of gap sequences. Fig. 3 shows the segmented domains with $s_0 = 5$. A major portion of the sequence (6.2 Mb) forms one domain at this level. This long homogeneous domain is in strong contrast to the first 2 Mb sequence, which contains many small domains with extreme G + C contents, both high and low. Fig. 3 illustrates one advantage of the recursive segmentation over the traditional moving window approaches: both large and small homogeneous domains can be delineated at the same time.

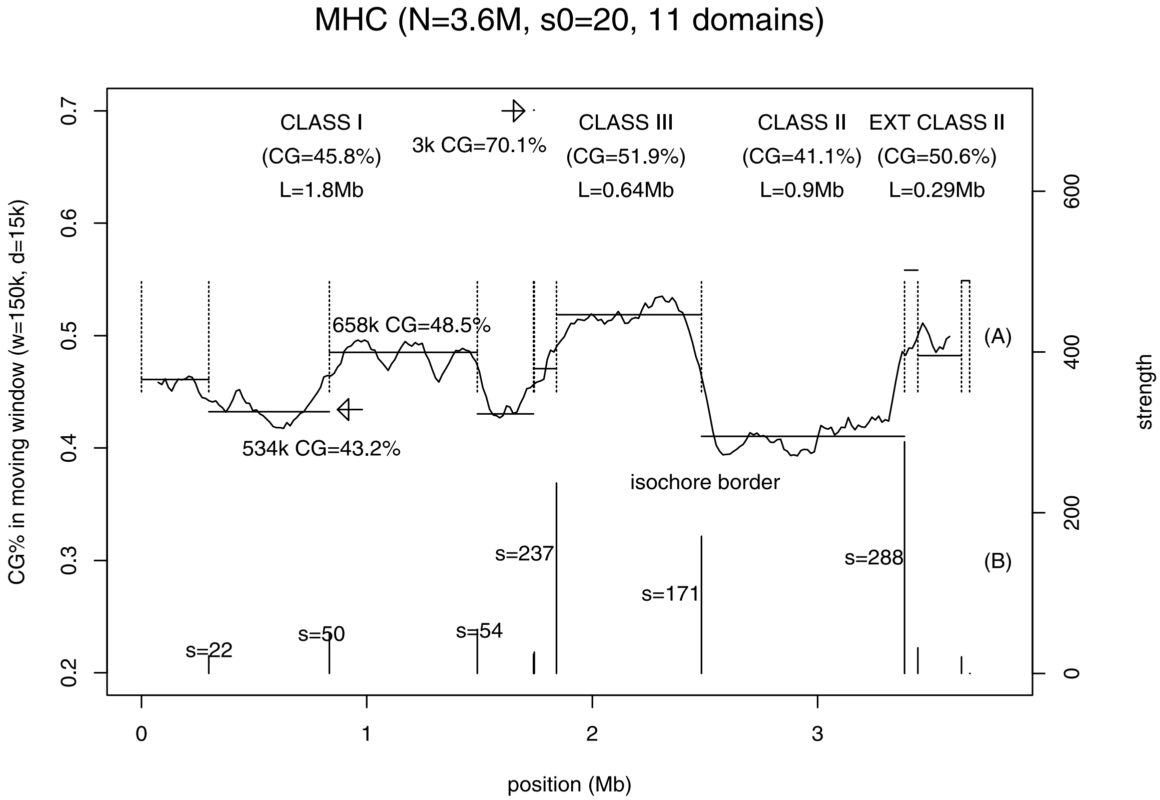

Fig. 1. Eleven domains segmented at $s_0 = 20$ for MHC sequence. Sequence length is $N = 3{,}673{,}778$ bases. (A) G + C% in a moving window (window size is 150 kb, and moving distance 15 kb); domain borders (vertical dashed lines); and G + C% in segmented domains (horizontal solid lines). (B) Segmentation strength s (vertical bars). The four known isochores (telomere is on the left side and centromere on the right side) class-I, -III, -II, and extended class-II, and the isochore border between class-III and -II are marked in the plot.

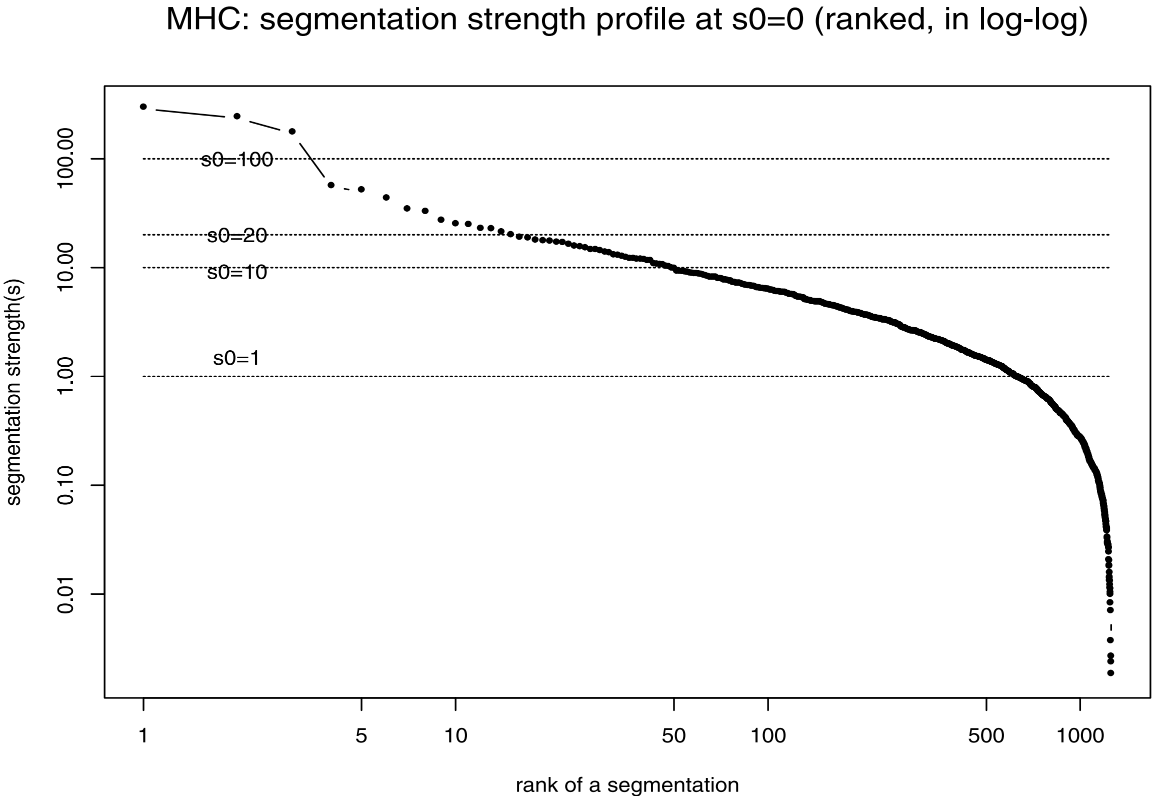

Fig. 2. Segmentation strength of all segmentations obtained at $s_0 = 0$ for MHC sequence. All segmentations are listed and ranked (total number is 1259). The segmentation strength is plotted against the rank, both in logarithmic scale. The number of segmented domains is the number of segmentations plus 1. With a raise of the s_0 (e.g. $s_0 = 1$, 10, 20, 100), the number of segments (and domains) is reduced, as shown in the plot.

4. Detection of CpG islands

CpG islands are short segments of unmethylated DNA sequences, usually (G + C)-rich and 5′-CG-3′ dinucleotide rich, that are dispersed throughout the comparatively (G + C)-poor genome (Cooper et al., 1983; Tykocinski and Max, 1984; Bird, 1986). CpG islands, generally 0.5–2 kb in length, have approximately 60–70% G + C content and are highly conserved throughout evolution. It is believed that mammalian chromosomes are organized into domains with characteristic CpG island density, where the distribution of the islands is correlated with the 5′ ends of genes such that the islands contain both the promoter and transcription unit (Cross et al., 2000). CpG islands have also been observed at the last exon and the 3′ untranslated region of genes (Gardiner-Garden and Frommer, 1987).

For computational detection of CpG islands, a standard sequence-based definition has been adopted which does not directly account for the methylation state of the sequence. This defines a CpG island as a region greater than a few hundreds of bases in length, with (G + C)% larger than 50%, and ratio of observed versus expected number of CpG dinucleotides, O/E = (CpG%)/(C%·G%), larger than 0.6, as proposed in Gardiner-Garden and Frommer (1987), Larsen et al. (1992). In the standard sliding window approach for detecting CpG islands, a moving average for the (G + C)% and O/E statistics are calculated by moving a window of length 100 bp and moving distance of 1 bp across the sequence (Gardiner-Garden and Frommer, 1987). Various threshold levels for these CpG based metrics and window sizes have been explored. For example, it was suggested in Venter et al. (2001) to use a more stringent value of 0.8 for O/E and Matsuo et al. (1993) proposed using a larger window size of 500 bp with moving distance of 10 bp for identification of CpG islands.

Matsuo et al. (1993) also proposed that CpG density (CpG%) may be used as a selection criterion for human CpG island, rather than the combined G + C content and ratio of observed versus expected CpG dinucleotides, since CpG density may be a predictor of methylation state. Their studies showed that a CpG dinucleotide count of greater than 6 per 100 bp might be an indicator that the region is unmethylated. This was proposed as a simple criterion for CpG islands. Here, we also chose to use the CpG density as the evaluation metrics for selection of candidate islands.

In our alternative method for delineating the location of CpG islands, we first preprocess the data, converting the DNA sequence into a binary sequence with two 1s corresponding to an observed CpG dinucleotide and 0 otherwise. In this way, CpG islands are characterized as regions with a high density of symbols coded as 1. Then, the recursive segmentation algorithm is applied to the binary sequence to identify subsequences with homogeneous 1 or 0 symbol composition, corresponding to potential CpG-rich and -poor domains, respectively.

In order to assess the accuracy of the computationally predicted CpG islands, we examined in detail the predictions of the segmentation method using a sample sequence with both putative and experimentally confirmed CpG islands. The sample DNA segment used was from human chromosome 22q13.2–13.3 with GenBank accession number AL022237. The GenBank annotation for this sequence contained three putative CpG islands with start and end base positions as follows: (#1) 4231–4718, (#2) 25125–26783 and (#3) 57881–58762. Although it is not specified in the annotation summary, these CpG islands were most likely derived from the sliding window method described in Larsen et al. (1992). A computer program implementing this algorithm predicted the three CpG islands exactly at the same locations, plus the fourth one at positions 3546–3852.

In Fig. 4, four CpG related statistics are calculated and displayed (i.e. CpG O/E, CpG%/GpC% ratio, (G + C)%, and CpG%) demonstrating that the CpG density (CpG%) is a valid indicator of CpG islands. The thresholds used for prediction of CpG islands for these statistics, given in the same order as above, are 0.6 (Gardiner-Garden and Frommer, 1987) and 0.8 (Venter et al., 2001), 0.6 (assuming GpC% ≈ C%·G%, as suggested by Bird (1986), 50% (Gardiner-Garden and Frommer, 1987), and 6% (Matsuo et al., 1993).

When the segmentation was applied to the converted binary version of this sequence, 30, 16 and 10 domains were observed for the corresponding segmentation strengths, $s_0 = 0$, 0.5, and 1, respectively. In Fig. 4, the 10 homogeneous domains obtained with $s_0 = 1$ are shown, along with the domain borders as well as the corresponding CpG densities. For the chosen level of significance for CpG density (CpG% > 6%), the first

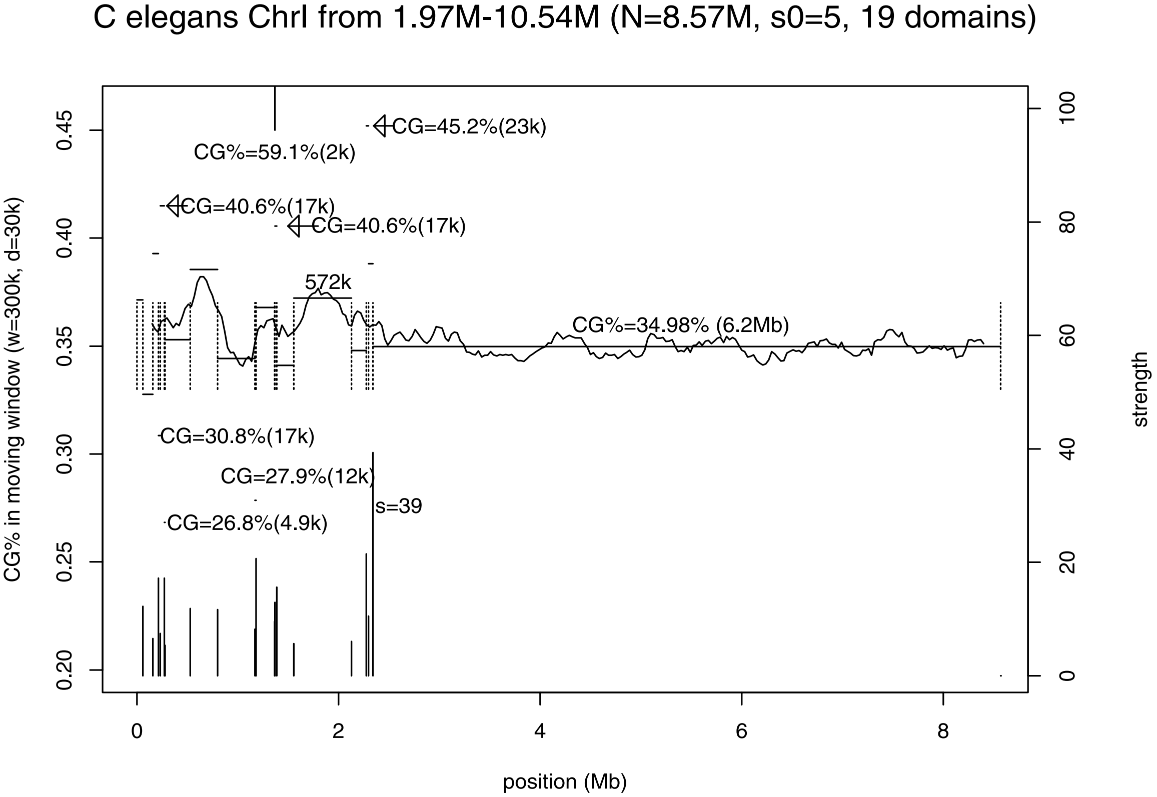

Fig. 3. Nineteen domains segmented at $s_0 = 5$ for the longest contig of chromosome 1 of *C. elegans* (see the caption of Fig. 1 for an explanation of the plot).

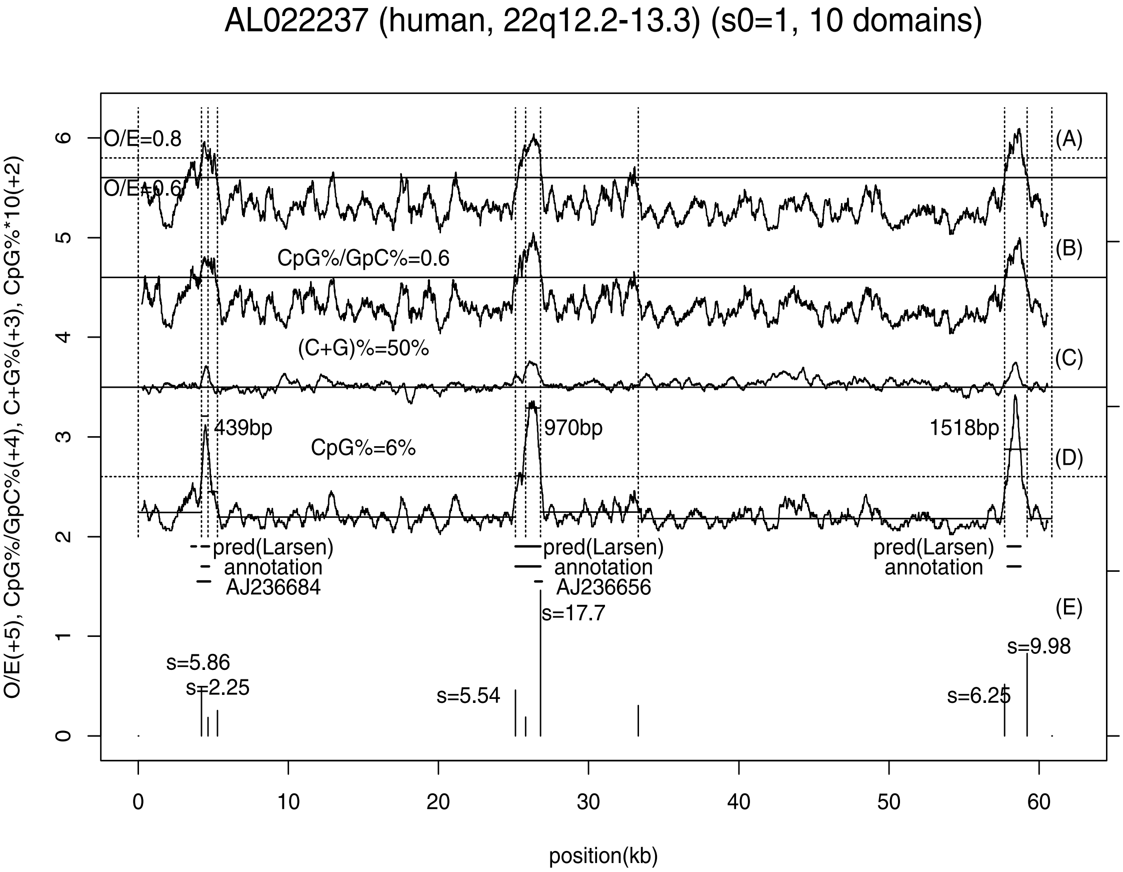

Fig. 4. CpG island detection for a human DNA sequence on 22q12.2–13.3 (AL022237). Several statistics are calculated in a moving-window (window length is 500 bp, moving distance is 10 bp). (A) Observed over expected occurrence of CpG dinucleotide: O/E = CpG%/(C%·G%). Two levels of O/E = 0.6 and 0.8 are also drawn. (B) Observed occurrence of CpG dinucleotide over that of GpC dinucleotide (CpG%/GpC%). The level of 0.6 is drawn. (C) G + C content. The level of 0.5 is drawn. (D) CpG content (number of CpG dinucleotide in a window divided by the window length). The level of 0.06 is drawn. A recursive segmentation (with $s_0 = 1$) is applied to the 1/0 (part of CpG/not part of CpG) sequence, and the domain borders (long vertical dashed line) and CpG% within the segmented domains are shown. The positions of three putative CpG island from the GenBank annotation, those of the four CpG islands as predicted by a C program based on the algorithm in Larsen et al. (1992), as well as two CpG-island-containing sequences (AJ236684 and AJ236656), are marked. (E) Segmentation strength of the nine segmentations.

annotated CpG island corresponds to one segmented domain (start–end: 4231–4669), the second corresponds to two segmented domains (start–end: 25125–25813, and 25814–26783), and the third corresponds to one segmented domain (start–end: 57682–59199) (see Fig. 4). Some domain borders match the annotated CpG islands exactly: these consist of the left border of CpG # 1 and, left and right borders of CpG # 2.

Two experimental CpG islands with GenBank accession numbers AJ236684 and AJ236656 were placed on the sequence at the following coordinates: start–end: 3953–4808 and 26422–26894. These experimentally derived CpG islands were obtained from the chromosome 22 CpG island libraries produced by Cross et al. (2000), and satisfied the CpG related threshold statistics for the assumed sequence-based definition (i.e. O/E = 0.77, (G + C)% = 62.0%, CpG% = 7.4% for AJ236684, and O/E = 0.94, (G + C)% = 63.4%, CpG% = 8.3% for AJ236656). AJ236684 spans a region larger than both the putative CpG island coordinate given in the GenBank annotation and our segmentation domain, whereas AJ236656 covers a smaller region (see Fig. 4). A likely explanation for AJ236684 being larger than the predicted size is that it may contain a core CpG island domain that may be flanked by other sequences. In fact, the experimental CpG island sequences are generated via MseI restriction enzyme digestion, that may either result in fragments with flanking sequences around a core CpG domain (as seen above), or fragments with split or partial CpG domains (Cross et al., 2000). This poses a challenge in evaluation of computational CpG island predictions relative to experimentally identified islands.

A possible explanation for AJ236656 being smaller than the predicted size is that there could be subdomains within a CpG island. In fact, our recursive segmentation method reveals that this is possible. The second predicted CpG island by our method corresponds to two segmented domains at segmentation strength threshold $s_0 = 1$, one with a borderline CpG% of 6.1% and another one with a much higher CpG% of 12.8%. The third CpG island on Fig. 4 provides another example: though it is one domain segmented at s_0, it becomes four domains segmented at $s_0 = 0.5$. These four domains are located at 57682–58130 (449 bases, CpG% = 6.2%), 58131–58699 (569 bases, CpG% = 1.39%), 58700–59191 (492 bases, CpG% = 4.7%), and 59192–59199 (eight bases, CGCGCCCG, with CpG% = 3/8 = 37.5%). It is clear that between the CpG-rich 449-bp subdomain and the last eight bases, there are CpG-poor subdomains. The recursive segmentation method is better suited in delineating these subdomains than a window-based approach.

5. Detection of replication origin and terminus

It is known that circular bacteria genomes are compartmented into two partitions by the replication origin and terminus (Lobry, 1996a,b). At the replication origin and terminus, (A − T)% or (G − C)% changes sign, in a violation of the global strand symmetry of A% ≈ T% and G% ≈ C% (Lin and Chargaff, 1967; Karkas et al., 1968; Rudner et al., 1968; Fickett et al., 1992). Possible mechanisms for strand symmetry and asymmetry have been discussed in Lobry (1995), Sueoka (1995), Francino and Ochman (1997), Frank and Lobry (1999). Clearly, the strand asymmetry can be used in a sequence analysis to locate the replication origin and terminus in bacteria sequences (Frank and Lobry, 2000).

If (A − T)% and (G − C)% change sign in the same direction (e.g. from positive to negative) at the replication origin or terminus, one can combine the two to use the (R − Y)% (R(purine, A/G), Y(pyrimidine, T/C)) to detect replication origin and terminus. If the two changes sign oppositely, one can use the (M − K)% (M(amino, A/C), K(keto, T/G)) for that purpose. Both situations are possible for bacteria sequences (McLean et al., 1998; Grigoriev, 1998b). Here, a sequence in the second situation, *H. influenzae* (Fleischmann et al., 1995), is used for an illustration.

When a recursive segmentation is applied to the amino(M)–keto(K) sequence (or R–Y sequence if (A − T)% and (G − C)% change sign in opposite directions) converted from a circular bacterium sequence, the two compartments separated by the replication origin and terminus are expected to have different M/K composition. Since we are only interested in the two partition points, the segmentations with the two highest segmentation strength are enough for matching the replication origin and terminus. Subsequent segmentation in a recursion and stopping criterion are not necessary.

Fig. 5 shows the result for *H. influenzae* bacterium genome. Skew statistics of (M − K)%/(M + K)% (solid lines), (A − T)%/(A + T)% (dotted line), and (G − C)%/(G + C)% (dash line) are shown in a moving window (window length = 5 kb, moving distance 500 bp). The two strongest segmentations ($s = 18.15$ at position 616,967 and $s = 14.41$ at position 1,503,627) are shown by the long vertical lines. These two segmentation points represent the 33.71 and 82.15% of the whole genome. The distance between the two segmentation positions is 48.44%, consistent with the typical distance of half of the genome size. The skews (A − T)% and (G − T)% both change sign at the segmentation point at 616 kb, in opposite direction, as expected for a replication origin.

In Grigoriev (1998b), the replication origin and terminus is reported to be at 32.9 and 80.1% of the genome, respectively. In Lobry (1996a), the replication origin and terminus for this bacterium is listed as at 32.95 and 82.94%. Both are close, but not exactly identical to our two largest segmentation positions. To test how robust out segmentation point is, we design the following experiment: rather than starting the sequence from the given end of the sequence, which is usually provided arbitrarily, we rotate the starting position around the circular bacterium genome. Fig. 6 shows the position of the first segmentation $I_{\max}$ (top plot) and the corresponding $2N\hat{D}_{\mathrm{JS}}$ (bottom plot) as a function of the sequence starting point *i*. If the result on the first segmentation position is robust, rotating the sequence starting point will more or less lead to the same point near replication terminus. Similarly, when the starting position is moved close to the replication terminus, the $I_{\max}$ will be close to the replication origin. Ideally, the $I_{\max}$ versus i plot should consist of two plateaus. It can be seen from Fig. 6 (top plot) that $I_{\max}$ does contain two major plateaus, but the level of the plateau may change slightly.

In order to determine which plateau value should be used as the candidate replication origin and terminus, we make the following assumption: the $2N\hat{D}_{\mathrm{JS}}$ will reach the highest value when the sequence starting point coincides with either the replication origin or terminus. In Fig. 6, when the starting point is rotated to 32.90%, a local (with respective to i) maximum $2N\hat{D}_{\mathrm{JS}}$ is reached, and the corresponding $I_{\max}$ is at 80.59%. Similarly, when the starting point is rotated to 80.60%, a local maximum $2N\hat{D}_{\mathrm{JS}}$ is reached, and the corresponding $I_{\max}$ is 32.78%. Clearly, the best segmentation is achieved (for M–K sequence) when the two partition points are at 32.78–32.90 and 80.59–80.60%, and the two points should be the best candidate for the replication origin and terminus.

These adjusted segmentation points are drawn in Fig. 5 (shorter vertical lines): they are clearly shifted away from the first two segmentations in a recursive approach. For replication origin, the adjusted position is almost identical with the previous reported positions. For replication terminus, the previous two reported positions are not consistent with each other (80.1% in Grigoriev (1998b) and 82.94% in Lobry (1996a)). Our adjusted position is 80.59–80.60%, in-between the two. In general, replication terminus may not be as well defined as the replication origin. But it will still be interesting to know whether our predicted position is better than previously reported ones.

If the recursion continues, the two segmentations with the next largest strength are $s = 3.50$ at position 155,088 and $s = 1.04$ at 451,651. The strengths of these segmentations are much smaller than those at the replication origin and terminus. The strong global signal at replication origin and terminus is in a clear contrast with the relative weak signal obtained from a moving window approach: it is not obvious from Fig. 5 that one can identify an unique position where (A − T)% and (G − T)% both change signs in opposite directions. In fact, the cumulative skew plot used in Freeman et al. (1998), Grigoriev (1998a,b) (also called ‘genomic landscapes’ in Lobry (1999)) takes the same advantage in using a global, cumulative method. Our recursive segmentation approach for detecting replication origin/terminus should be considered similar to the cumulative skew plot, except that our approach can be extended to skew plots of more than two symbols.

Table 1 lists the determination of replication origin and terminus in other bacteria genomes using the method described above. Since our segmentation method can only identify the two points with the strongest signal, but not on which one is the replication

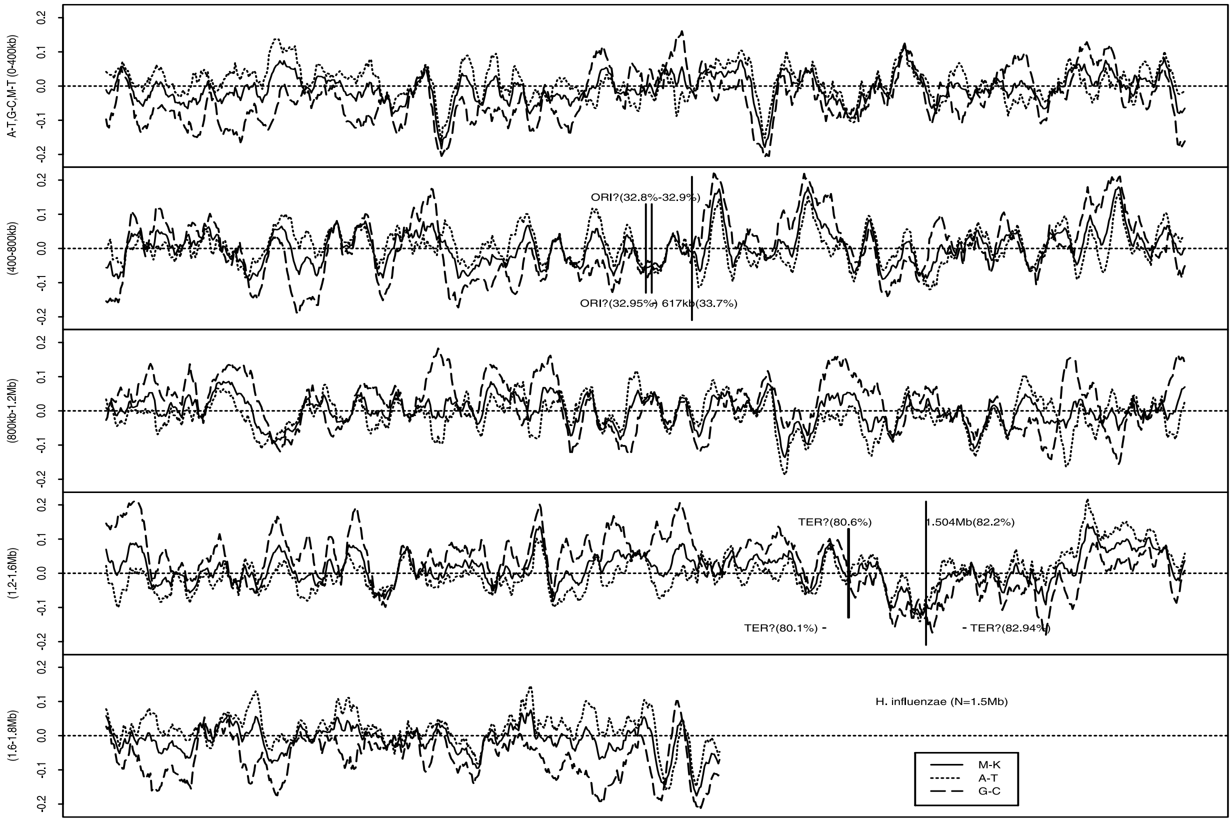

Fig. 5. Detection of replication origin and terminus in bacterium *H. influenzae* sequence. The first segmentation is obtained at position 1,503,627 (or 82.16%) with the strength $s = 14.4$, near the replication terminus. A second-stage segmentation is obtained at position 616,967 (or 33.71%) with the strength $s = 18.15$, near the replication origin. The shorter vertical lines are adjusted position for the segmentation points (see Fig. 6). The reported putative positions of the replication origin (32.9% in Grigoriev (1998b) and 32.95% in Lobry (1996a)) and replication terminus (80.1% in Grigoriev (1998b), and 82.94% in Lobry (1996a)). Several statistics in moving window (window length = 5 kb, moving distance = 500 bp) are also shown: (M − K)%/(M + K)% (M for amino, or A + C, K for keto, or T + G) in solid line, (A − T)%/(A + T)% in dotted line, (G − C)%/(G + C)% in dashed line.

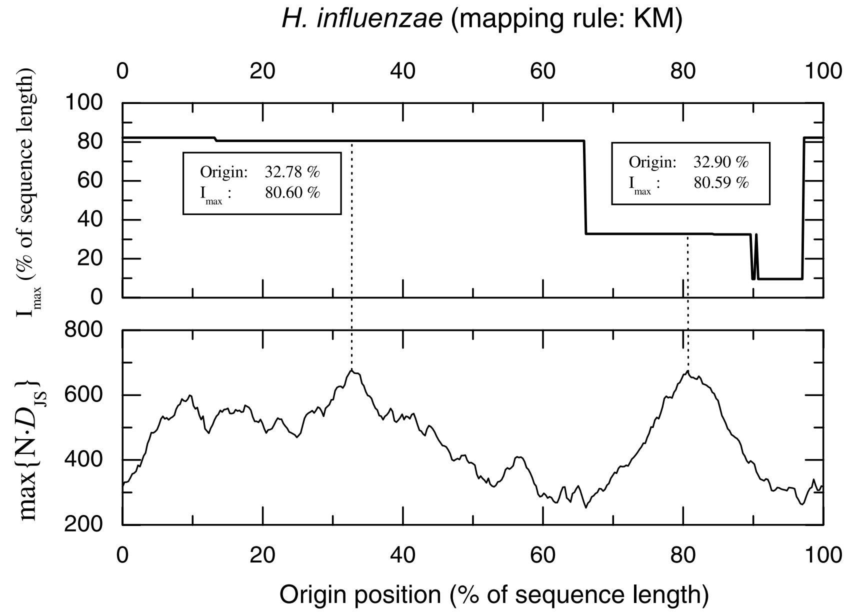

Fig. 6. Adjusted position for putative replication origin and terminus for bacterium genome *H. influenzae*. How the first segmentation result (no recursion) changes with the sequence starting point (rotating in the circular genome) is studied. Top: the first segmentation point (I_{max}) as a function of the starting point i (rotating to the 3′ direction in terms of percentage of the genome size) (solid line). Bottom: the $2N\hat{D}_{JS}$ as a function of i (solid line). Two local maxima (with respect to i) of $2N\hat{D}_{JS}$ are identified, and their corresponding I_{max}'s. This technique identifies the replication origin and terminus at 32.78–32.90 and 80.59–80.60%.

Table 1
Predicted locations of replication origin and terminus for 12 bacteria genomes

Genome	Origin position		Terminus position		Distance (%)
	bp	%	bp	%	
B. burgdorferi (KM)	458,750	*50.37*	910,606	*99.99*	49.61 or 50.39
E. coli (KM)	3,921,150	*84.52*	1,606,250	*34.62*	49.89 or 50.11
B. subtilis (RY)	350	*0.01*	1,941,675	*46.06*	46.06 or 53.94
C. trachomatis (KM)	720,350	*69.09*	209,250	*20.07*	49.03 or 50.97
H. influenzae (KM)	596,400	*32.58*	1,475.425	*80.62*	48.03 or 51.97
H. pylori (RY)	134,775?	8.08?	768,050	*46.05*	37.97?
M. genitalium (RY)	0	*0*	295,475	*50.94*	49.06 or 50.94
M. pneumoniae (RY)	113,900?	13.95?	418,100?	51.21?	37.26?
M. tuberculosis (KM)	1775	*0.04*	2,044,550	*46.35*	46.31 or 53.69
R. prowazekii (RY)	44,825	4.03	629,475	56.63	47.40 or 52.60
R. prowazekii (KM)	1,103,075	99.23	666,750	59.98	39.25 or 60.75?
T. maritima (KM)	156,250	*8.39*	1,156,800	*62.16*	46.23 or 53.77
T. pallidum (KM)	3300	*0.29*	556,300	*48.89*	48.59 or 51.41

It is indicated whether the purine–pyrimidine (R–Y) sequence or the amino–keto (M–K) sequence is used. The prediction locations are listed in both basepair (bp) and in percentage of the total genome size (%). The distance between the two positions in percentage of genome size is also given (in both direction). Problematic positions are labeled by question marks (?).

origin and which is the terminus, we rely on previously published information (for the first nine sequences in Table 1, the information provided by Grigoriev (1998b) is used; for *R. prowazekii*, the original sequence paper (Anderssen et al., 1998) is used; for *T. maritima* the prediction of the replication origin by a tetramer skew plot in Lopez et al. (2000) is used, though no information on the replication terminus is available; for *T. pallidum* the information given in Grigoriev (1998b) cannot be used because the starting point of the sequence had been shift since then, and we use the information on replication origin from the original paper (Fraser et al., 1998)).

Our segmentation procedure is applied to both the M–K and R–Y sequences for all these bacteria genomes. By the information provided in Grigoriev (1998b), it is mostly known whether the M–K or the R–Y sequence should be used. If this information is not known, we rely on two other pieces of information: whether the determined two positions differ by ~50% of the genome size, and whether the maximum segmentation strength is higher. For genome *R. prowazekii*, however, the R–Y sequence leads to the expected partition size, whereas K–M sequence leads to a higher segmentation strength (both are listed in Table 1). The replication origin of *H. pylori* should be close to position 0, but as pointed out in Grigoriev (1998b), a false positive signal is created at ~8% due to a large inversion. Our origin/terminus prediction of *M. pneumoniae* does not match the positions given in Grigoriev (1998b). Besides this sequence, prediction of replication origin/terminus for all other bacteria genomes is consistent with the previous results, though small differences exist. It should be interesting to confirm which method is more accurate.

6. Detection of complex patterns in telomeres

We have so far only discussed applications where the number of symbols in the filtered sequence is smaller than the original number of symbols. It is also possible to move in an opposite direction: to expand the symbol list. For example, all dinucleotide can be represented by a set of 16 symbols. If a region distinguishes itself by being abundant in certain dinucleotide, but not necessarily being abundant in certain nucleotide, a segmentation using dinucleotides may provide a better detection of this region than that based on nucleotide composition alone. As we have seen in Section 5, CpG-rich regions do not necessarily $(G+C)$-rich, even though most of the time the two are correlated.

To examine whether segmentation based on dinucleotides gives different result from that based on single nucleotides, we use the same sequence discussed in Li (2001a), the left telomere sequence (first 15 kb) of yeast *Saccharomyces cerevisiae* chromosome 12 (Johnston et al., 1997). It is known that yeast telomere contains both telomeric and subtelomeric elements (Olson, 1991). The telomeric element is the TEL sequence with a simple 5′-C_{1-3}A-3′ repeats at the tip of the telomere. The subtelomeric elements mainly refer to the X and Y′ elements which are conversed in yeast telomeres (Szostak and Blackburn, 1982; Chan and Tye, 1983; Louis and Haber, 1990, 1992; Louis et al., 1994; Wellinger and Sen, 1997).

Since TEL sequence is a simple repeat, it can be detected even by a visual inspection of the sequence. Subtelomeric elements X and Y′ are defined by a conservation among yeast telomere at different chromosomes, and they may not be compositionally distinct by itself. Fig. 7 shows the four base composition in a moving window (window size = 150 bp, moving distance = 50 bp) for left telomere of yeast chromosome 12. X element is G-poor and C-rich, but Y′ element does not seem to be compositionally distinct. The domains by segmenting the original four-symbol sequence and by segmenting amino(AC)–keto(TG) sequence are shown in Fig. 7. Both segmentations are able to detect the rough region of the X element, but both fail to detect the two Y′ elements.

Segmentation result on the converted 16-symbol sequence is very close to that of the four-symbol sequence, though with one less domain. If X and Y′ elements are not compositionally distinct themselves, we need to use other neighboring distinct sequences for the detection. For example, X elements usually contain, or locate nearby, the autonomous replication sequence (ARS) (Chan and Tye, 1983). As we have known from Section 5 that the mutation pattern near a replication origin tends to lead strand asymmetry, either M–K or R–Y sequence can be segmented to detect this position (though this may not be true for all ARS sequences, it is true for ARS's near the telomeres (Gierlik et al., 2000).

Y′ elements are known to contain, or locate nearby, a 36-bp repeat sequence (Horowitz et al., 1984). We have applied a tandem repeat program (Benson, 1999) to this sequence and the detected repeats are shown in Fig. 7. Indeed, the two 36-bp repeats are next to the two Y′ elements (whose position is estimated by a dot-matrix comparison with the known Y′ element using the DOTTER program (Sonnhammer and Durbin, 1995)). This 36-bp repeat is GT-rich (or AC-rich on the opposite strand). Why was this distinct repeat not detected by our recursive segmentation?

The reason for failing to detect the GT-rich 36-bp repeat by the recursive segmentation is that the distinct base composition in the repeat is confined to a local region. Its base composition may be significantly different locally, but not significant in a global scale. To solve this problem, we relax the segmentation stopping criterion first, then purge insignificant cuts afterwards. For example, when the s_0 is set between -0.7 and -1 for segmenting the 16-symbol sequence, many domains (most are of very small size) are created. We then keep only the segmentations that have $s > 0$. These segmentation points are shown in Fig. 7 $(+)$, and they do include borders separating the 36-bp repeats and their neighboring sequence (only one out of two border). This trick is similar to what is used in the binary tree or recursive partitioning (Breiman et al., 1984; Zhang and

Singer, 1999): one may not apply the stopping criterion in the tree branching stage, but apply a criterion for leave-merging.

7. Detection of coding–noncoding borders

Computational detection of protein-coding genes has been a long-standing topic of computational biology (Fickett, 1982; Staden and McLachlan, 1982). A typical gene recognition program contains one or all of these components: (1) a measurement of the coding potential in a moving window along the sequence (Fickett and Tung, 1992); (2) detection of 'signals' such as start/stop codon and splicing sites; (3) consultation to external information such as known protein sequences, cDNA, and ESTs. For reviews on gene prediction, see Fickett (1996), Claverie (1997), Burge and Karlin (1997). For more publications on this topic, see an online resource at http://linkage.rockefeller.edu/wli/gene/.

The coding potential measurement is obtained from within a coding or noncoding region (as versus from their borders). Such measurement can either be learned from the data or can be based on a known biological knowledge (the term 'measures dependent and independent of a model' is used in Guigo (1999)). If we know everything biologically about what makes a region coding or noncoding, the 'model' becomes a known knowledge, and the difference between the two disappears. However, the current biological knowledge about coding potential is still mainly limited to that of the codon structure. The fact that coding regions, and not the noncoding regions, consists of three-base unit, plus the fact that these units are not used with equal probability, provides a strong signal for coding potential. This 'periodicity-three' signal has been discussed in Fickett (1982), Tiwari et al. (1997), Yan et al. (1998), Li (1998), Grosse et al. (2000).

Due to the importance of codon position, we can combine base and position information by converting a

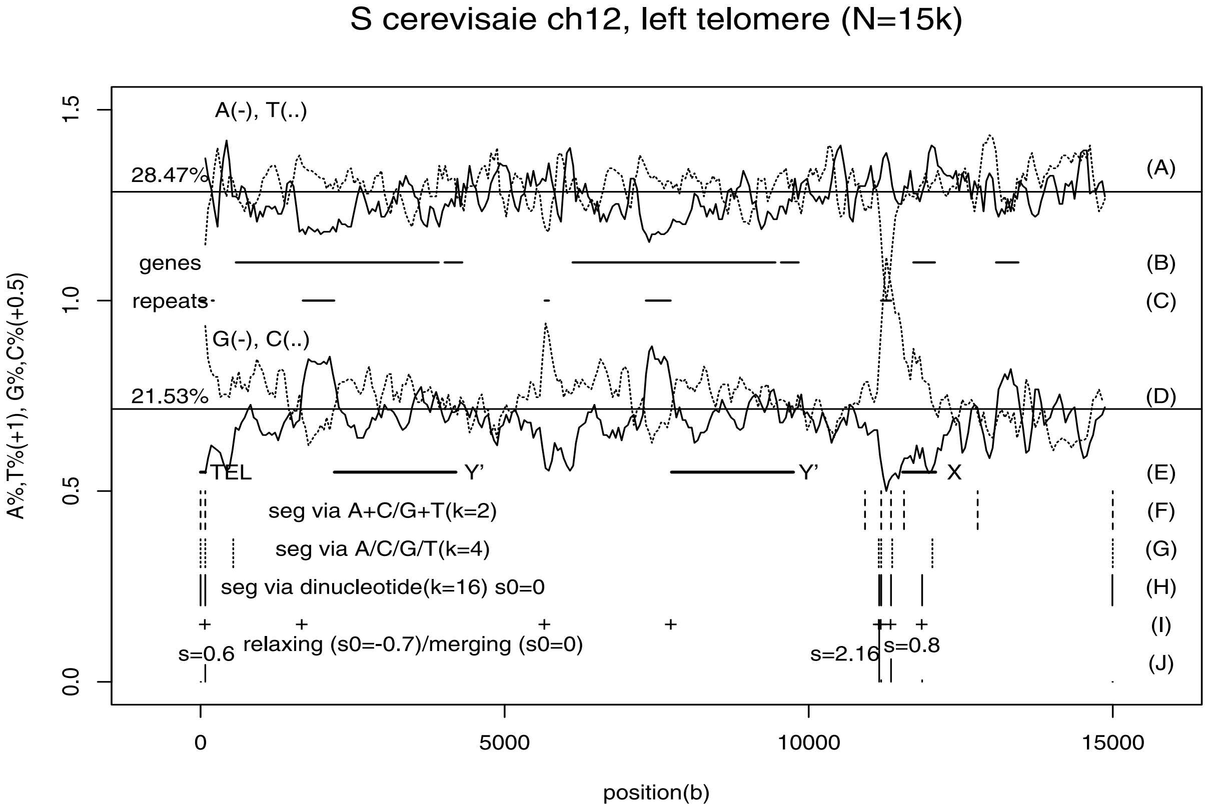

Fig. 7. The left telomere (15 kb) of the yeast *S. cerevisiae* chromosome 12 sequence. (A) A% (solid line) and T% (dotted line) in a moving-window (window size = 150 bp, moving distance = 50 bp). (B) Locations of four known genes in this region. (C) Regions of tandem repeats. (D) G% (solid line) and C% (dotted line) in a moving window. (E) Locations of conserved 'elements' (TEL sequence, two Y′ elements, one X element). (F) Borders of segmented domains using M–K sequence (at $s_0 = 0$). (G) Borders of segmented domains for the four-symbol (original) sequence (at $s_0 = 0$). (H) Borders of segmented domains for the 16-symbol (dinucleotide) sequence (at $s_0 = 0$). (I) Similar to (H), but the s_0 is initially set at -0.7, then only the segmentations with $s > 0$ are saved. (J) Segmentation strength corresponding to segmentations in (H).

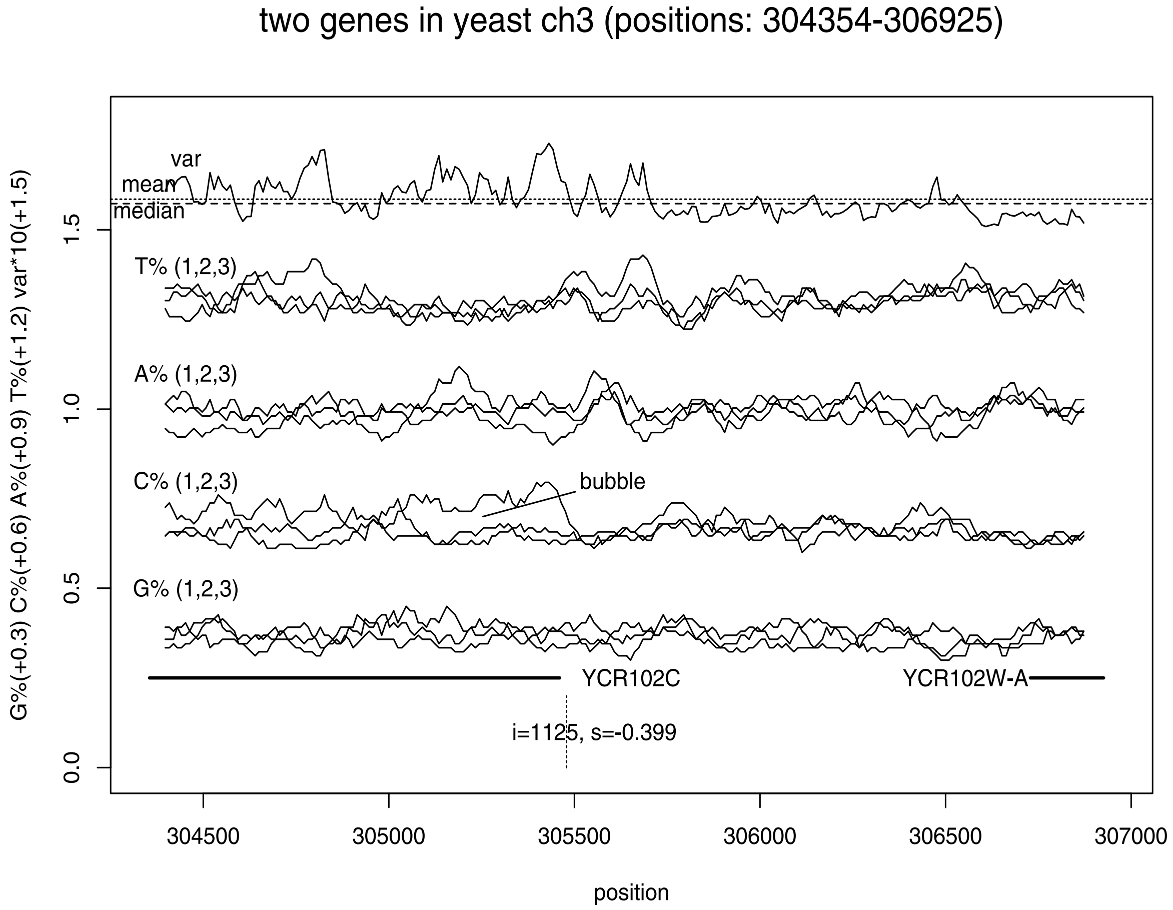

Fig. 8. Detection of coding/noncoding borders. A DNA sequence near the right telomere of yeast *S. cerevisiae* (positions 304,354–306,925). The location of two genes are marked. Base composition for each one of the three phases is calculated in a moving-window (window size = 87 bp, moving distance = 11 bp). Whenever the same base at three different phases exhibit different composition, three curves diverge to form a 'bubble'. One such bubble near the end of the first gene is marked. The variance (sum over four symbols) is also plotted, as well as the mean and the median values (over all windows). The first segmentation point is represented by a vertical line (not significant at $s_0 = 0$, but significant at $s_0 = -0.4$).

DNA sequence to a 12-symbol sequence (Bernaola-Galván et al., 2000): A_1 for nucleotide A in phase 1, A_2 for nucleotide A in phase 2, etc. If there is a difference between the 12-symbol composition in coding and noncoding regions, applying recursive segmentation may detect these domains. In most bacterial and yeast sequences, the first codon position is A + G-rich (purine-rich), second codon position is C-rich, and the third codon position is A-rich, etc. (Mrázek and Kypr, 1994; Li, 1999). In human sequences, the third codon position is C-rich, etc. (Wada et al., 1991). Suppose a gene starts from the phase 1, the above codon usage tendency will imply that this region in the 12-symbol sequence is rich in A_1, G_1, C_2, A_3, and C_3. On the other hand, nearby noncoding regions may not share this tendency.

We first illustrate this approach by a DNA sequence from yeast chromosome 3 (positions 304,354–306,925). This region is near the right telomere, and contains two genes (YCR102C, a gene coding for alcohol dehydrogenase, positions 304,354–305,460; and YCR102W-A, a gene similar to other membrane genes, positions 306,728–306,925). Base compositions for each base at three phases (within a moving window: window length = 87 bp, moving distance = 11 bp) are plotted as a group. Any divergence between the three compositions in the same group indicates a bias in three phases, most likely caused by the codon usage in a coding region. Such divergence forms a 'bubble' in the plot. On the other hand, if the three compositions in a group move together, there is no difference between the three phases, most likely indicating a noncoding region.

Fig. 8 indeed shows some bubbles being formed in the first gene (YCR102c), but not in the second gene (YCR102W-A). To summarize divergence/bubble for all four nucleotides, we define a sum of variance as:

$$\mathrm{var} = \sum_{i=1}^{3} [(\mathrm{A}_i\% - \bar{\mathrm{A}}\%)^2 + (\mathrm{C}_i\% - \bar{\mathrm{C}}\%)^2 + (\mathrm{G}_i\% - \bar{\mathrm{G}}\%)^2 + (\mathrm{T}_i\% - \bar{\mathrm{T}}\%)] \quad (6)$$

where $\bar{\mathrm{A}}\%$ is the average of base composition of A in three phases ($\mathrm{A}_1\%$, $\mathrm{A}_2\%$, and $\mathrm{A}_3\%$), etc. If all three compositions (e.g. for nucleotide A) move as one unit, $\Sigma_i\,(\mathrm{A}_i\% - \bar{\mathrm{A}}\%)^2 = 0$ and there is therefore no variance. We plot var by Eq. (6) in Fig. 8, and indeed it tends to be higher than the mean/median value in the first gene, but lower in the noncoding regions (though it is also lower in the second gene).

The stopping criterion based on BIC for the 12-symbol sequence is $2N\hat{D}_{\mathrm{JS}} > 10\log(N)$ instead of $2N\hat{D}_{\mathrm{JS}} > 12\log(N)$ (Li, 2001a). The reason for this is that there are three normalization conditions (constraints) for the 12 symbols, versus the one constraint in other cases. Consequently, the segmentation strength is defined as $s = (2N\hat{D}_{\mathrm{JS}} - 10\log(N))/(10\log(N))$. For this yeast sequence in Fig. 8, there is no segmentation that is significant at $s_0 = 0$. Nevertheless, if the segmentation stopping criterion is relaxed to $s_0 = -0.4$, one segmentation appears at position 305,478, close to the coding/noncoding border at 305,460. The second coding/noncoding border is not detected even if the s_0 is further reduced. One can see from Fig. 8 that the var is below the mean/median in the second gene, and it is unlikely that we will detect the second gene by this method. We will see later that segmenting a 12-symbol usually does not lead to a significant segmentation at $s_0 = 0$.

The second example is a human DNA sequence from chromosome 22 (total length = 50,823 bp) which contains one protein kinase *chk*2 gene (14 exons with

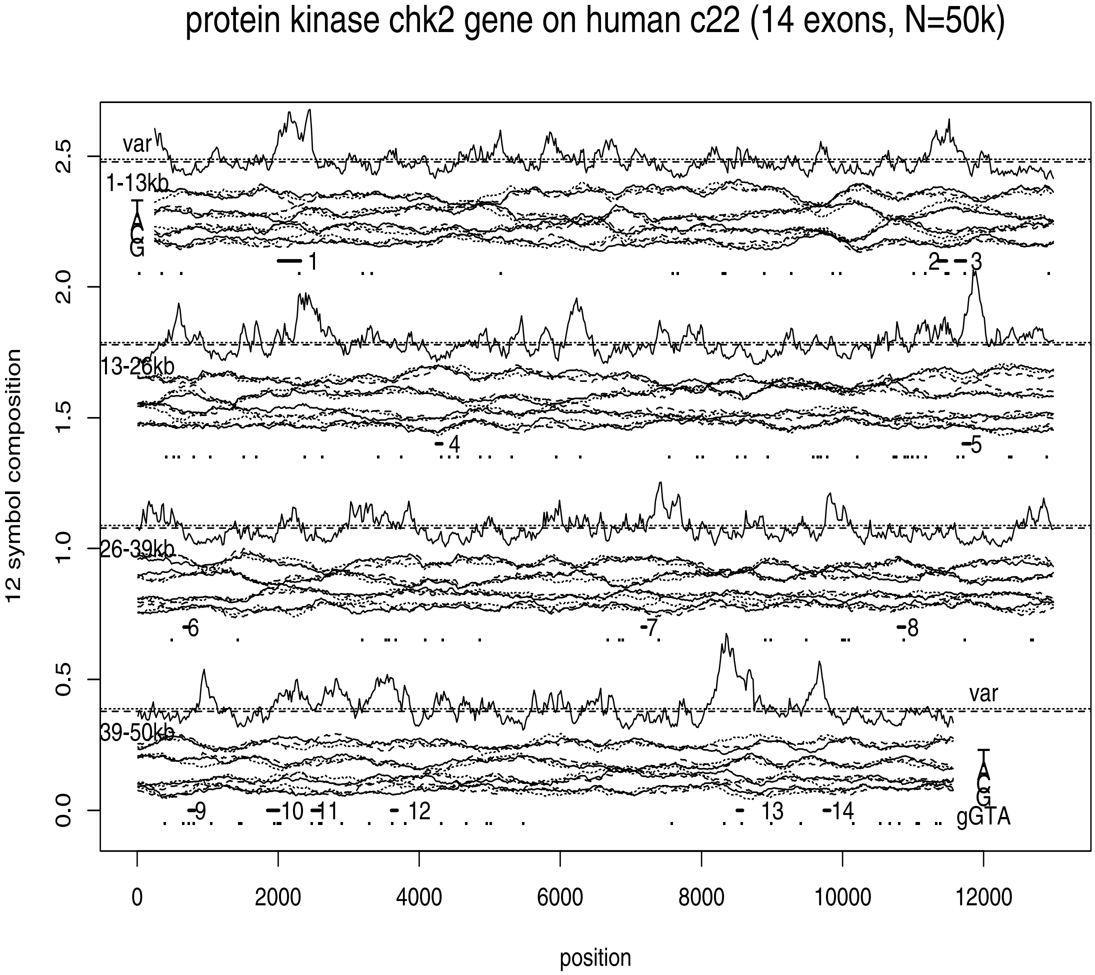

Fig. 9. The human protein kinase *chk*2 gene (on chromosome 22, 14 exons, sequence length $N = 50{,}824$ bases). The following information is shown: base compositions in three separate phases in a moving-window (window length = 500 bp, moving distance = 20 bp) G1%, G2%, G3%, etc.; variance by Eq. (6); location of 14 exons, matches with the pattern GGTA. Exon regions tend to have larger variances.

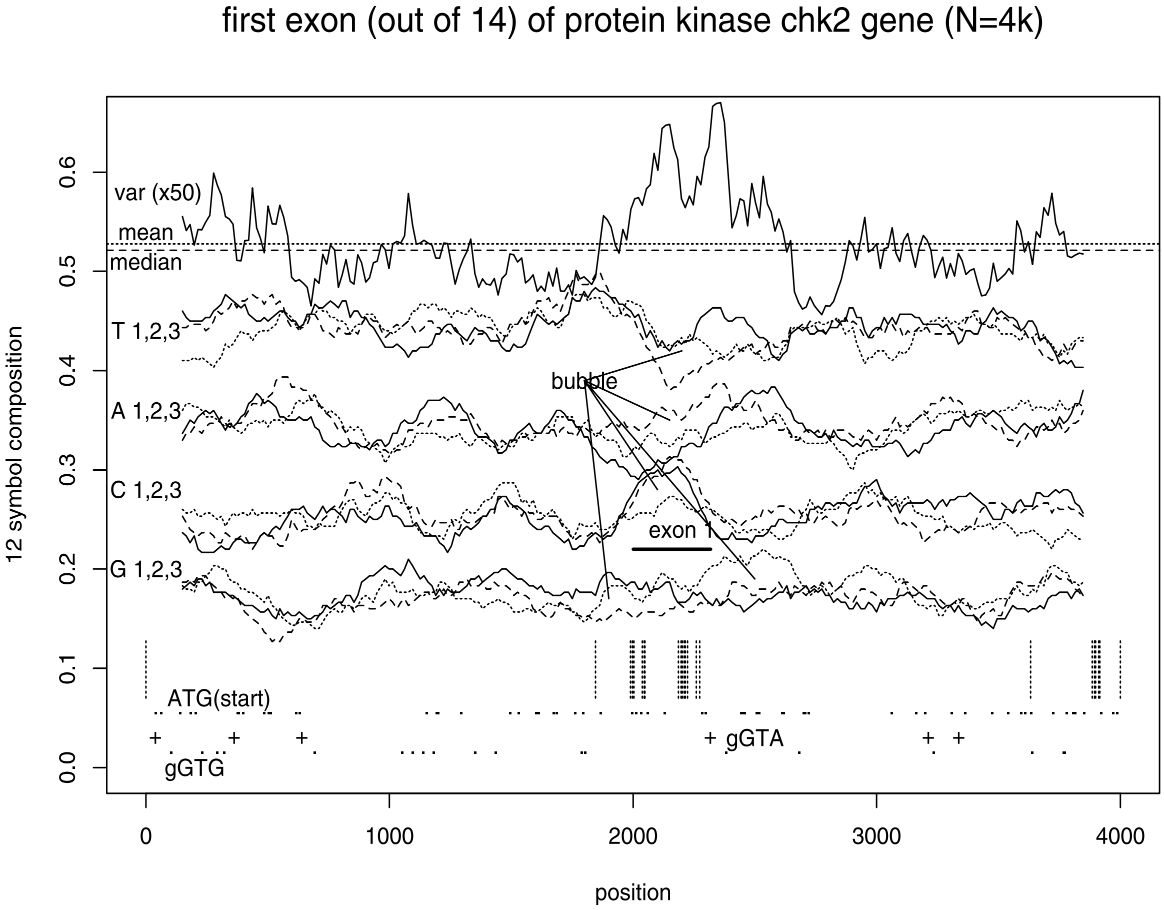

Fig. 10. The first exon of the gene sequence shown in Fig. 9. Again, base compositions in three different phases are calculated in moving windows (window length = 300 bp, moving distance = 16 bp). The variance calculated from Fig. 6 is shown, as well as its mean and median (averaged over all windows). There is clearly a peak in the variance near the first exon (which is marked by a solid bar). Matches to start codons (ATG), GGTA, and GGTG are shown in the plot. There is no significant segmentation at $s_0 = 0$. What are shown here are segmentations obtained at $s_0 = -0.7$ (total 22 domains).

cDNA length of 1632 bases). Fig. 9 shows 12-symbol composition in a moving window (window size = 500 bp, moving distance = 20 bp), with the same nucleotide in three different phases plotted together. The var defined by Eq. (6) is also plotted in Fig. 9. It is clear that exons 1, 2, 5, 7, 9, 10, 11, 12, 13 and 14 exhibit peaks in the var value, whereas exons 3, 4, 6, 8 do not seem to match any peaks. The segmentation result is not shown because the first segmentation is not significant at $s_0 = 0$. Even if the s_0 is reduced to a negative value, the resulting segmentation points do not match the exon borders.

Visual inspection of Fig. 9 seems to be able to identity regions with high variances, whereas recursive segmentation does not lead to any significant result. To reconcile the two, we focus on a shorter region—the first 4 kb sequence of the *chk*2-gene-containing sequence. Not only this region contain the largest exon (319 bases, while all other exons are around 100 bp or less), but also there is a clear peak in the variance plot. Once again, no segmentation is significant at $s_0 = 0$. Only with a relaxed stopping criterion $s_0 = -0.7$, does a recursive segmentation leads to 22 domains, half of them are at or near the exon 1 (see Fig. 10). The strongest segmentation is at position 2186 ($s = -0.37$), corresponds to a position inside the exon (positions 2000–2318).

Using segmentation or coding measure alone might be difficult to identify exon borders, and it is common to complement this approach by other signals such as putative splicing site. For example, we observed that for 10 out of 14 exons in this gene, the donor splicing site contains the pattern ‘G(− 1)G(+ 1)T(+ 2)A(+ 3)’ (dinucleotide GT at the first and second position of the intron, G at the last position of the exon, and A at the third position of the intron), two exons contain ‘T(−

1)G(+ 1)T(+ 2)A(+ 3)', one exon contains 'A(− 1)G(+ 1)T(+ 2)A(+ 3)', and one exon contains 'G(− 1)G(+ 1)T(+ 2)G(+ 3)'. Knowing this information, we can use the consensus pattern GGTA to allocate 3′ end of exons. When GGTA is searched in this sequence, there is only one match around the position 2000 (Fig. 10). Combining this information with the coding measure will point correctly to the first exon.

Note that both the locations of the start codon 'ATG' and those of the donor splicing site consensus pattern 'G(− 1)G(+ 1)T(+ 2)A(+ 3)' are too numerous to be much helpful (see Fig. 10). A success gene prediction program usually uses a combination of many pieces of information (Guigo et al., 1992; Solovyev et al., 1994; Uberbacher et al., 1996; Zhang, 1997; Burge and Karlin, 1997). It is interesting that when GENSCAN program (Burge and Karlin, 1997) is applied to this sequence, 10 exons are predicted exactly, but four other exons (10, 11, 12, 14) are missed completely.

8. Discussions

As can be seen in this paper, recursive segmentation is a very general approach for DNA sequence analysis. It is an interesting alternative to the traditional moving window approach. Admittedly, the moving window approach is simple, fast (O(N) computational complexity vs. the O(N log(N)) complexity for recursions), and usually provides an answer to questions of interest to investigators. Nevertheless, recursive segmentation approach can be more accurate; and it also avoids the common problem in a moving window approach to select a window size and a moving distance. We suggest to use recursive segmentation as a refinement of the moving window approach, or a second-stage analysis after a rough result is obtained from the moving window approach.

To apply the recursive segmentation in DNA sequence analysis, one only needs to construct a filter to highlight the feature to be segmented. With the filter, the four-symbol DNA sequence is converted to a sequence of a new set of symbols. For isochore, it is the G + C and A + T, for CpG island, it is the 5′-CG-3′ dinucleotide and others, for replication origin and terminus, it is either purine and pyrimidine, or amino and keto, for telomere sequences, it can either be the original four nucleotides or 16 dinucleotides, and for coding/noncoding detection, it is the 12-symbol that combines base and codon position information or any other symbol sets that include codon position information, etc. It is not difficult to generalize these five cases to any other application, once a filter is constructed.

One advantage of this unified framework for many DNA sequence analysis tasks is to simplify the computer program development: one core recursive segmentation plus a number of filter subroutines are enough for all these applications. Of course, each particular application needs its own parameter setting, stopping criterion, and explanation of the output. If the signal to be detected is regional instead of global, we need to relax the stopping criterion during the segmentation, then raise the criterion during the examination of the segmented result. This is what has been done in detecting local repeats in telomere sequences.

It has also been observed that we are more successful in detecting signal of interest when the number of symbol is reduced (isochore, CpG island, replication origin/terminus), and less successful when the number of symbols is increased (telomere, coding/noncoding borders). A direct explanation of this observation is that domains with more number of symbols may not exist on a global level. For example, none of the domain borders in Figs. 9 and 10 are significant by the BIC criterion (i.e. for $s_0 = 0$). It is thus not surprising that they do not match the exon borders for this gene. If the segmentation at the first few steps of the recursion is not significant, subsequent segmentations can be even more problematic.

The detection of isochore borders is the most natural application of the recursive segmentation. First, the domain structure in (G + C)% does exist and is common in DNA sequences (Bernardi, 1989, 1995). Second, the domains-within-domain phenomenon (Bernaola-Galván et al., 1996; Li, 1997a,b) makes the recursive segmentation a better choice than a moving window approach for detecting hierarchical patterns. Third, changing segmentation stopping criterion leads to examination of domains at different length scales, which is also convenient to investigators. We expect that recursive segmentation has a good chance to become the method of choice for isochore detection (Oliver et al., 2001).

Since the recursive segmentation works on a global scale (at least for the first few steps in the recursion) whereas isochore borders are determined by the (G + C)% change in a local region, there is an issue on the robustness of the segmentation point result. This issue can be addressed by selecting new starting and ending point of sequence (ideally they are preferably domain borders themselves) and examine whether the segmentation point changes. It can also be addressed by perturbing the existing segmentation point in the final result to see whether the likelihood (BIC) decreases (increases). There is another issue on whether the first few segmentations always have higher segmentation strengths than the subsequence segmentations. For isochore problem, we have observed that it is typically true, but exceptions do exist.

The detection of CpG island is perhaps relatively easier than the isochore problem, because the converted binary sequence (11 for 5′-CpG-3′, 0 otherwise) is mostly 0's, and clusters of 1's are rare. If there is no hierarchical pattern in these clusters, a moving window approach applied to this binary sequence should detect these clusters equally well. The success of our detecting of CpG island in the example in Fig. 4 is probably more a reflection of the observation in Matsuo et al. (1993) that CpG% is as good a predictor of CpG islands, at least for human sequences, as the standard prediction using a combination of CpG O/E and (G + C)%. Although CpG% criterion is simple and easy to explain, it is surprising that it is not commonly used.

The detection of replication origin and terminus in bacteria genomes is also relatively easier. The well-known strand asymmetry near the replication origin provides the basis for our segmentation design. Another simplification is that there is no need to segment the sequence more than twice. Here we are facing the same issue that segmentation is based on base/symbol composition in a global scale whereas the strand symmetry near the replication origin is local. For this purpose, we carried out a robustness test by rotating the sequence starting point. When the starting point falls on the replication origin (or the terminus), the best segmentation result is achieved at the replication terminus (or origin). This approach will determine the replication origin and terminus more accurately.

Detecting elements in telomere (or centromere) sequences using recursive segmentation works in an indirect way. Certain elements (usually defined by conserved regions among different telomere sequences) may be accompanied by other unusual sequence patterns, in particular the short repeat sequence. These repeat sequences tend to be biased in their base composition (unless the repeat is ACGTACGT...), so in principle detectable by a recursive segmentation. Our example in Fig. 7 shows, however, that the base composition difference between the repeats and its neighboring sequence can be small and regional. In this situation, a recursion design to relax the stopping criterion first, then merge domains by eliminating insignificant segmentations later, may achieve the goal of detecting repeats.

The most difficult but potentially more interesting application of the recursive segmentation is the coding/noncoding region detection. When a coding and a noncoding sequence with a similar length joints together, there is usually a clear border in the 12-symbol sequence (Bernaola-Galván et al., 2000), even though the segmentation at the border may not be significant by BIC criterion (see, Fig. 8). In a typical situation for human genes, however, the signal from the border is harder to detect for these reasons: (1) the initial sequence is not a 'black-and-white' type sequence with one half being coding and another half being noncoding; (2) coding sequence can be much shorter (e.g. less than 100 bp in Fig. 9) than the noncoding sequence; (3) the coding measure based on the 12 symbols may not change dramatically at the exon borders, but more gradually (see the var in Figs. 9 and 10).

For other applications, reasons #1 and #2 may not cause any problems since recursion will automatically take care of the uneven sizes of different domains and the initial heterogeneity. For the 12-symbol case, however, the main problem is that the first stage segmentation in a recursion is typically not significant (at $s_0 = 0$). This initial lack of significance leads difficulty in the subsequence segmentations. Interestingly, Fig. 9 shows that our coding measure based on the 12-symbol sequence (i.e. var) is able to visually pick 10 exons correctly out of total 14 exons (correct in the region, not correct for the border), with four false negatives (missing in a prediction) and perhaps three false positives (prediction does not correspond to a gene). But recursive segmentation on the 12-symbol sequence does not achieve this success rate. There are certainly rooms for improvements in this application, perhaps by a new filter with fewer number of symbols.

A final note is that besides the recursive segmentation discussed in this paper, there are other alternative approaches for segmentation (Liu and Lawrence, 1999; Ramensky et al., 2000; Guéguen, 2001). In particular, the Bayesien segmentation is a distinct method that does not address the question of whether the sequence is homogeneous or not, but it calculates the posterior probability of each position for being a change-point. The Bayesian approach to change-point problem has at least 25 years of history (Smith, 1975). A full discussion of this topic as well as a comparison between the Bayesian and recursive segmentation is outside the scope of this paper.

Acknowledgements

We would like to thank José Oliver for the initial participation of this project, Oliver Clay, Jean Lobry, William Grundy, Lijuan Fang for discussion and suggestions, and LaDeana Hillier for sharing a CpG island detection program. WL's work is supported by the NIH grant K01HG00024, PBG is partially supported by the grant BI099-0651-CO2-01 from the Spanish Government, FH's work is supported by grants from PhRMA foundation, NSF (DBI-0078523), and Sloan/DOE Fellowship in Computational Molecular Biology, and IG is supported by a fellowship from the Cold Spring Harbor Laboratory Association.

Appendix A. Source of the sequence used

- MHC sequence on human chromosome 6 was downloaded from Sanger Center: http://www.sanger.ac.uk/HGP/Chr6/MHC.shtml ('original consensus' version, May 1999, $N=3,673,778$ bases).
- Longest contig of *C. elegans* chromosome 1 was downloaded from Sanger Center: ftp://ftp.sanger.ac.uk/pub/C.elegans_sequences/CHROMOSOMES/CURRENT_RELEASE/ (last updated April 2001, contig $N=8,568,332$ bases, from position 1974932 to 10543263 out of 14,972,282 bases).
- A human CpG-island-containing DNA sequence on chromosome 22q13.2–13.3 was obtained from GenBank/NCBI: http://www.ncbi.nlm.nih.gov/Entrez/ (accession number: AL022237, $N=60,828$ bases).
- Bacterium *Haemophilus influenzae* sequence was obtained from GenBank/NCBI: http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html (accession number L42023, $N=1,830,138$, last updated May 1999). One hundred and fifteen undecided nucleotides were replaced randomly according to the actual base composition of the sequence.
- Yeast *S. cerevisiae* chromosome 12 sequence was obtained from GenBank/NCBI (accession number: NC-001144, $N=1,078,172$ bases, last updated April 2001).
- Yeast *S. cerevisiae* chromosome 3 sequence was also obtained from GenBank/NCBI.
- A human chromosome 22 genomic sequence which contains the kinase chk2 (RAD53) was obtained from http://www.sanger.ac.uk/HGP/Chr22/ ($N=$ 50,824 bases). The *chk*2 gene is consisted of 14 exons with a total length of 1632 bases.

Appendix B. Program used

The recursive segmentation is carried out by a *Perl* script written by us. To obtain a copy of the program, please send email to wli@linkage.rockefeller.edu or vic@cs.columbia.edu.

References

Ainscough, R., et al., (The C. elegans Sequencing Consortium) 1998. Genome sequence of the nematode *C elegans*: a platform for investigating biology. Science 282, 2012–2018.

Akaike, H., 1978. A Bayesian analysis of the minimum AIC procedure. Annals of the Institute of Statistical Mathematics 30 (Part A), 9–14.

Anderssen, S.G.E., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., et al., 1998. The genome sequence of Ricketettsia prowazekii and the origin of mitochondria. Nature 396, 133–140.

Beck, S., Geraghty, D., Inoko, H., Rowen, L., et al., (The MHC Sequencing Consortium) 1999. Complete sequence and gene map of a human major histocompatibility complex. Nature 401, 921–923.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acid Research 27, 573–580.

Bernaola-Galván, P., Román-Roldán, R., Oliver, J.L., 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. Physical Review E 53, 5181–5189.

Bernaola-Galván, P., Grosse, I., Carpena, P., Oliver, J.L., Roman-Roldán, R., Stanley, H.E., 2000. Finding borders between coding and noncoding DNA regions by an entropic segmentation method. Physical Review Letters 85, 1342–1345.

Bernardi, G., 1989. The isochore organization of the human genome. Annual Review of Genetics 23, 637–661.

Bernardi, G., 1995. The human genome: organization and evolutionary history. Annual Review of Genetics 29, 445–476.

Bird, A., 1986. CpG-rich islands and the function of DNA methylation. Nature 321, 209–213.

Braun, J.V., Müller, H.G., 1998. Statistical methods for DNA segmentation. Statistical Science 13, 142–162.

Braun, J.V., Braun, R.K., Müller, H.G., 2000. Multiple change-point fitting via quasi-likelihood, with application to DNA sequence segmentation. Biometrika 87, 301–314.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology 268, 78–94.

Carlstein, E., Müller, H.G., Siegmund, D. (Eds.), 1994. Change-Point Problems. Lecture Notes and Monograph Series, vol. 23. Institute of Mathematical Statistics, Hayward, CA.

Chan, C.S.M., Tye, B.K., 1983. Organization of DNA sequences and replication origins at yeast telomeres. Cell 33, 563–573.

Churchill, G.A., 1989. Stochastic models for heterogeneous DNA sequences. Bulletin of Mathematical Biology 51 (1), 79–94.

Churchill, G.A., 1992. Hidden Markov chains and the analysis of genome structure. Computer and Chemistry 16 (2), 107–115.

Claverie, J.M., 1997. Computational methods for the identification of genes in vertebrate genomic sequences. Human Molecular Genetics 6, 1735–1744.

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.

Cooper, D., Taggart, M., Bird, A., 1983. Unmethylated domains in vertebrate. Nucleic Acids Research 11, 647–658.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., 1990. Introduction to Algorithms. The MIT Press, Cambridge, MA.

Cross, S.H., Clark, V.H., Simmen, M.W., Bickmore, W.A., Maroon, H., Langford, C.F., Carter, N.P., Bird, A.P., 2000. CpG island libraries from human chromosomes 18 and 22: landmarks for novel genes. Mammalian Genome 11 (5), 373–383.

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. European Journal of Biochemistry 115, 227–233.

Elton, R.A., 1974. Theoretical models for heterogeneity for base composition in DNA. Journal of Theoretical Biology 45, 533–553.

Fickett, J.W., 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Research 10, 5303–5318.

Fickett, J.W., 1996. Finding genes by computer: the state of the art. Trends in Genetics 12, 316–320.

Fickett, J.W., Torney, D.C., Wolf, D.R., 1992. Base compositional structure of genomes. Genomics 13, 1056–1064.

Fickett, J.W., Tung, C.S., 1992. Assessment of protein coding measures. Nucleic Acids Research 20, 6441–6450.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., et al., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. Science 269, 496–521.

Francino, M.P., Ochman, H., 1997. Strand asymmetries in DNA evolution. Trends in Genetics 13, 240–245.

Frank, A.C., Lobry, J.R., 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238, 65–77.

Frank, A.C., Lobry, J.R., 2000. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics 16, 560–561.

Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O., et al., 1998. Complete genome sequence of Treponema pallidum, the Syphilis spirochete. Science 281, 375–388.

Freeman, J.M., Plasterer, T.N., Smith, T.F., Mohr, S.C., 1998. Patterns of genome organization in bacteria (technical comment, online). 279, 1827a.

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.

Gardiner-Garden, M., Frommer, M., 1987. CpG islands in vertebrate genomes. Journal of Molecular Biology 196, 261–282.

Gierlik, A., Kowalczuk, M., Mackiewicz, P., Dudek, M.R., 2000. Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome? Journal of Theoretical Biology 202, 305–314.

Grigoriev, A., 1998a. Genome arithmetic. Science 281, 1923a technical comment, online.

Grigoriev, A., 1998b. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Research 26, 2286–2290.

Grosse, I., Herzel, H., Buldyrev, S.V., Stanley, H.E., 2000. Species independence of mutual information in coding and noncoding DNA. Physical Review E 61, 5624–5629.

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. Physical Review E, in press.

Guéguen, L., 2001. Segmentation by maximal predictive partitioning according to composition biase. In: Gascuel, O., Sagot, M.-F. (Eds.), Computational Biology. In: Lecture Notes in Computer Science, vol. 2066. Springer-Verlag, Heidelberg.

Guigo, R., 1999. DNA composition, codon usage and exon prediction. In: Bishop, M. (Ed.), Genetic Databases. Academic Press.

Guigo, R., Knudsen, S., Drake, N., Smith, T., 1992. Prediction of gene structure. Journal of Molecular Biology 226, 141–157.

Häring, D., Kypr, J., 2001. No isochores in the human chromosomes 21 and 22? Biochemical and Biophysical Research Communication 280 (2), 567–573.

Horowitz, H., Thorburn, P., Haber, J.E., 1984. Rearrangements of highly polymorphic regions near telomeres of *Saccharomyces cerevisiae*. Molecular and Cellular Biology 4, 2509–2517.

Horvath, A.L., 1989. The limit distributions of likelihood ratio and cumulative sum tests for a change in a binomial probability. Journal of Multivariate Analysis 31, 148–159.

Johnston, M., Hillier, L., Riles, L., Albermann, K., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. Nature 387 (6632 Suppl.), 87–90.

Karkas, J.D., Rudner, R., Chargaff, E., 1968. Separation of B. subtilis DNA into complementary strands. II. Template functions and composition as determined by transcription by RNA polymerase. Proceedings of National Academy of Sciences 60, 915–920.

Lander, E.S., Waterston, R.H., Sulston, J., Collins, F.S., et al., (International Human Genome Sequencing Consortium) 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Larsen, F., Gundersen, G., Lopez, R., Prydz, H., 1992. CpG islands as gene markers in the human genome. Genomics 13, 1095–1107.

Li, W., 1992. Generating nontrivial long-range correlations and 1/f spectra by replication and mutation. International Journal of Bifurcation and Chaos 2, 137–154.

Li, W., 1997a. The study of correlation structures of DNA sequences—a critical review. Computer and Chemistry 21, 257–271.

Li, W., 1997b. The complexity of DNA. Complexity 3, 33–37.

Li, W., 1998. Comments on ‘simplicity and complexity in gene evolution’. Complexity 3, 10.

Li, W., 1999. Statistical properties of open reading frames in complete genome sequences. Computer and Chemistry 23, 283–301.

Li, W., 2001a. New stopping criteria for segmenting DNA sequences. Physical Review Letters 86, 5815–5818.

Li, W., 2001b. DNA segmentation as a model selection process. In: Proceedings of the Fifth Annual International Conference on Computational Biology, Association for Computing Machinery Press, New York, pp. 204–210.

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

Li, W., Marr, T.G., Kaneko, K., 1994. Understanding long-range correlations in DNA sequences. Physica 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 Research 8, 916–928.

Lin, H.J., Chargaff, E., 1967. On the denaturation of deoxyribonucleic acid. H. Effects of concentration. Biochimica Biophysics Acta 145, 398–409.

Liu, J., Lawrence, C.E., 1999. Bayesian inference on biopolymer model. Bioinformatics 15, 38–52.

Lobry, J.R., 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. Journal of Molecular Evolution 40, 326–330.

Lobry, J.R., 1996a. Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular Biology and Evolution 13, 660–665.

Lobry, J.R., 1996b. Origin of replication of Mycoplasma genitalium. Science 272, 745–746.

Lobry, J.R., 1999. Genomic landscapes. Microbiology Today 26, 164–165.

Lopez, P., Forterre, P., le Guyader, H., Philippe, H., 2000. Origin of replication of Thermotoga maritima. Trends in Genetics 16, 59–60.

Louis, E.J., Haber, J.E., 1990. The subtelomeric Y′ repeat family in *Saccharomyces cerevisiae*: an experimental system for repeated sequence evolution. Genetics 124, 533–545.

Louis, E.J., Haber, J.E., 1992. The structure and evolution of subtelomeric Y′ repeats in *Saccharomyces cerevisiae*. Genetics 1331, 559–574.

Louis, E.J., Naumova, E.S., Lee, A., Naumov, G., Haber, J.E., 1994. The chromosome end in yeast: its mosaic nature and influence on recombinational dynamics. Genetics 136, 789–802.

Macaya, G., Thiery, J.-P., Bernardi, G., 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. Journal of Molecular Biology 108, 237–254.

Matsuo, K., Clay, O., Takahashi, T., Silke, J., Schaffner, W., 1993. Evidence for erosion of mouse CpG islands during mammalian evolution. Somatic Cell and Molecular Genetics 19, 535–543.

McLean, K.J., Wolfe, K.H., Devine, K.M., 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. Journal of Molecular Evolution 47, 691–696.

Mrázek, J., Kypr, J., 1994. Biased distribution of Adenine and Thymine in gene nucleotide sequences. Journal of Molecular Biology 39, 439–447.

Nekrutenko, A., Li, W.H., 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Research 10, 1986–1995.

Oliver, J.L., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276, 47–56.

Oliver, J.L., Román-Roldán, R., Perez, J., Bernaola-Galván, P., 1999. SEGMENT: identifying compositional domains in DNA sequences. Bioinformatics 15, 974–979.

Olson, M.V., 1991. Genome structure and organization in *Saccharomyces cerevisiae*. In: The Molecular and Cellular Biology of the Yeast Saccharomyces: I. Genome Dynamics, Protein Synthesis, and Energetics. Cold Spring Harbor Press, New York, pp. 1–39.

Pettitt, A.N., 1980. A simple cumulative sum type statistic for the change-point problem with zero-one variables. Biometrika 67, 79–84.

Raftery, A.E., 1995. Bayesian model selection in social research. In: Marsden, P.V. (Ed.), Sociological Methodology. Blackwells, Oxford, UK, pp. 185–195.

Ramensky, V.E., Makeev, V.Ju., Roytberg, M.A., Tumanyan, V.G., 2000. DNA segmentation through the Bayesian approach. Journal of Computational Biology 7, 215–231.

Rudner, R., Karkas, J.D., Chargaff, E., 1968. Separation of *B. subtilis* DNA into complementary strands. III. Direct analysis. Proceedings of National Academy of Sciences 60, 921–922.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Tech. J. 27, 379–423 623–656.

Smith, A.F.M., 1975. A Bayesian approach to inference about a change-point in a sequence of random variables. Biometrika 62, 407–416.

Solovyev, V.V., Salamov, A.A., Lawrence, C.B., 1994. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Research 22, 5156–5163.

Sonnhammer, E.L.L., Durbin, R., 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167, GC1–GC10.

Staden, R., McLachlan, A.D., 1982. Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucleic Acids Research 10, 141–156.

Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proceedings of the National Academy of Sciences 48 (4), 582–592.

Sueoka, N., 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. Journal of Molecular Biology 40, 318–325.

Szostak, J.W., Blackburn, E.H., 1982. Cloning yeast telomeres on linear plasmid vectors. Cell 29, 245–255.

Thiery, J.-P., Macaya, G., Bernardi, G., 1976. An analysis of eukaryotic genomes by density gradient centrifugation. Journal of Molecular Biology 108, 219–235.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R., 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Computer Applications in Biosciences 13, 263–270.

Tykocinski, M., Max, E., 1984. Methylation of cytosine in CG dinucleotide clusters in MHC genes and in 5′ demethylated genes. Nucleic Acids Research 12, 4385–4396.

Uberbacher, E.C., Xu, Y., Mural, R.J., 1996. Discovering and understanding genes in human DNA sequence using GRAIL. Methods in Enzymology 266, 259–281.

Venter, J.C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

Wada, K., Wada, Y., Doi, H., Ishibashi, F., Gojobori, T., Ikemura, T., 1991. Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Research 19, 1981–1986 Suppl.

Wellinger, R.J., Sen, D., 1997. The DNA structures at the ends of eukaryotic chromosomes. European Journal of Cancer 33, 735–749.

Yan, M., Lin, Z.S., Zhang, C.T., 1998. A new Fourier transform approach for protein coding measure based on the format of the Z curve. Bioinformatics 14, 685–690.

Zhang, H., Singer, B., 1999. Recursive Partitioning in the Health Sciences. Springer-Verlag, New York.

Zhang, M.Q., 1997. Identification of protein coding regions in the human genome based on quadratic discriminant analysis. Proceedings of National Academy of Sciences 94, 559–564.