New Stopping Criteria for Segmenting DNA Sequences

Wentian Li
Laboratory of Statistical Genetics, Box 192, Rockefeller University, 1230 York Avenue, New York, New York 10021
(Received 15 September 2000)

We propose a solution on the stopping criterion in segmenting inhomogeneous DNA sequences with complex statistical patterns. This new stopping criterion is based on Bayesian information criterion in the model selection framework. When this criterion is applied to telomere of *S. cerevisiae* and the complete sequence of *E. coli,* borders of biologically meaningful units were identified, and a more reasonable number of domains was obtained. We also introduce a measure called segmentation strength which can be used to control the delineation of large domains. The relationship between the average domain size and the threshold of segmentation strength is determined for several genome sequences.

DOI: 10.1103/PhysRevLett.86.5815 PACS numbers: 87.10.+e, 02.50.Tt, 87.15.Cc

DNA sequences are usually not homogeneous. Regions with high concentrations of *G* or *C* bases alternate with regions which lack *G* or *C* [1]; stretches of sequences with an abundance of *CG* dinucleotide (*CpG* island) interrupt regular sequences; coding regions distinguish themselves from noncoding regions by the strong periodicity-of-three pattern, etc. The alternation of long (e.g., $>$ 300 kilobases) $G + C$ rich and $G + C$ poor regions (also known as "isochores" [1]) is shown to be related to chromosome bands, gene density, and perhaps chromosomal structure [1]. The concepts of inhomogeneity and domains can also be generalized recursively to different length scales, and such domains-within-domains phenomena have indeed been observed in DNA sequences [2,3]. These hierarchical patterns are the cause of the fractional long-range correlations and $1/f$ spectra observed in DNA sequences [4]. There have been discussions of the possible biological meaning of this hierarchical pattern [5] and its connection to other complex systems [6].

Computational methods used to identify homogeneous regions are called segmentation procedures [2,7] which are important for many DNA sequence analysis tasks: detecting the existence of isochores, identifying complicated repeat patterns within telomeres and centromeres, determining coding-noncoding borders [8], etc. Segmentation procedures can also be applied to any inhomogeneous/disorder media (e.g., one-dimensional solid, spin glass chain) or nonstationary time series (e.g., symbolic dynamics) to determine the domain borders or turning points. An application of the segmentation procedure to determine the mobility edge of vibrational states in disordered materials can be found in [9]. The segmentation procedure and the physical fragmentation [10] are highly reminiscent of each other [11]. The ease of a segmentation procedure directly affects the scaling exponent of the size distribution in a fragmentation [11].

In the segmentation procedure proposed in [2], one crucial step—the stopping criterion—is arbitrarily determined. This is because this criterion is presented within the framework of hypothesis testing. It is common in this framework to reject or accept the null hypothesis based on a chosen significance level, typically, 0.01 or 0.001. Not choosing other levels, say, 0.025 or 10^{-6}, is to some extent arbitrary. Another practical problem of the criterion in [2] is that it is extremely hard to halt the recursion at a large length scale even with a very small significance level, whereas many biologically interesting domains such as isochores are large. We solve these problems here by discussing segmentation in a new framework—the model selection framework. As a result, an alternative meaning of segmentation is proposed, and a minimum requirement for choosing one model over another is introduced.

In the model selection framework, basic 1-to-2 segmentation is carried out as a comparison of two stochastic models of the DNA sequence: before the segmentation, the sequence is modeled by a homogeneous random sequence (with three base composition parameters); after the segmentation, by *two* homogeneous random sequences separated by a partition point (with seven parameters). Whether a 1-to-2 segmentation should be continued or not is determined by whether the two-random-subsequence model is better than the one-random-sequence model. In model selection, the answer to this question is determined by two factors: (i) the model's ability to fit the data and (ii) the model's complexity. Overfitting and underfitting models are not considered to be good, either because of high model complexity or because of poor fitting performance. The Bayesian information criterion (BIC) is a proposal for balancing the two factors, defined as [12]

$$\begin{aligned} \mathrm{BIC} &= -2\log(\hat{L}) + \log(N)K + O(1) + O\left(\frac{1}{\sqrt{N}}\right) + O\left(\frac{1}{N}\right) \\ &\approx -2\log(\hat{L}) + \log(N)K, \end{aligned} \tag{1}$$

where $\hat{L}$ is the maximum likelihood [13], K is the number of parameters in the model, and N is the number of data points. BIC is an approximation of the logarithm of integrated likelihood of a model multiplied by -2 [12]. The integrated likelihood represents the overall performance of a model. The better the model, the larger the integrated likelihood, and thus the smaller the BIC. A similar

concept is the Akaike information criterion (AIC) [14], with the $\log(N)$ term in Eq. (1) replaced by 2. BIC penalizes complex models more severely than AIC.

We show here that the entropy-based segmentation in [2] can be recast in the likelihood framework [13], which in turn can be generalized to a model selection framework [15]. The likelihoods of the random-sequence model and the two-random-subsequence model (before and after a 1-to-2 segmentation) are $L_1(\{p_\alpha\}) = \prod_\alpha p_\alpha^{N_\alpha}$, $L_2(\{p_\alpha^l\},\{p_\alpha^r\},N_l) = \prod_\alpha (p_\alpha^l)^{N_\alpha^l} \prod_\beta (p_\beta^r)^{N_\beta^r}$, where $\{p_\alpha\}$, $\{p_\alpha^l\},\{p_\alpha^r\}$ $(\alpha = A, C, G, T)$ are the base composition parameters for the whole sequence, left and right subsequence, respectively; $\{N_\alpha\},\{N_\alpha^l\},\{N_\alpha^r\}$ are the corresponding base counts; and N_l is the size of the left subsequence. The maximum likelihood estimation of the parameters is simply $\hat{p}_\alpha = N_\alpha/N$, and the maximum log likelihoods before and after segmentation are $\log\hat{L}_1 = -NE$ and $\log\hat{L}_2 = -N^lE^l - N^rE^r$, where E, E^l, E^r are the entropies for the whole, left, and right sequences. The segmentation position N_l is also a parameter in the model and is determined by the position that maximizes the likelihood (though this parameter is discrete and its range changes with N). The increase of log-likelihood is $\log(\hat{L}_2/\hat{L}_1) = NE - (N^lE^l + N^rE^r) = N \cdot \hat{D}_{JS}$, where $\hat{D}_{JS}$ is the maximum of Jensen-Shannon divergence $D_{JS} = E - (N^lE^l + N^rE^r)/N$ [2,16].

We require that the BIC be reduced by the segmentation for the procedure to continue, i.e., $\Delta\text{BIC} < 0$, which leads to (note $K_2 = 7$ and $K_1 = 3$):

$$2N\hat{D}_{JS} > 4\log(N)\,. \tag{2}$$

Equation (2) is our new stopping criterion.

Lower (relaxed) bound of the significance level.—The stopping criterion in Eq. (2) differs from the criterion in [2] in that the significance level cannot be arbitrarily relaxed. The criterion in [2] compares the maximum D_{JS} with that of a random sequence. If the sequence is indeed random, $2N\hat{D}_{JS}$ typically follows a χ^2 distribution, and the tail area under this distribution is the corresponding significance level [17]. Numerical simulation and studies on the change-point problem in statistics have shown that the distribution may not be a χ^2, but an N-dependent "extreme value" distribution [18]. The dependence of Eq. (2) on the sequence length N is consistent with these studies. It also has important practical implications: the stopping criterion in Eq. (2) is not fixed but adjustable. It is particularly important for a long sequence, when the criterion in [2] may not be able to stop segmentations with large $2N\hat{D}_{JS}$.

In Fig. 1, we illustrate the new criterion for the left telomere of chromosome 12 of yeast *Saccharomyces cerevisiae* [19]. It is known that telomere sequences are compositionally complex. There is a highly repetitive sequence called TEL at the tip of the telomere (for yeast, it is $5'$-$C_{1-3}A$-$3'$). There are also subsequences that are conserved among different yeast chromosomes: the Y' and

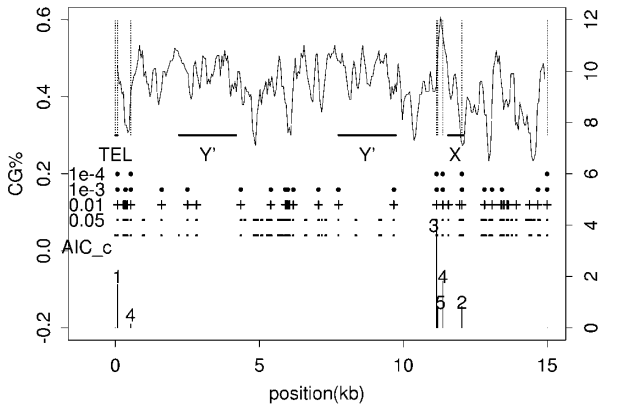

FIG. 1. Partition points determined by the segmentation with the stopping criterion Eq. (2) for the left telomere of yeast *S. cerevisiae* chromosome 12 (dashed vertical lines). The partition points determined by AIC (dot) (with the high-order term included), hypothesis testing framework with significance level of 0.05 (dot), 0.01 (cross), 0.001 and 0.0001 (solid dot) are shown for comparison. Also shown is the $G + C$ content in moving windows (window size = 150 bases; moving distance = 51 bases). The location of the telomeric sequence (TEL) and subtelomeric sequences (Y' and X) are marked. The lower plot shows the segmentation strength s of a 1-to-2 segmentation. The numbers are the order in which the segmentation is carried out.

X subtelomeric sequence [20]. A segmentation procedure can be applied to telomere sequences to identify some compositionally distinct elements [21]. It can be seen from Fig. 1 that the criterion in Eq. (2) manages to delineate the borders for TEL and X elements [22]. Although Eq. (2) missed the two Y' elements, an indication that Y' elements are not compositionally distinct, it is the cost of avoiding many false positives.

Segmentation strength.—Although a lower (relaxed) bound of the significance level is set in Eq. (2), no limit on the upper (stringent) bound is possible. We introduce a measure for segmentation strength s [15]:

$$s = \frac{2N\hat{D}_{JS} - 4\log(N)}{4\log(N)}\,, \tag{3}$$

and the stringency level can be raised by choosing a nonzero value of the threshold s_0: $s > s_0 > 0$. Equation (2) is equivalent to $s_0 = 0$. The prominence of TEL and X elements is indicated by their large segmentation strength ($s = 1.7$, 0.85, and 4.16; see the lower plot of Fig. 1). These segmentations are also chosen earlier in the recursive segmentation (being first, second, and third).

Minimum domain size.—To test a model on a data set, the number of samples must be larger than the number of parameters in the model. Since we compare two models with three and seven parameters, respectively, the sequence has to contain at least seven bases before the segmentation and three bases after the segmentation. Unlike the criterion in [2], these minimum size requirements are not set arbitrarily.

Binary and 12-symbol sequences.—For many practical applications of the segmentation procedures, DNA texts

are converted to symbolic sequences with less (or more) than four symbols. For example, the two-symbol sequence with symbols S (for strong, G and C) and W (for weak, A and T), is frequently used for studying large-scale homogeneous domains. The stopping criteria for binary sequences can be modified easily: with $K_1 = 1$ and $K_2 = 3$, the right-hand side of Eq. (2) becomes $2\log(\mathrm{N})$. For coding region recognition, it is proposed in [8] that a DNA sequence can be converted to a 12-symbol sequence: each symbol contains information on both the base and the codon position (i.e., $A_1, C_1, G_1, T_1, A_2, \cdots$). With $K_1 = 9$ and $K_2 = 19$, the stopping criteria in Eq. (2) become $2N\hat{D}_{JS} > 10\log(\mathrm{N})$.

Threshold for segmentation strength and domain sizes.—Since Eq. (2) does not provide an upper (stringent) limit on the significance level, there is still some degree of subjectivity in our segmentation procedure. If one is interested in largest domains, or the strongest segmentation signals, the threshold for segmentation strength s_0 should be set larger than zero. Taking the complete sequence of *Escherichia coli* genome [23], for example, the replication origin and the replication terminus present the two most significant segmentation signals (see Fig. 2). If the s_0 is set to 20, only these two 1-to-2 segmentations will make the cut.

The larger the s_0, the larger the domain sizes in the final configuration. The relationship between the two is empirically determined by segmentations on several genome sequences, shown in Fig. 3. It can be seen that the relationship is not universal for all sequences: with the same s_0, sequences with high compositional complexity (e.g., MHC sequence) contain smaller domain sizes in the final configuration than sequences with lower complexity (e.g., yeast). It can also be seen that in order to reach the average size of isochore (300 kilobases), s_0 should be set as large as 5.

Domain size distribution.—Another indirect evidence that our new stopping criterion is more reasonable than the one in [2] (with a typical significance level) can be seen by examining the domain size distribution in the final configuration. The 281 domains in the *Escherichia coli* genome in Fig. 3 are ranked by size. These sizes are plotted against the rank (Zipf's plot) in Fig. 4. The Zipf's plot for sizes from rank 4 to rank 180 approximately exhibit a power law $1/r^{1.21}$ (Fig. 4). This is similar to the power-law behavior in Zipf's plot of many other natural and social phenomena (known as Zipf's law [24,25] when the scaling exponent is close to -1).

When a more relaxed stopping criterion is used, there is a lack of large domains. We illustrate this by a AIC-based segmentation which is equivalent to the criterion in [2] with the significance level of 0.091 578 (using $\chi^2_{df=4}$). The Zipf's plot for domains derived from the AIC-based segmentation is not a power-law function. Even a forced curve fitting by a power-law function leads to a slope merely ~ -0.5. This indicates that the size distribution by criterion Eq. (2) is more self-similar, more balanced between the small and large domains than those by the AIC-based segmentation.

In summary, this paper solves a problem encountered in [2] that recursive segmentation is not easy to stop even when a stringent significance level is used (the most stringent significance level in the SEGMENT program [26] is 10^{-6}). This solution allows us to investigate much larger domains and longer-range hierarchical correlation

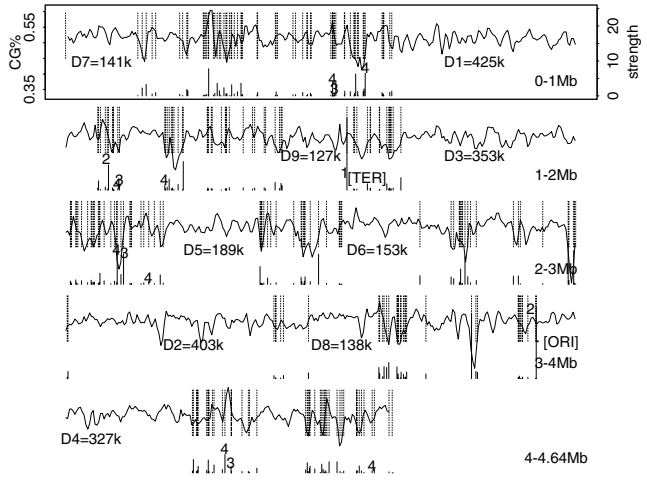

FIG. 2. Segmentation points determined by Eq. (2) for *E. coli* genome (dashed vertical lines). Also shown are the $G + C$ content in moving windows (window size = 9000 bases; moving distance = 3571 bases), and the segmentation strength s. The replication origin (ORI), replication terminus (TER), and the nine largest domains (D1, D2, ...) are marked in the plot. Each one of the subplots represents 1 megabase of the sequence (total length is 4.639 megabases).

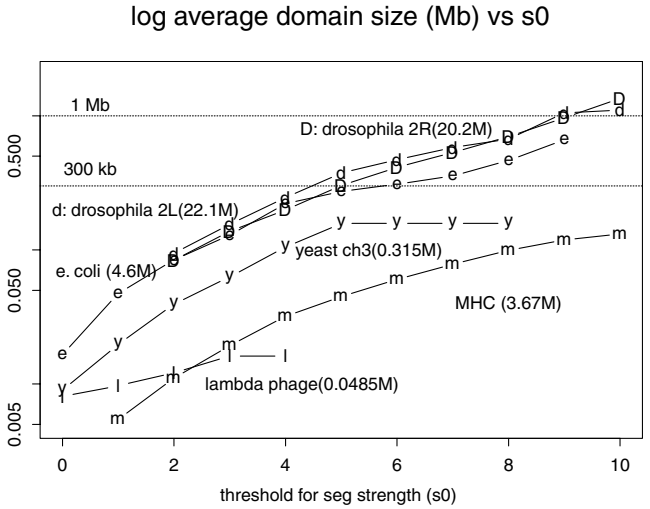

FIG. 3. Average domain size vs segmentation strength s_0 for these sequences: human major histocompatibility complex (MHC), λ bacteriophage, chromosome 3 of *S. cerevisiae*, *E. coli*, left and right arms of chromosome 2 of *Drosophila melanogaster*.

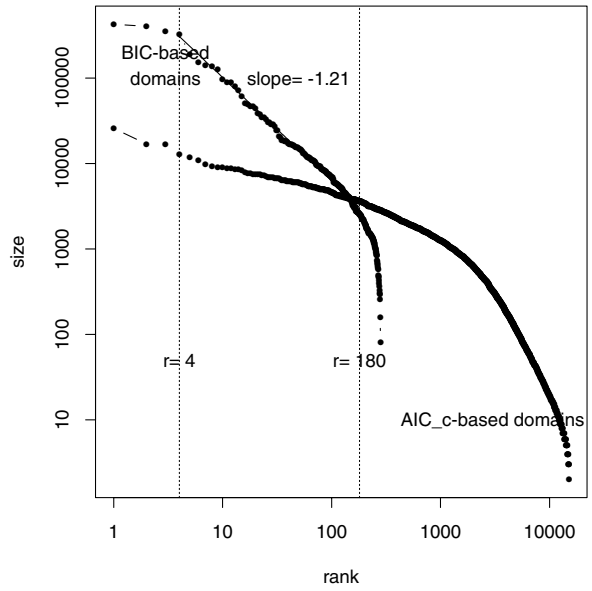

FIG. 4. Size-rank plot (Zipf's plot) of domains obtained by segmentation with the stopping criterion in Eq. (2). Those obtained by the AIC-based segmentation are also shown.

in DNA sequences. The framework from which our solution is derived is also ideal for generalizations to other more complicated situations. Determining the number of domains in a DNA sequence, as any other descriptions of the sequence, is relative—it is relative to the length scale of interests, relative to the model used. By changing the segmentation strength, we essentially change the level of description of the sequence.

The work is supported by Grant No. K01HG00024 from NIH. I thank José Oliver for sending me the partition points used in Fig. 1 produced by the SEGMENT program [26], and Yaning Yang, David Siegmund, Ivo Grosse, and Pedro Bernaola-Galván for discussions.

[1] G. Bernardi, Annu. Rev. Genet. **29**, 445–476 (1995).
[2] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996); R. Román-Roldán, P. Bernaola-Galván, and J. L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).
[3] W. Li, Complexity **3**, 33–37 (1997).
[4] W. Li, Int. J. Bifurcation Chaos **2**, 137–154 (1992); W. Li and K. Kaneko, Europhys. Lett. **17**, 655–660 (1992); C. K. Peng *et al.,* Nature (London) **356**, 168–170 (1992); R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992); W. Li, T. G. Marr, and K. Kaneko, Physica (Amsterdam) **75D**, 392–416 (1994); W. Li, Comput. Chem. (Oxford) **21**, 257–271 (1997).
[5] A. Audit, C. Thermes, and C. Vaillant, Y. d'Aubenton-Carafa, J. F. Muzy, and A. Arneodo, Phys. Rev. Lett. **86**, 2471 (2001).
[6] More publications on correlations in DNA sequences can be found at http://linkage.rockefeller.edu/wli/dna_corr/.
[7] J. V. Braum and H. G. Müller, Stat. Sci. **13**, 142–162 (1998); V. E. Ramensky, V. Ju. Makeev, M. A. Roytberg, and V. G. Tumanyan, J. Comput. Biol. **7**, 1–46 (2000).
[8] P. Bernaola-Galván *et al.,* Phys. Rev. Lett. **85**, 1342 (2000).
[9] P. Carpena and P. Bernaola-Galván, Phys. Rev. B **60**, 201 (1999).
[10] S. Redney, in *Statistical Models for the Fracture of Disordered Media,* edited by H. J. Herrmann and S. Roux (Plenum, New York, 1990).
[11] P. L. Krapivsky, I. Grosse, and E. Ben-Naim, Phys. Rev. E **61**, R993 (2000).
[12] H. Jeffreys, *Theory of Probability* (Clarendon Press, Oxford, 1961), 3rd ed.; G. Schwartz, Ann. Stat. **6**, 461–464 (1978); A. E. Raftery, in *Sociological Methodology,* edited by P. V. Marsden (Blackwells, Oxford, U.K., 1995), pp. 185–195.
[13] A. W. F. Edwards, *Likelihood* (Cambridge University Press, Cambridge, U.K., 1972).
[14] K. P. Burnham and D. R. Anderson, *Model Selection and Inference* (Springer-Verlag, New York, 1998).
[15] W. Li, in *RECOMB'01: Proceedings of the Fifth Annual International Conference on Computational Molecular Biology* (ACM Press, New York, 2001), pp. 204–210.
[16] J. Lin, IEEE Trans. Inf. Theory **37**, 145–151 (1991).
[17] The term "significance level" used here is consistent with the typical usage in statistics literature (e.g., $s = 0.01$). Note that in [2], $1 - s$ is called the significance level.
[18] B. James, K. L. James, and D. Siegmund, Biometrika **74**, 71–83 (1987).
[19] M. Johnson *et al.,* Nature (London) (Suppl.) **387**, 87–90 (1997).
[20] M. V. Olson, in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Genome Dynamics, Protein Synthesis, and Energetics,* edited by James R. Broach, John R. Pringle, and Elizabeth W. Jones (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1991), pp. 1–39.
[21] D. Kessler and W. Li (to be published).
[22] The location of the Y' and X elements is determined by comparing the telomere sequence with the Y' (2022 bases) and X (550 bases) sequences. The comparison is carried out with the dot-matrix method by the DOTTER program at www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html.
[23] F. R. Blattner *et al.,* Science **277**, 1453–1462 (1997).
[24] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Reading, MA, 1949).
[25] More information on Zipfs law can be found at http://linkage.rockefeller.edu/wli/zipf.
[26] J. L. Oliver, R. Román-Roldán, J. Perez, and P. Bernaola-Galván, Bioinformatics **15**, 974–979 (1999).