# Identify Borders of Biological Meaningful Units

**Qing-jun Wang**

## Introduction

DNA sequences of a genome is inhomogeneous. The fractional long-range correlations (domains-within-domains phenomena, Figure 1) have been observed in non-coding, but not coding DNA sequences (Figure 2). However, the origin of the long-range correlations is controversial.

Segmentation procedures are the statistical approaches for identification of homogeneous regions, such as the different compositional patches or DNA domains in a sequence. The proposed segmentation procedure is based on Jensen-Shannon divergence. In brief, the segmentation procedures include three stages: (1) sequence splitting using a sliding window; (2) Jensen-Shannon divergence measure directly dealing with sequence compositional complexity (SCC); (3) stopping criteria.

## Theory

Segmentation procedures extract information from DNA sequences of a genome based on the information theory. The Jensen-Shannon divergence measure is closely related to Shannon entropy. S={$a_1$, $a_2$ …, $a_N$} is a sequence compos of N symbols from the alphabet $\mathcal{A}$ ={A, T, C, G} for DNA. $f_j^A$, $f_j^T$, $f_j^C$, $f_j^G$ are the relative proportions of the nucleotides in subsequence $S_j$. $F_j$={$f_j^A$, $f_j^T$, $f_j^C$, $f_j^G$} is the respective vectors of relative nucleotide frequencies for subsequence $S_j$. The Jensen-Shannon divergence measure for a sequence containing n domains is:

$$\mathrm{JS_n}(s) = \mathrm{H[S]} - \sum_{\mathrm{i}=1}^{n} \frac{l_{\mathrm{i}}}{\mathrm{L}} H[\mathrm{S_i}]$$

where s is a given significance level, L=$\Sigma_i l_i$, S=concatenation of all subsequences $S_i$'s, and Shannon entropy H[]=-p$\log_2$p. The segmentation is called complete (optimum) with divergence $JS_n^*(s)$ if there is no other with higher divergence at the same s. $JS_n^*(s)$ increases monotonically as s decreases. $JS_n(s)$ increases with complexity. The long-range fractal correlations are associated with higher levels of SCC.

The stopping criteria is arbitrary if the significant level s is used. The model selection framework is introduced to build a new stopping criteria. In this framework, basic 1-to-2 segmentation is carried out recursively. Whether or not the 1-to-2 segmentation should be continued is determined by whether the two-random-subsequence model is better than the one-random-sequence model in terms of the model's ability to fit the data and the model's complexity. The Bayesian information criterion (BIC) is a measure of the balancing of the two factors.

$$\mathrm{BIC} = -2\log(\hat{\mathrm{L}}) + \log(\mathrm{N})\mathrm{K} + \mathrm{O}(1) + \mathrm{O}(\frac{1}{\sqrt{\mathrm{N}}}) + \mathrm{O}(\frac{1}{N}) \approx -2\log(\hat{\mathrm{L}}) + \log(\mathrm{N})\mathrm{K}$$

where $\hat{\mathrm{L}}$ is the maximum likelihood, K is the number of parameters in the model, and N is the number of data points. The better the model, the larger the integrated likelihood, and thus the smaller the BIC. Therefore, whether or not the 1-to-2 segmentation should be continued is determined by whether ΔBIC<0 or not. This leads to the new stopping criteria

$$2N \bullet \hat{D}_{JS} > 4\log(N)$$

which sets the lower (relaxed) bound of the significance level, where $\hat{D}_{JS}$ is the maximum of Jensen-Shannon divergence (i.e. $JS_n^*(s)$).

To set the upper (stringent) bound, a measure for segmentation strength s (not the same s as the previous significance level s) is introduced:

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

The stringency level can be raised by choosing a positive threshold $s_0$: $s>s_0>0$. Empirically, the larger the $s_0$, the larger the domain sizes (Figure 3).

## Applications

It is a great idea to apply information theory to genomic sequences. The segmentation procedures are very successful and helpful in understanding the genome sequences.

By applying the SCC measure to real sequences with different correlation structures (HUMTCRADCV, a human DNA sequence whose long-range correlations and complexity are well known; ECO110K, a bacterial uncorrelated sequence), we confirmed that the long-range fractal correlations are associated with higher levels of SCC: the more complexity, the greater $JS_n(s)$ (Figure 4 (b)).

Differences in $JS_n(s)$ between coding and noncoding regions of *E. coli* complete genome (Figure 5) indicate that the noncoding regions have higher complexity than the coding regions.

The complexity profile of myosin heavy-chain genes in different species (Figure 6) indicate that SCC is consistent with biological complexity, i.e. the higher species have higher $JS_n(s)$ values.

With the new stopping criteria, the genomic sequence (e.g. the left telomere of yeast chromosome 12) can be segmented with high accuracy, comparing with known segmentations (Figure 7).

## References

1. Wentain Li. (2001) New stopping criteria for segmenting DNA sequences. Physical Review Letters. 86. p.5815-5818.
2. Roman-Roldan, Bernaola-Galvan, and Oliver (1998) Sequence compositional complexity of DNA through an entropic segmentation method. Physical Review Letters. 80. p1344-1347.
3. Bernaola-Galvan, Roman-Roldan, and Oliver (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. Physical Review E. 53. p5181-5189.

4. Buldyrev et al. (1995) Long-range properties of coding and noncoding DNA sequences: GeneBank analysis. Physical Review E. 51. p5084-5091.
5. Voss. (1992) Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Physical review letters. 68. p3805-3808.

Figure 1.

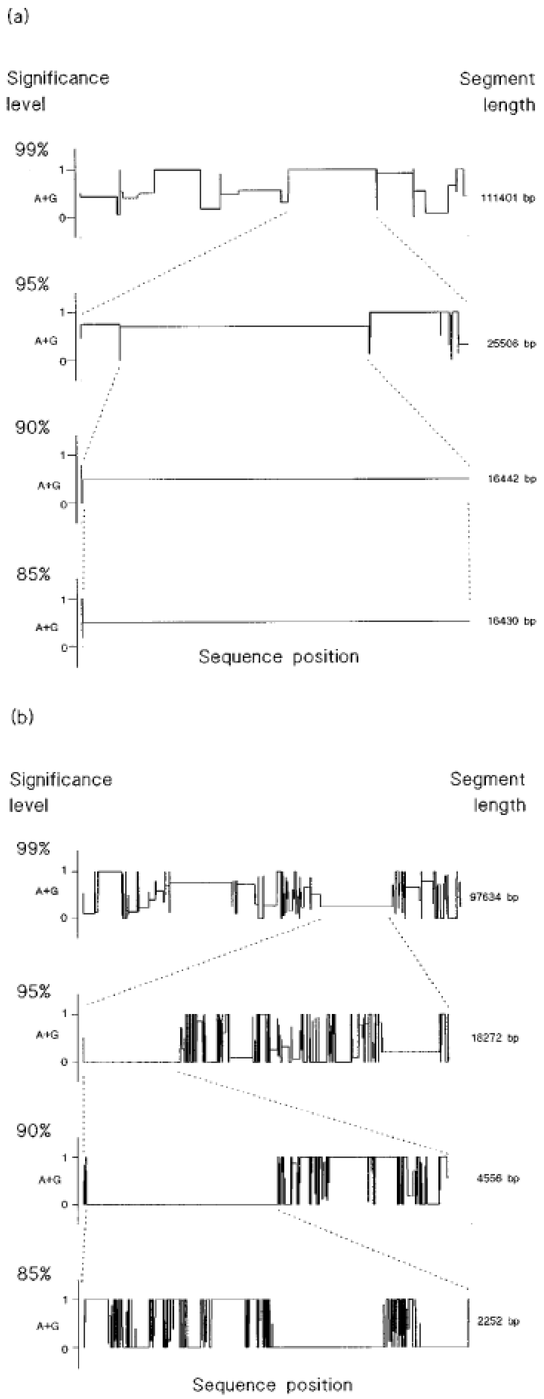

FIG. 3. Recursive segmentation of both the noncorrelated bacterial sequence ECO110K (a) and the long-range correlated human one HUMTCRADCV (b). The longer segment obtained at a given significance step was recursively segmented at a lower significance level. The proportion of purines $(A+G)$ in each segment is represented on the ordinates. The binary alphabet $\{R(A \text{ or } G), Y(C \text{ or } T)\}$ was used.

Figure 2.

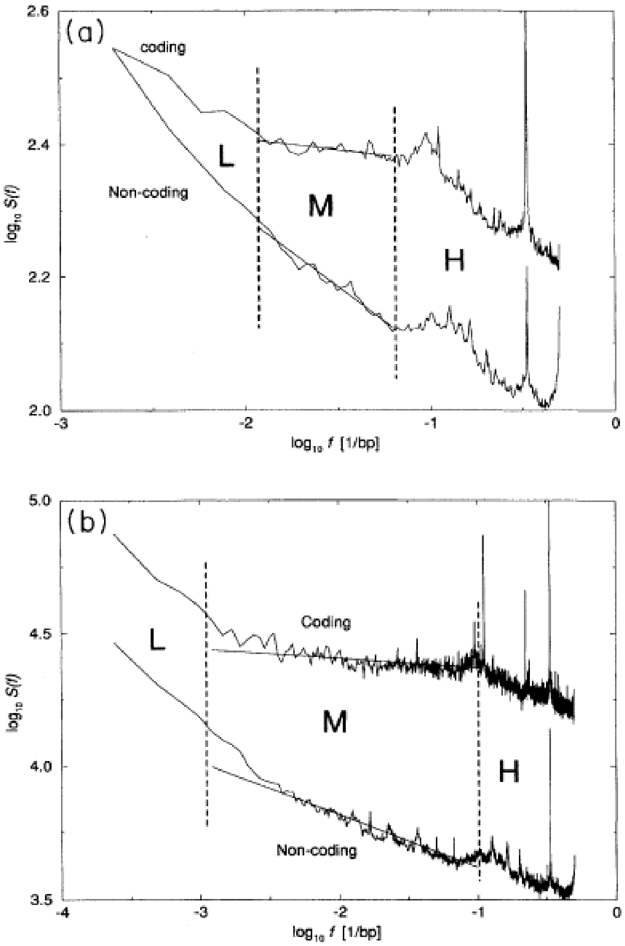

FIG. 1. (a) Power spectra averaged over all eukaryotic sequences longer than 512 bp, obtained by a FFT with a window size of 512. The upper curve is the average over 29 453 coding sequences; the lower curve is the average over 33 301 noncoding sequences. The straight lines are least-squares fits for the second decade (region $M$). The values of $\beta$ measured as the slopes of the fits are 0.03 and 0.21, respectively. (b) Same data for all sequences larger than 4096 bp, obtained by a FFT with a window size of 4096. The average is computed over 874 coding and 1157 noncoding sequences. Note that for high frequencies, the power spectra for both window sizes practically coincide. In the region of frequencies $f < 1/100$ bp$^{-1}$ [region $H$ of (a)], the power spectra in (a) bend upward from the apparent straight line. For (b) (larger windows) the $S(f)$ spectra have a constant slope over more than one decade (region $M$). The fits are the same for both (a) and (b): for coding, $\beta = 0.04$, while for noncoding, $\beta = 0.21$.

Figure 3.

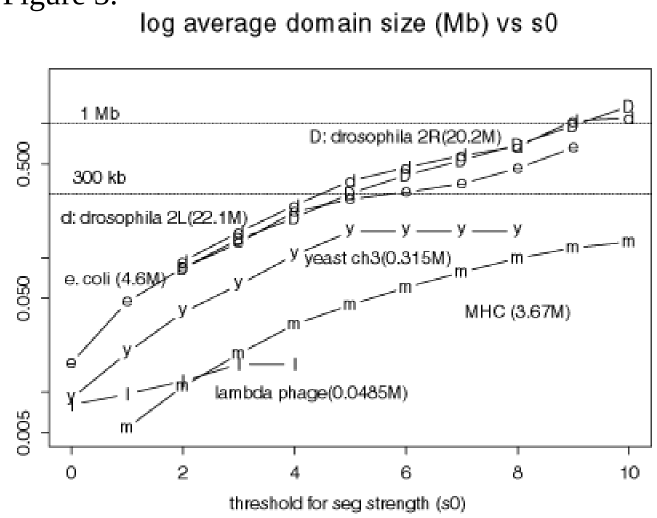

FIG. 3. Average domain size vs segmentation strength $s_0$ for these sequences: human major histocompatibility complex (MHC), $\lambda$ bacteriophage, chromosome 3 of *S. cerevisiae*, *E. coli*, left and right arms of chromosome 2 of *Drosophila melanogaster*.

Figure 4.

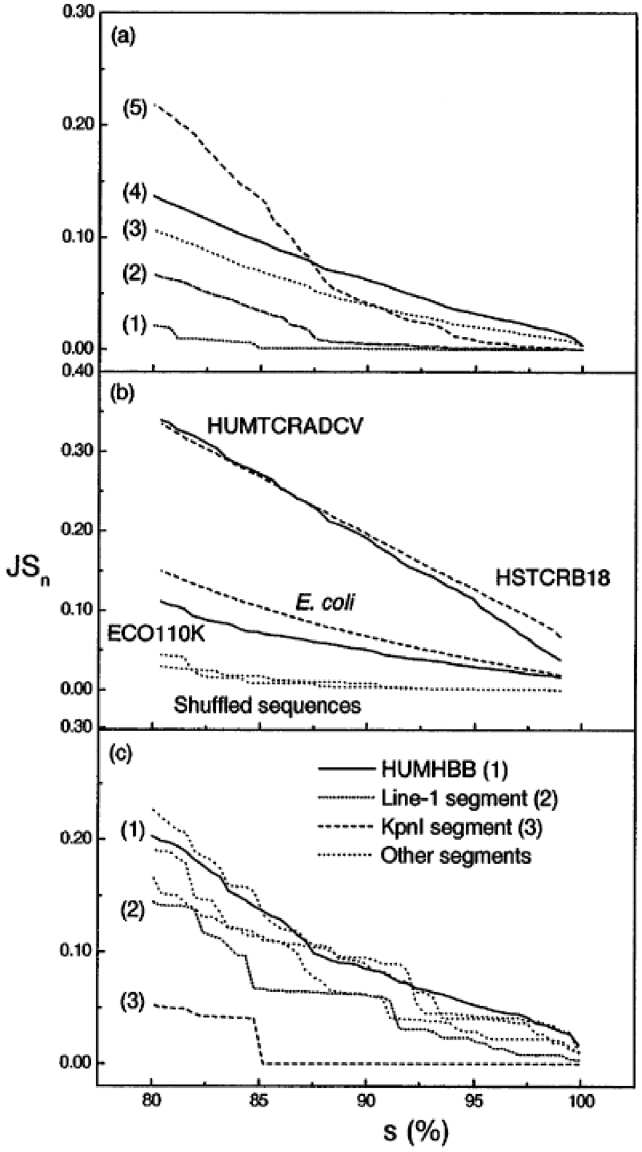

FIG. 1. (a) Complexity profiles of computer-generated sequences with increasing degrees of built-in complexity: (1) A pseudochromosome simulating mutual-information properties of yeast chromosomes but without taking into account compositional heterogeneity [13]; (2) a first-order Markov chain with the same transition probabilities as *HUMTCRADCV*; (3) a generalized Lévy-walk sequence with the parameters described in Ref. [18]; (4) a sequence produced by the insertion-deletion model [14]; (5) a sequence obtained by means of the expansion-modification system [19]. (b) Complexity profiles of *HUMTCRADCV* [97 634 base pairs (bp)], *ECO*110*K* (111401 bp), and their corresponding shuffled sequences. For comparison, two larger sequences are included: the longest human sequence *HSTCRB*18 (684 973 bp) and the complete genome of *E. coli* (4 638 858 bp); the $\{A, T, C, G\}$ alphabet was used in elaborating this plot. (c) Complexity profiles of *HUMHBB* (73 326 bp) and some of the longer fragments resulting from segmenting it ($s = 99\%$). Line-1 and KpnI are two families of repetitive DNA.

Figure 5.

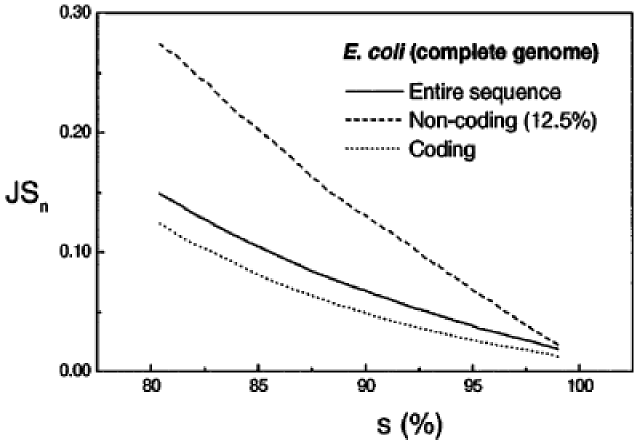

FIG. 3. Differences in SCC between coding and noncoding regions of the *E. coli* complete genome (4 638 858 bp). The quaternary $\{A, T, C, G\}$ alphabet was used.

Figure 6.

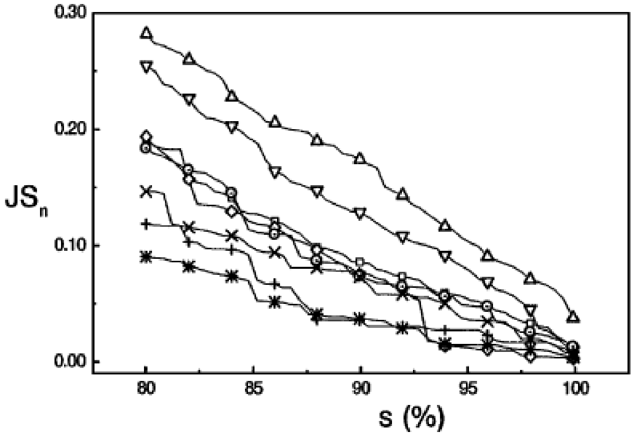

FIG. 4. Complexity profiles of myosyn heavy-chain genes in different species (total length, percentage of introns): (△) Human (28 438 bp, 74%), (▽) rat (25 759 bp, 77%), (○) chicken (31 111 bp, 74%), (◇) *Caenorhabditis* (10 780 bp, 14%), (⊙) *Brugia* (11 766 bp, 32%), (+) yeast (6108 bp, 0%), (×) *Acanthamoeba* (5894 bp, 10%), and (*) *Drosophila* (22 663 bp, 66%).

Figure 7.

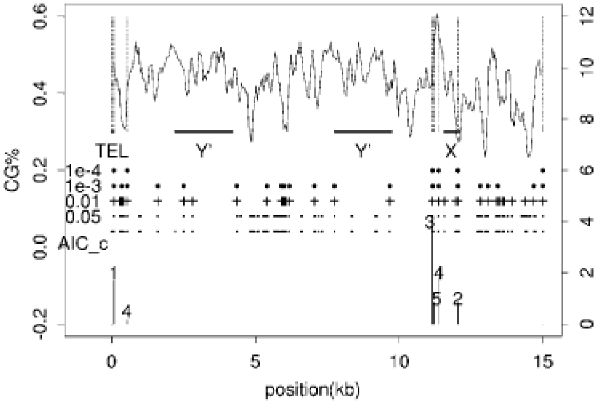

FIG. 1. Partition points determined by the segmentation with the stopping criterion Eq. (2) for the left telomere of yeast *S. cerevisiae* chromosome 12 (dashed vertical lines). The partition points determined by AIC (dot) (with the high-order term included), hypothesis testing framework with significance 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.