Segmentation of genomic DNA through entropic divergence: Power laws and scaling

Rajeev K. Azad,[1,*] Pedro Bernaola-Galván,[2] Ramakrishna Ramaswamy,[3,†] and J. Subba Rao[1]
[1]*School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110 067, India*
[2]*Departamento de Fisica Aplicada II; Universidad de Málaga, Málaga E-29071, Spain*
[3]*School of Physical Sciences, Jawaharlal Nehru University, New Delhi 110 067, India*

(Received 8 October 2001; revised manuscript received 22 January 2002; published 8 May 2002)

Genomic DNA is fragmented into segments using the Jensen-Shannon divergence. Use of this criterion results in the fragments being *entropically homogeneous* to within a predefined level of statistical significance. Application of this procedure is made to complete genomes of organisms from archaebacteria, eubacteria, and eukaryotes. The distribution of fragment lengths in bacterial and primitive eukaryotic DNAs shows two distinct regimes of power-law scaling. The characteristic length separating these two regimes appears to be an intrinsic property of the sequence rather than a finite-size artifact, and is independent of the significance level used in segmenting a given genome. Fragment length distributions obtained in the segmentation of the genomes of more highly evolved eukaryotes do not have such distinct regimes of power-law behavior.

DOI: 10.1103/PhysRevE.65.051909 PACS number(s): 87.14.Gg, 87.10.+e

I. INTRODUCTION

With the genomes of a number of organisms being completely sequenced now, it has become possible to apply a variety of mathematical and statistical tools to analyze DNA sequences in order to study various features and underlying patterns. DNA, for the purpose of such analysis, is typically considered as a symbolic string of length N bases, the bases being the nucleotides denoted A, T, G, and C. The sequences are very heterogenous, typically having a *mosaic* structure [1–4]. This is a consequence both of function—different parts of the DNA have different biological actions—and history—different parts of a genomic DNA evolved at different times in different environments. Thus the base composition of DNA is nonuniform along the chain with both biochemical and physical consequences; nucleotide densities and purine-pyrimidine (A or G and C or T) ratio, etc., differ significantly in different portions of a DNA sequence.

This mosaic organization has been the focus of a number of studies that attempt to delineate the different functional parts of a DNA sequence based on some physical or statistical measures. Through this process, generally termed "segmentation," the aim is to find mutually distinctive portions of the DNA, which are, however, homogenous with respect to a given criterion. A number of different criteria can therefore be used to segment DNA [5–10]. For instance, the Shannon entropy [11] appears useful both in describing the statistical properties of DNA sequences and in having some correlation with biological aspects of the DNA. Thus some studies have used an entropic measure, the *Jensen-Shannon* divergence [5–8,12] as a quantitative criterion for segmentation, the intention being to break a given DNA string into substrings such that the intrasubstring entropic variation is small, while the intersubstring entropic variation is large.

In the present paper we use the Jensen-Shannon entropic divergence [5,7] to fragment several entire genomes. Our motivation is to examine the properties of the segmentation procedure, and to examine the statistical properties of the segments themselves. One feature that we wish to explore is whether the fragments obtained by entropic segmentation have evidence of scale invariance; numerous examples of similar processes appear to be scale invariant [13,14]. Indeed, physical fragmentation also leads to a mass distribution, which may be fractal [15,16].

A complementary direction that has been pursued extensively in understanding DNA organization is the examination of long-range correlations in sequences. A number of studies have found evidence (with varying levels of certainty) for long-range fractal correlations [1–5,17–27] although this is both controversial, and in the end, apparently of questionable biological significance [26–29]. Both noncoding and coding sequences have been seen to give some evidence of power-law correlations [4,21].

In the following section of this paper, we briefly describe the segmentation procedure [5,7,8], as well as the different methods of analysis that we have employed. Section III contains the results for the segmentation of the five genomes studied here. These include species from the three kingdoms commonly used to classify life on earth, namely, archaebacteria, eubacteria, and eukaryota. In addition to the usual four-letter alphabet usually used to treat DNA as a symbolic string, there are alternate encoding protocols, ranging from a two-letter (purine-pyrimidine code) to a 12-letter (frame position specific) alphabet. Results for the distributions using the 12-letter alphabet are also presented in Sec. III. The paper concludes with a discussion and summary in Sec. IV.

II. SEGMENTATION PROCEDURE

The segmentation of a genomic sequence is accomplished by the application of entropic measure, Jensen-Shannon divergence [11]. Given two symbolic sequences built from an alphabet of k symbols, the Jensen-Shannon divergence

[*] Present address: School of Biology, Georgia Institute of Technology, Atlanta, GA 30332.

[†] Email address: rama@vsnl.com

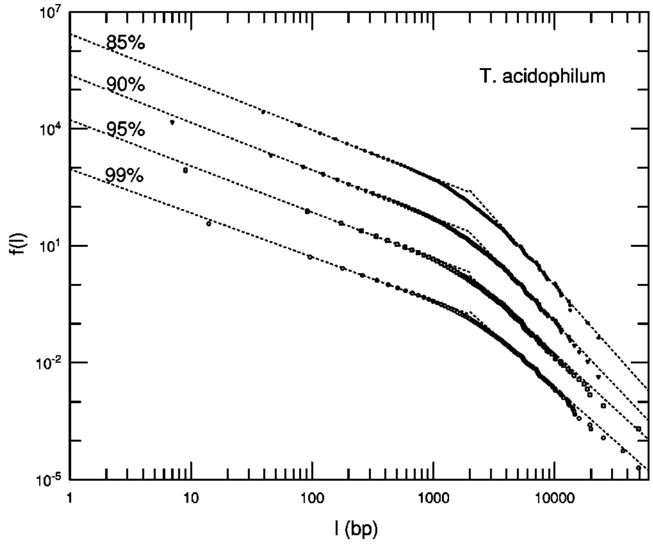

FIG. 1. The patch length distribution $f(l)$ of homogeneous segments obtained by segmenting complete genome of *Thermoplasma acidophilum* represented by four-symbol alphabet A,T,C,G of bases, at 99%, 95%, 90%, and 85% levels of statistical significance. The genomic sequence has been segmented using a recursive segmentation method as described in the text. Each distribution from below has been separated by a decade for clarity. (bp denotes base pairs.)

$$\mathbf{J}(\mathcal{F}^{(1)},\mathcal{F}^{(2)}) = H(\mathcal{F}) - \frac{n^{(1)}}{N}H(\mathcal{F}^{(1)}) - \frac{n^{(2)}}{N}H(\mathcal{F}^{(2)}) \tag{1}$$

is a measure of the compositional difference between them. Here $n^{(i)}$ and $\mathcal{F}^{(i)}=\{f_1^{(i)},f_2^{(i)},\ldots,f_k^{(i)}\}$ $i=1,2$ are the lengths and relative frequency vectors, respectively, of the two sequences. By concatenating the sequences to get a single sequence of length $N=n^{(1)}+n^{(2)}$ with $\mathcal{F}$ the corresponding frequency vector, the corresponding Shannon entropy is

$$H(\mathcal{F}) = -\sum_{i=1}^{k} f_i \log_2 f_i . \tag{2}$$

$\mathbf{J}$ may be generalized to study the divergence among m sequences; the process of segmentation suggested in [5,7] calculates the difference for $m=2$ sequences.

The procedure to segment a given sequence into homogeneous subsequences (domains or patches) is as follows. A sequence of length N is partitioned into two subsequences of lengths $n^{(1)}$ and $n^{(2)}=N-n^{(1)}$, respectively, by varying the partition, namely, all choices of $n^{(1)}$ so as to maximize the divergence $\mathbf{J}$. This procedure is carried out recursively.

To determine whether the partitioning point that maximizes $\mathbf{J}$ is statistically significant or not, two potential subsegments are compared with those from random fluctuations. If $\mathbf{J}_{\max}$ is the maximum value of $\mathbf{J}$ among all possible cutting points, the statistical significance of this maximum is determined by obtaining the probability of getting this value or less in a random sequence. The significance level is thus defined as

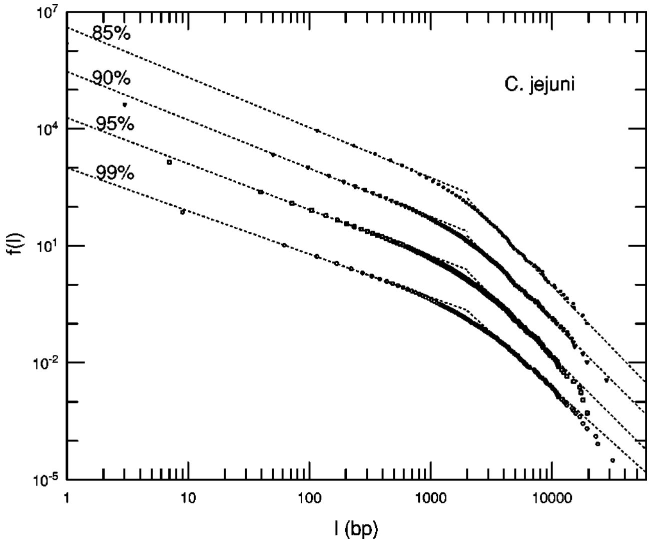

FIG. 2. As in Fig. 1, for the bacterium *Campylobacter jejuni*.

$$s_{\max}(x) = \mathrm{Prob}\{\mathbf{J}_{\max} \leqslant x\}. \tag{3}$$

An approximate analytic expression for the probability distribution of $\mathbf{J}_{\max}$ has been found [7,8],

$$s_{\max}(x) = [F_\nu(\beta 2N(\ln 2)x)]^{N_{\mathrm{eff}}}, \tag{4}$$

where F_ν is the χ^2 distribution function with $\nu=(k-1)(m-1)$ degrees of freedom, β is a scale factor largely independent of N and k and for each k, $N_{\mathrm{eff}}=a\ln N+b$ $(a,b$ are constants). The values of a, b, and β are obtained from Monte Carlo simulations by fitting the empirical distributions to the above expression [7,8].

A sequence is segmented at a preassigned *significance level* s_0 as follows. If $s_{\max}$, determined as discussed above, exceeds s_0, the sequence is segmented at this point. The procedure is continued recursively for each of the two resulting segments. It is necessary to ensure that at each stage, the resulting subsequences maintain their distinction (vis-a-vis the Jensen-Shannon divergence) from their neighbors formed at the previous segmentation steps. The process is terminated when *all* segments thus obtained either have $s_{\max}\leqslant s_0$, or if a possible partition will lead to segments that are not compositionally distinct from their neighbors.

III. APPLICATION AND RESULTS

The genomic DNA sequences studied here are represented by the usual four-symbol alphabet $\{A,T,C,G\}$ where A,T,C,G represent the bases adenine, thymine, cytosine, guanine respectively, and a 12-symbol alphabet $\{A_i,T_i,C_i,G_i\}$, $i=1,2,3$, where the subscripts indicate the positions of bases within a codon [8]. The representative genomes taken from the GenBank [30] are *Thermoplasma acidophilum* (archaebacteria, 1.56 Mbp), *Campylobacter jejuni* (eubacteria, 1.64 Mbp), *Saccharomyces cerevisiae* chromosome IV (eukaryota, 1.53 Mbp), *Arabidopsis thaliana* chromosome II (eukaryota, 1.5 Mbp) and human chromosome 22 (eukaryota, 1.52 Mbp), where bp means base pairs. The chromosomes of *A. thaliana* and human are very long and incomplete; we take a long string of 1.5 Mbp, which is

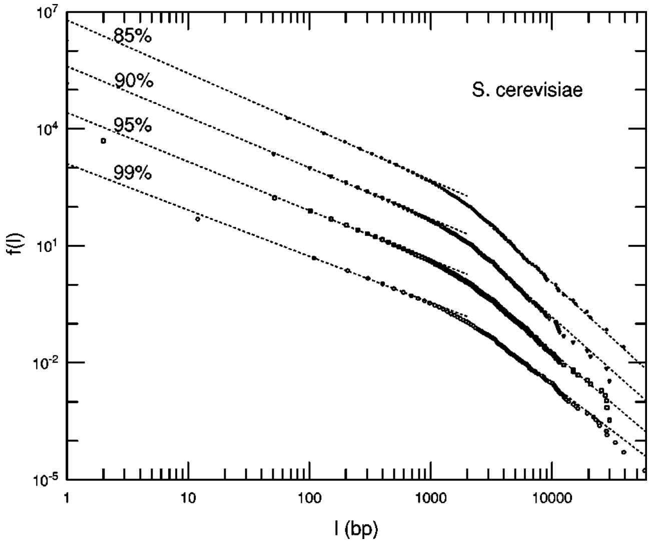

FIG. 3. As in Fig. 1, for *Saccharomyces cerevisiae* chromosome IV.

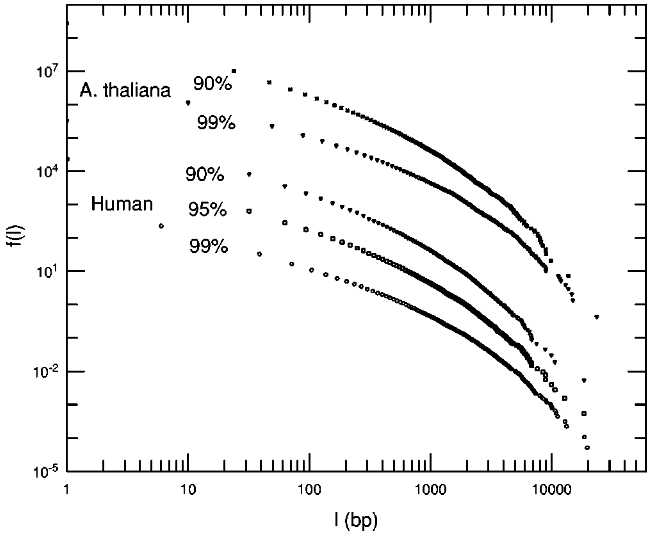

FIG. 4. As in Fig. 1, for *Arabidopsis thaliana* chromosome II and a human chromosome 22 contig.

completely specified for *A. thaliana* and a contig $(gi|10880022|ref|NT\_011522.1|)$ of human chromosome 22.

Segmentation, as discussed in the preceding section was carried out. In order to search for scaling laws in such a fragmentation, we observe the probability distribution of segments obtained for different genomes. We denote by $n(l)$ the number of segments of length *l*, and consider the distribution

$$F(l)=\sum_{l'=l}^{\infty} n(l')\approx\int_{l'=l}^{\infty} n(l')dl', \tag{5}$$

which corresponds to the number of segments with length greater than or equal to *l*. If the distribution of lengths follows a power law, namely, $n(l)\propto l^{-\alpha}$, $F(l)$ also follows a power law, with exponent $1-\alpha$.

Figures 1–4 show $f(l)=l^{-1}F(l)$ for the above genomes using the four-symbol alphabet, segmented at four different significance levels (99%, 95%, 90%, and 85%). Note that, for $k=4$ and $m=2$, the number of degrees of freedom $\nu=3$, and from Monte Carlo simulation, $a=2.44$, $b=-6.15$, and $\beta=0.79$ [7]. The number of segments obtained at 99% significance level are 516 for *T. acidophilum*, 654 for *C. jejuni*, 584 for *S. cerevisiae*, 1122 for *A. thaliana*, and 1360 for human chromosome 22 contig. At a given significance level, a heterogeneous DNA sequence yields more segments than a relatively homogeneous one [5,7]. As may be expected, therefore, the human chromosome has larger number of segments than a bacterial sequence of comparable length. Lowering the significance level increases the number of segments obtained for an organism; similar trends observed as above for the five species can also be observed at other significance levels.

Patch length distributions in Figs. 1–3 show a characteristic scale separating two clear regimes of power laws. This scale is nearly constant from archaeal genome to early eukaryotic and lies in the range of $\approx$1300–1500 bp. Thus, $f(l)\propto l^{-\alpha_i}$, $i=1,2$, α_1 is the exponent below the characteristic scale and α_2 above. The exponents are obtained from a fit to the data, and $\alpha_1<\alpha_2$. The values of α_1 and α_2 from Figs. 1–3 are given in Table I. The segment distributions of higher eukaryotes, for example, *A. thaliana* and human (see Fig. 4) appear to be smooth and are conspicuous by the absence of a characteristic scale.

This agrees with recent results obtained using the autocorrelation function [31]. For prokaryotes and early eukaryotes the autocorrelation function usually presents a characteristic length scale (around a few kilobases) beyond which it essentially drops to zero, while for higher eukaryotes and especially for human DNA, the autocorrelation function has a power-law behavior extending in some cases to more than 5 decades, indicating the absence of a characteristic length scale. This last result has been also obtained for human chromosome 22 using mutual information [22].

Distribution profiles for the genomes encoded in 12-symbol alphabet and segmented at 99%, 95%, and 90% levels of significance are shown in Figs. 5(a)–5(d). For this code, the number of degrees of freedom is $\nu=9$ [7,8] and a,b,β are $2.34,-3.69,0.84$, respectively. The distributions show

TABLE I. The scaling exponents α_1 and α_2 observed below and above the characteristic scale that partitions the patch length distribution into two power-law regimes for the genomes of three different representative organisms. The scaling exponents given correspond to the genomic sequences coded in $k=4$ symbol alphabet and at four different significance levels as described in the text.

	99%		95%		90%		85%	
Genome	α_1	α_2	α_1	α_2	α_1	α_2	α_1	α_2
T. acidophilum	1.12	2.76	1.18	2.85	1.21	3.27	1.23	3.48
C. jejuni	1.1	2.81	1.18	3.1	1.24	3.1	1.28	3.21
S. cerevisiae	1.18	2.37	1.25	2.65	1.3	2.77	1.36	2.86

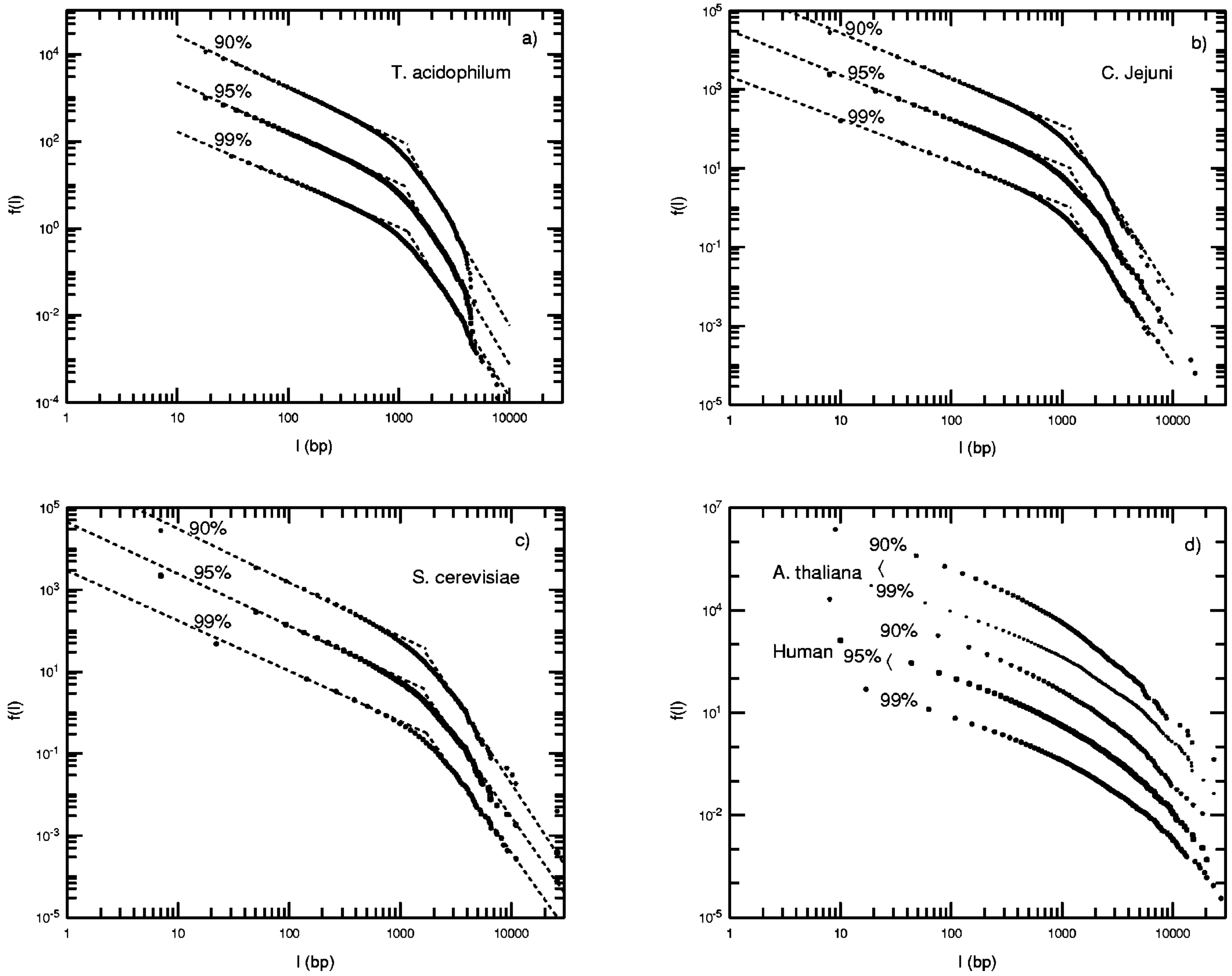

FIG. 5. The patch length distribution $f(l)$ as in Fig. 1 of five genomes, using a 12-symbol alphabet that takes into account both the base and the codon position $\{A_i, T_i, C_i, G_i\}$, $i=1,2,3$ (see text). The segmentation has been done at 99%, 95%, and 90% levels of statistical significance.

similar trends in bending profiles as for the four-symbol alphabet, though the characteristic scales now show differences from archaea to early eukaryotes; see Figs. 5(a)–5(c). The characteristic scale for *T. acidophilum* and *C. jejuni* is ≈ 800 bp and for *S. cerevisiae* ≈ 1100 bp. We find the two scaling regimes to be most clearly distinguished at 99% significance level. Segmentation of a nucleotide sequence coded in 12-symbol alphabet delineates the coding and noncoding regions; we find these characteristic lengths almost the same as the average size of coding segments of the respective genomes. The power-law exponents α_1 and α_2 are given in Table II.

The characteristic length, which is observed as separating the two regimes of the power-law behavior does not appear to be an artifact either of the statistical significance level or of finite size (length). This is evident from our results shown for a wide range of significance levels in Figs. 1–3 and Figs. 5(a)–5(c). The scaling exponents on either side of the characteristic length vary with the significance level used to segment a genome, but the length itself remains unchanged. A higher significance level results in reduced segmentation and thus larger patch sizes, while a lower significance level causes larger patches to segment further. Even when only three-fourths or half of the genome is used, two scaling regimes result with the same characteristic length, as shown in Fig. 6 for *T. acidophilum* (90% significance level).

TABLE II. The scaling exponents α_1 and α_2 observed below and above the characteristic scale for the genomes of three different representative organisms. The scaling exponents given correspond to the genomic sequences coded in $k=12$ symbol alphabet and at three different significance levels as described in the text.

	99%		95%		90%	
Genome	α_1	α_2	α_1	α_2	α_1	α_2
T. acidophilum	1.09	4.12	1.15	4.28	1.19	4.4
C. jejuni	1.08	4.13	1.13	4.49	1.17	4.4
S. cerevisiae	1.22	3.81	1.26	3.88	1.31	4.14

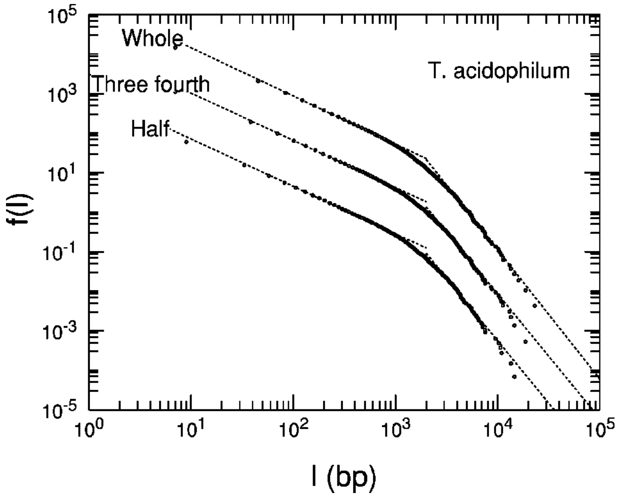

FIG. 6. The patch length distribution $f(l)$ of *Thermoplasma acidophilum* using four-symbol alphabet for half, three-fourth, and whole genome size at 90% significance level.

Note that the present sets of segment lengths of bacterial and primitive eukaryotic genomes do not appear to follow the lognormal distribution that arises in the Kolmogorov theory of physical fragmentation [15]. Shown in Fig. 7(a) are data from fragmenting *T. acidophilum* at 90% significance level (recall that the patch length l is lognormally distributed if $\ln l$ is normally distributed). The corresponding distribution of the human sequences superficially seems somewhat closer to the lognormal distribution [see Fig. 7(b)] although we have found it possible to fit the distributions for human or *A. thaliana* to the lognormal probability density function only over limited ranges. It has not been possible to fit the distributions to the lognormal form over the entire range using a single set of fitting parameters.

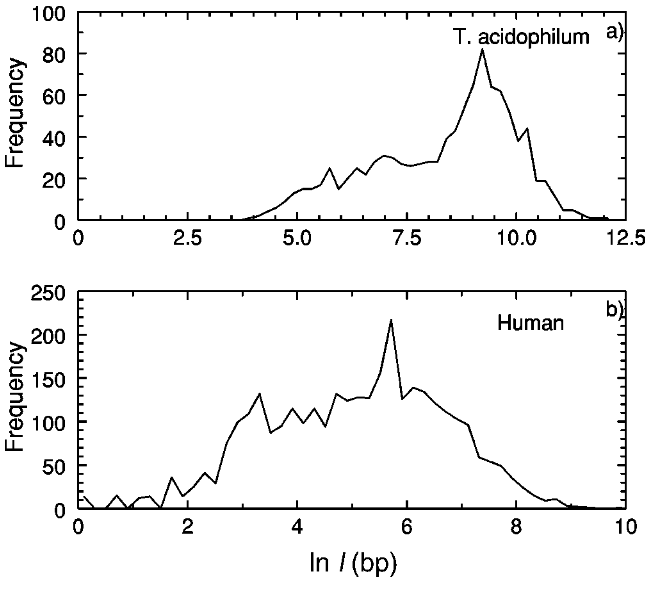

FIG. 7. Distribution of the (natural) logarithm of the patch lengths of (a) *T. acidophilum* and (b) human at 90% significance level.

IV. DISCUSSION AND SUMMARY

Unraveling the history of a given genome is a complex task, and one that needs a variety of different approaches. Identification of the different structural features within the DNA—the exons and introns, repetitive DNA, telomeres, isochores, for instance, is one objective. However, the DNA of any organism itself has a complicated evolutionary history with a variety of different selection pressures acting on it at different times. To uncover this aspect of DNA evolution, it may be essential to go beyond an analysis of the functional parts of the DNA, and segmentation study, like the one presented in this paper, is one approach to understand the genome organization.

The principle behind segmentation is simple: break up a complex object into its ‘‘constituent’’ parts, and thereby attempt to understand how the organization comes about in the first place.

This is the motivation, for instance, in studying the fragmentation of physical objects—rocks or gypsum molds of different shapes, for instance. Distribution profiles of masses of fragments so obtained have been seen to follow a power-law behavior [16]. There is, however, a characteristic size so that the actual distribution of fragment masses m follows the law $am^{-b}\exp(-m/m_0)$, where m_0 is the characteristic finite-size cutoff mass, and a and b are constants. Below m_0 the distribution is close to a power-law behavior with exponent b. Physical fragmentation is dictated by the breaking of strong bonds between molecules that determine the structure and tensile strengths of solids; the mass distribution follows a power law for the entire range of masses (minus the cutoff). Such fragments are held together in a solid by the same type of adhesion so that a smaller fragment breaks in a similar manner as a larger one.

DNA sequences are heterogeneous at various levels of description and thus the fragmentation of DNA into entropically homogeneous segments is, in principle, very different. We have studied the segment length distributions for the genomic DNAs of representative organisms spanning the three classified kingdoms. The segments that a given sequence is divided into are such that within a given domain, the composition is uniform (in terms of the Shannon entropy); thus these domains could reflect the evolution of a given genome. Based on this premise, one would expect that an organism further along the evolutionary tree will have a more complex genomic organization.

This is borne out in our studies: for bacterial genomes, the domains appear to have a power-law distribution with evidence of two separate regimes of scaling behavior. Although we have analyzed all available complete genomes of archaebacteria and eubacteria, only representative data have been presented here and compared with chromosome wide data for three eukaryotes (again, not all the examples analyzed are presented here). Our results, which are consistent across the kingdoms, show two regimes of power-law scaling in the bacterial genomes as well as primitive eukaryotes like yeast. Application of the segmentation algorithm to other bacterial genomes shows similar features, displaying the general fractal organization of nucleotides that make up such genomes.

This is in contrast to the segment length distribution of highly evolved eukaryotes, such as human or plant; here the distribution shows a smooth transition across the entire range of segment lengths, thus lacking distinct characteristic scales. The nucleotide composition of higher eukaryotes is very complex; this is attributed to the abundance of noncoding sequences or introns in DNA sequences of such organisms. Of as yet undetermined function, noncoding DNA may well be crucial vis-a-vis evolution. Such regions are more prone to alteration by different evolutionary processes, e.g., duplication, mutation, insertion, deletion, etc. For example, certain repetitive elements (closely related to such processes) have been recently identified as responsible for the long-range correlations observed in human DNA [22]. The fragments resulting from the segmentation algorithm carry the imprint of these different processes and it may be anticipated that the scaling features would be more complex than for bacterial sequences.

ACKNOWLEDGMENTS

We thank Mark Borodovsky for comments on the manuscript. R.K.A. and R.R. were supported by a grant from the Department of BioTechnology, India. P.B.G. was partially supported by Grant No. BIO99-0651-CO2-01 from the Spanish Government.

[1] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).
[2] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).
[3] W. Li, T. G. Marr, and K. Kaneko, Physica D **75**, 392 (1994).
[4] R. F. Voss, Fractals **2**, 1 (1994).
[5] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).
[6] W. Li, Phys. Rev. Lett. **86**, 5815 (2001).
[7] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. L. Oliver, and H. E. Stanley, Phys. Rev. E (to be published).
[8] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).
[9] V. E. Ramensky, V. Ju Markeev, M. A. Roytberg, and V. G. Tumanyan, J. Comput. Biol. **7**, 1 (2000).
[10] J. V. Braum and H. G. Müller, Stat. Sci. **13**, 142 (1998).
[11] J. Lin, IEEE Trans. Inf. Theory **37**, 145 (1991).
[12] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley, Phys. Rev. E **61**, 5624 (2000).
[13] P. L. Krapivsky, I. Grosse, and E. Ben-Naim, Phys. Rev. E **61**, R993 (2000).
[14] P. Bernaola-Galván, P. Ch. Ivanov, L. A. N. Amaral, and H. E. Stanley, Phys. Rev. Lett. **87**, 168105 (2001).
[15] J. Aitchison and J. A. C. Brown, *The Lognormal Distribution* (Cambridge University Press, Cambridge, 1957).
[16] L. Oddershede, P. Dimon, and J. Bohr, Phys. Rev. Lett. **71**, 3107 (1993).
[17] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[18] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[19] C. A. Chatzidimitriou-Dreismann, and D. Larhammar, Nature (London) **361**, 212 (1993).
[20] S. Karlin and V. Brendel, Science **259**, 677 (1993).
[21] A. Arneodo, Y. d'Aubenton-Carafa, B. Audit, E. Bacry, J. F. Muzy, and C. Thermes, Eur. Phys. J. B **1**, 259 (1998).
[22] D. Holste, I. Grosse, and H. Herzel, Phys. Rev. E **64**, 041917 (2001).
[23] Z.-G. Yu, V. V. Anh, and B. Wang, Phys. Rev. E **63**, 011903 (2000).
[24] X. Lu, Z. Sun, H. Chen, and Y. Li, Phys. Rev. E **58**, 3578 (1998).
[25] M. de Sousa Vieria, Phys. Rev. E **60**, 5932 (1999).
[26] B.-L. Hao, H. C. Lee, and S.-Y. Zhang, Chaos, Solitons Fractals **11**, 825 (2000).
[27] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J. F. Muzy, and A. Arneodo, Phys. Rev. Lett. **86**, 2471 (2001).
[28] W. Li, Int. J. Bifurcation Chaos Appl. Sci. Eng. **2**, 137 (1992).
[29] W. Li and K. Kaneko, Nature (London) **360**, 635 (1992).
[30] All sequences have been retrieved from GenBank at NCBI (http://ncbi.nlm.nih.gov).
[31] P. Bernaola-Galván, P. Carpena, R. Román-Roldán, and J. L. Oliver, Gene (to be published).