Simplifying the mosaic description of DNA sequences

Rajeev K. Azad*, J. Subba Rao, Wentian Li[1], and Ramakrishna Ramaswamy † [2]

School of Environmental Sciences, [2] *School of Physical Sciences*

Jawaharlal Nehru University, New Delhi 110 067, India

[1] *Center for Genomics and Human Genetics, North Shore - LIJ Research Institute, Manhasset, NY 11030, USA*

(February 2, 2008)

Abstract

By using the Jensen-Shannon divergence, genomic DNA can be divided into compositionally distinct domains through a standard recursive segmentation procedure. Each domain, while significantly different from its neighbours, may however share compositional similarity with one or more distant (non–neighbouring) domains. We thus obtain a coarse–grained description of the given DNA string in terms of a smaller set of distinct domain labels. This yields a *minimal* domain description of a given DNA sequence, significantly reducing its organizational complexity. This procedure gives a new means of evaluating genomic complexity as one examines organisms ranging from bacteria to human. The mosaic organization of DNA sequences could have originated from the *insertion* of fragments of one genome (the parasite) inside another (the host), and we present numerical experiments that are suggestive of this scenario.

*Present address: School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA

†For correspondence, email:rama@vsnl.com

I. INTRODUCTION

One of the major goals in DNA sequence analysis is in gaining an understanding of the overall organization of the genome. Beyond identifying the manifestly functional regions such as genes, promoters, repeats, etc., it has also been of interest to analyse the properties of the DNA string itself. One set of studies has been directed towards examining the nature of correlations between the bases. There is some evidence for long-range correlations which give rise to *1/f* spectra in genomic DNA [1,2,3]; this feature has been attributed to the presence of *complex heterogeneities* in nucleotide sequences [3]. These result in hierarchical patterns in DNA, the mosaic or 'domain within domain' picture [4]. This structure is most conveniently explored through segmentation analysis based on information theoretic measures [4,5,6,7], although other schemes to uncover the correlation structure over long scales, such as detrended fluctuation analysis of DNA walks [8] or wavelet tranform technique [9] have also been applied. There have been some attempts to decode the biological implications of such complexity [9,10,11], but these are incompletely understood as of now. On shorter length scales there is a prominent 3-base correlation in coding regions of DNA; this offers a means of locating and identifying genes [12]. There are other short–range correlations as well [13,14] corresponding to structural constraints on the DNA double helix.

Segmentation analysis is a powerful means of examining the large–scale organization of DNA sequences [4,5,6,15,16,17,18]. The most commonly used procedure [4,5,6] is based on maximization of the Jensen-Shannon (J-S) *divergence* through which a given DNA string is recursively separated into compostionallly homogeneous segments called domains (or patches). This results in a coarse-grained description of the DNA string as a sequence of distinct domains. The criterion for continuing the segmentation process is based on statistical significance (this is equivalent to hypothesis testing) [4,5] or, alternatively, within a model selection framework based on the Bayesian information criterion [7]. This criterion can be extended and used to detect isochores [7], CpG islands, origin and terminus of replication in bacterial genomes, complex repeats in telomere sequences, etc. [19]. Segmentation

using a 12-symbol alphabet derived from codon usage has been shown recently to delineate the border between coding and noncoding regions in a DNA sequence [6].

In the present work, we analyse the segmentation structure of genomic DNA for a class of genomes ranging in (evolutionary) complexity from bacteria to human. Our motivation is to understand the complexity of genome organization in terms of the domains obtained. We further aim to correlate the domain picture with evolutonary biological processes.

By construction a given domain is heterogenous with respect to its neighbours, but it may nevertheless be compositionally similar to other domains. Based on this premise, we attempt to draw a larger domain picture by obtaining 'domain sets'. These consist of a set of domains which are homogeneous when concatenated. A domain set may thus be interpreted as a larger homogeneous sequence, parts of which are scattered nonuniformly in a genomic sequence. The number of domain sets constructed thus is found to be much fewer than the domains obtained upon segmentation [4,5,6,7]. We propose here an optimal procedure, starting from the domains found from one of the above segmentation methods, and building up a domain set by adding together all its components. We then use standard complexity measures to show that this gives a superior model in as much as the complexity is reduced.

This paper is organised as follows. In the next section, we briefly review the segmentation methods based on the J-S divergence. Section III contains our main results. We first segment a given genome to reveal the primary domain structure that derives from the J-S divergence. We then show how the domain sets are constructed, and analyse the attendant decrease in complexity. In Section IV, we speculate that such domain organization ocurred during genomic evolution when there was lateral gene and/or DNA transfer between species. To that end, we present the results of numerical experiments based on a host-parasite model, where we artificially insert fragments of one genome inside another, and demonstrate that this process can be uncovered via segmentation. Section V concludes the paper with a summary and discussion of our results.

II. SEGMENTATION METHODS

In this section we briefly review the segmentation methodology that is used here in order to fragment a genome into homogeneous domains. Consider a sequence $\mathcal{S}$ as a concatenation of two subsequences $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$. The Jensen–Shannon divergence [20] of the subsequences is

$$\mathbf{D}(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}) = H(\pi^{(1)}\mathcal{F}^{(1)} + \pi^{(2)}\mathcal{F}^{(2)}) - [\pi^{(1)}H(\mathcal{F}^{(1)}) + \pi^{(2)}H(\mathcal{F}^{(2)})], \tag{1}$$

where $\mathcal{F}^{(i)} = \{f_1^{(i)}, f_2^{(i)}, ..., f_k^{(i)}\}, i = 1, 2$ are the relative frequency vectors, and $\pi^{(1)}$ and $\pi^{(2)}$ their weights. In Eq. (1), H is the Shannon entropy (in unit of bits)

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

although, as can be appreciated, a variety of other functions on the f_i's can also be used as a criterion for estimating the divergence of two sequences.

The algorithm proposed by Bernaola-Galván *et al.* [4,5] proceeds as follows. A sequence is segmented in two domains such that the J-S divergence $\mathbf{D}$ is maximum over all possible partitions. Each resulting domain is then further segmented recursively.

The main issue with regard to continual segmentation is that unless the significance of a given segmentation step is properly assessed, it is possible to arrive at segments which have no great significance. This question is also related to a second issue, namely when one should stop the recursion. Since we consider finite DNA sequences, it is again possible to keep segmenting until the segments are very short. Both these questions can be answered through one of two possible approaches which we now describe.

A. Hypothesis testing framework

The statistical significance of the segmentation is determined by computing the maximum value of the J-S divergence for the two potential subsegments, $\mathbf{D}_{max}$, and estimating the

probability of getting this value or less in a random sequence. This defines the significance level, $s(x)$, as

$$s(x) = Prob\{\mathbf{D}_{max} \leq x\}. \tag{3}$$

The probablility distribution of $\mathbf{D}_{max}$ has an analytic approximation [5,6] and

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

where F_ν is the chi–square distribution function with ν degrees of freedom, N is the sequence length, β is a scale factor which is essentially independent of N and k and for each k, $N_{eff} = a\ \ln N + b$. The values of β and N_{eff} (and thus the constants a and b) are found from Monte Carlo simulations by fitting the empirical distributions to the above expression [5,6].

Within the hypothesis testing framework, then, the segmentation is allowed if and only if $s(x)$ is greater than a preset level of statistical significance. It is possible to segment a given sequence initially at a (usually very high) significance level, and these domains are further segmented at lower levels of significance to detect the inner structure or other patterns [15].

B. Model selection framework

A different criterion can be evolved for stopping the recursive segmentation within the so-called model selection framework [7]. This is based on the Bayesian information criterion [21,22,23], denoted **B** below,

$$\mathbf{B} = -2\log(\hat{L}) + \log(N)K + O(1) + O(\frac{1}{\sqrt{N}}) + O(\frac{1}{N}), \tag{5}$$

where $\hat{L}$ is the maximum likelihood of the model, N is the sample size and K is the number of parameters in the model.

A potential segmentation based on the J-S divergence **D** is deemed acceptable if **B** is reduced after segmentation. From the above equation, this condition is [7]

$$2N\mathbf{D} > (K_2 - K_1)\log(N), \tag{6}$$

where K_1 and K_2 are the number of free parameters of the models before and after the segmentation. This is the lower bound of the significance level; an upper bound can be preset by using a measure of *segmentation strength* [7],

$$s = \frac{2N\mathbf{D} - (K_2 - K_1)\log(N)}{(K_2 - K_1)\log(N)}. \tag{7}$$

Eq. (6) is equivalent to the condition $s > 0$.

III. APPLICATIONS AND ANALYSIS

In the present work we consider DNA sequences as strings in a 4–letter alphabet (A, T, C, G). In the model selection framework discussed above, therefore, the relevant parameters are $K_1 = 3$ (since only 3 of the 4 nucleotides are independent) and $K_2 = 7$ (the 3 free parameters from each of the two subsegments, and in addition, the partition point which is another independent parameter) [7]. The importance of this segmentation approach in detecting some of the structural and functional units in DNA sequences has been demonstrated recently [19]. The results that follow have been obtained by the application of this approach.

A. Labeling the domains

The complete genome of a bacterium *Ureaplasma urealyticum* (751719 bp) and a contig of human chromosome 22 ($gi \mid 10879979 \mid ref \mid NT\_011521.1 \mid$, 767357 bp) were segmented at the lower bound of the stopping criterion, namely Eq. (6). The number of segments obtained by this procedure is 86 for the bacterium and 248 for human chromosome 22 contig. Labeling each of these segments by a unique symbol gives a coarse–grained view of the entire sequence, say $S_1 \cdot S_2 \cdots S_N$.

While each segment S_k is heterogeneous with respect to its neighbours, $S_{k\pm1}$, it need not be compositionally distinct from a non–neighbouring segment, S_j. Therefore, we now

examine the *inter se* heterogeneity of all segments with respect to each other. Segments S_k and S_j are concatenated, and if this 'supersegment' cannot be segmented by the same criterion, then both S_k and S_j are assigned the same domain symbol. This is done recursively and exhaustively, so that within the model selection framework of segmentation, all domains that cannot be distinguished from one another are assigned the same symbol. This gives a reduced and further coarse–grained view of the domain structure of a DNA sequence.

To ensure that the above procedure is as complete and self–consistent as possible, we examine each segment S_k by concatenating it with S_j and <u>all</u> preceding distinct segments that share the same domain symbol as S_j, and examine whether this larger sequence can be segmented. Explicitly, if segments S_i and S_j have the same symbol (following the procedure given above) we examine the supersegment $S_i \cdot S_j \cdot S_k$ to determine whether segment S_k should share the same domain symbol or not. It is further required to to consider all possible subsets ($S_i \cdot S_k$, $S_j \cdot S_k$, etc.) to ensure that all segments that are deemed to share a given domain symbol do indeed belong to one class, namely that such superdomains do not undergo further segmentation.

Following the above, the 86 domains obtained from the segmentation of *U. urealyticum* are reduced to a total of 17 distinct domain types:

$\underline{S_1}\ S_2\ S_3\ S_4\ S_5\ S_3\ \underline{S_1}\ S_2\ \underline{S_1}\ S_6\ S_4\ \underline{S_1}$

$S_6\ S_7\ S_2\ \underline{S_1}\ S_6\ S_4\ S_8\ S_9\ S_4\ S_9\ S_{10}\ S_4$

$S_9\ S_4\ S_{11}\ S_{12}\ S_6\ S_4\ S_{10}\ S_6\ S_{10}\ S_6\ S_{11}\ S_6$

$S_7\ S_6\ S_{11}\ S_7\ S_3\ S_{11}\ S_3\ S_{10}\ S_6\ S_3\ S_9\ S_{11}$

$S_{10}\ S_4\ S_{11}\ S_{10}\ S_{13}\ S_4\ S_{13}\ S_9\ S_{11}\ S_4\ S_6\ S_4$

$S_{11}\ S_4\ S_{14}\ S_6\ S_8\ S_6\ S_{14}\ S_4\ S_6\ S_{15}\ \underline{S_1}\ S_9$

$S_4\ S_{16}\ S_9\ S_{17}\ S_{15}\ S_6\ S_{17}\ S_7\ S_{17}\ \underline{S_1}\ S_{17}\ S_8$

$S_{16}\ S_{14}$

The 248 segments of human chromosome 22 also undergo simplification, to a total of 53 distinct domain types:

S_1 S_2 S_3 S_4 S_5 S_4 S_3 S_6 S_4 S_6 S_7 S_4

S_8 S_4 S_9 S_{10} S_6 S_4 S_7 S_1 S_4 S_7 S_6 S_4

S_7 S_{11} S_4 S_{12} $\underline{S_{13}}$ S_4 S_{14} S_{12} S_4 S_{15} S_{16} S_{14}

S_6 S_9 S_{10} S_{17} S_{16} S_{10} S_{16} S_6 S_{12} S_{18} S_{12} S_{10}

S_3 S_1 S_3 S_1 S_{10} S_9 S_6 S_3 S_{12} S_{16} S_3 S_{12}

S_{14} S_1 S_7 S_6 S_{12} S_7 S_1 S_6 S_{19} S_6 S_{20} S_{17}

S_7 S_{21} S_7 S_{22} S_{21} S_{22} S_{23} S_7 S_{23} S_{24} S_{17} S_{21}

S_7 S_{21} S_1 S_{21} S_7 S_{21} S_7 S_{16} S_{25} S_1 S_{16} S_{15}

S_{26} S_8 S_{15} S_8 S_{21} S_8 S_{21} S_{27} S_{16} S_{12} S_1 S_{28}

S_{21} S_{28} S_{21} S_{12} S_{21} S_{16} S_{12} S_{16} S_{12} S_{28} S_{16} S_{19}

S_{17} S_{27} S_{28} S_{16} S_{20} S_{21} S_{29} S_{25} S_{30} S_{25} S_{31} S_{25}

S_{28} S_8 S_{25} S_{29} S_{32} S_3 S_{25} S_{31} S_{33} S_8 S_{31} S_{34}

S_{31} S_{29} S_{30} S_{31} S_{35} S_{36} S_{21} S_{36} S_{37} S_{36} S_2 S_{36}

S_9 S_1 S_9 $\underline{S_{13}}$ S_{38} $\underline{S_{13}}$ S_{39} S_{29} S_{34} S_{37} S_2 S_{29}

S_{40} S_{41} S_{31} S_{37} S_{31} $\underline{S_{13}}$ S_{35} S_{42} S_9 S_5 S_9 S_{42}

S_7 S_{41} S_1 S_{43} S_{44} S_{45} S_{46} S_{42} S_{45} S_{47} S_{45} S_{44}

S_{32} S_{44} S_{45} S_{44} S_{48} S_{43} S_{25} S_{45} S_{11} S_{49} $\underline{S_{13}}$ S_{49}

S_{11} S_{49} S_{47} S_{50} S_{47} $\underline{S_{13}}$ S_{26} $\underline{S_{13}}$ S_{44} $\underline{S_{13}}$ S_{45} $\underline{S_{13}}$

S_8 S_9 S_{45} S_{50} S_9 S_{51} S_5 S_{52} S_{32} S_{51} S_5 S_{51}

S_{45} S_9 S_{21} S_2 S_9 S_{21} S_9 S_{39} S_9 S_{43} $\underline{S_{13}}$ S_{53}

S_{39} $\underline{S_{13}}$ S_{43} $\underline{S_{13}}$ S_{49} $\underline{S_{13}}$ S_{47} $\underline{S_{13}}$

This gives a maximally coarse–grained view of the DNA squence, in terms of "domain sets": these are the elements of a given domain type which may be scattered over the entire genome. Examples above are domains like S_1 in bacterium or S_{13} in human which are widely dispersed (these are underlined for visual clarity above), suggesting that these fragments possibly had a common origin, or that they were inserted at the same time during evolution. Expansion–modification [24,25] and insertion–deletion [26] are thought to play

major role in evolution: the former ensures duplication accompanied by point mutations in genomes and the latter results in insertion of a part of chromosome inside a nucleotide sequence or deletion of base pairs from a nucleotide sequence. An initial homogeneous sequence may thus become heterogeneous by insertions/deletions that consistently go on with the evolution. Insertions may cause the pieces of a homogeneous sequence to spread.

B. Insertion–deletion and heterogeneity

The process of insertion–deletion [26] has played an important role in increasing the complexity of genomes. Motivated by the simplification of domain description as above, we perform the following numerical experiment in order to examine the increase in complexity by such processes. Fragments of the *U. urealyticum* bacterial sequence of total length 80 Kbp are inserted at N random positions in the human chromosome 22 contig ($gi|10879979|ref|NT\_011521.1|$). The heterogeneity will naturally increase because of such insertions.

Prior to the insertion of bacterial fragments, the total number of domains in the human chromosome 22 contig is 248; after inserting the fragments at random positions, in a typical realization, the number of segments obtained is 375. The results of such experiments can be quantified through the sequence compositional complexity [18,27], denoted $\mathbf{S}$,

$$\begin{aligned}\mathbf{S} &= H(S) - \sum_{i=1}^{n} \frac{n_i}{N} H(S_i) \\ &= \sum_{i=1}^{n} \frac{n_i}{N} [H(S) - H(S_i)], \end{aligned} \tag{8}$$

where S denotes the whole sequence of length N and S_i is the ith domain of length n_i. This measure, which is independent of the length of sequence quantifies the difference or dispersion among the compositions of the domains. The higher the $\mathbf{S}$, the more heterogeneous the DNA sequence.

When fragments of very different composition are inserted into a given DNA sequence, the complexity will necessarily increase. We compute $\Delta_S = \mathbf{S}' - \mathbf{S}$ for domains obtained

after and before the insertion for the example as above and also for a number of genomes. In all cases $\Delta_S > 0$: the compositional complexity increases after insertion. If deletion is also introduced, say by removing a fragment of random length from a random position (the range of lengths being deleted is kept same as that of the 'inserts') in general Δ_S increases further.

C. Measuring the complexity

We quantify the simplication of domain description of the two representative genomes by considering a complexity measure within the model selection framework, namely the Bayesian information criterion (**B**). Within standard statistical analyisis, one model is superior in comparison with another if it has a lower **B**. For the case of *U. urealyticum* , where the segmentation procedure gives 86 domains,

$$\mathbf{B} = -2\log(\hat{L}) + 343\log(N) \tag{9}$$

where $K = 343$ parameters correspond to 86×3 base compositions and 85 borders. These are reorganized into 17 domain sets, and thus

$$\mathbf{B}' = -2\log(\hat{L}') + 136\log(N) \tag{10}$$

$(136 = 17 \times 3 + 85)$. The maximum likelihood can be expressed as

$$L(p_\alpha) = \prod_\alpha p_\alpha^{N_\alpha}, \tag{11}$$

where $\{p_\alpha\}$ and $\{N_\alpha\}$ are the base composition parameter and the base counts respectively corresponding to alphabet $\{\alpha = A, T, G, C\}$ of a sequence. $\Delta_B = \mathbf{B}' - \mathbf{B}$ depends on the relative contribution of both terms; typically $L > L'$ since the first segmentation uses a more accurate measurement of base composition. The reduction in this measure comes from the second term through the drastic reduction in the number of domains which reduces the model complexity.

For *U. urealyticum* and human, $\Delta_B = -1709$ and -4884 respectively which shows that the model representative of the domain set is better than the original one (we use the lower bound i.e. $\Delta_B < 0$ for determining the statistical significance [7]). As another example, we found Δ_B for *Thermoplasma acidophilum* (archaeabacteria, 1564906 bp) and another contig of human chromosome 22 (*gi* | 10880022 | *ref* | *NT*_011522.1 |, 1528072 bp) to be -2808 and -10420 respectively. We repeated this procedure for different available genomes and found the above results to be consistent. Note that the simplication can also be quantified in terms of $\mathbf{S}$ and we observe $\Delta_S < 0$ in all cases.

IV. *IN SILICO* EXPERIMENTS ON DOMAIN INSERTION: A HOST–PARASITE PERSPECTIVE

It is tempting to speculate that the heterogeneity that is uncovered by the segmentation procedures discussed above is a reflection of the evolutionary history of the given sequence, and in particular, that the different domains arise from insertion processes acting at different evolutionary times. For instance, it is well–known that the human genome contains a small fraction of bacterial genome which have most likely arisen from processes such as viral insertion or lateral gene transfer.

To what extent can the segmentation process determine the exact pattern of insertions? Here we describe some simple experiments that are designed to explore this question. Starting with a homogeneous fragment of human DNA, we insert fragments from (a homogeneous segment of) bacterial genomes; this increases the heterogeneity. We then apply the segmentation algorithm followed by the labeling procedure and compare the results with the (known) control.

Experiments were done on a homogeneous domain set from the human genome, of total length 100139 bp. Into this, fragments from a homogeneous segment of length 17584 bp from the genome of *U. urealyticum* were inserted. In a representative case, we took 3 fragments (of lengths 5000, 7000 and 5584 bp respectively) and inserted them at locations 10000, 50000

and 92000 in the human genome domain.

Upon segmentation, all seven segments were identified, with the boundaries between the bacterial and human DNA sequences determined as follows: 9984 (10000), 15000 (15000), 49751, 50060 (50000), 56968 (57000), 91636 (92000) and 97575 (97584), (the exact values are given in brackets). There is thus one false positive, but otherwise all the boundaries are determined to fairly high precision. The domain sets can also be reconstructed, and the seven segments, $\mathcal{S}_1\mathcal{S}_2\mathcal{S}_1\mathcal{S}_2\mathcal{S}_1\mathcal{S}_2\mathcal{S}_1$ conform to two sets.

Shown in Fig. 1(a) is the insertion process for a case where fragments from two bacterial genomes, *Ureaplasma urealyticum* and *Thermoplasma acidophilum* are randomly inserted in the human genome segment. Carrying out segmentation at varying strength s gives a greater number of segments compared to the correct value of 13. With $s = 0.2$, one gets 18 segments (see Fig. 1(b)) which is the best reconstruction possible within the present framework. On obtaining domain sets, we find that up to about 85% of human and *U. urealyticum* genomes are properly identified, the errors affecting the reconstruction of *T. acidophilum* which is only 67% accurate.

To summarize, our results from several numerical experiments show that the reconstruction of the fragmentation process can be done to high accuracy so long as the inserted fragments are sufficiently long and widely separated.

V. DISCUSSION AND SUMMARY

Segmentation offers a novel view of the compositional heterogeneity of a DNA sequence. In the present work we have applied the segmentation analysis to genomic sequences from several organisms.

Our main focus has been on understanding the organization and to this end we have applied a number of different analytical tools. Our main analysis has been directed towards obtaining a coarse–grained representation of DNA as a string of minimal domain labels. Complexity measures indicate that the reduced model in terms of domain sets is superior

to a model where each domain is treated as independent.

Insofar as the different domains are considered, our main hypothesis is that these arise when fragments of one (possibly homogenous) DNA sequence get randomly inserted into another (also possibly homogenous) sequence. A controlled set of (numerical) experiments give support to this hypothesis: we are able to identify domain boundaries to high accuracy so long as inserted domains are not very short. The accuracy could be further increased by improving the segmentation process, for example, using 1 to 3 segmentation rather than the binary or 1 to 2 segmentation used here: binary segmentation is only one of several possible segmentation procedures (see Ref. [17]).

A consequence of this analysis, and one that we are currently exploring, is that different domains (or domain sets) in one genome can have arisen via insertion from another organism. Homology analysis (say by the use of standard tools such as BLAST or FASTA) can help to unravel the origins of the domains. Thus segmentation analysis can possibly help in reconstructing the evolutionary history of the genome.

ACKNOWLEDGMENT:

RR is supported by a grant from the Department of BioTechnology, India.

REFERENCES

[1] W. Li and K. Kaneko, Europhys. Letts.**17**, 655 (1992).

[2] W. Li, T. G. Marr, and K. Kaneko, Physica D **75**, 392 (1994).

[3] W. Li, Complexity **3**, 33 (1997).

[4] P. Bernaola-Galván, R. Román-Roldán, and J.L. Oliver, Phys. Rev. E **53**, 5181 (1996).

[5] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. L. Oliver and H. E. Stanley, Phys. Rev. E **65**, 041905 (2002).

[6] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).

[7] W. Li, Phys. Rev. Lett. **86**, 5815 (2001).

[8] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 168 (1992).

[9] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J. F. Muzy, and A. Arneodo, Phys. Rev. Lett. **86**, 2471 (2001).

[10] W. Li, Int. J. Bifurcation and Chaos **2**, 137 (1992).

[11] W. Li and K. Kaneko, Nature **360**, 635 (1992).

[12] S. Tiwari, S. Ramachandran, S. Bhattacharya, A. Bhattacharya and R. Ramaswamy, CABIOS, **13**, 263 (1997).

[13] J. Widom, J. Mol. Biol., **259**, 579 (1996).

[14] E. N. Trifonov, Physica A **249**, 511 (1998).

[15] R. K. Azad, P. Bernaola-Galván, R. Ramaswamy, and J. S. Rao, Phys. Rev. E **65**, 051909 (2002).

[16] V. E. Ramensky, V. Ju Markeev, M. A. Roytberg, and V. G. Tumanyan, J. Comp. Biol. **7**, 1 (2000).

[17] J. V. Braun and H. G. Müller, Stat. Sci. **13**, 142 (1998).

[18] R. Román-Roldán, P. Bernaola-Galván, and J. L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).

[19] W. Li, P. Bernaola-Galván, F. Haghighi, and I. Grosse, Comput. Chem., to appear.

[20] J. Lin, IEEE Trans. Inf. Theor. **37**, 145 (1991).

[21] H. Jeffreys, *Theory of Probability*, (Clarendon Press, Oxford, 1961).

[22] A. E. Raftery, in *Sociological Methodology*, edited by P. V. Marsden (Blackwell, Oxford, 1995), pp. 185-195.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* (Springer-Verlag, New York, 2001).

[24] W. Li, Phys. Rev. A **43**, 5240 (1991).

[25] S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, New York, 1970).

[26] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M. H. R. Stanley, and S. Simon, Biophys. J. **65**, 2673 (1993).

[27] P. Bernaola-Galván, J. L. Oliver, and R. Román-Roldán, Phys. Rev. Lett. **83**, 3336 (1999).

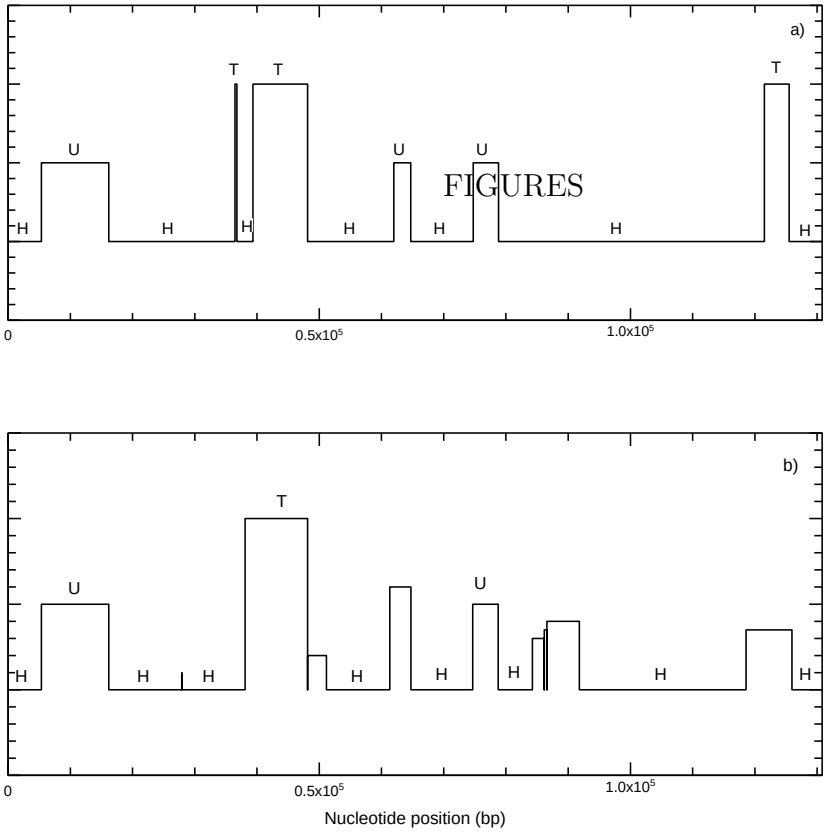

FIG. 1. (a) Representation of a DNA sequence obtained by random insertion of fragments of two bacterial sequences *T. acidophilum* (T) and *U. urealyticum* (U) into a human sequence (H) (see text). (b) The domain structure as uncovered by the procedure of segmentation and labeling (as described in the text).