Compositional segmentation and long-range fractal correlations in DNA sequences

Pedro Bernaola-Galván
Department of Applied Physics II, University of Málaga, Spain

Ramón Román-Roldán
Department of Applied Physics, University of Granada, Spain

José L. Oliver*
Department of Genetics and Institute of Biotechnology, University of Granada, Spain

(Received 28 September 1995)

A segmentation algorithm based on the Jensen-Shannon entropic divergence is used to decompose long-range correlated DNA sequences into statistically significant, compositionally homogeneous patches. By adequately setting the significance level for segmenting the sequence, the underlying power-law distribution of patch lengths can be revealed. Some of the identified DNA domains were uncorrelated, but most of them continued to display long-range correlations even after several steps of recursive segmentation, thus indicating a complex multi-length-scaled structure for the sequence. On the other hand, by separately shuffling each segment, or by randomly rearranging the order in which the different segments occur in the sequence, shuffled sequences preserving the original statistical distribution of patch lengths were generated. Both types of random sequences displayed the same correlation scaling exponents as the original DNA sequence, thus demonstrating that neither the internal structure of patches nor the order in which these are arranged in the sequence is critical; therefore, long-range correlations in nucleotide sequences seem to rely only on the power-law distribution of patch lengths. [S1063-651X(96)05905-3]

PACS number(s): 87.10.+e

I. INTRODUCTION

The finding in 1992 of long-range power-law correlations extending across more that 10^4 nucleotides [1–4] has provoked intense controversy between the stance that these have a far-reaching nature [5–7], given the implied fractal structure, and the contention that long-range correlations can be trivially caused by simple variations in nucleotide composition along DNA sequences [8–10]. It has been argued, however, that the way in which patches are organized, and not the mere existence of patchiness, would be the true source for long-range correlations [11,12]. For the evaluation of such alternative proposals, the length distribution of compositional patches in DNA sequences needs first to be determined unambiguously.

The identification of the different compositional patches or DNA domains in a sequence is an important issue in current computational molecular biology, since it may be one of the key steps in understanding the large-scale structure of the genome. Some preliminary results on the distribution of such patches have been obtained [13], but, to our knowledge, a mathematically rigorous definition of a *patch* is still missing. In simple, not long-range correlated DNA sequences, such as those predominantly integrated by coding regions, compositional patches can simply be identified by eye (see, for example, [9]). However, for complex, long-range correlated DNA sequences, such a method would be useless, given the lack of a characteristic patch size [12], and, therefore, the presence of subsequences covering the entire range of possible lengths. What is needed, therefore, is a statistical approach capable of estimating, with a given confidence level, the location of borders separating adjacent compositional patches in a sequence. Here, we introduce a segmentation algorithm, based on the Jensen-Shannon divergence [14], which can be used to segment DNA sequences with long-range correlations into statistically significant, compositionally homogeneous patches. The shuffling of such segments in a way that preserves the original statistical distribution of patch lengths, raises important questions about the origin and significance of long-range correlations in nucleotide sequences.

II. METHODS

A. Segmentation procedure

Our aim is to divide a sequence into segments in such a way as to maximize the compositional divergence between the resulting DNA domains. We therefore need a method capable of detecting shifts in sequence composition, thus locating the possible borders or edges between adjacent domains. The segmentation procedure is described in three stages: (1) the general strategy of segmentation: sequence splitting; (2) the splitting decision: an entropic measure; and (3) the halt criterion: statistical confidence.

1. Sequence splitting

Most current segmentation methods use a sliding window along the sequence [15]. However, since our aim is to ana-

*Corresponding author. Mailing address: Departamento de Genética, Instituto de Biotecnología, Facultad de Ciencias, E-18071 Granada, Spain. FAX +34 58 24 40 73. Electronic address: oliver@ugr.es

lyze long-range correlations, an overall (not local) view of the whole sequence may better serve a decision about segmentation. On the other hand, the detection of diffuse borders of variable extent betrays an undesirable dependence on the sliding-window size used to scan the sequence [15]. We performed a splitting procedure one at a time, iteratively, which is computationally simple and allows a halt at any point. As a result, any sequence considered for segmentation is scanned by a sliding border, and the location that optimizes an appropriate measure is selected.

2. The entropic decision

Different measures may be used to quantify the distance between probability distributions (variance, Kullback divergence, etc.). For our purposes (discussed below), we choose the Jensen-Shannon divergence measure [14], used recently by some of us [16] for segmenting textured images.

a. Subsequence probability distributions. Let $S=\{a_1,a_2,\ldots,a_N\}$ be a sequence composed of N symbols from the alphabet $\mathcal{A}=\{A_1,\ldots,A_k\}$; take a given position $n(1\leqslant n<N)$ and consider the two resulting subsequences:

$$S^{(1)}=\{a_1,a_2,\ldots,a_n\},\quad S^{(2)}=\{a_{n+1},a_{n+2},\ldots,a_N\}$$

and let

$$\mathcal{F}^{(1)}=\{f_1^{(1)},\ldots,f_k^{(1)}\},\quad \mathcal{F}^{(2)}=\{f_1^{(2)},\ldots,f_k^{(2)}\}$$

be the respective vectors of relative nucleotide frequencies, i.e., $f_i^{(1)}$ is the relative proportion of the symbol A_i in $S^{(1)}$ and $f_i^{(2)}$ is the relative proportion of the symbol A_i in $S^{(2)}$.

For DNA sequences, we can consider the alphabet $\mathcal{A}=\{A,T,C,G\}$, with $k=4$, but also the binary alphabets $\{R(A \text{ or } G), Y(C \text{ or } T)\}$, or $\{S(G \text{ or } C), W(A \text{ or } T)\}$, with $k=2$. Other mapping rules [4] also could be considered.

b. The Jensen-Shannon divergence. The difference between $\mathcal{F}^{(1)}$ and $\mathcal{F}^{(2)}$ is quantified by the Jensen-Shannon divergence measure D_{JS} [14]. For two distributions,

$$D_{\mathrm{JS}}(\mathcal{F}^{(1)},\mathcal{F}^{(2)})=H(\pi_1\mathcal{F}^{(1)}+\pi_2\mathcal{F}^{(2)})-(\pi_1 H\mathcal{F}^{(1)}+\pi_2 H\mathcal{F}^{(2)}),\tag{1}$$

where

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

is Shannon's entropy of the distribution $\mathcal{F}$, and $\pi_1,\pi_2\geqslant 0,\pi_1+\pi_2=1$ are the weights of $\mathcal{F}^{(1)}$ and $\mathcal{F}^{(2)}$, respectively.

From the Jensen inequality, it is easy to prove that $D_{\mathrm{JS}}(\mathcal{F}^{(1)},\mathcal{F}^{(2)})\geqslant 0$, with equality if, and only if, $\mathcal{F}^{(1)}=\mathcal{F}^{(2)}$. Other relevant properties of this measure are as follows: (1) there is symmetry with respect to the arguments; (2) $\mathcal{F}^{(1)}$ and $\mathcal{F}^{(2)}$ are not required to be absolutely continuous; (3) a reachable upper bound exists [17], which is simplified to $\min\log_2$, $\log_2$ (number of distributions)] (in general, not reachable); (4) the ability to be generalized to any number of distributions; and (5) the distributions can be weighted.

c. Application of the Jensen-Shannon divergence for segmenting nucleotide sequences. Besides the above convenient general properties, the D_{JS} measure has three advantages for our particular application. First, the weight property makes it possible to quantify the influence of the two subsequence lengths being compared. These different lengths must contribute differently to the D_{JS} measure; we therefore take $\pi_1=n/N$ and $\pi_2=(N-n)/N$. Second, the low value of k, besides the fact of having only two distributions, prevents the risk of handling D_{JS} values close to the upper bound, which would otherwise lower the discrimination power. Third, a simple expression may be obtained for the significance level of D_{JS}, which makes it possible to halt the splitting process when convenient (see below).

The value of n corresponding to the maximum D_{JS} along a given sequence segment is assumed to separate subsegments with the maximum compositional differentiation between them. Therefore, the compositional homogeneity is higher within each of the two resulting subsegments than in the parental segment. A more detailed description of the algorithm, generalizing it to non-DNA alphabets, will be reported elsewhere [18].

3. Halting the segmentation

A problem with this segmentation method is that, except for equal-symbol strings, at least one position in the sequence always maximizes the difference between the two resulting subsequences. Therefore, the segmentation process will continue until the number of segments virtually reaches the number of nucleotides in the sequence, which would be useless. This problem can be overcome in two ways: (1) by taking a minimum segment length beyond which further division of a given segment is not allowed; and (2) by establishing a minimum significance level for the Jensen-Shannon entropy below which segmentation cannot take place. The first bound can be easily established heuristically, i.e., taking a minimum segment length of 10 nucleotides for which short-range correlations are known to be in effect [4]. However, such a bound leads to biased distributions of lengths.

In establishing the second type of bound, we need to distinguish statistically significant differences between two potential subsegments from purely random fluctuations. Let us consider the original sequence as an outcome of a source consisting in a series of N independent identically distributed (i.i.d.) random variables; then, we can ask for the probability of obtaining a divergence value equal to or higher than the observed one. If such a probability is lower than a given value $1-r$ (with r usually close to 1), it is inferred that the new border establishes a difference between $S^{(1)}$ and $S^{(2)}$ at a significance level r. Choosing r is a critical question —low values would lead to an exhaustive segmentation of the sequence, while high values would result in no segmentation at all. The division of the sequence is continued while significant differences between potential subsegments are found and, therefore, until all significant borders in the sequence are identified.

The statistical distribution of D_{JS} is unknown, being difficult to ascertain analytically (even for binary alphabets). In the Appendix, we present an approximation to this problem, showing that the quantiles of the distribution are independent of both the source probability distribution (thus allowing the

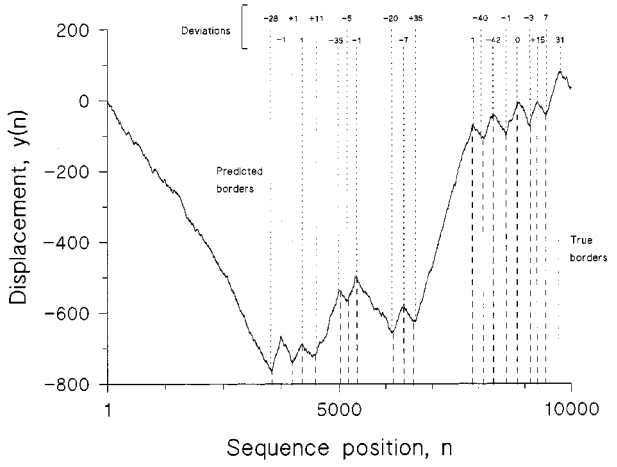

FIG. 1. Segmentation of a computer-generated sequence of length 10 164 bp conforming to a Lévy walk. A generalized Lévy walk was generated as described in detail in the Ref. [13], except that successive strings were taken in alternating directions. The parameters for generating the Lévy walk were set as follows: $\mu=2.45$ (the Lévy-walk parameter), $l_c=150$ (the lower cutoff characteristic length), and $\epsilon=0.3$ (the bias parameter). The borders for different segments are shown superimposed over a random walk of the sequence. True (preestablished) borders in the Lévy walk are indicated by broken lines (below the random walk), and those predicted by the segmentation algorithm at 99.95% significance level by dotted lines (above the random walk). Border prediction deviations (expressed in nucleotides) are indicated by numbers on the top of the respective dotted lines.

comparison of DNA sequences of any base compositions) and the splitting point intended. This conclusion was confirmed numerically through Monte Carlo experiments. A series of pseudorandom sequences with lengths ranging from 50 to 100 000 symbols were obtained by using the RAN3 random number generator [19]. Some experiments have been replicated using the Acarry [20] and RANECU [21] random generators, with similar results. Using $k=2,4$ and different nucleotide compositions, we computed the D_{JS} values for all the possible segmentations.

B. Monitoring long-range correlations

DNA walks were constructed as described in [2]. In brief, a one-dimensional walker dictated by the nucleotide sequence descends one step when there is a purine $[u(i)=-1]$ and rises one step when there is a pyrimidine $[u(i)=1]$. The displacement of the walker after n steps, $y(n)$, is defined as

$$y(n)\equiv\sum_{i=1}^{n} u(i), \tag{3}$$

which is displayed on a graph of y vesus n as in Fig. 1. For monitoring long-range correlations, we used the detrended fluctuation analysis (DFA), as described in [12]. This method enables the detection of long-range correlations embedded in a patchy landscape, also avoiding the spurious detection of apparent long-range correlations that are an artifact of patchiness. DFA involves the following steps: (1) Divide the entire sequence of length N into N/l nonoverlapping boxes, each containing l nucleotides, and define the local trend in each box to be the ordinate of a linear-squares fit for the DNA walk displacement in that box. (2) Define the detrended walk, denoted by $y_l(n)$, as the difference between the original walk $y(n)$ and the local trend. Calculate the variance about the detrended walk for each box, and calculate the average of these variances over all the boxes of size l, denoted $F_d^2(l)$. Recently, this method was slightly modified by using a sliding box [4].

III. RESULTS

Since the long-range correlations found in DNA sequences are mostly based on purine-pyrimidine strand biases [2,12], we used mainly this alphabet throughout this work, except where indicated.

A. Segmenting Lévy-walk sequences

A generalized Lévy walk is an ensemble of many uncorrelated biased random walks that are spliced together, where the length of these biased random walks follows a power-law distribution [13]. Figure 1 shows the results of assaying our segmentation algorithm on a computer-generated sequence conforming to a Lévy walk. The segmentation is shown superimposed over a random walk of the sequence in order to emphasize the power of our algorithm in uncovering the borders for purine-pyrimidine strand-biased regions. The 20 preestablished borders in the Lévy-walk were all identified by segmenting the sequence at the 99.95% significance level. True and predicted border positions showed good agreement, taking into account the statistical nature of the segmentation method used.

B. Segmenting DNA sequences

The segmentation algorithm was then applied to both a human DNA sequence (HUMTCRADCV, GenBank accession number M94081, 97634 bp) and a bacterial one (ECO110K, accession number D10483, 111401 base pair) at different significance levels. It is known that the scaling exponents for these sequences, as determined by DFA, are 0.61 and 0.51, respectively, thus indicating long-range correlations in the first but not in the second sequence [12]. Such a difference may rely on the existence of long-range correlations in non-coding, and its absence from coding DNA [2,22,23], given the differential abundances of both types of regions in these sequences.

The basic statistics for the obtained distributions of segment lengths in both sequences, using three different DNA alphabets, are shown in Table I. For a given significance level, the number of compositional segments was higher and the mean segment length was lower in the human sequence than in the bacterial one. The distributions of segment lengths in both sequences were widespread (standard deviation $\gg$ mean) and always strongly skewed to the right (mean $\gg$ median), the skewness coefficients being clearly higher for the human sequence (Table I). However, when the $\{S,W\}$ alphabet was used, the skewness differences between correlated and noncorrelated sequences were not as apparent (in fact, at 95% significance level, the skewness coefficient

TABLE I. Results of segmenting DNA sequences with (HUMTCRADCV) and without (ECO110K) long-range correlations at different significance levels, and using three different DNA alphabets.

Sequence	Significance level	No. of segments	Mean length	Median	σ	Skewness
		Alphabet $\{R,Y\}$				
HUMTCRADCV	99	189	517	60	1887	7.4
	95	1335	73	10	300	9.9
	90	4945	20	5	82	17.9
ECO110K	99	35	3183	609	5424	2.4
	95	181	615	14	1905	4.9
	90	691	161	6	858	12.2
		Alphabet $\{S,W\}$				
HUMTCRADCV	99	121	807	226	1505	3.0
	95	580	168	24	470	6.4
	90	1663	59	6	226	10.2
ECO110K	99	72	1547	282	3076	2.9
	95	401	278	9	1089	6.9
	90	1294	86	5	532	14.3
		Alphabet $\{A,T,C,G\}$				
HUMTCRADCV	99	181	539	157	1114	3.7
	95	1399	70	19	174	8.0
	90	5306	18	8	51	17.7
ECO110K	99	75	1475	553	2002	1.9
	95	388	287	22	832	5.0
	90	1995	56	7	312	14.1

was slightly higher for the length distribution derived from the bacterial sequence). This result agrees with both the lack of long-range correlations observed in the HUMTCRADCV sequence when the $\{S,W\}$ mapping rule was applied ($\alpha=0.52$, $l=4$ to 100), and previous observations [24] noting that $\{S,W\}$ landscapes did not exhibit a power-law correlation as robust as with the $\{R,Y\}$ alphabet.

Figure 2 shows a double logarithmic plot with the frequency distributions of lengths obtained after segmenting both the human and the bacterial sequences. The best fit to a straight line, indicating the presence of a power-law distribution of patch lengths, was found by segmenting the human sequence at 80% significance level.

C. Recursive segmentation

Another way to demonstrate the structural differences between sequences with and without long-range correlations is to segment the sequences recursively, each time to a deeper level. In such analyses, the longer segment obtained at a given step was again divided using a lower significance level, the proccess being repeated over several steps. A clear difference appears: while in the bacterial sequence compositionally homogeneous, as well as relatively long segments were soon found [Fig. 3(a)], significant segmentations continued to appear at each step in the human sequence [Fig. 3(b)].

As expected, each of the regions resulting from segmenting Lévy-walk sequences Fig. 1 lacked long-range correlations (it should be remembered that the various random walks spliced together to construct the Lévy walk were all uncorrelated). The segments from the bacterial sequence were also uncorrelated, a result consistent with the observation that coding sequences lack long-range correlations [2]. However, most of the regions resulting from segmenting the human sequence at 99% significance level show scaling-

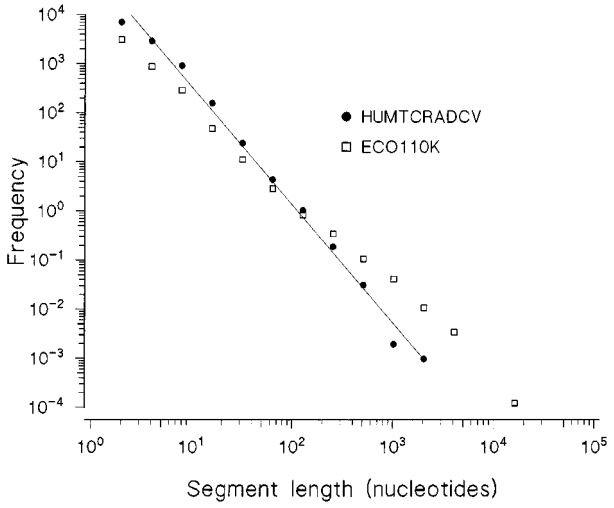

FIG. 2. Double logarithmic plot of frequency distributions of segment lengths in HUMTCRADCV and ECO110K after segmenting at 80% significance level. The binary alphabet $\{R(A \text{ or } G), Y(C \text{ or } T)\}$ was used. The regression line for the length distribution obtained in the human sequence is shown ($r^2=0.99, P=5.81\times10^{-11}$).

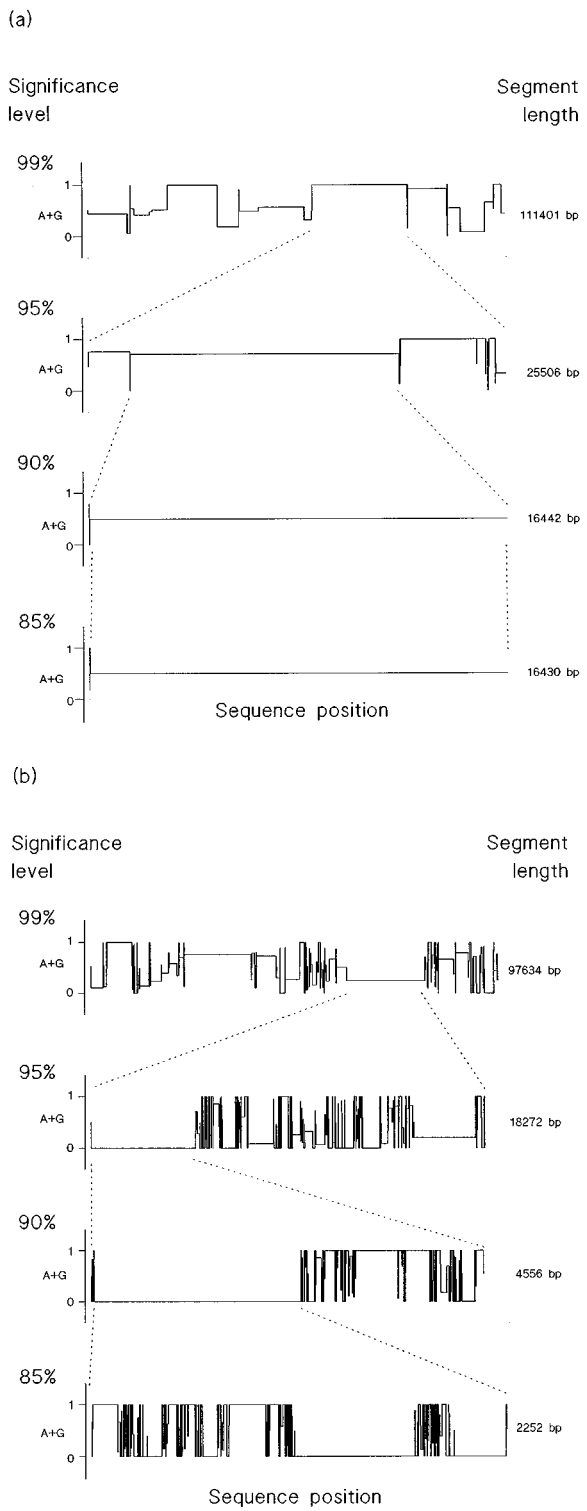

FIG. 3. Recursive segmentation of both the noncorrelated bacterial sequence ECO110K (a) and the long-range correlated human one HUMTCRADCV (b). The longer segment obtained at a given significance step was recursively segmented at a lower significance level. The proportion of purines $(A+G)$ in each segment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

exponent values ≈ 0.6, as occurs in the entire sequence [Fig. 4(a)]. When the longer segment in the previous step (nucleotide positions 59,754-78,025) was again segmented at 95% significance level, all the resulting subsegments also showed scaling exponents ≈ 0.6 [Fig. 4(b)]. This remained true in subsequent steps, until the segment length was short enough (<200 bp) to prevent the safe application of DFA.

Given the lack of long-range correlations in coding DNA [2,22,23], the few uncorrelated segments identified by our algorithm in the human sequence were searched for the presence of genes or exons. None of the 15 exons and 55 possible gene fragments known in this sequence (see sequence annotation) map within the coordinates of the only identified segment with a scaling exponent below 0.50 [nucleotide positions 94,141-94,811, see [Fig. 4(a)]. Instead, what we found there were 21 repeats of the decamer GCCTGTGGAG. In other long-range correlated sequences, we identified some longer uncorrelated segments. This occurs, for example, in the human sequence for the β-globin on chromosome 11 (HUMHBB, accession number J00179, 73326 bp), where two long segments without long-range correlations were identified. Again, both segments included long stretches of repetitive DNA. The first one, from nucleotide position 22 896 to 29 407, included the LINE-1*c* region, and the second one, from nucleotide position 67 089 to 73 213, corresponded to the KpnI family repeat.

D. Shuffling the resulting segments

The higher-level organization proposed for sequences with long-range correlations [5–7] may rely on a specific distribution of compositional patches [11,12]. Therefore, once a sequence is divided into compositionally homogeneous segments, these can be manipulated in different ways in order to deduce some clues about the principles governing long-range correlations. The segments identified in a given DNA sequence were shuffled by two different procedures: (1) by randomly reordering the nucleotides within each individual segment (thus obtaining a shuffled sequence thereafter called type I), and (2) by randomly rearranging the order in which the different segments occur within the sequence (then obtaining a shuffled sequence of type II). While sequences of type I conserve both the original patchiness and its spatial distribution along the DNA sequence (it should be remembered that for a given significance level, the compositional differences within a segment are not statistically significant), shuffled sequences of type II conserve the original patchiness, but not the location of the different patches along the sequence. However, the two shuffling procedures share an important feature: both conserve the estimated distribution of patch lengths in the natural sequence.

Figure 5 shows the detrended fluctuation analyses of HUMTCRADCV and two random sequences derived from it, each obtained by segmenting the original sequence at the 95% significance level, and then separately applying each of the shuffling procedures mentioned above. The scaling exponents for the two shuffled sequences, as determined by DFA, were similar ($\alpha=0.62$ for both sequences) to that observed in the natural one ($\alpha=0.61$), a result confirmed by standard Fourier analysis, as implemented by [1] (not shown). The significance level used for segmenting the DNA sequence was not a critical factor, as figures ranging from 50 to 99% equally led to shuffled sequences with unchanged values for the scaling exponents (not shown). For comparative purposes, we also analyzed random sequences derived from

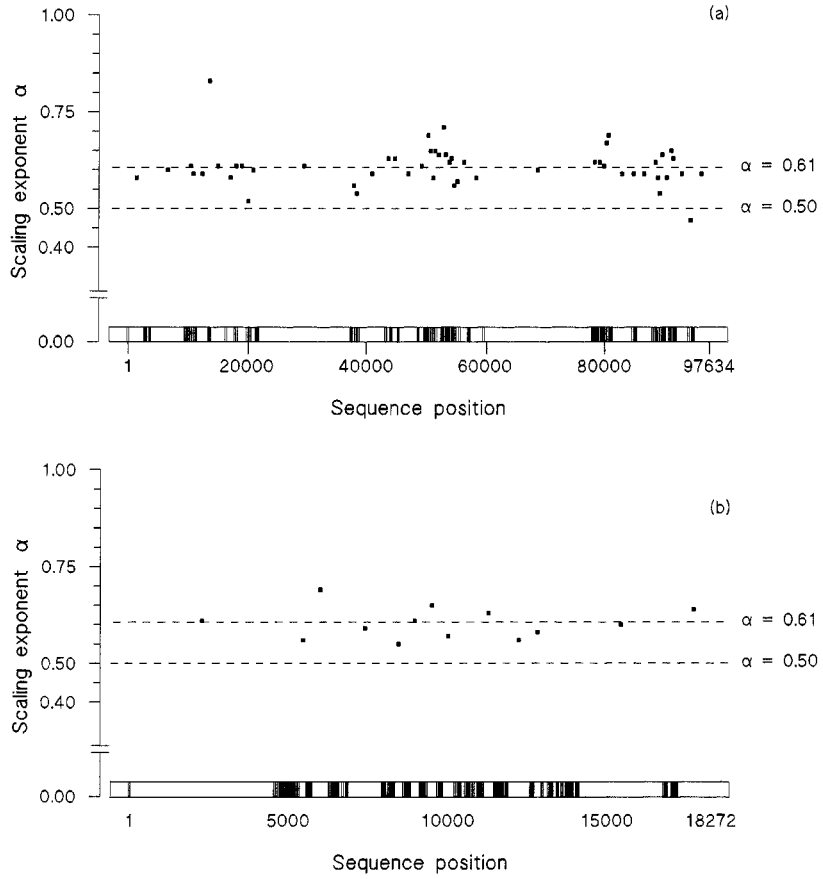

FIG. 4. (a) Scaling-exponent values in segments longer than 200 bp, after segmenting the human sequence HUMTCRADCV at 99% significance level. The corresponding values are plotted at the middle of each segment. (b) The longer resulting domain (18 272 bp) was again segmented at 95% significance level, then computing the scaling-exponent value for each subsegment longer than 200 bp. The segmentation borders are indicated by vertical tick marks on the horizontal bar below the figure. Broken horizontal lines mark scaling exponents of 0.61 (corresponding to the entire sequence) and 0.50 (corresponding to the absence of long-range correlations).

HUMTCRADCV by separately shuffling (or randomly rearranging) equal-length segments of a fixed length (e.g., 500 or 1000 bp). In this third type of random sequences, the original distribution of lengths was destroyed; consequently, the long-range correlations were also lost ($\alpha = 0.52$, Fig. 5).

IV. DISCUSSION

A. The statistical distribution of patch lengths

When our segmentation algorithm was applied to DNA sequences using the $\{R,Y\}$ or $\{A,T,C,G\}$ alphabets, the resulting distributions of lengths were more strongly skewed to the right for long-range correlated sequences than for noncorrelated ones (Table I). Such a difference, clearly revealed in Fig. 2, is probably due to the power-law distribution of lengths proposed for long-range correlated sequences [11,12]. This figure also shows another difference in the distribution of lengths displayed by sequences with and without long-range correlations —the shorter segments were more frequent in the human than in the bacterial sequence, while the contrary occurred for the longer patches.

B. Compositional scale invariance

Our segmentation algorithm provides an additional way to look for structural differences between sequences with and without long-range correlations; Fig. 3 shows that, after recursive segmentation, only the human sequence displayed similar distributions of lengths at different segmentation steps. This means that compositional variations in long-range correlated sequences show scale invariance, a feature typical of fractal structures. The distributions of lengths shown by the human sequence at different segmentation levels are in fact reminiscent of fractal sets with statistical self-similarity [25].

While some authors [8,9] have claimed that long-range correlations can be fully accounted for by compositional differences between different subregions of DNA sequences, others [13] assimilated long-range correlated DNA sequences to generalized Lévy walks. A prediction shared by both proposals is that a sequence could be decomposed into statistically stationary subregions with no internal structure. This was the result we obtained when segmenting either Lévy walks (Fig. 1) or the bacterial uncorrelated sequence [Fig. 3(a)]. However, when we analyzed the human sequence the results looked very different—except for domains including repetitive DNA, most of the segments obtained at a given significance level continued to show long-range correlations [Fig. 4(a)]. When correlated segments were recursively divided, the resulting subsegments again showed long-range correlations [Fig. 4(b)], until a point was reached where the segment length was so short that DFA could not be safely applied.

It has been claimed that compositional heterogeneity may be present at all scales in all classes of DNA sequences [8]. With the help of a tool such as the segmentation algorithm presented here, we could attempt to distinguish statistically significant heterogeneities in DNA nucleotide composition from purely random fluctuations. Our results revealed that statistically significant, and probably biologically meaningful, compositional heterogeneities at all scales appear only in

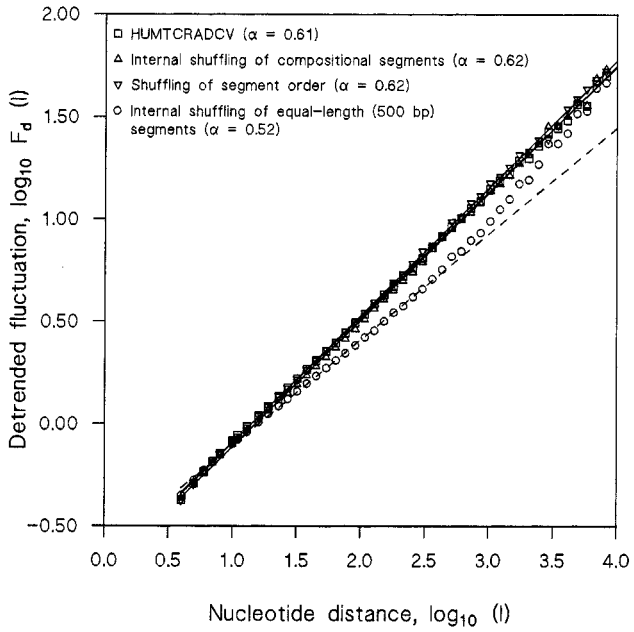

FIG. 5. Detrended fluctuation analysis of HUMTCRADCV and two shuffled sequences derived from it after segmenting at 95% significance level, and then separately shuffling the resulting segments in the two ways described in the text. The solid lines, the best fits from $l = 4$ to 8192, have similar slopes ($\alpha \approx 0.6$) in the three sequences. The analysis of a random sequence obtained by separately shuffling equal-length segments of 500 bp is also shown for comparison; the broken line is the best fit from $l = 4$ to 362 ($\alpha = 0.52$). The binary alphabet $\{R, Y\}$ was used for the analysis of all four sequences.

long-range correlated DNA sequences [Fig. 3(b)], and not in the noncorrelated ones [Fig. 3(a)]. In this way, in agreement with previous observations [11,12], only long-range correlated sequences exhibit a complex, statistically significant, multi-length-scaled structure.

C. The organization of compositional patches

Type I and type II shuffled sequences displayed scaling exponents similar to that of the DNA sequence from which they were derived (Fig. 5). This means that the segmentation algorithm described here allows us to decompose a complex, long-range correlated sequence into segments in such a way that, when shuffled separately or when rearranged at random, the original correlation structure is retained. However, when the same shuffling procedures were applied to segments of a fixed length (say equal-length segments of 500 or 1000 bp), the scaling exponent dropped to ≈ 0.5.

These results first of all demonstrate that our segmentation algorithm estimated reasonably well the distribution of lengths present in long-range correlated sequences—once identified, the segments could be reordered at random without altering the correlation scaling exponent of the entire sequence, at least as measured by DFA. The reason the shuffled sequence retained long-range correlations might be, therefore, that the length distribution in the original sequence was not destroyed by the shuffling procedure we used. Secondly, our results also prove that the internal structure of patches is irrelevant to the scaling exponent computed by DFA—each segment can be independently shuffled without changing the value of the scaling exponent for the entire sequence. Again, the length distribution in the original sequence was not altered by this second shuffling procedure. Thus, neither the internal structure of patches nor the order in which they were arranged in the sequence seems to be relevant for the obtained scaling exponent value. We conclude, therefore, that the long-range correlated structure in a sequence is mainly dependent on a specific distribution of patch lengths. Some authors [11,12] have stressed that the way in which compositional patches are organized determines long-range correlations. According to the results presented here, such organization seems to reduce to a power-law distribution of patch lengths. This is not to say that in biological sequences the order of the different patches or their internal structure would be unimportant, but only that for the particular measure of long-range correlations we are using—the correlation scaling exponent derived from DFA—both of these features appear to be not critical.

D. The source for long-range correlations

Searching for the source of long-range correlations in DNA sequences is equivalent, therefore, to seeking the evolutionary mechanisms behind a power-law distribution of patch lengths. Two models have been proposed to account for the generation of long-range correlated sequences: The expansion modification system [26] and the insertion-deletion model [24]. Using the premises of these models, we generated long-range correlated sequences, which were then segmented by means of our algorithm. In both instances, all the resulting segments also showed long-range correlations; the recursive segmentation of these segments led to shorter subsegments that were again correlated (not shown). Therefore, the sequences generated by both models displayed the characteristic multi-length-scaled structure seen in DNA sequences.

Li's model is based on duplication with modification (mutation), two well-documented processes in the evolution of DNA sequences [11,27]. The second model attributed the origin of long-range correlations to repeated cycles of deletions and/or insertions of sequence segments, probably corresponding to retroviral insertions, partial gene duplications, or transpositions. The recent demonstration that the size distribution of deletions and/or insertions follows a power-law distribution [28] supports such a model. Given the higher rate at which natural selection accepts all these mutations in noncoding DNA, as compared to coding genome regions [29], the presence of long-range correlations in the first but not in the second ones [2] can be readily explained [11,24].

ACKNOWLEDGMENTS

We are most grateful to Wentian Li for kindly providing access to unpublished work and for earlier encouragement on this subject, and to Chun-Kang Peng for providing the DFA source code. The help with the manuscript of David Nesbitt is also appreciated. This work was partially supported by Grants No. TIC94-535 to R.R.R. and No. PB93-1152-CO2-01 to J.L.O. from the DGICYT of the Spanish Government.

APPENDIX: AN ANALYTICAL APPROXIMATION TO THE PROBABILITY DISTRIBUTION OF D_{JS}

The probability distribution of D_{JS} is unknown. As a first approximation, we estimated the characteristic value ΔD_{JS} in segmenting a sequence of N i.i.d. random variables $a_i \in \{A_1, A_2, \ldots, A_k\}$ with a probability distribution $\mathscr{P}$.

Let $\mathscr{F}$ be the relative frequency vector corresponding to a certain sequence; generally, $H(\mathscr{F}) \leqslant H(\mathscr{P})$ [30]. Therefore, the sample entropy systematically underestimates the source entropy. In this first approximation, such an error can be expressed as [31]

$$\Delta H(N,k) = H(\mathscr{P}) - H(\mathscr{F}) \approx \frac{k}{2N\ln 2}. \tag{A1}$$

For the two subsequences $S^{(1)}$ and $S^{(2)}$, and the whole sequence S,

$$H(\mathscr{F}) \approx H(\mathscr{P}) - \Delta H(N,k), \tag{A2}$$

$$H(\mathscr{F}^{(1)}) \approx H(\mathscr{P}) - \Delta H(n,k), \tag{A3}$$

$$H(\mathscr{F}^{(2)}) \approx H(\mathscr{P}) - \Delta H(N-n,k), \tag{A4}$$

where $\mathscr{F}^{(1)}$ and $\mathscr{F}^{(2)}$ are the relative vector frequencies of the two resulting subsequences when segmenting at position n. Since the original sequence corresponds to a series of i.i.d., the same source entropy is considered in all three cases.

Taking into account that the weights are $\pi_1 = n/N$ and $\pi_2 = (N-n)/N$, and by replacing in Eq. (1), we get

$$\Delta D_{JS}(N,K) \approx \Delta H(N,k) \approx \frac{k}{2N\ln 2}. \tag{A5}$$

We assume that, if ΔD_{JS} is independent of n and $\mathscr{P}$, then the distribution of D_{JS} is too, at least in a first approximation. Therefore, we expect that

TABLE II. Significance levels (r) and d_r values obtained numerically from Monte Carlo experiments.

Significance level (r)	d_r	Standard error
99	2.41	0.0030
98	1.97	0.0010
97	1.72	0.0008
96	1.54	0.0006
95	1.41	0.0005
90	0.99	0.0002
85	0.76	0.0001

$$P\{D_{JS} \leqslant x\} = f'\left(\frac{x}{D_{JS}}\right) = f\left(\frac{Nx}{k}\right) \tag{A6}$$

and, therefore, the quantiles of the distribution may be expressed as

$$D_r(N,k) = \frac{k}{N} d_r, \tag{A7}$$

where $D_r(N,k)$ is the value of D_{JS} for which lower or equal divergences occur with a probability r, d_r being a factor depending only on r.

The Monte Carlo experiments mentioned in the main text allowed us to estimate d_r for different significance levels (Table II). Note that D_r is independent of both $\mathscr{P}$ (thus allowing the comparison of DNA sequences of whatever base compositions) and n (regardless of the splitting point intended).

[1] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[2] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, Nature **356**, 168 (1992).
[3] R. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[4] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E **51**, 5084 (1995).
[5] J. Amato, Science **257**, 747 (1992).
[6] J. Maddox, Nature **358**, 103 (1992).
[7] P. Yam, Sci. Am. **267**, 23 (1992).
[8] S. Karlin and V. Brendel, Science **259**, 677 (1993).
[9] D. Larhammar and C.A. Chatzidimitriou-Dreissman, Nucl. Acids Res. **21**, 5167 (1993).
[10] C.A. Chatzidimitriou-Dreismann, R.M.F. Streffer, and D. Larhammar, Biochim. Biophys. Acta **1217**, 181 (1994).
[11] W. Li, T.G. Marr, and K. Kaneko, Physica D **75**, 392 (1994).
[12] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, and A.L. Goldberger, Phys. Rev. E **49**, 1685 (1994).
[13] S.V. Buldirev, A.L. Goldberger, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E **47**, 4514 (1993).
[14] J. Lin, IEEE Trans. Inf. Theor. **37**, 145 (1991).
[15] W.K. Pratt, *Digital Image Processing*, 2nd ed. (Wiley, New York, 1991).
[16] V. Barranco-López, P. Luque-Escamilla, J. Martínez-Aroza, and R. Román-Roldán, Electron. Lett. **31**, 867 (1995).
[17] A. Robles-Pérez, H. Ben-Haniza, J. Martínez-Aroza, and R. Román-Roldán (unpublished).
[18] P. Bernaola-Galván, J.L. Oliver, and R. Román-Roldán (unpublished).
[19] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes*, 2nd ed. (Cambridge University Press, New York, 1992).
[20] F. James, Comput. Phys. Commun. **60**, 329 (1990).
[21] P. L'Ecuyer, Commun. ACM **31**, 742 (1988).
[22] S.M. Ossadnik, S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, C.-K. Peng, M. Simons, and H.E. Stanley, Biophys. J. **67**, 64 (1994).

[23] A. Arneodo, E. Bacry, P.V. Graves, and J.F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).
[24] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, M.H.R. Stanley, and M. Simons, Biophys. J. **65**, 2673 (1993).
[25] M. Schroeder, *Fractals, Chaos, Power Laws. Minutes from an Infinite Paradise* (Freeman, New York, 1991).
[26] W. Li, Phys. Rev. A **43**, 5240 (1991).
[27] S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, New York, 1970).
[28] X. Gu and W.-H. Li, J. Mol. Evol. **40**, 464 (1995).
[29] Nei. M, *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).
[30] It can be rigorously proven that $E(H(\mathscr{F}))\leqslant H(\mathscr{P})$ [see, for example, M. Mansuripur, *Introduction to Information Theory*, (Prentice-Hall, Englewood Cliffs, NJ, 1987)].
[31] H. Herzel, A.O. Schmitt, and W. Ebeling, Chaos Solitons Fractals **4**, 97 (1994).