Sequence Compositional Complexity of DNA through an Entropic Segmentation Method

Ramón Román-Roldán,[1] Pedro Bernaola-Galván,[2] and José L. Oliver[3,*]

[1]*Department of Applied Physics, University of Granada, Spain*
[2]*Department of Applied Physics II, University of Málaga, Spain*
[3]*Department of Genetics and Institute of Biotechnology, University of Granada, Spain*

(Received 12 May 1997)

A new complexity measure, based on the entropic segmentation of DNA sequences into compositionally homogeneous domains, is proposed. Sequence compositional complexity (SCC) deals directly with the complex heterogeneity in nonstationary DNA sequences. The plot of SCC as a function of significance level provides a profile of sequence structure at different length scales. SCC is found to be higher in sequences with long-range correlation than those without, and higher in noncoding sequences than coding sequences. Furthermore, a general agreement is found between the SCC of the DNA sequence, on one hand, and the biological complexity of the organism, on the other, attributable to an increasingly complex organization of noncoding DNA over the course of evolution. [S0031-9007(97)05210-1]

PACS numbers: 87.10.+e

The analysis of sequence correlation structure, in both the spatial and the frequency domains, resulted in the finding of short-range [1] and long-range [2] correlations in nucleotide sequences, thus uncovering a complex fractal structure for DNA.

Sequence structure can be adequately revealed through segmentation algorithms. One of these, conceptually simple and computationally efficient, was proposed last year by our group [3]. With such a method, a DNA sequence can be decomposed into homogeneous subsequences (patches or domains). By varying an appropriate threshold, we obtain different partitions of the sequence at different statistical significance levels. When segmenting sequences with simple domain structures, homogeneous domains can be consistently found (if purely random fluctuations are excluded). However, when the method is applied to more complex, long-range correlated sequences, such homogeneity vanishes: by relaxing the threshold value, we find new domains within other domains, previously taken as homogeneous under a higher threshold value. This domains-within-domains phenomenon points to complex compositional heterogeneity in DNA sequences, which is consistent with the hierarchical nature of biological complexity [4].

One implication of this complex heterogeneity is that conventional methods to estimate the complexity of DNA sequences (i.e., Shannon-Gatlin measures or DNA spectra), while adequate mostly for stationary stochastic processes, in this case can be problematic [5], since they are able to distinguish sequences of different complexities only on an indirect basis. Consequently, a measure directly dealing with complex heterogeneity is needed. Such a measure would be useful in modeling DNA sequences [6], as well as in molecular evolutionary studies or comparative genomics. A good definition of complexity must be objective and mathematically tractable, yet consistent with the intuitive notion of what the complexity is about [7]. Neither the algorithmic complexity [8] (maximum for randomness) nor other derived measures [9] nor those based on mutual information (statistical dependence) [10] are completely satisfactory. Instead, the number of domains and their compositional heterogeneity should be internally accounted for—i.e., without any reference to environments or ensembles.

Compositional domains.—These are defined as subsequences with a different base composition in comparison to the two adjacent subsequences, at a given level of statistical confidence. The domains so defined are neighborhood dependent. Some measure of the difference between compositions is needed to compare adjacent subsequences and decide whether they are different domains. For this task, we use the Jensen-Shannon divergence measure [11]. For two subsequences S_1 and S_2 of lengths l_1 and l_2, $\mathrm{JS}_2(S_1, S_2) = \mathrm{H}[S] - (\frac{l_1}{L}\,\mathrm{H}[S_1] + \frac{l_2}{L}\,\mathrm{H}[S_2]) \geq 0$, where $L = l_1 + l_2$, $S = S_1 \oplus S_2$ (concatenation) and $\mathrm{H}[\cdot] = -\sum p \log_2 p$ is the Shannon's entropy of the probability distribution $\{p\}$ obtained from base frequencies in the corresponding subsequence. Given that most relevant differences in correlation structure have been found with the $\{R(A \text{ or } G),\ Y(C \text{ or } T)\}$ mapping rule, we used this binary alphabet throughout this study, except where indicated. Similar results can be obtained with the quaternary alphabet $\{A, T, C, G\}$, while the $\{S(C \text{ or } G),\ W(A \text{ or } T)\}$ mapping rule allows for less marked differences between the DNA sequences analyzed here.

Statistical confidence is established by calculating the probability that the given divergence value (or lower) appears in a random sequence (with the same length and base composition), once it has been randomly split. For short subsequences, the probability is exactly computed from the hypergeometric distribution; for large ones, an approximation has been derived in the appendix of Ref. [3].

Given a significance level s, a sequence can be partitioned in many different ways. To define and select the best partition, the following optimization criterion is established based on the partition divergence $\mathrm{JS}_n(s)$, which is an extension of JS_2 to the n domains obtained:

$$\mathrm{JS}_n(s) = \mathrm{H}[S] - \sum_{i=1}^{n} \frac{l_i}{L}\mathrm{H}[S_i] = \sum_{i=1}^{n} \frac{l_i}{L}(\mathrm{H}[S] - \mathrm{H}[S_i]) . \qquad (1)$$

This equation is applicable to any partition of a set of objects, but it makes sense in this context only if applied to a sequence segmented in domains. We call a partition complete (optimum) with divergence $\mathrm{JS}_n^*(s)$, if there is no other with higher divergence, at the same s. In particular, it follows that for a complete partition, no domain can be further divided into two new domains; otherwise, $\mathrm{JS}_n^*(s)$ would increase, in which case the preceding partition would not be complete.

A heuristic segmentation method.—Here, we describe how to segment a sequence. Computing $\mathrm{JS}_n^*(s)$ requires us to have the sequence completely segmented. This implies that all possible partitions of the sequence must be checked, and thus an *NP*-complete problem must be solved.

To circumvent the extreme complexity of this problem, we shall design an appropriate heuristic procedure that would render a good approximation to complete segmentation. Our published method [3] performs the task by iteratively splitting the sequence, $S \rightarrow S_1 \oplus S_2$ with $\max\{\mathrm{JS}_2(S_1, S_2)\}$, once the new cut satisfies the significance condition, while maintaining the location of preceding cuts. The procedure ends when no further legal splits can be made. We now slightly improve the above procedure, in which every new intended split can lead to the loss of significance at previous adjacent cuts. Now, we verify that the statistical significance of these cuts is preserved before accepting the new split. This refinement ensures that all the cuts remain significant throughout the process and in the final partition. The $\mathrm{JS}_n(s)$ value thus obtained is a lower bound of $\mathrm{JS}_n^*(s)$.

A measure of sequence compositional complexity (SCC).—The $\mathrm{JS}_n(s)$ measure fits the requirements for representing the compositional complexity of the sequence. In fact, the last expression in Eq. (1) is the weighted sum of histogram entropy deviations of the domains from the average, thus being a measure of both the number and compositional biases of domains. Moreover, this information-theoretic divergence has the following three major properties: (1) It is defined within the same framework that the criteria for both segmenting the sequence and optimizing the partition, and therefore the SCC is the maximum reachable divergence—i.e., that corresponding to the optimum partition. (2) It is not dependent on the total length L of the sequence being analyzed. (3) It may be used for nonstationary sequences—a major advantage over many other statistics that are currently, even controversially [5], used.

We now explore the dependency of $\mathrm{JS}_n(s)$ on s. For all sequences, $\mathrm{JS}_n^*(s)$ increases monotonically as s decreases, since all previous cuts in a complete partition remain significant for any $s' < s$, therefore $n' \geq n$. This behavior holds for the heuristic segmentation, as can easily be seen by deriving an alternative expression for $\mathrm{JS}_n(s)$ given in terms of the splitting process. This progressive expression, which can be proven by induction and by using the branching property of the entropy function, is

$$\mathrm{JS}_n(s) = \sum_{k=1}^{n-1} \frac{l_k}{L}\mathrm{JS}_2(S_{k_1}, S_{k_2}) , \qquad (2)$$

where the sum extends over all the splits produced along the heuristic segmentation and l_k is the length of the subsequence to be split (not the length of the k domain) into S_{k_1} and S_{k_2}; note that $\sum l_k \neq L$. Although in Eq. (2) $\mathrm{JS}_n(s)$ seems to be dependent explicitly on the path followed by the splitting algorithm, Eq. (1) clearly shows that $\mathrm{JS}_n(s)$ is path invariant—i.e., it depends only on the final partition. JS_n as a function of s is more informative than a single JS_n value. The representation of a good sample of points constitutes the $\mathrm{JS}_n(s)$ profile. The profiles shown here are always below and close to those corresponding to the theoretical, complete segmentation.

Any profile is limited to the rectangle $0\% \leq s \leq 100\%$, $0 \leq \mathrm{JS}_n(s) \leq H[S]$. The profile of a sequence depends on the splitting rate as s decreases. Two basic patterns are (1) a step function (a burst of domains is produced at a critical s; e.g., in a periodically biased sequence), and (2) a plateau function (no splits are made within a given Δs interval; e.g., in a compositionally homogeneous sequence). Any profile can be described as a certain combination of steps and plateaus, either exact or approximate. On this basis, an interpretation concerning sequence complexity versus s can then be derived: a constant rate of variation for $\mathrm{JS}_n(s)$ seems to reveal a self-similar-like structure.

Profiles for pure random sequences begin to depart appreciably from zero for low s values ($<80\%$), meaning that the profiles are being contaminated by spurious complexity, due to statistical fluctuations. Therefore, we choose the range $80\% < s < 100\%$ for all sequence comparisons.

Sequence analysis.—We first analyzed the profiles of several artificially generated sequences [Fig. 1(a)] with different degrees of *a priori* complexity. As expected, $\mathrm{JS}_n(s)$ increased with the intended complexity built into the sequence. There were clear differences between the profiles—a marked improvement on previous methods, which were unable to distinguish between many of these sequences [12].

Second, we apply the complexity measure to real DNA sequences with different correlation structure. Figure 1(b) shows that the values obtained for *HUMTCRADCV*, a human DNA sequence whose long-range correlations [12] and complex heterogeneity [3] are both well known, are higher than those for the uncorrelated bacterial sequence *ECO*110*K*, regardless of the s value considered. The corresponding two shuffled sequences exhibit the lowest profiles. The profiles of two larger sequences (the human sequence *HSTCRB*18 and the complete genome of

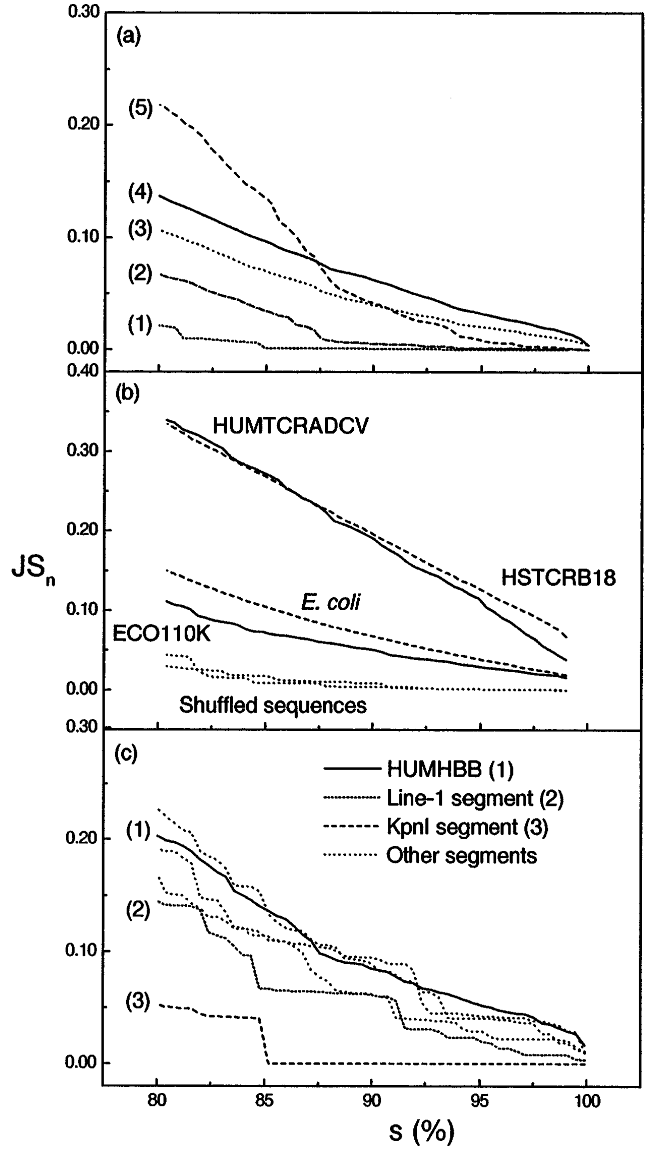

FIG. 1. (a) Complexity profiles of computer-generated sequences with increasing degrees of built-in complexity: (1) A pseudochromosome simulating mutual-information properties of yeast chromosomes but without taking into account compositional heterogeneity [13]; (2) a first-order Markov chain with the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

E. coli) confirms that long-range fractal correlations are associated with higher levels of SCC [Fig. 1(b)].

Third, we performed a quantification of the "domains-within-domains" phenomenon (see Fig. 3 of Ref. [3]). Most of the segments from a sequence with long-range base-base correlation obtained at a given s value showed complexity profiles similar to the entire sequence when segmented at decreasing s [Fig. 1(c)]. The only exceptions were some uncorrelated segments, such as those containing simple repetitive DNA sequences, which showed lower complexity profiles [Fig. 1(c)].

Fourth, we focused on correlation-structure differences between coding and noncoding sequences [2,13]. Complexity profiles enabled a clear distinction between these two types of DNA: Noncoding regions consistently showed a higher SCC than did the coding ones, irrespective of the proportion of noncoding segments in the sequence (Fig. 2). This was true for both eukaryotic introns [Fig. 2(a)] and prokaryotic intergenic regions [Fig. 2(b)]. This greater SCC for noncoding segments agrees with the observations that intron sequences show long-range correlations as robust as those of entire genes [14]. Complexity differences between coding and noncoding regions are also apparent when the quaternary $\{A, T, C, G\}$ alphabet is used (Fig. 3). Several explanations can be given for the observed higher SCC of noncoding DNA: (1) Given the evolutionary conservatism of some intergenic sequences [15], noncoding ("junk") DNA could perform some important function in the cell, the exact nature of which is as yet undetermined; (2) some type of hidden language could be present in noncoding

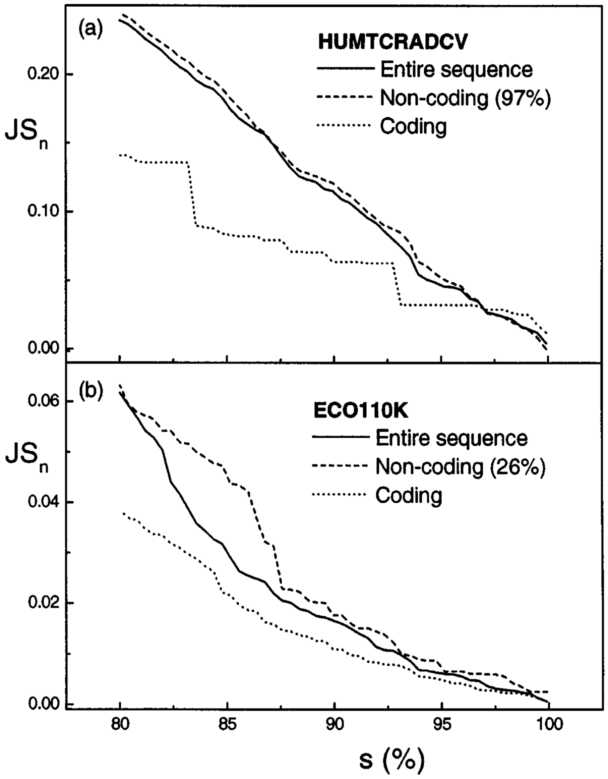

FIG. 2. Differences in SCC between coding and noncoding regions of the sequences (a) *HUMTCRADCV* (97 634 bp) and (b) *ECO*110*K* (111 401 bp). The percentages of noncoding DNA at each sequence are given in parentheses.

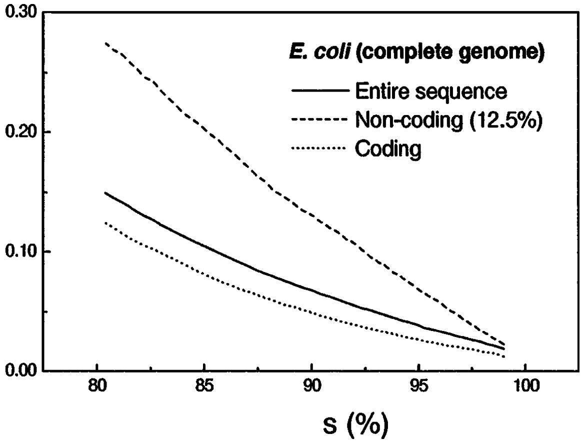

FIG. 3. Differences in SCC between coding and noncoding regions of the *E. coli* complete genome (4 638 858 bp). The quaternary $\{A, T, C, G\}$ alphabet was used.

sequences, although previous evidence for this [16] has been disputed [17]; (3) a simpler alternative explanation would be that noncoding regions have a higher evolutionary tolerance to mutations as duplications and insertions, thus developing a less random, more redundant structure [6]. Nevertheless, the mere presence of DNA repeats does not lead to higher values for SCC [Fig. 1(c)].

Finally, we address the comparison of SCC over evolution. It has been shown that the scaling exponents of the random-walk representation of homologous DNA sequences increases with evolution [14]. We determined the SCC of the four gene families analyzed by Buldyrev *et al.* Myosin profiles generally agreed with biological complexity (Fig. 4), and the same was true for lysozyme genes (not shown). However, some exceptions to this rule, mostly in short sequences, were found for actins and cytochrome *C* (not shown). The observed increase in SCC with evolution cannot be exclusively attributed to increasing percentages of introns in the corresponding sequences (compare the profiles and the percentages of introns for human, rat, and chicken in Fig. 4); therefore, an increasingly complex organization of noncoding regions over evolution may also play a role in this process.

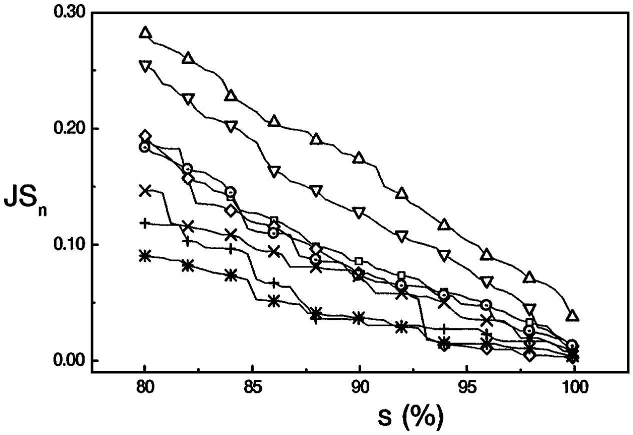

FIG. 4. Complexity profiles of myosyn heavy-chain genes in different species (total length, percentage of introns): (△) Human (28 438 bp, 74%), (▽) rat (25 759 bp, 77%), (○) chicken (31 111 bp, 74%), (◇) *Caenorhabditis* (10 780 bp, 14%), (⊙) *Brugia* (11 766 bp, 32%), (+) yeast (6108 bp, 0%), (×) *Acanthamoeba*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

This work was supported by Grants No. TIC94-535 (CICYT) and No. PB93-1152-C02-01 (DGICYT) from the Spanish Government. We acknowledge the critical reading and helpful comments of Wentian Li.

*Corresponding author: Dpto. de Genetica, Fac. de Ciencias, Universidad de Granada, E-18071-Granada, Spain.
Electronic address: oliver@ugr.es

[1] L. Gatlin, *Information Theory and the Living System,* (Columbia University Press, New York, 1972); J.L. Oliver *et al.*, J. Theor. Biol. **160**, 457 (1993); H. Herzel, W. Ebeling, and A.O. Schmitt, Phys. Rev. E **50**, 5061 (1994); R. Román-Roldán, P. Bernaola-Galván, and J.L. Oliver, Pattern Recognit. **29**, 1187 (1996).

[2] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992); C-K. Peng *et al.*, Nature (London) **356**, 168 (1992); R. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

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

[4] P. Schuster, Complexity **2**, 22 (1996).

[5] S. Karlin and V. Brendel, Science **259**, 677 (1993).

[6] W. Li, Comput. Chem. **21**, 257 (1997).

[7] C.H. Benneth, in *Complexity, Entropy and the Physics of Information,* edited by W.H. Zurek (Addison-Wesley Press, Cambridge, MA, 1990).

[8] G.J. Chaitin, *Algorithmic Information Theory* (Cambridge University Press, New York, 1987); M. Li and P.M.B. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications* (Springer-Verlag, New York, 1997), 2nd ed.

[9] M. Gell-Mann and S. Lloyd, Complexity **2**, 44 (1996).

[10] P. Grassberger, Int. J. Theor. Phys. **25**, 907 (1986); C. Adami and N.J. Cerf, Physica D (Amsterdam) (to be published).

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

[12] C-K. Peng *et al.*, Phys. Rev. E **49**, 1685 (1994).

[13] H. Herzel and I. Große, Phys. Rev. E **55**, 800 (1997).

[14] S.V. Buldyrev *et al.*, Biophys. J. **65**, 2673 (1993).

[15] D. Yaffe *et al.*, Nucl. Acids Res. **13**, 3723 (1985); B.F. Koop *et al.*, Genomics **13**, 1209 (1992); L. Duret, F. Dorkeld, and Ch. Gautier, Nucl. Acids Res. **21**, 2315 (1993).

[16] R.N. Mantegna *et al.*, Phys. Rev. Lett. **73**, 3169 (1994); R.N. Mantegna *et al.*, Phys. Rev. E **52**, 2939 (1995); R.N. Mantegna *et al.*, Phys. Rev. Lett. **76**, 1979 (1996).

[17] N.E. Israeloff, M. Kagalenko, and K. Chan, Phys. Rev. Lett. **76**, 1976 (1996); S. Bonhoeffer *et al.*, Phys. Rev. Lett. **76**, 1977 (1996); R.F. Voss, Phys. Rev. Lett. **76**, 1978 (1996); C.A. Chatzidimitriou-Driesmann, R.M.F. Streffer, and D. Larhammar, Nucl. Acids Res. **24**, 1676 (1996); W. Li, Complexity **1**, 6 (1996); C. Martindale and A.K. Konopka, Comput. Chem. **20**, 35 (1996).

[18] S.V. Buldyrev *et al.*, Phys. Rev. E **47**, 4514 (1993).

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