Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method

Pedro Bernaola-Galván,[1,2,*] Ivo Grosse,[1] Pedro Carpena,[2,3] José L. Oliver,[4]
Ramón Román-Roldán,[5] and H. Eugene Stanley[1]

[1]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*
[2]*Departamento de Física Aplicada II, Universidad de Málaga, E-29071, Spain*
[3]*Theoretical Physics, Oxford University, 1 Keble Road, OX1 3NP Oxford, England*
[4]*Departamento de Genética e Instituto de Biotecnología, Universidad de Granada, E-18071, Spain*
[5]*Departamento de Física Aplicada, Universidad de Granada, E-18071, Spain*

(Received 2 August 1999)

We present a new computational approach to finding borders between coding and noncoding DNA. This approach has two features: (i) DNA sequences are described by a 12-letter alphabet that captures the differential base composition at each codon position, and (ii) the search for the borders is carried out by means of an entropic segmentation method which uses only the general statistical properties of coding DNA. We find that this method is highly accurate in finding borders between coding and noncoding regions and requires no "prior training" on known data sets. Our results appear to be more accurate than those obtained with moving windows in the discrimination of coding from noncoding DNA.

PACS numbers: 87.15.Cc, 87.10.+e, 87.14.Gg

The entropic segmentation process partitions a heterogeneous DNA sequence into homogeneous subsequences, which we term compositional domains [1–3]. Although all the alphabets or mapping rules conventionally used in describing DNA sequences had been tried, we had not been able so far to assign any biological function to the obtained domains. Here we introduce a new alphabet that takes into account the differential base composition at each codon position, and we find that the compositional domains correlate to either coding or noncoding DNA regions. This finding suggests the possibility of using the entropic segmentation process for computational gene finding.

The computational recognition of genes is one of the challenges in the analysis of newly sequenced genomes, which is fundamental for modern functional genomics (the goal of which is the search for the different functional elements which make up the DNA sequences [4]).

The predictive power of current computational methods, although adequate in genomes for which a significant fraction of genes are previously known [5], remains rather low when faced with the large anonymous sequences now being generated by genome projects. An example is provided by the DNA sequence of human chromosome 22, where only 20 per cent of annotated genes have all exons predicted exactly [6]. A critical limitation of most current methods is the need of "prior training," given the lack of experimentally annotated DNA sequences in most of the recently sequenced genomes. Hence, new strategies for computational gene finding that do not require prior training on organism-specific data sets are needed, and the entropic segmentation process may serve as a first step in this direction.

Current methods of gene finding use a variety of biological information as potential sequence signals involved in gene specification or sequence similarity database searches. These signals are mainly used to obtain probable borders of coding regions, and they should be obtained from elaborated biological information, which is highly dependent on the particular genome considered.

Here we offer an entropic segmentation method to detect borders between coding and noncoding DNA. Although previous works based on the Shannon's entropy deal with the problem of finding patterns in DNA or protein sequences [7], their approaches are of local character, instead of the global segmentation we address here. Our method uses only the known statistical general properties of coding DNA. In this way, the prior training on known data sets is avoided; furthermore, the search for additional biological information (such as splice sites or termination signals) can also be avoided; however, such additional information could be easily incorporated and exploited in a concrete implementation of the algorithm.

One of the most relevant and well-known statistical features of coding regions is nonuniform codon usage [8]. This means that, inside coding regions, not all triplets of nucleotides (called codons) occur with the same probability. In particular, the probability of appearance of a nucleotide is different in each of the three positions of the triplets [9–11]. This may be due to the restrictions imposed by the genetic code and also probably to some kind of preferences in the synonymous codon usage; but no matter what its origin is, this feature is not present in noncoding DNA, so this property can be used to distinguish between coding and noncoding DNA. In fact, based on these differences, the first generation of gene prediction programs, designed to identify approximate locations of coding regions in genomic DNA, were developed [10].

To take into account this statistical property of coding DNA, we develop a segmentation algorithm based on a 12-symbol alphabet. We define the phase of position,

 0031-9007/00/85(6)/1342(4)$15.00

i, of a nucleotide to the number $j = i \bmod 3$, where $j \in \{0, 1, 2\}$. So, each of the nucleotides of the DNA sequences can be substituted by one of the following symbols: $\mathcal{A}_{12} = \{A_0, A_1, A_2, T_0, T_1, T_2, C_0, C_1, C_2, G_0, G_1, G_2\}$, where, for example, T_2 means that we have found a nucleotide T with phase $= 2$.

Our aim is to divide a DNA sequence into segments in such a way as to maximize the difference in composition between them, and where the composition is measured by a 12-dimensional frequency vector based on this 12-symbol alphabet. We hope these segments will correspond to alternating coding and noncoding regions. Our method improves on that of Ref. [1] in several fashions. This method has been used to define a measure of DNA sequence compositional complexity [2,12] and, recently, to determine statistically the mobility edge of one-dimensional disordered materials [13].

To compute the difference in composition between two regions of DNA, in order to decide whether they are different domains or not, we use what we call "contrast function": a comparative function which reaches low values when the DNA regions being compared are both coding or noncoding and high values in other case (coding-noncoding or vice versa).

As a contrast function we use the Jensen-Shannon measure [14]. Consider a DNA sequence composed of symbols belonging to $\mathcal{A}_{12}$, and define the 12-symbol frequency vector $\mathcal{F} \equiv \{f_{\ell,j}\}$, where $\ell \in \{A, T, C, G\}$ and $j \in \{0, 1, 2\}$, where $f_{\ell,j}$ is the relative number of nucleotides of type ℓ with phase j. Given two sequences of lengths n_1 and n_2 with frequency vectors $\mathcal{F}_1$ and $\mathcal{F}_2$, the Jensen-Shannon divergence is defined as

$$C(\mathcal{F}_1, \mathcal{F}_2) = 2\ln 2[N\mathrm{H}(\mathcal{F}) - n_1\mathrm{H}(\mathcal{F}_1) - n_2\mathrm{H}(\mathcal{F}_2)], \tag{1}$$

where $N = n_1 + n_2$, $\mathcal{F} = (n_1/N)\mathcal{F}_1 + (n_2/N)\mathcal{F}_2$ is the frequency vector of the entire sequence obtained concatenating both subsequences, and $\mathrm{H}(\mathcal{F})$ is the Shannon entropy, defined by $\mathrm{H}(\mathcal{F}) = -\sum_{\ell,i} f_{\ell,j} \log_2 f_{\ell,j}$. Among other interesting properties [1], $C(\mathcal{F}_1, \mathcal{F}_2)$ is almost not affected by the different sizes of the sequences being compared.

To test the ability of C to separate coding from noncoding DNA, we do the following control experiments. First we take a known coding DNA sequence and a known noncoding one, concatenate them, and go along the resulting sequence with a moving pointer, computing C for the subsequence to the left and the subsequence to the right of the pointer. The results are shown in Fig. 1(a) (solid line). Note that the maximum is clearly obtained in the boundary between both regions (vertical dashed line). We also test the effect of inserting one and two nucleotides between the original sequences. This does not affect the global composition, but changes the phase of the nucleotides of the second sequence, and hence the resulting frequency vector $\mathcal{F}$ of the right hand side sequence can be considerably different from the original. As can be seen, the maximum value of C is again obtained in the boundary between the two subsequences and the values are very close to the ones obtained without shift. This is due to the fact that in noncoding DNA all three phases are almost indistinguishable.

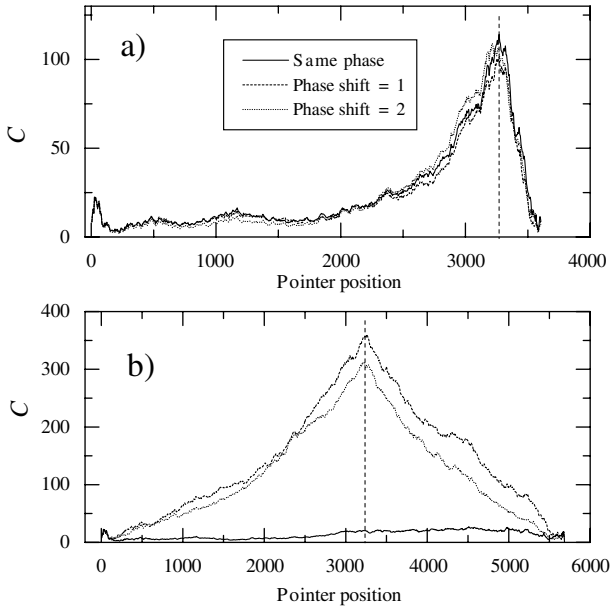

FIG. 1. (a) C vs cutting position for a sequence obtained by joining a coding region (gene *carB* of bacterium *E. coli,* 3222 bp long) and a noncoding region (intergenic region between genes *leuO* and *ilvI* of *E. coli,* 389 bp long); the dashed vertical line is the border between both regions. (b) C vs cutting position for a sequence obtained by joining two coding regions: genes *carB* (3222 bp) and *polB* (2463 bp) of *E. coli*. The dashed vertical line is the boundary between the two regions

Sometimes, especially in prokaryotic genomes, the coding regions are separated by a very small noncoding region, too small to be separately identified on a statistical basis. C is able to distinguish such coding regions, provided they are in different phases, even if they are in consecutive subregions. The only drawback is that if the regions are in the same phase they would be identified as only one coding region, but they could be easily separated by other methods. To show this, we analyze in Fig. 1(b) two coding DNA regions, following the same as in Fig. 1(a). The solid line (the two regions are in phase) reaches very low values and does not seem to present a maximum in the boundary of both regions (vertical dashed line). On the other hand, when we introduce a phase shift (dashed and dotted line), we obtain very high values of C and the maximum is clearly reached in the vicinity of the boundary.

To partition a natural DNA sequence which, in general, will be composed of several coding and noncoding regions, we search for the partition that maximizes the compositional difference between segments, as measured by C. If the number of such regions is large, the problem presents

high complexity, so we can use a heuristic algorithm [2] which works as follows. We move a sliding pointer along the sequence which divides at each position the sequence into two subsequences and we compute C, and select the point at which C reaches its maximum value ($C_{\max}$) and compute its statistical significance (see below). If this significance exceeds a given threshold s, then the sequence is cut at this point. Otherwise the sequence remains undivided. The procedure continues recursively for each of the two resulting subsequences created by each cut. Before a new cut is accepted, we check that the subsequences formed by the cut remain significantly different from their neighbors. The process stops when none of the possible cutting points has a significance level exceeding s. We say that such a sequence is segmented at the "s significance level."

The significance level $s_{\max}(x)$ of a possible cutting point with $C_{\max} = x$ is defined as the probability of obtaining this value or lower within a random sequence: i.e., $s_{\max}(x) = \text{Prob}\{C_{\max} \leq x\}$. As $s_{\max}(x)$ does not seem to admit an easy analytical expression, we have obtained an approximation.

It is known that C has an asymptotic chi-square distribution when the null hypothesis is true. So, given any fixed cutting point (not necessarily the point where the maximum is achieved), the probability of obtaining a given value $C \leq x$ within a random sequence verifies:

$$\text{Prob}\{C \leq x\} \approx F_9(x)\,, \tag{2}$$

where F_ν is the chi-square distribution function with ν degrees of freedom. Here $\nu = 9$ because, although we use a 12-symbol alphabet, there are three constraints: $\sum_\ell f_{\ell,j} = 1/3$ for $j = 0, 1, 2$—i.e., the number of nucleotides in each phase is $1/3$ of the total. If all values of C in all possible cutting points along the sequence were independent,

$$s_{\max}(x) = [F_9(x)]^N, \tag{3}$$

which is not the case because the value of C at a given cutting point is strongly affected by the preceding values—the frequency vectors are minimally affected by the change of only one nucleotide. We have observed that the empirical distribution obtained for $C_{\max}$ by means of Monte Carlo simulations [15], seems to be very similar to (3) but replacing N by an effective length, N_{eff}, and introducing a factor $\alpha < 1$ multiplying x,

$$s_{\max}(x) = [F_9(\alpha x)]^{N_{\text{eff}}}. \tag{4}$$

N_{eff} can be understood as the effective number of independent cutting points and the factor α deals with the fact that the limits of variation of $C_{\max}$ are also reduced due to the correlations. The fitting of the empirical distributions obtained by means of Monte Carlo simulations to this model gives $N_{\text{eff}} = 2.45 \ln N - 9.87$ and $\alpha = 0.84$ independent of N [16].

In Fig. 2 we show the results of the segmentation of a region of the genome of the bacterium *Rickettsia prowazekii*. The shaded areas correspond to the coding regions obtained from annotations (GenBank acc. AJ235269 [17]), and the vertical dotted lines are the positions of the cuts produced by the segmentation algorithm. Note the good agreement between cuts and known coding region borders. Note also that, as can be inferred from Fig. 1(b), the algorithm does not detect the border between two very close coding regions in the same phase (marked with an arrow in Fig. 2).

In order to quantify the coincidence between cuts obtained using the segmentation algorithm and known borders between coding and noncoding regions, we introduce the following quantity:

$$D \equiv \frac{1}{2}\left[\sum_i \frac{\min_j |b_i - c_j|}{N_T} + \sum_j \frac{\min_i |b_i - c_j|}{N_T}\right], \tag{5}$$

where $\{b_i\}$ is the set of all borders between coding and noncoding regions, and $\{c_j\}$ is the set of all cuts produced by the segmentation, and N_T the total length of the sequence. The first summation measures the discrepancy between cuts and borders by adding for each real border the distance to the closest cut. The second summation performs the same operation, but now including for each cut the distance to the closest real border. Both summations are required to take into account not only the correctness in the position of the cuts (D would be zero just when cuts and borders coincide), but also the difference between the number of borders and cuts. D can be viewed as an average of the error in the determination of the correct boundaries between coding and noncoding regions, so $(1 - D)$ is a reasonable measure of the accuracy of the method.

Figure 3 plots $100(1 - D)$ for the segmentations of three bacterial complete genomes at several significance levels. The accuracy of the method is reasonably good

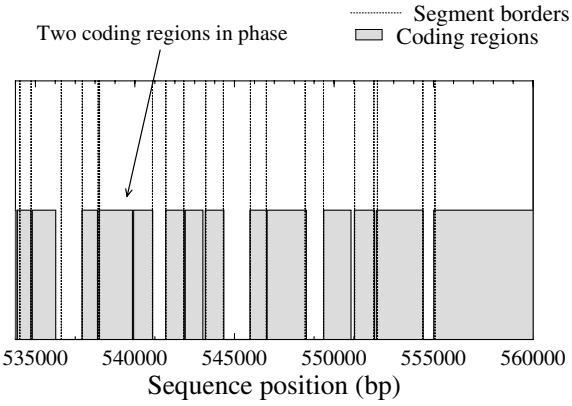

FIG. 2. Comparison between the known coding regions of *Rickettsia* (shaded areas) and the cuts obtained at significance level $s = 99\%$ (dotted lines).

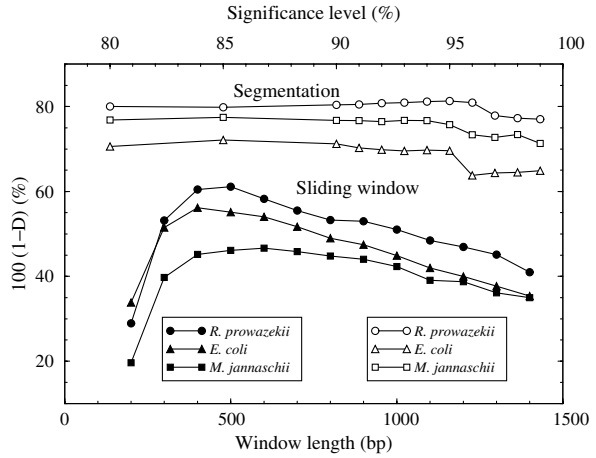

FIG. 3. Comparison of the accuracy of segmentation (open symbols) and sliding window (closed symbols) approaches in finding borders between coding and noncoding regions for three complete bacterial genomes: *Rickettsia prowazekii* (○), *Escherichia coli* (△), and *Methanococcus jannaschii* (□); we find the best results when the training of the windows is carried out using the same sequence as the one analyzed.

(between 70%–80%), especially since the method cannot separate adjacent phase-coding regions (see Fig. 2).

For the sake of comparison with other methods, we also include results obtained for the same bacteria with a sliding window, which moves along the sequence and, at each position, some discriminant function is evaluated [18]. The central nucleotide of the window is considered to be coding when the value of the discriminant function is above a certain threshold, and noncoding when it is below. The positions where the discriminant function equals the threshold are proposed to be borders between coding and noncoding regions. The main problem with this method is the determination of the threshold: the only way to obtain it is to perform a *prior training,* i.e., to analyze a sequence for which coding and noncoding regions are known and to choose the value which maximizes the number of matches for each window size.

In Fig. 3 we also include the values of $100(1 - D)$ obtained with the sliding window approach. These values are always below those obtained using the segmentation algorithm. One advantage of our method is that the segmentation algorithm is not very sensitive to a change of significance level. In fact, any segmentation with a significance level within the range 90%–95% (the usual range) gives similar results. On the other hand, the choice of the window size seems to be critical, and the optimal values are different for each bacterium.

The work of P. B. G., P. C., J. L. O., and R. R. R. is partially supported by Grants No. PR98-0025095592 and No. BIO99-0651-CO2-01 from the Spanish Government, and I. G. and H. E. S. are supported by the NIH.

*Electronic address: rick@uma.es

[1] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).
[2] R. Román-Roldán, P. Bernaola-Galván, and J. L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).
[3] J. L. Oliver, R. Román-Roldán, J. Pérez, and P. Bernaola-Galván, Bioinformatics **15**, 974 (1999).
[4] V. A. McKusick, Genomics **45**, 244 (1997); Special Genome Issue [Science **282** (1998)].
[5] M. Burset and R. Guigó, Genomics **34**, 353 (1996).
[6] I. Dunham *et al.,* Nature (London) **402**, 489 (1999).
[7] G. D. Stormo and G. W. Hartzell, Proc. Natl. Acad. Sci. U.S.A. **86**, 1183 (1989); C. E. Lawrence *et al.,* Science **262**, 208 (1993).
[8] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier, Nucleic Acids Res. **9**, R43 (1981).
[9] J. C. W. Shepherd, Proc. Natl. Acad. Sci. U.S.A. **78**, 1596 (1981).
[10] R. Staden and A. D. McLachlan, Nucleic Acid Res. **10**, 141 (1982); J. W. Fickett, Nucleic Acid Res. **10**, 5303 (1982).
[11] H. Herzel and I. Grosse, Physica (Amsterdam) **216A**, 518 (1995).
[12] P. Bernaola-Galván, P. Carpena, R. Román-Roldán, and J. L. Oliver, Comput. Phys. Commun. **121–122**, 136 (1999); W. Li, G. Stolovitzky, P. Bernaola-Galván, and J. L. Oliver, Genome Research **8**, 916 (1998).
[13] P. Carpena and P. Bernaola-Galván, Phys. Rev. B **60**, 201 (1999).
[14] J. Lin, IEEE Trans. Inf. Theory **37**, 145 (1991).
[15] To obtain empirically the distribution followed by $C_{\max}$ we generate, for each sequence length, an ensemble of 100 000 random sequences, and for each of them we compute $C_{\max}$ following the procedure described in the text.
[16] Note that, instead of using F_9 to fit $s_{\max}(x)$, we could have used a normal distribution which is known to be very close to F_ν for large enough values of ν. Nevertheless, we have observed that the approach used here is valid for all values of ν ranging from 1 to 9. The coefficients in the logarithmic fit of N_{eff} as a function of N depend on the particular value of ν, but the value $\alpha = 0.84$ remains unchanged.
[17] S. G. E. Anderson *et al.,* Nature (London) **396**, 133 (1998).
[18] A discriminant function is any function that can be computed on a region of a DNA sequence which reaches different values for coding and noncoding DNA. In particular, we use here $\sum_{\ell,j} |f_\ell - 3f_{\ell,j}|$, where $f_{\ell,j}$ is the relative number of nucleotides of type ℓ with phase j, and f_ℓ is the relative number of nucleotides of type ℓ in any of the three phases. See R. Staden, Nucl. Acid Res. **21**, 551 (1984).