Computer Physics Communications 121–122 (1999) 136–138

Computer Physics Communications

www.elsevier.nl/locate/cpc

Compositional complexity of DNA sequence models

P. Bernaola-Galván [a,1], P. Carpena [a], R. Román-Roldán [b], J.L. Oliver [c]

[a] *Departamento de Física Aplicada II, Universidad de Málaga, Málaga, Spain*
[b] *Departamento de Física Aplicada, Universidad de Granada, Granada, Spain*
[c] *Departamento de Genética e Instituto de Biocomputación, Universidad de Granada, Granada, Spain*

Abstract

Recently, we proposed a new measure of complexity for symbolic sequences (Sequence Compositional Complexity, SCC) based on the entropic segmentation of a sequence into compositionally homogeneous domains. Such segmentation is carried out by means of a conceptually simple, computationally efficient heuristic algorithm. SCC is now applied to the sequences generated by several stochastic models which describe the statistical properties of DNA, in particular the observed long-range fractal correlations. This approach allows us to test the capability of the different models in describing the complex compositional heterogeneity found in DNA sequences. Moreover, SCC detects clear differences where conventional standard methods fail.

1. Introduction

DNA sequences are formed by patches or domains of different nucleotide composition; given the huge spatial heterogeneity of most genomes, the identification of compositional patches or domains in a sequence is a critical step in understanding large-scale genome structure. Moreover, in sequences from higher organisms, these domains are organized in very complex structures (with fractal properties in many cases), and therefore domains need to be defined on a statistical basis.

2. Sequence compositional complexity

To obtain the partition of a given sequence into domains we proposed a segmentation method, based on the Jensen–Shannon entropic divergence (JS_m) [1]. We search for the partition that maximizes JS_m, defined as:

$$\mathrm{JS}_m = H[S] - \sum_{i=1}^{m} \frac{l_i}{L} H[S_i], \tag{1}$$

where $H[S]$ is the Shannon entropy of the sequence of length L, and $H[S_i]$ is the Shannon entropy of ith segment of length l_i. As the segmentation is carried out by means of a statistical criterion, a significance level (s) must be established, so the final result depends critically on this parameter. If s is close to 100% a small number of domains is obtained, but with a very significant difference between them; on the contrary, if s is lower the number of domains increases but the difference between them is less significative. In other words: for high values of s only the big scale details of the sequences are revealed, meanwhile by lowering s the small scale structure of the sequence emerges.

Since searching for the partition that maximizes (1) requires the solution of a NP-complete prob-

[1] E-mail: rick@ctima.uma.es.

0010-4655/99/$ – see front matter
PII: S0010-4655(99)00298-2

lem, we introduced a computationally efficient heuristic algorithm which implements such a segmentation method [1]. Once a sequence is partitioned, in order to measure its complexity we define the sequence compositional complexity (SCC) as the JS_m value obtained in the maximization procedure [1]. This measure accounts for both the number and compositional differences between the domains. The plot of SCC as a function of s (*complexity profile*) provides a view of the sequence structure at different scales.

3. DNA models

In recent years, several models to explain and describe the fractal properties and long-range correlations of DNA sequences have been proposed (see [2, 3] for reviews). An important question to be addressed is whether these models deal properly with the complex heterogeneity present in natural DNA sequences. Since SCC has revealed as a useful magnitude in DNA sequence analysis [1,4,5], we are going to compute SCC for several artificial sequences generated with these models, comparing the results with those obtained in DNA.

3.1. First order Markov chains

This model generates artificial DNA sequences by using the transition matrix observed in natural sequences. In Fig. 1(a) we show the complexity profile of a human sequence (HUMTCRAD) and a bacterial one (ECO110K), and the corresponding artificial sequences obtained from their transition probabilities. This model only produces short-range correlations and the resulting sequence is stationary. The plot indicates that this model does not provide the complexity observed in natural sequences.

3.2. Mutation-duplication model

This kind of models [6] generate binary sequences of increasing order by using an iterative substitutional rule (probabilities in brackets):

$$0_t \to (00)_{t+1}[1-p], \quad 0_t \to (1)_{t+1}[p],$$
$$1_t \to (11)_{t+1}[1-p], \quad 1_t \to (0)_{t+1}[p].$$

This model produces long-range fractal correlations, as in natural sequences, but, as can be seen in Fig. 1(b), it only provides adequate values of complexity for low values of significance level. The proposed substitutional rule may mimic several processes that lead to repeated DNA. In fact, the complexity profiles obtained with this model are very similar to the ones obtained with repeated DNA [7].

3.3. Insertion-deletion model

The starting point of the model [8] is a biased random binary sequence of length L (thus imitating natural fragments of coding DNA). Then the sequence

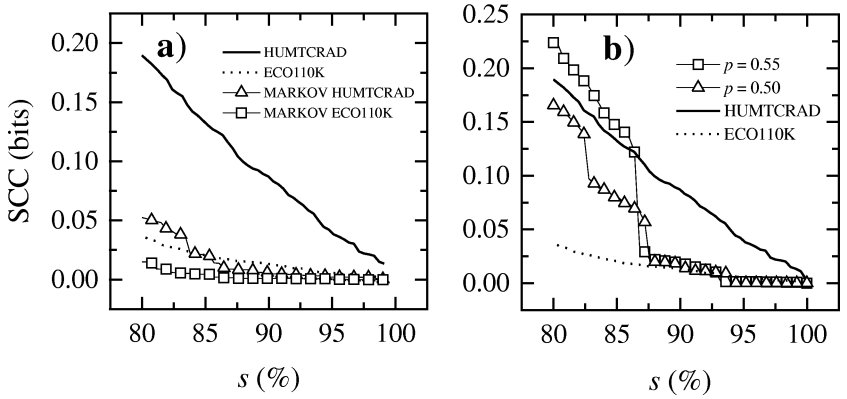

Fig. 1. (a) Complexity profiles of natural DNA sequences as compared to artificial ones generated by a Markov model. (b) Complexity profiles of natural DNA sequences as compared to artificial ones generated by the mutation-duplication model.

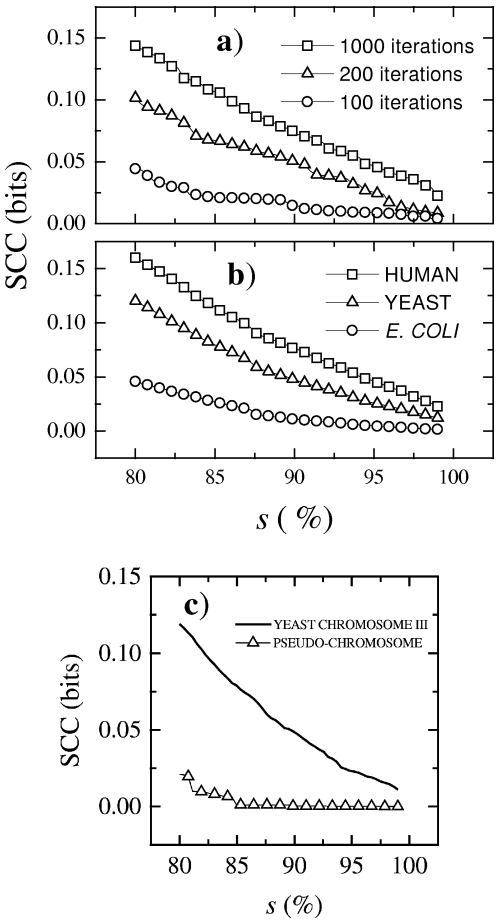

Fig. 2. (a) Evolution of the SCC with the number of iterations for the insertion-deletion model. (b) SCC for natural DNA sequences from organisms with different degrees of biological complexity. (c) Complexity profiles of the yeast natural sequence as compared to the artificial one obtained from the pseudo-chromosome model.

evolves following certain rules which mimic the insertion/deletion processes provoked by retrovirus. Successive iterations increase the sequence complexity. In Fig. 2(a) we show the complexity profiles of sequences obtained for three different numbers of iterations (we use the values of the parameters suggested by the authors). The increase in complexity with the process seems to imitate the increase in complexity with evolution (Fig. 2(b)).

3.4. Pseudo-chromosomes

This model proposes that the complex structure observed in DNA can be explained for several DNA sequences (v.g. in the yeast genome) in terms of the correlations introduced by nonuniform codon usage in coding regions. The model deals properly with the correlations present in the sequence, as measured by the mutual information [9]. Nevertheless, the complexity profiles of the sequences generated with this model are very different to those corresponding to natural sequences, as shown in Fig. 2(c).

Despite its extreme simplicity, the models of genome dynamics reviewed above – incorporating point mutations, tandem duplications and insertion/deletion mechanisms – are able to generate sequences with self-similar long-range correlation and $1/f$ power spectra. SCC profiles show, however, that neither one of these models lead to sequences with the complex heterogeneity characterizing DNA sequences [1]. The most successful was the insertion/deletion model, which generates sequences of similar complexity to DNA sequences, at least for high s values. We conclude that new models, perhaps incorporating genome-wide mechanisms, such as polyploidy – and the subsequent diploidization process – or inter-chromosomal exchanges [5], are required to embrace all the heterogeneity built into the genome.

References

[1] P. Bernaola-Galván, R. Román-Roldán, J.L. Oliver, Phys. Rev. E 53 (1996) 5181;
R. Román-Roldán, P.Bernaola-Galván, J.L. Oliver, Phys. Rev. Lett. 80 (1998) 1344.
[2] W. Li, Comput. Chem. 21 (1997) 257.
[3] H.E. Stanley et al., Physica A 205 (1994) 214.
[4] W. Li, Complexity 3 (1997) 33.
[5] W. Li, G. Stolovitzky, P. Bernaola-Galván, J.L. Oliver, Genome Res. 8 (1998) 916.
[6] W. Li, Int. J. Bifurcation Chaos 2 (1992) 137.
[7] P. Bernaola-Galván, PhD. Dissertation (1997).
[8] S.V. Buldyrev et al., Biophys. J. 65 (1993) 2673.
[9] H. Herzel, I. Große, Phys. Rev. E 55 (1997) 800.