BIOINFORMATICS

Vol. 15 no. 12 1999
Pages 974–979

SEGMENT: identifying compositional domains in DNA sequences

José L. Oliver[1,*], Ramón Román-Roldán[2], Javier Pérez[3] and Pedro Bernaola-Galván[4,5]

[1]Department of Genetics and Institute of Biotechnology, Faculty of Sciences, University of Granada, E-18071-Granada, Spain, [2]Department of Applied Physics, University of Granada, Spain, [3]Institute of Biotechnology, University of Granada, E-18071-Granada, Spain and [4]Department of Applied Physics II, University of Málaga, Spain

Received on March 30, 1999; revised on June 18, 1999; accepted on August 4, 1999

Abstract

Motivation: *DNA sequences are formed by patches or domains of different nucleotide composition. In a few simple sequences, domains can simply be identified by eye; however, most DNA sequences show a complex compositional heterogeneity (fractal structure), which cannot be properly detected by current methods. Recently, a computationally efficient segmentation method to analyse such nonstationary sequence structures, based on the Jensen–Shannon entropic divergence, has been described. Specific algorithms implementing this method are now needed.*

Results: *Here we describe a heuristic segmentation algorithm for DNA sequences, which was implemented on a Windows program (SEGMENT). The program divides a DNA sequence into compositionally homogeneous domains by iterating a local optimization procedure at a given statistical significance. Once a sequence is partitioned into domains, a global measure of sequence compositional complexity (SCC), accounting for both the sizes and compositional biases of all the domains in the sequence, is derived. SEGMENT computes SCC as a function of the significance level, which provides a multiscale view of sequence complexity.*

Availability: *SEGMENT is available on the Web at http://www.ugr.es/local/oliver/segment/*

Contact: *oliver@ugr.es*

Introduction

With the steadily increasing lengths for DNA-sequence contigs and complete genome sequences in nucleotide databases, the age of large-scale analysis of genome structure is coming. Given the pervasive spatial heterogeneity in base composition found in most genomes (for a review see Bernardi, 1995), the identification of compositional patches or domains in a sequence is a prerequisite to understand genome structure.

Compositional variability in DNA sequences is currently revealed through moving windows. The window length and the step used thus become critical parameters in revealing sequence structure. Unfortunately, there are no criteria to choose appropriate values for these two parameters. Whether a larger window is more or less statistically significant than a shorter one cannot be decided, and the same is true for the step. Consequently, both parameters are fixed arbitrarily, which introduces an unavoidable load of subjectivity in the analysis. On the contrary, compositional domains, identified through an entropic segmentation method developed by our group (Bernaola-Galván *et al.*, 1996), are defined on a statistical basis. With such a method, a DNA sequence can be decomposed into homogeneous subsequences (patches or domains) at a given significance level. Such domains may advantageously replace moving windows in computing compositional statistics along nonstationary DNA sequences. A heuristic algorithm implementing this segmentation method is described here. By varying an appropriate threshold, the partition of the sequence at different confidence levels can be obtained. In addition, once a sequence is partitioned into domains, a global measure of sequence compositional complexity (SCC) can be derived; our program computes SCC at different significance levels, thus providing a multiscale view of sequence structure (Román-Roldán *et al.*, 1998).

Another segmentation algorithm has been previously described both for proteins and for DNA (Wootton and Federhen, 1993, 1996, respectively). It includes a log-likelihood definition of local complexity analogous to informational entropy. The main purpose is the location

*To whom correspondence should be addressed.

[5]Present address: Center for Polymer Studies and Department of Physics, Boston University, Massachusetts, USA.

© Oxford University Press 1999

of low-complexity sequence regions, in order to filter them before scanning databases searching for homologies. However, the aim of the algorithm described here is to identify statistically significant, compositionally homogeneous DNA domains which may used to compute a variety of statistics useful to reveal the spatial compositional structure in DNA sequences. Segmentation methods have been recently reviewed (Braun and Müller, 1998).

Systems and methods

The program code was written in C++, and the Borland C++ Builder was used to compile the program for Windows 98/NT systems. Also, a complete Graphical User Interface (GUI) for input/output is incorporated. A Unix version is under development.

Heuristic segmentation algorithm

Our heuristic algorithm divides a DNA sequence into compositionally homogeneous domains by iterating a local optimization procedure at a given statistical significance. Compositional domains are defined (Bernaola-Galván *et al.*, 1996; Román-Roldán *et al.*, 1998) as subsequences with a different base composition in comparison to the left and right adjacent ones, at a given level of statistical confidence, s. To decide whether two adjacent subsequences are domains, the Jensen–Shannon divergence measure is used. For two subsequences S_1 and S_2 of lengths l_1 and l_2 we have

$$JS_2(S_1, S_2) = H[S] - \left(\frac{l_1}{L}H[S_1] + \frac{l_2}{L}H[S_2]\right) \geq 0$$

where $L = l_1 + l_2$, $S = S_1 \oplus S_2$ (concatenation) and

$$H[\wp] = -\sum p \log_2 p$$

is the Shannon's entropy of the probability distribution $\wp$ obtained from base frequencies in the corresponding subsequence.

Quaternary $\{A, T, C, G\}$ as well as binary ($\{R/Y\}$, $\{S/W\}, \ldots$) alphabets may be used to compute JS_2. 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, i.e. given a particular value of $JS_2 = x$, its statistical confidence is defined as: $s(\mathrm{x}) = \mathrm{Prob}\{JS_2 \geq x\}$ in a random sequence. For short subsequences, the probability is exactly computed from the hypergeometric distribution; for large ones, we found that JS_2 follows the Chi-square distribution with three degrees of freedom for quaternary alphabets and one degree of freedom for binary ones (Bernaola-Galván, 1997).

The heuristic algorithm implemented here works as follows. A sliding border is moved along the initial sequence, computing at each point the JS_2 divergence between the left and right subsequences defined by the border; the partition $S \rightarrow S_1 \oplus S_2$ with $\max\{JS_2(S_1, S_2)\}$ is chosen, provided that it satisfies the significance condition. The procedure is iterated on the resulting subsegments until no further splits satisfying the significance condition can be made. A further caution is that before a new cut is accepted the statistical significance of the potential left and right segments with respect to previous adjacent domains should be checked. If the new cut would provoke the loss of significance with any of the two adjacent domains, it is rejected; otherwise, it is accepted. This ensures that all the segments in the final partition remain significant (for further details see Román-Roldán *et al.*, 1998).

Once the partition of a sequence at a given a significance level s has been accomplished, the sequence compositional complexity (SCC) is given by:

$$\mathrm{SCC}(S) = H[S] - \sum_{i=1}^{n} \frac{l_i}{L} H[S_i] = \sum_{i=1}^{n} \frac{l_i}{L}(H[S] - H[S_i])$$

which is a global complexity measure taking into account both the number and compositional biases of all the domains in the sequence. The plot of SCC as a function of s constitutes the sequence complexity profile, which shows sequence structure at different resolution scales. A constant slope in the profile reveals a self-similar-like sequence structure (Román-Roldán *et al.*, 1998).

Implementation

Two main tasks are performed by the program: (1) the segmentation of a sequence at a given significance level; and (2) the generation of a complexity profile within a given range of significance values, equivalent to a multiscale view of sequence structure.

Input

The following parameters should be provided (see screen input in Figure 1):

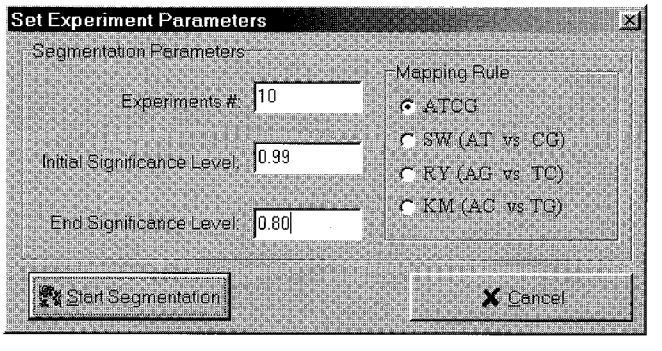

Fig. 1. SEGMENT window to input segmentation parameters.

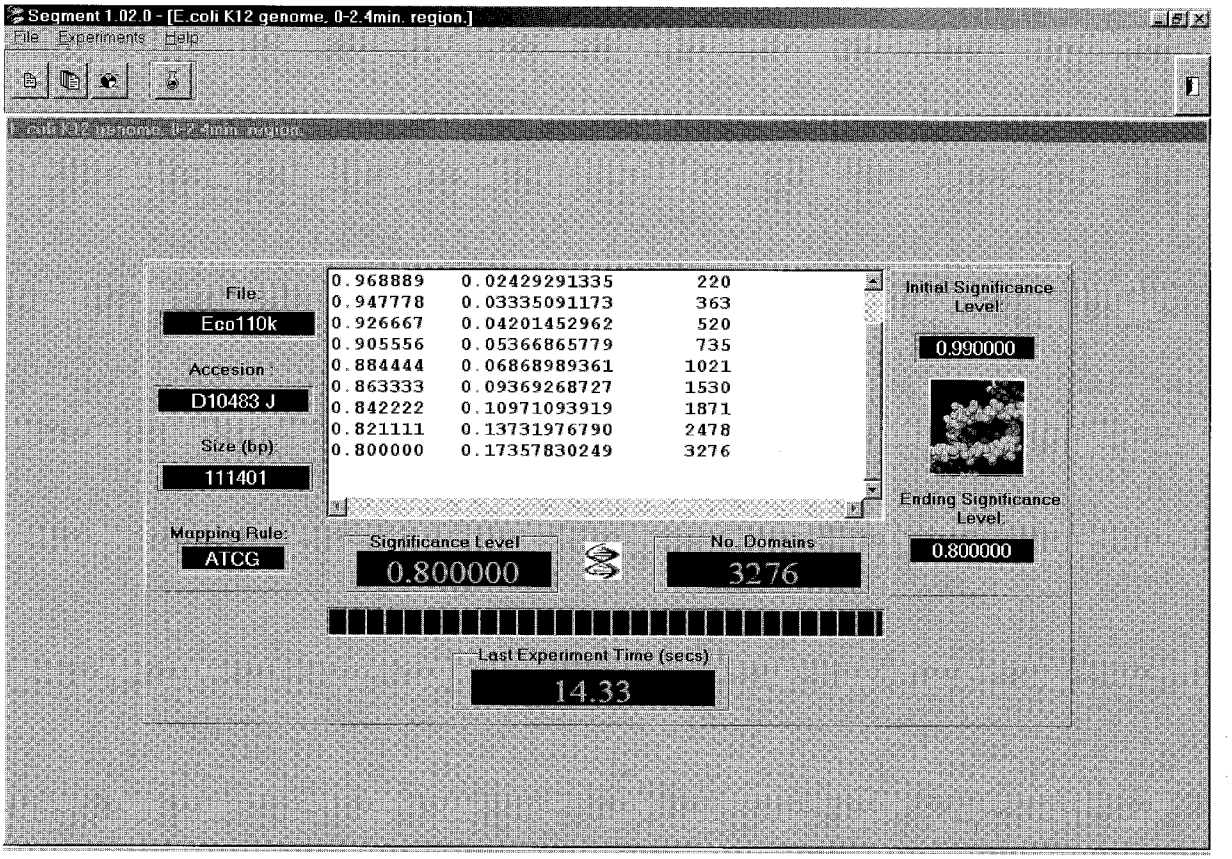

Fig. 2. SEGMENT main window.

- **Sequence name:** Any valid filename in the Windows standard format. DNA sequences of any length (only limited by the available physical memory of the computer) can be handled. Standard EMBL and GenBank formats are accepted.
- **Alphabet:** The user may choose among several quaternary and binary alphabets or mapping rules (see the Heuristic segmentation algorithm section).
- **Range of significance levels:** If the complexity profile is desired, two numbers, $s_h > s_\ell$, each in the range $0.1 \leq s \leq 0.99999$ should be provided, indicating the significance range for the complexity profile. If the segmentation at only one significance level is intended, enter the same s value in the two fields.

Output

The results of the segmentation analysis are first shown on the computer screen (Figure 2). Domain coordinates and composition are shown in a graphical viewer by means of the Map View program option (Figure 3). Nucleotide composition is numerically displayed at the top-right corner of the graphical screen, as the mouse is moved along the sequence window. Either colour or grey-scale can be used to visualize the compositional biases at different domains. A more variegated (complex) structure for the human sequence HSTCRADCV, as compared with the bacterial sequence ECO110K, can be readily visualized on this figure. In addition, two plain files are output to the default directory:

(a) **sequence_name.seg:** the coordinates for all the domains found in the sequence. The nucleotide composition of each domain is also given in this file.

(b) **sequence_name.pro:** numerical results to trace the sequence compositional profile (as that shown in Figure 4).

More details for input/output are provided in the program's on-line help.

Discussion

Here we present a heuristic segmentation algorithm able to decompose a DNA sequence into statistically significant compositional patches or domains. The final partition of a sequence into compositional domains is made here de-

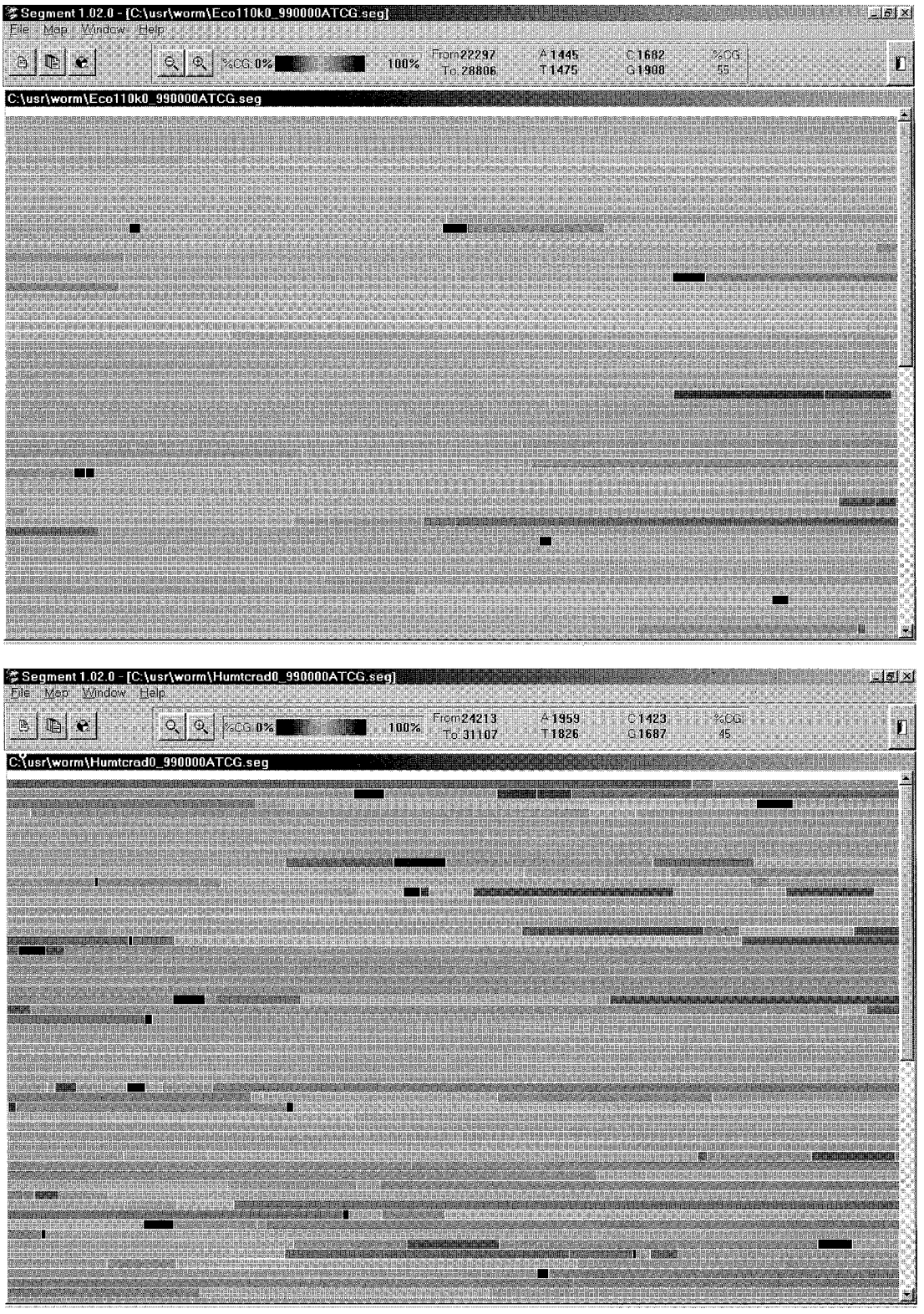

Fig. 3. Compositional domains in the *E.coli* sequence ECO110K (accession D10483, above) and the human sequence HSTCRADCV (accession M94081, below) after segmentation at $s = 99\%$, as shown by the Map View SEGMENT option.

pendent on a single parameter: the statistical significance level. The higher the statistical confidence, the higher the differences in nucleotide composition, and the lower the number of domains. Lowering the statistical confidence is then equivalent to using a magnifying glass: more details (domains), but less pronounced differences among them, are seen. We exploited this dependence on significance level to derive a multiscale measure of sequence complexity, the SCC profile (Figure 4).

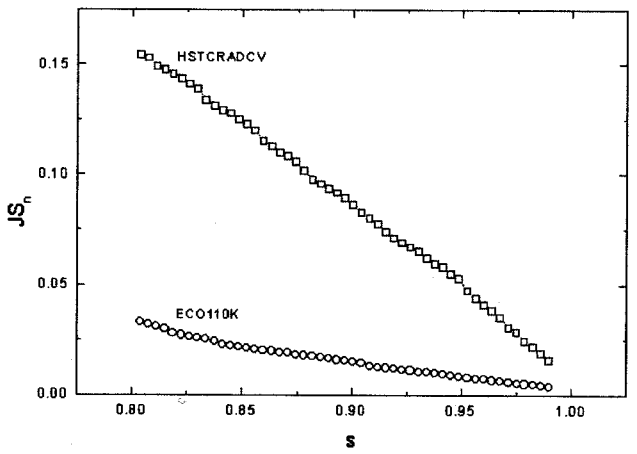

Fig. 4. SCC profiles of the *E.coli* sequence ECO110K and the human sequence HSTCRADCV after segmentation between $s_h = 0.99$ and $s_\ell = 0.80$. The purine/pyrimidine (R/Y) mapping rule was used.

The general-purpose tool described here allows many potential applications in extracting biologically relevant knowledge from DNA sequences, such as the following:

(1) Compositional heterogeneity in a DNA sequence can be simple or complex (Li *et al.*, 1994; Li, 1997a,b). Examples of simple sequences are bacterial or phage genomes, where compositional patches can be identified simply by eye (Karlin and Brendel, 1993). However, most eukaryotic sequences show complex heterogeneity, and the only way to identify domains is on a statistical basis. SEGMENT uses this approach, allowing a clear distinction between sequences with simple or complex correlation structure (Bernaola-Galván *et al.*, 1996). Figure 3 shows that, at the 99% confidence level, HSTCRADCV is more complex than ECO110K. In addition, Figure 4 shows that such higher complexity of the human sequence is extended to all the scales. See Román-Roldán *et al.* (1998), for further discussion on this topic.

(2) Another achievement of the segmentation method presented here was the uncovering for the first time of 'domains-within-domains' (see Figure 3 in Bernaola-Galván *et al.*, 1996) in complex, fractal DNA sequences. This feature was a prediction based on theoretical grounds (Li *et al.*, 1994), but only the recursive segmentation of

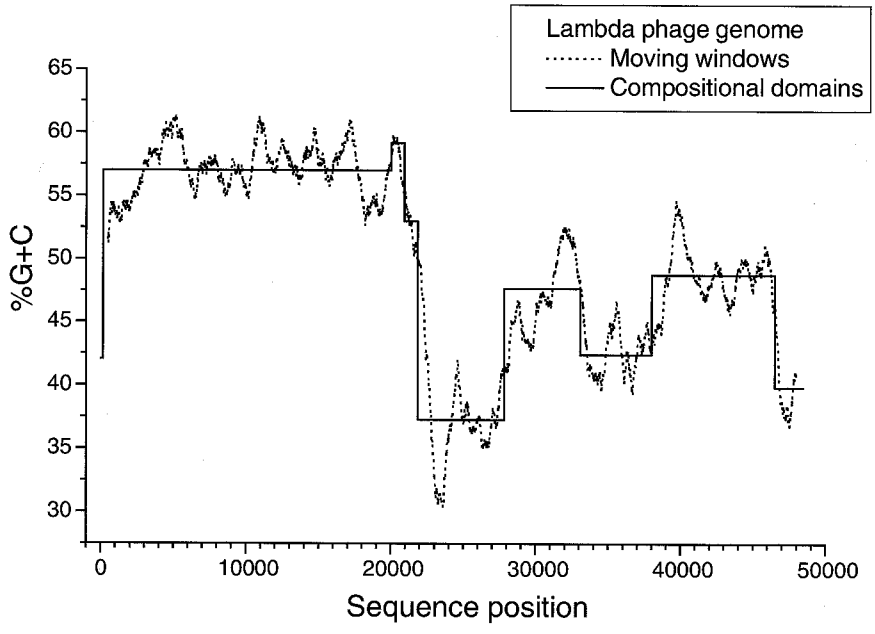

Fig. 5. %GC variation along the DNA sequence of the lambda phage genome (accession J02459). The broken line shows the %GC variation in a moving window (length = 1000, step = 10), while the solid line corresponds to the %GC variation at statistically significant compositional domains ($s = 99.99\%$, $\{S, W\}$ mapping rule).

a sequence at different levels of detail allowed by our algorithm proved capable of uncovering this important feature of complex DNA sequences (for further discussion see Li, 1997a).

(3) The segmentation method described here has been recently applied to the entire set of yeast chromosomes (Li *et al.*, 1998), revealing that both the domain length distributions, as well as the compositional complexity profiles for the 16 chromosomes of this unicellular eukaryote, are all strikingly similar.

(4) Finally, we are now exploring the use of compositional domains as an alternative to moving windows in grasping the spatial compositional heterogeneity along DNA sequences (J.L.Oliver *et al.*, for publication). As a first example, let us mention the variation in GC content along a sequence. Usually, this quantity is computed in a moving window that slides along the sequence; however, it would be advantageous to compute it at compositional domains (Figure 5). The main advantage is that the compositional differences between domains are all statistically significant at the specified confidence level, whereas the statistical significance of different windows is unknown. A second potential application of compositional domains is in the field of functional genomics. For example, compositional domains, instead of moving windows, would be used to compute a variety of coding discriminant statistics intended for gene finding in the large anonymous sequences now being generated by genome projects (P.Bernaola-Galván *et al.*, submitted for publication). The homogeneous base composition of domains would warrant more discriminant power for these statistics.

Acknowledgements

This work was supported by grants PB96-1414-CO2-01 from the Spanish Government. We thank David Nesbitt for linguistic help with the manuscript.

References

Bernaola-Galván,P. (1997) PhD Dissertation, University of Granada, (Spain), (in Spanish).

Bernaola-Galván,P., Román-Roldán,R. and Oliver,J.L. (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys. Rev. E*, **53**, 5181–5189.

Bernardi,G. (1995) The human genome: organization and evolutionary history. *Annu. Rev. Genetics*, **29**, 445–476.

Braun,J.V. and Müller,H.G. (1998) Statistical methods for DNA sequence segmentation. *Stat. Sci.*, **13**, 142–162.

Karlin,S. and Brendel,V. (1993) Patchiness and correlations in DNA sequences. *Science*, **259**, 677–680.

Li,W. (1997a) The complexity of DNA: the measure of compositional heterogeneity in DNA sequences and measures of complexity. *Complexity*, **3**, 33–37.

Li,W. (1997b) The study of correlation structures of DNA sequences–a critical review. *Comput. Chem.*, **21**, 257–272.

Li,W., Marr,T.G. and Kaneko,K. (1994) Understanding long-range correlations in DNA sequences. *Physica D*, **75**, 392–416.

Li,W., Stolovitzky,G., Bernaola-Galván,P. and Oliver,J.L. (1998) Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. *Genome Res.*, **8**, 916–928.

Peng,C-K., Buldyrev,S.V., Goldberger,A.L., Havlin,S., Sciortino,F., Simons,M. and Stanley,H.E. (1992) Long-range correlations in nucleotide sequences. *Nature*, **356**, 168–170.

Román-Roldán,R., Bernaola-Galván,P. and Oliver,J.L. (1998) Sequence compositional complexity of DNA through an entropic segmentation method. *Phys. Rev. Lett.*, **80**, 1344–1347.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino-acid-sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.

Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol.*, **266**, 554–571.