ELSEVIER

Physica A 249 (1998) 581–586

PHYSICA A

Long-range correlation measures for quantifying patchiness: Deviations from uniform power-law scaling in genomic DNA

G.M. Viswanathan [a,*], S.V. Buldyrev [a], S. Havlin [a,b], H.E. Stanley [a]

[a] *Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA*

[b] *Gonda-Goldschmied Center and Department of Physics, Bar Ilan University, Ramat Gan, Israel*

Abstract

We introduce and develop new techniques to quantify DNA patchiness, and characteristics of their mosaic structure. These techniques, which involve calculating two functions, $\alpha(\ell)$ and $\beta(\ell)$, measure correlation exponents at length scale ℓ and detect distinct characteristic patch sizes embedded in scale invariant patch size distributions. Using these methods, it is possible to address a number of issues relating to the mosaic structure of genomic DNA. We find several distinct characteristic patch sizes in yeast, human, and prokaryotic sequences. We also find that the distinct patch sizes in all 16 yeast chromosomes are similar.

PACS: 87.10.+e

Keywords: DNA; Long-range correlations

1. Introduction

It is well known that DNA polymer sequences have a mosaic structure, which is characterized by "patches" with an excess of one type of nucleotide [1–3]. Patchiness is usually associated in the biological literature with the phenomenon of isochores, which are DNA regions having homogeneous base compositions and typical scales of about 1 Mbp [4]. Here, we extend the concept of patchiness to include nonuniformities on scales smaller than 1 Mbp [5].

It is found that for many DNA sequences the nucleotide concentration fluctuation σ grows not with the square root of ℓ, but as another power law: $\sigma \sim \ell^{\alpha}$, where

* Corresponding author.

PII S0378-4371(97)00519-0

$\alpha \neq 1/2$ is a scaling exponent which describes the "roughness" of the fluctuations [6–9]. In such cases the DNA sequences have patches of all length scales, i.e., there exists no characteristic patch size because power-law behavior is the signature of scale invariance [10]. Indeed, recently a direct measurement of patchiness was performed and a power-law distribution was found for the patch sizes [11]. The basic premise behind our newly developed methods is that deviations from power-law behavior can be related to characteristic scales [5]. The degree to which the mosaic structure of DNA is related to such long-range correlation properties of DNA sequences has been discussed [11–16]. However, long-range correlation measures have never been used to quantify patchiness or to identify characteristic patch sizes in DNA sequences. Here we adapt and extend methods [8,17] for studying patchiness in DNA by developing techniques to quantify *departures* from power-law behavior and to estimate distinct characteristic DNA patch sizes embedded in power-law distributions of patch sizes. These techniques, which involve the calculation of two functions, $\alpha(\ell)$ and $\beta(\ell)$, measure correlation exponents at length scale ℓ and detect distinct characteristic patch sizes embedded in scale invariant domain size distributions. For ideal power-law correlations, the two exponents are related by $\alpha(\ell) = [1 + \beta(\ell)]/2 = \text{constant}$ [9,17].

2. Methods and controls

In order to apply numerical methods to a DNA sequence $\{n_i\}$ consisting of the four nucleotides A, C, T and G, we generate a binary sequence $\{u_i\}$ for each DNA sequence [17]. We use the following three binary mapping rules:

(i) *Purine-pyrimidine (RY) rule.* If n_i is a purine (A or G) then $u_i = 1$; if n_i is a pyrimidine (C or T), then $u_i = -1$.

(ii) *Hydrogen bond energy (SW) rule* [18,19]. For strongly bonded pairs (G or C) $u_i = 1$ while for weakly bonded pairs (A or T) $u_i = -1$.

(iii) *Hybrid (KM) rule.* For A and C $u_i = 1$, while for T and G $u_i = -1$.

Each of these rules probes a different aspect of the mosaic structure of DNA [5], e.g., the SW rule is related to the energy balance of strand separation, while the RY rule is related to strand chemical bias.

First, we develop techniques for detecting and examining characteristic scales of patchiness by studying a "control sequence" of $+1$ and -1 with patches of 3 different characteristic scales. The control sequence is constructed by concatenating uncorrelated patches of fixed sizes of 200, 2000 and 20 000 bp. For each patch j of length L_j, we randomly assign $P_j(1)$, the concentration of "+1", to be either $P_j(1) = 0.3$ or $P_j(1) = 0.7$ with equal probability, i.e., each patch has randomly assigned biases $b = P_j(1) - P_j(-1) = \pm 0.40$. Then we concatenate these patches to make a sequence of length $N = 10^6$ bp or more. We use the following rule for generating long-range correlations [5]. For the patch j,

(i) A random number x_j is chosen in the interval $[0, 1]$.

(ii) A preliminary length quantity ℓ_j is computed as $\ell_j = 200/x_j$.

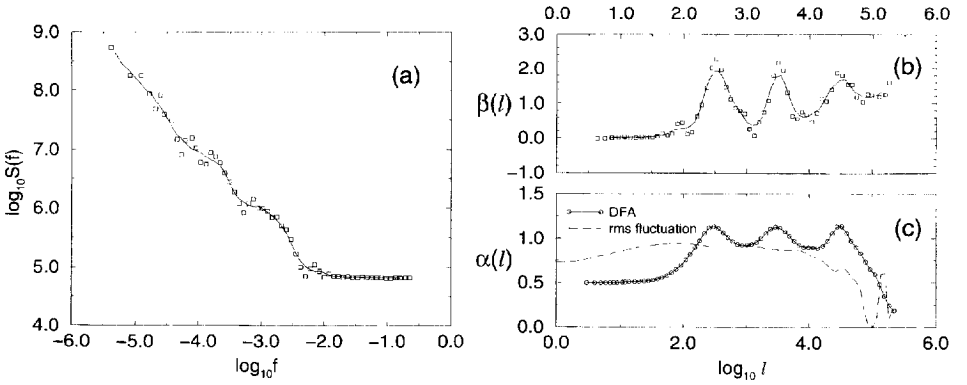

Fig. 1. (a) Double log plot of the power spectrum $S(f)$ of an artificial control DNA sequence of length 2^{19} bp. The solid line shows the spectrum after further smoothing using a subjective smoothing criterion [5]. The 3 characteristic scales are not readily discernible in the spectrum. (b) Log-linear plot of the power spectrum correlation exponent $\beta(\ell)$ for the same sequence, where $\ell \equiv 1/f$. The solid line is $\beta(\ell)$ after further smoothing. (c) DFA correlation exponent $\alpha(\ell)$ for the same sequence. The exponents $\alpha(\ell)$ and $\beta(\ell)$ peak at three locations corresponding to the three characteristic patch sizes. The location of the peaks is always about 1.5 multiplied by the patch sizes. Also shown is $\alpha(\ell)$ found from the "DNA walk" rms fluctuation method (dashed line) [8], which is unable to detect the three characteristic patch sizes.

(iii) If ℓ_j is less than 2000 then a patch of size $L_j = 200$ bp is chosen. Otherwise if ℓ_j is less than 20 000 then a patch of size $L_j = 2000$ bp is chosen. Otherwise a patch of size $L_j = 20\,000$ is chosen.

The power spectrum $S(f)$ for this control sequence is defined as the modulus squared of the discrete Fourier transform $\tilde{u}_f$ of u_i: $S(f) \equiv |\tilde{u}_f|^2$. We find that $S(f)$ resembles a "$1/f$-type" spectrum, as shown in Fig. 1a. The spectrum scales approximately as $S(f) \sim f^{-\beta}$, where $\beta \approx 1$ for this sequence. However, there are important deviations from pure power-law behavior which indicate the presence of characteristic scales. We define the correlation exponent $\beta(\ell)$ as

$$\beta(\ell) \equiv -\left| d \log S(f) d \log f \right|_{f=1/\ell} \tag{1}$$

where $\ell = 1/f$ has dimensions of length, i.e., $\beta(\ell)$ represent successive slopes of the double log plot of $S(f)$. We find that after additional smoothing $\beta(\ell)$ displays three local maxima which correspond to the three scales of patchiness of the control sequence (Fig. 1b).

The estimation of $\beta(\ell)$ requires an arbitrary amount of smoothing by visual inspection, making it susceptible to human judgement. However, *detrended fluctuation analysis* (DFA) [16] does not suffer from these disadvantages. We use the variant of the DFA method described in Ref. [17]. The net displacement $y(n)$ of the sequence u is defined by $y(n) \equiv \sum_{i=1}^{n} u_i$, which can be thought of graphically as a one-dimensional random walk. The sequence $y(n)$ is then divided into a number of overlapping subsequences of length ℓ, each of which is shifted with respect to the previous subsequence by a single nucleotide. For each subsequence, linear regression is used to calculate

an interpolated "detrended" walk $y'(n) \equiv a + b(n - n_0)$. Then we define the "DFA fluctuation" by $F_D(\ell) \equiv \sqrt{\langle (\delta y)^2 \rangle}$, where $\delta y \equiv y(n) - y'(n)$, and the angular brackets denote averaging over all points $y(n)$. We use a moving window to obtain better statistics. The DFA exponent $\alpha(\ell)$ is defined by

$$\alpha(\ell) \equiv \frac{d \log F_D(\ell)}{d \log(\ell + 3)} \tag{2}$$

where the $+3$ term is a correction important for small ℓ [17]. As we mention above, for an ideal power law $\alpha(\ell) = [1 + \beta(\ell)]/2 =$ constant [9,17]. (See also Ref. [5].) To present both $\alpha(\ell)$ and $\beta(\ell)$ on approximately the same scale, we can plot $[1 + \beta(\ell)]/2$ instead of $\beta(\ell)$ [9,17]. Fig. 1c shows $\alpha(\ell)$ for the artificial control model described above.

The functions $\alpha(\ell)$ and $\beta(\ell)$ are measures of how correlated a sequence is on different length scales. Since peaks in $\alpha(\ell)$ and $\beta(\ell)$ correspond to higher correlations, therefore, by studying peaks in $\alpha(\ell)$ and $\beta(\ell)$, we can estimate distinct characteristic DNA patch sizes embedded in a sequence with an apparent $1/f$ power spectrum. *We emphasize that such peaks corresponding to a given size do not imply the existence or absence of domains of that size, but rather imply an abundance of patches with that size relative to a power-law distribution of patch sizes.* It can be shown [5] that the peaks should occur at scales of the patch size multiplied by a factor a, where $a = 1/\ln 2 \approx 1.44$. This is numerically close to the measured value $a = 1.5$ obtained from simulations.

3. Results for known DNA sequences

We next apply these methods for detecting and examining characteristic scales of patchiness to the sixteen chromosomes of *Saccharomyces cerevisiae*. Fig. 2a shows the DFA exponent for each of the 16 yeast chromosomes individually for the RY rule and Fig. 2b for the SW rule. Note the similarity of $\alpha(\ell)$ for different chromosomes [5]. See Ref. [5] for the power spectrum exponents $\beta(\ell)$.

Next, we estimate characteristic patch sizes for several eukaryotic sequences longer than 10^5 bp, as well as for some *E. coli* bacterial sequences, as shown in Fig. 3. We used the peaks in $\alpha(\ell)$ divided by the factor $a = 1.5$ to evaluate the actual patch sizes. We find that similar patch sizes appear in several sequences, and some even appear on sequences from different species.

4. Discussion

Our study shows that the yeast chromosomes have similar $\alpha(\ell)$. We find that for $\ell < 10^3$ bp the yeast chromosomes have almost identical mosaic structure and correlation properties. This suggests that unique mechanisms organize all yeast chromosomes and that these mechanisms may be significantly different in higher eukaryotes and

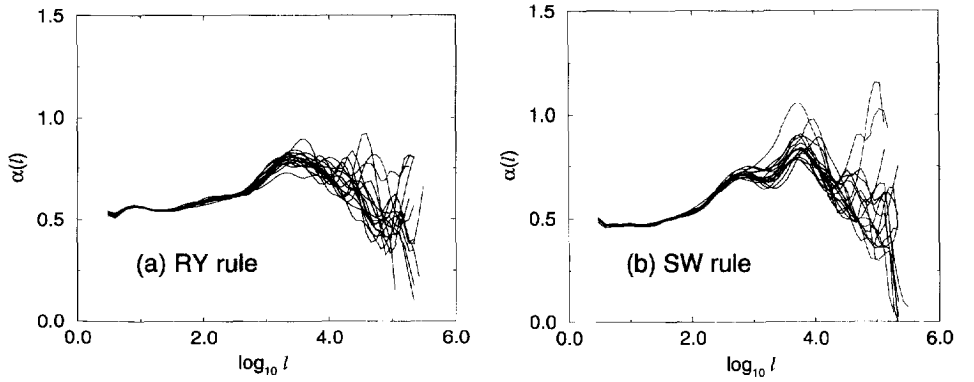

Fig. 2. DFA exponent $\alpha(\ell)$ for yeast chromosomes using (a) the RY rule and (b) the SW rule. We note that the general shape of $\alpha(\ell)$ is similar for all 16 chromosomes.

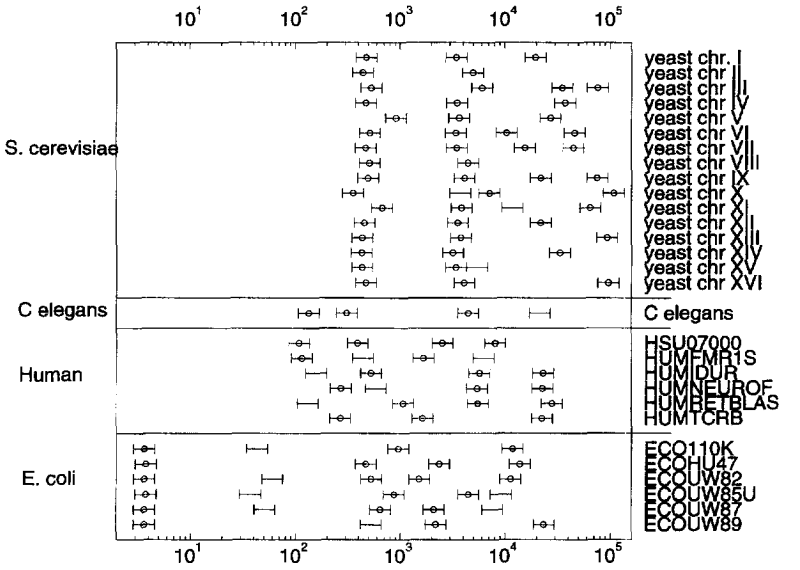

Fig. 3. Characteristic patch sizes for the 16 yeast chromosomes and other sequences estimated using the SW rule. Patch sizes which could only estimated by visual inspection of the peaks are indicated by error bars without circles.

prokaryotes. We also find distinct characteristic patch sizes in genomic DNA sequences. As seen in Fig. 3, similar patch sizes appear in several sequences, and some even appear in sequences from different species. The patchiness in eukaryotic DNA could be due partially to the elaborate organization and folding of DNA by proteins into nucleosomes and higher-order structures of chromatin. Note that the yeast sequences do not show patchiness on scales from 50 to 200 bp. Perhaps this is due to the absence in yeast of the normal H1 histones which help pack nucleosomes together [20].

In summary, we find distinct characteristic DNA patch sizes embedded in scale invariant patch size distributions by applying the new techniques developed here for quantifying DNA patchiness.

Acknowledgements

We wish to thank A.L. Goldberger, I. Große, P. Ivanov, C.-K. Peng, R. Mantegna and M. Simons for significant help at the initial stages of this work. We thank NIH for support. We also thank the organizers of the conference.

References

[1] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, F. Rodier, Science 228 (1985) 953.
[2] G.A. Churchill, Bull. Math. Biol. 51 (1989) 79.
[3] J.W. Fickett, D.C. Torney, D.R. Wolf, Genomics 13 (1992) 1056.
[4] G. Bernardi, Annu. Rev. Genet. 23 (1989) 637.
[5] G.M. Viswanathan, S.V. Buldyrev, S. Havlin, H.E. Stanley, Biophys. J. 72 (1997) 866.
[6] A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, Phys. Rev. Lett. 74 (1995) 3293.
[7] W. Li, K. Kaneko, Europhys. Lett. 17 (1992) 655.
[8] C.-K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature 356 (1992) 168.
[9] R. Voss, Phys. Rev. Lett. 68 (1992) 3805.
[10] H.E., Stanley, Introduction to Phase Transitions and Critical Phenomena, Oxford University Press, London, 1971.
[11] P. Bernaola-Galvan, R. Roman-Roldan, J.L. Oliver, Phys. Rev. E 53 (1996) 5181.
[12] S. Nee, Nature 357 (1992) 450.
[13] S. Karlin, V. Brendel, Science 259 (1993) 677.
[14] P.J. Munson, R.C. Taylor, G.S. Michaels, Nature 360 (1992) 636.
[15] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E 47 (1993) 4514.
[16] C.K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Phys. Rev. E 49 (1994) 1685.
[17] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E 51 (1995) 5084.
[18] M.Ya. Azbel, Y. Kantor, L. Verkh, A. Vilenkin, Biopolymers 21 (1982) 1687.
[19] M.Ya. Azbel, Phys. Rev. Lett. 31 (1973) 589.
[20] F. Thoma, G. Cavalli, S. Tanaka, in: P.M.A. Broda, S.G. Oliver, P.F.G. Sims (Eds.), The Eukaryotic Genome: Organization and Regulation, Cambridge University Press, Cambridge, 1993, pp. 43–52.