# Characteristic Patch Sizes in DNA Sequences

G. M. Viswanathan,[1] S. V. Buldyrev,[1] S. Havlin[1,2] and H. E. Stanley[1]
[1] *Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215*
[2] *Department of Physics, Bar Ilan University, Ramat Gan, Israel*
(**gandhi**.tex — April 29, 1999— **draft**)

We develop techniques for studying characteristic length scales in DNA sequences and apply these to the analysis of long genomic sequences including all available human sequences longer than 100000 bp and the nine sequenced yeast chromosomes. We find evidence suggesting the existence of a hierarchy of characteristic length scales in all the genomic DNA sequences analyzed. In particular, we find similar patch sizes in all nine yeast chromosomes, and some patch sizes exist in several organisms. We examine the possibility that in yeast the patchiness is caused by the alternation of coding and noncoding DNA sequences. We also examine that in human sequences the patchiness is related to repetitive sequences. However, we conclude that neither repetitive sequences nor the alternation of coding and noncoding DNA can fully explain the the mosaic structure of DNA.

## I. INTRODUCTION

It is well known that DNA nucleotides have a mosaic structure, in which there are "patches" with an excess of one type of nucleotide (Bernardi et al., 1985; Churchill, 1989; Fickett et al., 1992). It is also known that the mosaic structure of DNA may affect the correlation properties of DNA sequences (Nee, 1992; Karlin and Brendel, 1993;Peng et al., 1994). Although the effects of patchiness on correlation have been studied, correlation measures have never been used to study patchiness and to identify characteristic patch sizes in DNA sequences. Here, we develop techniques for studying characteristic patch sizes in DNA sequences and then apply these to the analysis of long genomic DNA sequences, including the nine sequenced yeast chromosomes, several human sequences, and some prokaryotic sequences.

In order to apply numerical methods to a DNA sequence $\{n_i\}$ consisting of base pairs A (adenine), C (cytosine), T (thymine) and G (guanine), we generate a numerical sequence $\{u_i\}$ for each DNA sequence using the following 3 ways of 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, then $u_i = -1$.

(ii) *Hydrogen bond energy (SW) rule* (Azbel, 1973). 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 $u_i = -1$.

Other mappings can also be studied (Buldyrev et al., 1995).

## II. METHODS AND RESULTS

First, we develop techniques for detecting and examining characteristic scales of patchiness by studying a control sequence of +1's and -1's with patches of 3 different characteristic scales. The control sequence is constructed by concatenating uncorrelated patches of fixed sizes of 200 bp, 2000 bp, and 20000 bp. The patches have randomly assigned biases $b = P(1) - P(-1) = \pm 0.40$. To obtain an approximate power law distribution of patch sizes, the smallest patch size is chosen with the highest probability and the largest patch size is chosen with the smallest probability according to the following rule. For the $j$th patch,

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

(ii) A preliminary length quantity $L$ is computed as $L = 200/x$.

(iii) If $L$ is less than 2000 then a patch of size 200 bp is chosen. Otherwise if $L$ is less than 20000 then a patch of size 2000 bp is chosen. Otherwise a patch of size 20000 is chosen.

The power spectrum $S(f)$ for this control sequence, defined as the modulus squared of the discrete Fourier transform of $u_i$, is shown in Fig. 1(a). The spectrum resembles a "1/f type" power-law spectrum. Studying the unsmoothed spectrum alone can lead to the erroneous conclusion that the sequence is scale invariant (scale-free), since we find that the correlation exponent

$$\beta(\ell) = -\frac{d \log S(f)}{d \log f}, \qquad (1)$$

where $\ell = 1/f$ is a length, displays three bumps which correspond to the three scales of patchiness. See Fig. 1(b).

Whereas the estimation of $\beta(\ell)$ requires some smoothing and filtering, making it susceptible to human error, *Detrended fluctuation analysis* (DFA) (Peng et al., 1993) does not suffer from this disadvantage. We use the variant of the DFA method described in (Buldyrev

et al., 1995). 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 subsequences of length $\ell$. 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 averaging is over all points $y(n)$. We use a moving window to obtain better statistics. The DFA exponent $\alpha(\ell)$ is defined by

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

where the "+3" is a correction for small $\ell$ (Buldyrev *et al.*, 1995). Fig. 1(c) shows $\alpha(\ell)$ for the artificial control model described above. Note that unlike DFA, the original "DNA walk" fluctuation analysis (Peng, 1992) is unable to detect the patchiness.

The functions $\alpha(\ell)$ and $\beta(\ell)$ are thus a measure of how correlated a sequence is on different length scales. Peaks in $\alpha(\ell)$ and $\beta(\ell)$ correspond to characteristic patch sizes. It can be shown that the peaks should occur at scales of the patch size multiplied by a factor $\eta \approx 1.5$, which depends on quantities like the bias and length of the patches. Therefore, by looking for peaks in $\alpha(\ell)$ and $\beta(\ell)$, we can estimate characteristic DNA patch sizes. See the appendix for an approximate derivation of $\eta$.

Having developed the techniques for detecting and examining characteristic scales of patchiness in model sequences, we next apply these methods to real data. Fig. 2 shows the DFA exponent $\alpha(\ell)$ for the nine sequenced chromosomes of *Saccharomyces cerevisiae* using the purine-pyrimidine rule and the hydrogen bond energy rule. Note how similar $\alpha(\ell)$ is for different chromosomes. For $\ell < 1000$ bp the different chromosomes have almost identical $\alpha(\ell)$. This beautiful similarity indicates that the correlation properties of the different chromosomes are very similar for $\ell < 1000$ bp. Note also how the first couple of peaks in $\alpha(\ell)$ roughly coincide for the different chromosomes in Fig. 2(b). This indicates that the nine chromosomes have similar patch sizes, because peaks in $\alpha(\ell)$ correspond to characteristic patch sizes.

To test the idea that the correlation properties and patchiness in yeast chromosomes may be due simply to the alternation of coding and noncoding DNA (Nee, 1992), we study the length distribution of introns and exons in yeast chromosome III. We then generate control "introns" and control "exons" of uncorrelated DNA with the same length distributions and the same nucleotide concentrations as those found in yeast chromosome III. By alternating these control exons and introns, we construct a control sequence for which the introns and exons have similar length distributions and nucleotide concentrations as yeast chromosome III. We find the DFA exponent $\alpha(\ell)$ for this control sequence, and compare it to yeast chromosome III, as shown in Fig. 3(c) and (d).

Next, we estimate the characteristic patch sizes for several eukaryotic sequences longer than 100000 bp, as well as for some *E. coli* bacterial sequences, as shown in Fig. 4. We used the peaks in $\alpha(\ell)$ divided by the factor $\eta = 1.5$ to calculate the patch sizes. Similar patch sizes appear in several sequences, and some even appear on sequences from different species.

Finally we test the hypothesis that the patchiness could simply arise from the abundance of repetitive sequences in genomic DNA (Buldyrev et al., 1993). If this is true, then a control sequence constructed from repetitive sequences should be able to reproduce the patchiness and the correlation properties of genomic DNA sequences. Specifically, we study the DFA exponent $\alpha(\ell)$ of a control sequence composed 7.5% of Alu repeats and 15% of Line-1c repeats interspersed with uncorrelated sequences with average nucleotide concentrations estimated from all available human sequences larger than 50 bp. The uncorrelated spacer-sequences have a bias of $b = P(AT) - P(CG) = 0.15$, and have an exponential length distribution. The parameters we use for the repetitive sequences are typical for human DNA sequences (Bell, 1992; Bell 1993). The model is able to account for some features found in the real data (cf: Fig. 5).

## III. DISCUSSION

Recently, long-range power-law correlations were found to exist in some genomic DNA sequences (Arneodo et al., 1995; Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992). Tentative explanations have been put forward to explain this phenomenon involving 3D structure (Grosberg et al., 1993), repetitive sequences (Buldyrev et al., 1993), and point mutation and duplication (Li, 1991; Li and Kaneko 1992). There have also been several attempts to explain long-range correlations by the presence of the patch sizes of the fixed size (Azbel, 1995; Karlin and Brendel, 199; Azbel 1973), or as due to alternation of coding and noncoding sequences of certain characteristic sizes (Nee, 1992). But the origin of such long-range correlations in DNA is still regarded as an open question.

The values of $\alpha(\ell)$ for yeast chromosome III and for the coding-noncoding model of yeast chromosome III described above shows that the alternation of coding and noncoding DNA indeed contributes to the long-range correlation properties of yeast chromosome III. But Nee's hypothesis cannot explain all the correlation properties of the chromosome. Specifically, as seen in Fig.3, whereas for the hydrogen bond energy rule there are some similarities between the model and the real chromosome, for the purine-pyrimidine rule there is no resemblance at all.

Our results also go against the hypothesis that the known long-range correlations in DNA sequences are due to repetitive sequences. Specifically, the results in Fig. 5 show that although repetitive sequences are able to explain some features of the patchiness found in real data,

there are qualitative differences between the model and the real data which are unlikely to disappear by increasing the number of types of repetitive elements. Therefore we conclude that repetitive sequences cannot fully explain the known mosaic structure and the known long-range correlation properties of DNA sequences.

We comment on the finding that the yeast chromosomes have similar $\alpha(\ell)$. For $\ell < 1000$ bp the yeast chromosomes have almost identical mosaic structure and correlation properties. This suggests that the mechanisms which organize the yeast genome affect every chromosome in similar ways.

Our finding that there exists a hierarchy of characteristic patch sizes in genomic DNA sequences may shed some light on this observation. As seen in Fig. 4, similar patch sizes appear in several sequences, and some even appear on sequences from different species, suggesting that the complex global structure of genomic DNA may have some universal properties. 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. Nucleosome structure may be responsible for strong correlations near $\ell \approx 200$ bp, while the packaging of DNA into higher order structures like looped domains might lead to correlations on larger length scales (Alberts et al., 1994). Note that the yeast sequences do not show patchiness on scales from 50 bp to 200 bp. Perhaps this is due to the absence in yeast of the normal H1 histones which help pack nucleosomes together(Thoma et al., 1993).

In summary, we have shown that long-range correlated DNA sequences have a series of characteristic scales. We also have shown that neither the alternation of introns and exons nor the known existence of repetitive sequences can fully explain these findings. Since these characteristic scales may be related not only to biological function but also to genomic organization, evolution, and dynamics, we feel that this problem should be studied further.

We wish to thank A.L. Goldberger, I. Große, P. Ivanov, C.-K. Peng, and M. Simons, for help at the initial stages of this work. We also wish very much to thank the scientists who have made public the newly sequenced yeast chromosomes not yet incorporated into the GenBank database (Bussey et al., 1995; Feldmann et al.,1995; Galibert et al., unpublished 1995; Dietrich et al., 1994, unpublished) and NIH for support.

## IV. APPENDIX

Consider a sequence represented by a "square wave" function $y_1(x)$ of period $2L$ which consists of alternating patches of length $L$. Consider another sequence $y_2(x)$ which is white noise. Then

$$y(x) \equiv A y_1(x) + y_2(x) \tag{3}$$

is uncorrelated with pacthes of size $L$. Let $y_1(k)$ and $y_2(k)$ be the Fourier transforms of $y_1(x)$ and $y_2(x)$ respectively, where the wave number $k$ is the Fourier conjugate of $x$. Since $y_2$ is white noise, it is orthogonal to $y_1$ so that the power spectrum $S(k)$ of $y(x)$ is simply the sum of the power spectra of $y_1(x)$ and $y_2(x)$ :

$$S(k) = |y_1(k)|^2 + |y_2(k)|^2 \tag{4}$$

Since $y_2(x)$ is white noise, we can set $|y_2(k)|^2 \equiv C$, where $C$ is some constant proportional to the variance of $y_2(x)$. The function $y_1(k)$ peaks strongly at $k = \pi/L$ so that $y(k)$ also peaks at $k = \pi/L$. The peak in $\beta(k) \equiv -\mathrm{d}|y(k)|^2/\mathrm{d}k$ occurs very close to $k = \pi/L$. It does not occur exactly at $k = \pi/L$ partially because of the constant white noise term $C$, and partially because taking the log of the spectrum shifts the location of the peak. The peak location can shift by upto 10% below $k = \pi/L$, depending on the values of the bias and the sequence length. Taking this 10% into account, we get $L \leq \eta \leq 1.1L$.

However, if we assume that in real sequences the biases are not strictly alternating in sign but can take on positive and negative values with equal probabilities—as in the models—then we have to take into account an extra factor of $1/\ln 2 \approx 1.44$, because several patches with the same bias can become joined to form larger patches. The mean patch size becomes $L/\ln 2$. Thus we are left with the numerical estimate of

$$1.44L \leq \eta \leq 1.58 \tag{5}$$

which is in agreement with the measured value of $\eta \approx 1.5$.

## V. REFERENCES

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. 1994. Molecular Biology of the Cell, Third Edition, Garland Publishing, New York.

Arneodo, A., E. Bacry, P. V. Graves, and J. F. Muzy. 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.* 74:3293–3296.

Azbel, M. Ya. 1973. Random two-component, one-dimensional Ising model for heteropolymer melting. *Phys. Rev. Lett.* 31:589–593.

Azbel, M. Ya. 1995. Universality in a DNA statistical structure. *Phys. Rev. Lett.* 75:168–171.

Bell, G. I. 1993. Repetitive DNA sequences: some considerations for simple sequence repeats. *Computers and Chemistry* 17:185–190.

Bell, G. I. 1992. Roles of repetitive sequences. *Computers and Chemistry* 16[2]:135–143.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley. 1993. Generalized Lévy Walk model for DNA nucleotide sequences. *Phys. Rev. E* 47:4514–4523.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley. 1995. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev. E* 51:5084–5091.

Bussey, H., D. B. Kaback, W. Zhong, D. T. Vo, M. W. Clark, N. Fortin, J. Hall, B. F. F. Ouellette, T. Keng, A. B. Barton, Y. Su, C. K. Davies and R. K. Storms. 1995. The nucleotide sequence of chromosome I of Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.* 92:3809–3813.

Churchill, G. A. 1989. *Bull. Math. Biol.* 51:79.

Feldmann, H., M. Aigle, G. Aljinovic, B. Andre, M. C. Baclet, C. Barthe, A. Baur, A.-M. Becam, N. Biteau, E. Boles, T. Brandt, M. Brendel, M. Brueckner, F. Bussereau, C. Christiansen, R. Contreras, M. Crouzet, C. Cziepluch, N. Demolis, T. Delaveau, F. Doignon, H. Domdey, S. Duesterhus, E. Dubois, B. Dujon, M. El Bakkoury, K.-D. Entian, M. Feuermann, W. Fiers, G. M. Fobo, C. Fritz, H. Gassenhuber, N. Glansdorff, A. Goffeau,, L.A.Grivell, M. de Haan, C. Hein, C. J. Herbert, C. P. Hollenberg, K. Holmstrvm , C. Jacq, M. Jacquet, J. C. Jauniaux, J.-L. Jonniaux, T. Kallesoe, P. Kiesau, L. Kirchrath, P. Kötter, S. Korol, S. Liebl, M. Logghe, A. J. E. Lohan, E. J. Louis, Z. Y. Li, M. J. Maat, L. Mallet, G. Mannhaupt, F. Messenguy, T. Miosga, F. Molemans, S. Mueller, F. Nasr, B. Obermaier, J. Perea, A. Pierard, E. Piravandi, F. M. Pohl, T. M. Pohl, S. Potier, M. Proft, B. Purnelle, M. Ramezani Rad, M. Rieger, M. Rose, I. Schaaff-Gerstenschlaeger, B. Scherens, C. Schwarzlose, J. Skala, P. P. Slonimski, P. H. M. Smits, J. L. Souciet, H. Y. Steensma, R. Stucka, A. Urrestarazu, Q. J. M. van der Aart, L. van Dyck, A. Vassarotti, I. Vetter, F. Vierendeels, S. Vissers, G. Wagner, P. de Wergifosse, K. H. Wolfe, M. Zagulski, F.K. Zimmermann, H. W. Mewes, and K. Kleine. 1995. Complete DNA sequence of yeast chromosome II. *EMBL J.* 13:5795–5809.

Fickett, J. W., D. C. Torney, and D. R. Wolf. 1992. Base compositional structure of genomes. *Genomics* 13:1056–1064.

Grosberg, A., Y. Rabin, S. Havlin and A. Nir. 1993. Self-similarity in DNA structure. *Europhys. Lett.* 23:373–377.

Johnston, M., S. Andrews, R. Brinkman, J. Cooper, H. Ding, J. Dover, Z. Du, A. Favello, and L. Fulton. 1994. Complete nucleotide sequence of S. cerevisiae chromosome VIII. *Science* 265:2077–2082.

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

Li, W. 1991. Expansion-modification systems: a model for spatial $1/f$ spectra. *Phys. Rev. A* 43:5240–5260.

Li, W., and K. Kaneko. 1992. Long-range correlations and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys. Lett.* 17:655–660.

Nee, S. 1992. Uncorrelated DNA walks. *Nature* 357:450–450.

Peng, C.-K., S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons and H. E. Stanley. 1992. Long-range correlations in nucleotide sequences. *Nature* 356:168–171.

Peng, C. K., S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley and A. L. Goldberger. 1994. Mosaic organization of DNA nucleotides. *Phys. Rev. E* 49:1685–1689.

Thoma, F., G. Cavalli, and S. Tanaka. 1993. In: The Eukaryotic Genome: Organization and Regulation. Edited by P. M. A. Broda, S. G. Oliver, and P. F. G. Sims. Cambridge University Press, Cambridge. 43–52.

Voss, R. 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Lett.* 68:3805–3808.

FIG. 1. (a) Double log plot of the power spectrum of an artificial control DNA sequence of length $2^{19}$ bp. The spectrum is an average over a moving window of size $2^{18}$ bp with shifts of $2^{16}$ bp, and is plotted using logarithmic binning. The solid line shows the spectrum after smoothing. We note that the spectrum scales approximately as a power-law over 3 decades. The 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 $l \equiv 1/f$. The exponent $\beta(\ell)$ is estimated by smoothing the power spectrum and taking the negative of the local slope of the log-log plot of the spectrum. The solid line is $\beta(\ell)$ after further smoothing, showing three clear maxima. The degree of smoothing is to some extent arbitrary. (c) DFA correlation exponent $\alpha(\ell)$ for the same sequence. The exponent $\alpha(\ell)$ is found by calculating the local slope of the double-log plot of the DFA function. No smoothing or filtering is required. The exponents $\alpha(\ell)$ and $\beta(\ell)$ peak at three locations corresponding to the three characteristic patch sizes. The peaks occur at approximately 300 bp, 3000 bp, and 30000 bp, showing that the location of the peaks is always about 1.5 multiplied by the patch sizes. Also shown is $\alpha(\ell)$ found from Peng et al.'s original "DNA walk" rms flucutation method, which is unable to detect the three characteristic patch sizes.

FIG. 2. DFA exponent $\alpha(\ell)$ for the yeast chromosomes using (a) the purine-pyrimidine rule and (b) the hydrogen-bond energy rule. We note that the general shape of $\alpha(\ell)$ is similar for all four chromosomes. In particular, $\alpha(\ell)$ is almost identical for all nine chromosomes for $\ell < 1000$ bp, and the peaks and valleys (i.e. extrema) are close to each other for the hydrogen-bond energy rule, suggesting that there are similar characteristic patch sizes present in all chromosomes.

FIG. 3. Comparison of $\alpha(\ell)$ for yeast chromosome III and the model described in the text of alternating "coding" and "noncoding" patches of uncorrelated DNA for (a) the purine-pyrimidine rule and (b) the hydrogen bond energy rule. Whereas for the hydrogen bond energy rule there are some similarities between the model and the real chromosome, for the purine-pyrimidine rule there is no resemblance at all.

FIG. 4. Characteristic patch sizes in 6 *E. coli* sequences, 4 yeast sequences, 1 *C. elegans* sequence, and 6 human sequences estimated using the hydrogen bond energy rule. Only sequences larger than 100000 bp were used. The patch sizes were estimated by locating the peaks in $\alpha(\ell)$ and dividing the position of the peaks by 1.5. Patch sizes which could only estimated by visual inspection of the peaks are indicated by error bars without circles. Similar patch sizes are found in several sequences, suggesting that the complex global structure of genomic DNA may have some universal characteristics. In eukaryotic sequences the patchiness may be a result of 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 bp to 200 bp, possibly due to the absence in yeast of H1 histones which help pack nucleosomes together. The bacterial sequences have a patch size which is absent in the other sequences. The loci names of the human and E. coli sequences are as they appear in the figure. Except for some yeast sequences, all sequences are found in the GenBank database.

FIG. 5. Comparison of DFA exponent $\alpha(\ell)$ for a human sequences and an artificial control sequence composed of interspersed LINE-1c repeats, ALU repeats and uncorrelated sequences for (a) purine-pyrimidine rule, and (b) hydrogen bond energy rule, and (b) hybrid rule. The RMS of the DFA fluctuation of the human sequences is used to estimate $\alpha(\ell)$ for human sequences. The maxima for the hybrid rule occur at the same scale for human sequences and the model, suggesting that long-range correlations may be partially due to repetitive sequences. However, we note that this artificial control sequence gives rise at most to two characteristic patch sizes, and cannot reproduce the plateau in $\alpha(\ell)$ for the hydrogen bong energy rule. For the purine-pyrimidine rule the model disagrees with the data badly. So this simple model cannot explain the correlation roperties and the patchiness found in DNA. A larger variety of repeats is unlikely to remove the qualitative differences between the data and the model. We used the LINE-1c region in HUMHBB starting at 23137 and ending at 29515, and the ALU region starting at 66776 and ending at 67042.