Physica A 191 (1992) 25–29
North-Holland

PHYSICA A

Fractal landscape analysis of DNA walks

C.-K. Peng[a], S.V. Buldyrev[a], A.L. Goldberger[b], S. Havlin[a,c], F. Sciortino[a], M. Simons[b,d] and H.E. Stanley[a]

[a]*Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA*
[b]*Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, MA 02215, USA*
[c]*Department of Physics, Bar-Ilan University, Ramat-Gan, Israel*
[d]*Biology Department, MIT, Cambridge, MA 02139, USA*

By mapping nucleotide sequences onto a "DNA walk", we uncovered remarkably *long-range* power law correlations [Nature 356 (1992) 168] that imply a new scale invariant property of DNA. We found such long-range correlations in intron-containing genes and in non-transcribed regulatory DNA sequences, but not in cDNA sequences or intron-less genes. In this paper, we present more explicit evidences to support our findings.

The DNA walk (see ref. [1a])[#1] is defined by the rule that the walker steps up ($u(i) = +1$) if a pyrimidine occurs at position a linear distance i along the DNA chain, while the walker steps down ($u(i) = -1$) if a purine occurs at position i. The trajectories of the DNA walk, defined as $y(l) \equiv \Sigma_{i=1}^{l} u(i)$, produce a contour reminiscent of the irregular *fractal landscapes* (see fig. 1a) that have been widely studied in physical systems [2].

A quantitative characterization of such a "landscape" is the mean fluctuation function $F(l)$, defined as

$$F^2(l) \equiv \overline{[\Delta y(l) - \overline{\Delta y(l)}]^2} = \overline{[\Delta y(l)]^2} - \overline{\Delta y(l)}^2 , \tag{1}$$

of a quantity $\Delta y(l)$ defined by

$$\Delta y(l) \equiv y(l_0 + l) - y(l_0) . \tag{2}$$

Here the bars indicate an *average* over all positions l_0 in the gene.

If the nucleotides are uncorrelated or only short-range correlated (i.e., with a characteristic correlation length), then $F(l) \sim l^{1/2}$ (as expected for a *normal*

[#1] Long-range correlations in non-coding regions of DNA were reported independently by Li and Kaneko [1b].

0378-4371/92/$05.00

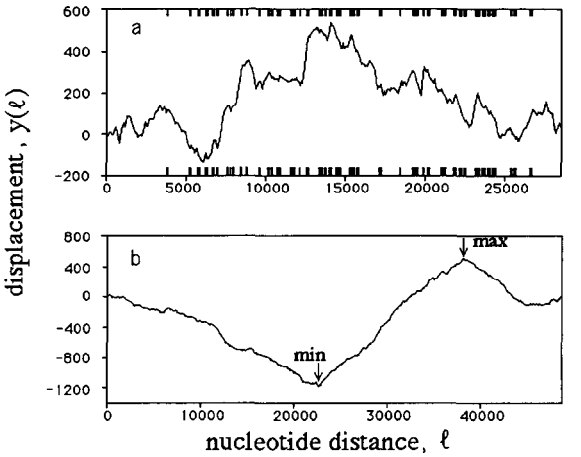

Fig. 1. The DNA walk representations of (a) intron-rich human β-cardiac myosin heavy chain gene sequence and (b) the intron-less bacteriophage λ DNA sequence. Heavy bars correspond to the coding regions of the gene. In order that the graphical representation are not affected by the global differences in concentration between purines and pyrimidines, we plot the DNA walk representation such that the end point has the same vertical displacement as the starting point (for the statistical analysis, we use the original definitions (1)–(3), without any adjustment of vertical displacement). The minimum (min) and maximum (max) points on the landscape are denoted by arrows in (b), and their application in the analysis is described in the text. Note that for almost all intron-less DNA sequences, there appear regions with one strand bias, followed by regions of a different strand bias (as shown in (b)). The fluctuation on either side of the overall strand bias is found to be random, a fact that is plausible by looking at these DNA walk representations. The bias introduced by the change in concentration of purine and pyrimidine would not be eliminated by the average term in eq. (1) if pieces of different bias would be analyzed together.

random walk). In contrast, if there is a long-range (scale-free) correlation in the purine–pyrimidine sequence, then

$$F(l) \sim l^{\alpha} \tag{3}$$

with $\alpha \neq 1/2$.

As noted previously [1] there are two different types of DNA walks (based on whether the walks contain non-coding sequences or not, compare figs. 1a and 1b). A systematic "min–max" procedure was applied to the DNA walks analysis in order to treat all DNA sequences on equal footing without applying any a priori knowledge of the sequence itself. The technical reason for using this min–max procedure is the following:

We find that for almost all intron-less genes and cDNA sequences studied (total of over 40 sequences) that there appear regions with one strand bias (rich in purines or pyrimidines), followed by regions of a different strand bias (fig. 1b). The log–log plot of $F(l)$ vs l (fig. 2a) shows a typical crossover behavior as can be demonstrated by progressive increase of the *local slope*, denoted as

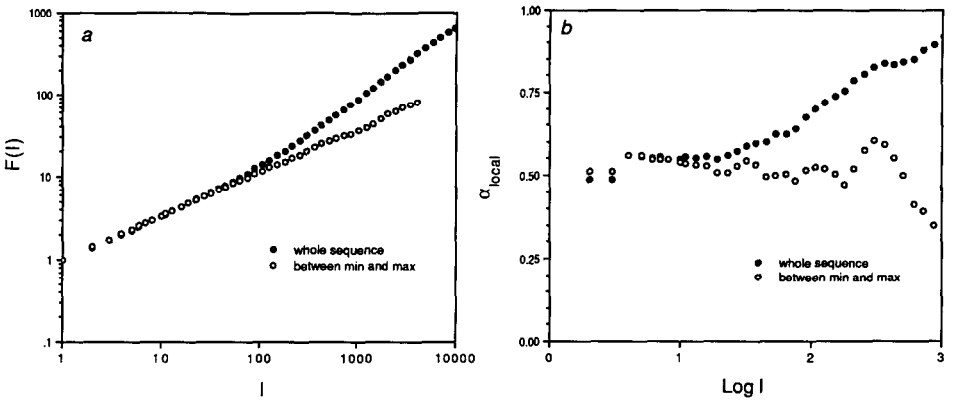

Fig. 2. (a) The log–log plot of $F(l)$ vs l for the intron-less LAMCG sequence and (b) plot of the local slope, α_{local}, the successive slope of log $F(l)$ vs log l. The solid circles, results of analysing the entire sequence, show a crossover behavior due to the global bias, while the open circles, results of using the min–max procedure, show a broad range of plateau around $\alpha = 0.5$.

α_{local}, from 0.5 toward 1 (fig. 2b). In order to study the scaling behavior of $F(l)$, we need to remove the effect due to such large regions of bias. Therefore, instead of analyzing the whole sequence, we can partition the sequence into two to three shorter sequences which have the same overall bias. This can be simply accomplished by cutting the sequence at the global maximum and minimum of the DNA walks (the min–max procedure [1]). Analysis of the sub-sequence, showing a broad range of $\alpha_{local} = 0.5$ (fig. 2b), implies $F(l) \sim l^{0.5}$. The above analysis reveals the fact that there is a characteristic length scale for the strand bias; this length scale may be determined by the function of the protein. Furthermore, we find that the fluctuation of purine/pyrimidine on a length scale smaller than this characteristic scale, i.e., on either side of the overall strand bias shown in fig. 2b, is random.

The DNA walks for intron-containing sequences do not show any apparent length scale of strand bias (fig. 1a). In fact, they seem to have a broad distribution of lengths of strand bias. In principle, for this type of DNA walk, the min–max procedure is not necessary since there is no characteristic length that needs to be taken into account. Indeed, for many intron-containing sequences, our analysis shows that there is a broad range of scaling region (with $\alpha > 0.5$, see fig. 3) when we study the entire sequence as a whole. Nevertheless, we find that, for certain intron-containing genes, the min–max procedure can extend the scaling region, i.e., a broader range of constant α. Most of the nucleotide sequences in GenBank [3] are associated with coding regions. We are, therefore, analyzing a very selective subset of the entire genomic sequences. Since there is a characteristic length scale in most cDNA sequences and intron-less genes, we expect that this length scale will also appear in some intron-containing DNA sequences chosen from GenBank.

To eliminate the possible dominating length scale that is hidden in the samples due to the chosen guideline of the samples, further analysis is necessary. The scale-free long-range correlation properties in non-coding sequences are supported by the following studies:

(1) We obtain the same value of α by iterating the min–max procedure several times – i.e., by cutting each sequence into even smaller pieces (provided that the final pieces are still large enough, ~1000 base pairs, to give statistically meaningful results). This result demonstrates the scale invariance properties of the DNA walks.

(2) We obtain the same value of α when analysing several randomly chosen subsequences of a genomic sequence that contains mostly non-coding sequences (see fig. 4). This result indicates that long-range correlations in DNA are indeed robust.

Finally, we would like to mention some recent works [4–7] stimulated by our findings. Buldyrev et al. [4] studied the evolution of the myosin heavy chain (MHC) gene family (and also two other gene families) by analysing the fractal landscapes of the DNA walks. They found that during the course of evolution, MHC genes increase the fractal complexity as described by the increasing value of α but not in direct proportion to the number of introns. They noted that while early on there is an increase in the length of introns, later an increase in α occurs even though the intron length remains the same. For the origin of the

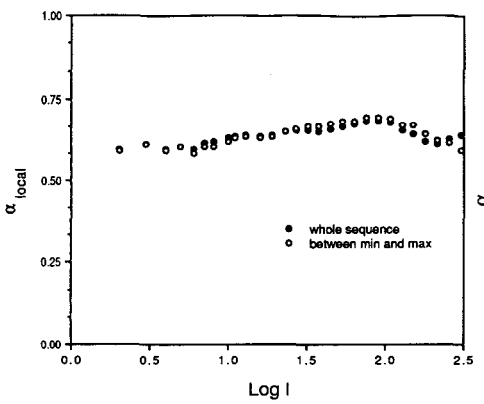

Fig. 3.

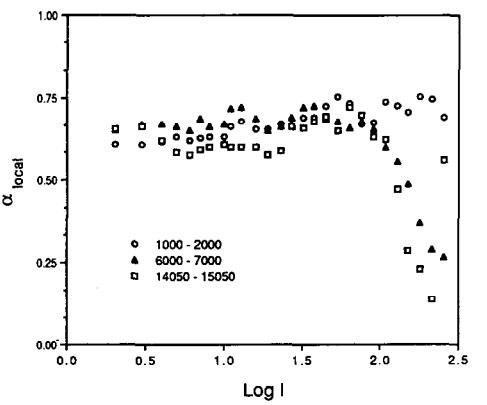

Fig. 4.

Fig. 3. Plot of α_{local} vs log l for intron-containing RATMHCG. The solid circles are results from analysing the entire sequence, while the open circles are results obtained by using the min–max procedure.

Fig. 4. Plots of local slopes, α_{local}, for three randomly chosen sub-sequences (1000 base pairs of each). Note that the three sequences have roughly the same scaling exponent.

correlation, Grosberg et al. [5] proposed a possible model to relate the exponent α with the fractal structure of a self-similar globular unknotted polymer. Their prediction $\alpha = 2/3$ is very close to our results. Although there are people still skeptical about the long-range correlation in DNA sequences [6], our finding has recently been confirmed by a massive calculation by Voss [7] on 25 000 GenBank sequences (containing a total of 50 million nucleotides); see also the news reported by Maddox [8], Amato [9] and Yam [10].

We wish to thank J. Hausdorff for important contributions in the initial stages of this project, and C. Cantor, C. DeLisi, R.D. Rosenberg, M. Schwartz and E. Trifonov for valuable discussions. Partial support was provided to ALG by the G. Harold and Leila Y. Mathers Charitable Foundation, the National Heart, Lung and Blood Institute and the National Aeronautics and Space Administration, to MS by the American Heart Association, to CKP by the NIH Graduate Traineeship Award, and to HES by the NSF.

References

[1] (a) C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley, Nature 356 (1992) 168; (b) W. Li and K. Kaneko, Europhys. Lett. 17 (1992) 655.
[2] M. Shlesinger, Random Walks (World Scientific, Singapore), in press.
[3] H.S. Bilofsky and C. Burkes, Nucl. Acids. Res. 16 (1988) 1861.
[4] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.K. Peng, H.E. Stanley and M. Simons, preprint.
[5] A. Grosberg, S. Havlin. A. Nir and Y. Rabin, preprint.
[6] S. Nee, Nature 357 (1992) 450.
[7] R. Voss, Phys. Rev. Lett. 68 (1992) 3805.
[8] J. Maddox, Nature 358 (1992) 103.
[9] I. Amato, Science 257 (1992) 747.
[10] P. Yam, Sci. Am. (Sept. 1992) 23.