LONG OR SHORT RANGE CORRELATIONS IN DNA SEQUENCES?

Yannis Almirantis and
Spyros Papageorgiou

National Research Center for Physical
Sciences "Demokritos",
15310 Aghia Paraskevi, Athens, Greece.

Introduction

Peng and co-workers [1] have analyzed DNA sequences looking for long range correlations using the root mean square fluctuation $F(l)$ about the average of the displacement, where l is a correlation length. They introduced first a 1:1 map of the DNA sequence to a so-called "DNA walk", defined by the rule that the walker steps up $(u(i)=+1)$ if a pyrimidine occurs at the i^{th} step along the DNA chain, whereas $u(i)=-1$ for a purine occuring at position i. Then, $F(l)$ is defined by the expression:

$$F(l) \equiv \left[\overline{[\Delta y(l)]^2} - [\overline{\Delta y(l)}]^2 \right]^{1/2} \qquad (1)$$

with

$$\Delta y(l) \equiv y(l_0+l)-y(l_0) \qquad (2)$$

and

$$y(l) \equiv \sum_{i=1}^{l} u(i) \qquad (3)$$

The bars represent an average over all l_0 positions in the DNA sequence.

They calculated the statistical quantity (1) for a variety of DNA samples and for different correlation lengths l and they derived the following conclusions: i) $F(l)$ obeys a power law:

$$F(l) \simeq l^{\alpha} \qquad (4)$$

ii) The slope $\alpha(l)=\frac{\log(F(l))}{\log(l)}$ remains constant for a wide range of l in all examined cases. iii) Exons or c-DNA sequences give $\alpha \cong 0.5$ which reflects a random walk and reveals absence of long-range correlations. iv) Introns, as well as non transcribed DNA sequences give again constant but higher than 0.5 value for $\alpha(l)$, indicating, as these authors conclude, a scale-invariant property of DNA.

The above hypothesis has provoked several comments and scepticism [2],[3],[4]. We have undertaken an investigation on the lines of the work of Peng et al. trying to determine the structure and properties of DNA sequences behaving as in eq.(4).

Some properties of the root mean square fluctuation function.

I. No significant difference is found between exon and intron sequences, in agreement to refs.[3],[4]. However other approaches lead to some degree of such a systematic difference [5].

II. The slope $\alpha(l)$ is not always l-independent [3,4].

III. We investigated the behavior of $\alpha(l)$ for several artificial DNA sequences. Our results are helpful in understanding what happens in natural DNA sequences:

(a). Any random sequence of purines and pyrimidines with equal probabilities ($P_{Pu}=P_{Py}$) gives $\alpha(l)=0.5$ (as expected).

(b). When $P_{Pu}\neq P_{Py}$, $\alpha(l)$ remains constant again for a wide range of l but smaller than 0.5.

(c). When P_{Pu} (or equivalently P_{Py}) is linearly position-dependent $\alpha(l)$ reaches approximately constant values higher than 0.5.

(d) When sequences with different and position independent probabilities (case (b)) are put together, $\alpha(l)$ in the enlarged sequence increases for low l and for high l it stabilizes to a value clearly higher of 0.5.

We see that in cases (c) and (d) the resulting $\alpha(l)$ is approximately constant and higher than 0.5. However, by construction, these sequences do not possess any inherent self-similarity or scale invariance in the clustering of their bases. We found that the genome of the lambda phage is such a naturally occurring case, since it consists of four regions with clear-cut properties from the point of view of purines-pyrimidines probabilities (see Fig.1a). Each one of the first two regions A and B present position independent but unequal probabilities. Each one by itself has a value of $\alpha(l)$ slightly higher than 0.5 (Fig 1b, curves A,B). However, when put together (A-B) or with the rest of the phage's genome (A-B-C-D) it becomes a typical case of the behavior (c) described above (see also ref.[3]). Moreover, the region D presents a graded distribution of bases and as a consequence, the corresponding $\alpha(l)$ (if taken alone) is relatively high.

n-dupletes' occurence and root mean square fluctuation function. In many cases of natural exon or intron sequences, $\alpha(l)$ is found to be remarkably high-valued (and l-independent) even if P_{Pu}, P_{Py} are approximately equal and position independent. In order to extract some information about the particularities of the structure of such sequences, we calculate (using a suitable computer program) the probabilities of appearance of the 2^n n-duplets in it. Notice that the calculation includes all possible reading frames (that is, partially superimposed n-duplets are counted separately). It is found that in cases with $P_{Pu} \cong P_{Py}$ (a common situation in DNA sequences) when $\alpha(1)>0.5$ not justified by the above cases, the n-duplet's probabilities differ significantly from the expected value, which is $1/2^n$.

In order to test the direct dependence of the high $\alpha(l)$ on the n-duplet's probability distribution we constructed a computer program forming an artificial nucleotide sequence by means of a random number generator and a set of 2^n probability values for the n-duplets of the resulting sequence. These 2^n numbers is convenient to be the corresponding probabilities of n-duplets of a naturally occuring DNA sequence. The algorithm, for every next pur. or pyr. added during the chain elongation, "pulls" a random number taking also into account the relative probabilities of the two possible n-duplets formed by the last n-1 nucleotides combined with the new. This is a purely "local" procedure which cannot endow the chain with any long-range structure formation, provided that the random number generator is unbiased. We used the Linear Congruential Method described by Knuth [8], and in order to test its suitability we verified that when the probability for every n-duplet is required to be $1/2^n$ the resulting sequence is characterized by $\alpha(l) \cong 0.5$.

In Fig.2 is presented with the line denoted by g the $\alpha(l)$ of the sequence of the Human homolog of Drosophila female sterile homeotic mRNA from the EMBO Databank (Code HSFSHG) while the line denoted by r presents a random sequence with Pur./Pyr. ratio equal to the gene's corresponding ratio. Curves 2, 5, 10, correspond to artificial sequences produced by our program and using the same random numbers we used for r and the probabilities of the 2-duplets, 5-duplets and 10-duplets of the g sequence respectively. Without going into details we state the conclusion:

The consideration of the n-duplet's probabilities for increasing n approximates better the curve of the initial sequence g. It is essestial that a considerable fraction of the correlation-measure [$\alpha(1)-0.5$] for the sequence g, can be reduced to n-duplets' probabilities for n less or equal to 10. This is a kind of behavior more complex than simple repetivity. However this situation too, does not imply self-similarity, fractal clustering, 1/f noise-type structure or other "authentic" long-range correlations, at least as long as this is mimicked by artificial sequences.

Conclusion

The above discussed results (see also refs.[2,6,7]) seem to indicate that the behavior of the correlation function (1) introduced by Peng et al. [1], may be significantly reducible to short-range interactions. It remains open the question whether there is a residual component with an "authentic" long-range structure of the DNA sequence. Another point for further elaboration is the consideration of 4 nucleotides in DNA sequences and the resulting n-duplet probabilities.

References

1.- Peng C.-K., Buldyrev S.V., Goldberger A.L., Havlin S., Sciortino F., Simons M., & Stanley H.E. Long-range correlations in nucleotide sequences. *Nature*, 356, p168-170, (1992).

2.- Nee S. Uncorrelated DNA walks. *Nature*, 357, p450, (1992).

3.- Prabhu V.V. & Clavert J.-M. Correlations in intronless DNA. *Nature*, 359, p782, (1992).

4.- Chatzidimitriou-Delsmann & Larhammar. Long-range correlations in DNA. *Nature*, 361, p782, (1992).

5.- Li W. & Kaneko K. Long Range correlation and Partial $1/f^{a}$ Spectrum in a Noncoding DNA Sequence. *Europhysics Letters*, 17, p655-660, (1992).

6.- Voss R.F. Evolution of Long-Range Correlations and 1/f noise in DNA base sequences. *Phys.Rev.Lett.* 68, p3805-3808, (1992).

7.- Munson P.J., Taylor R.C. & Mickaels G.S. DNA correlations. *Nature*, 360, p636, (1992).

8.- Knuth D.E. The art of computer programming. *Addison-Westley Publishing Compagny*, 1981.

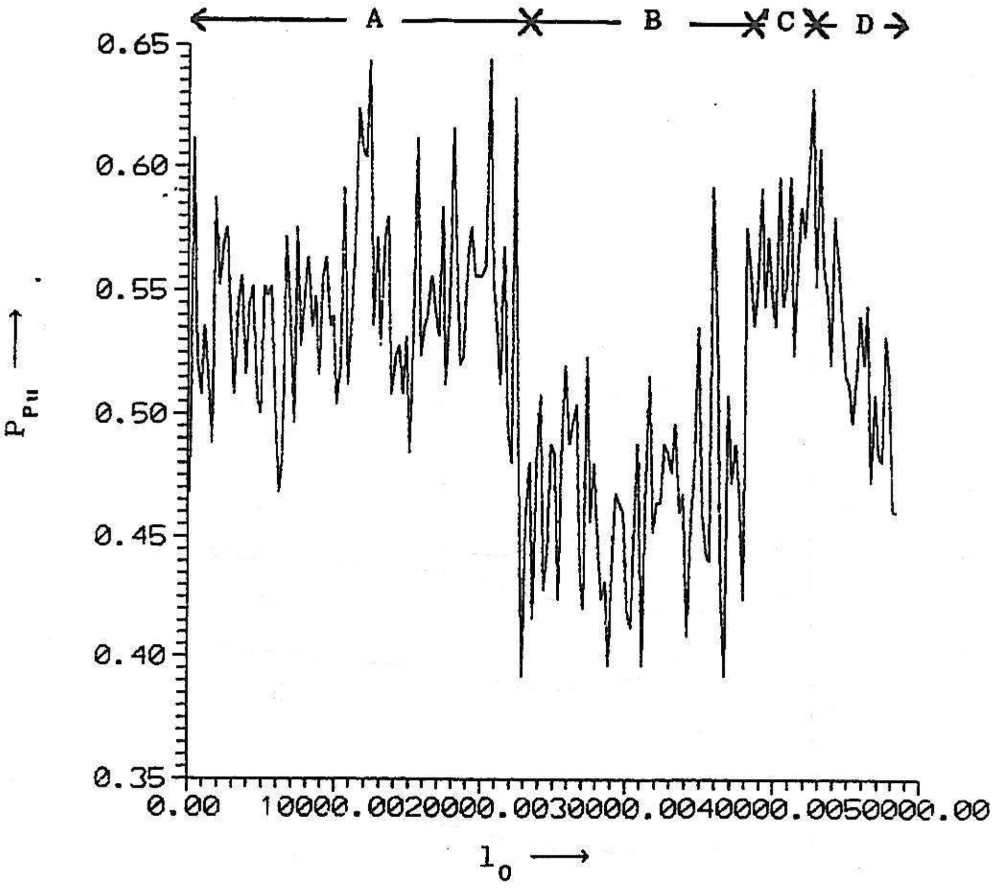

Figure 1a

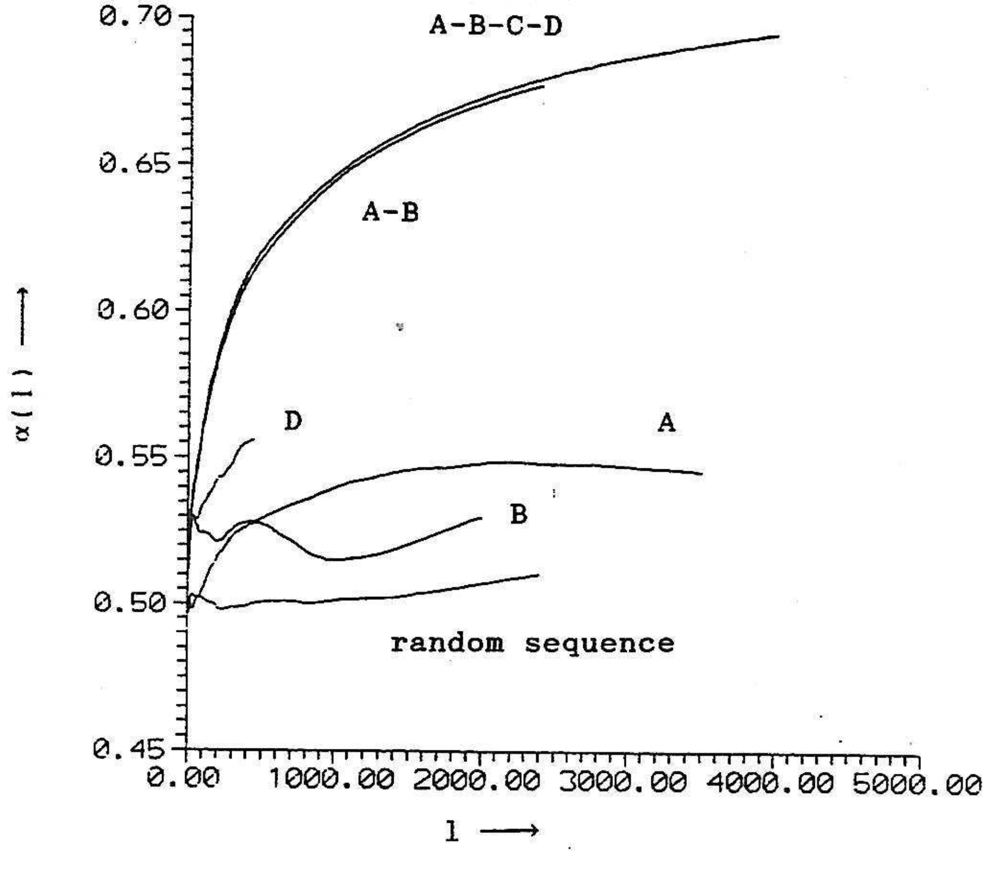

Figure 1b

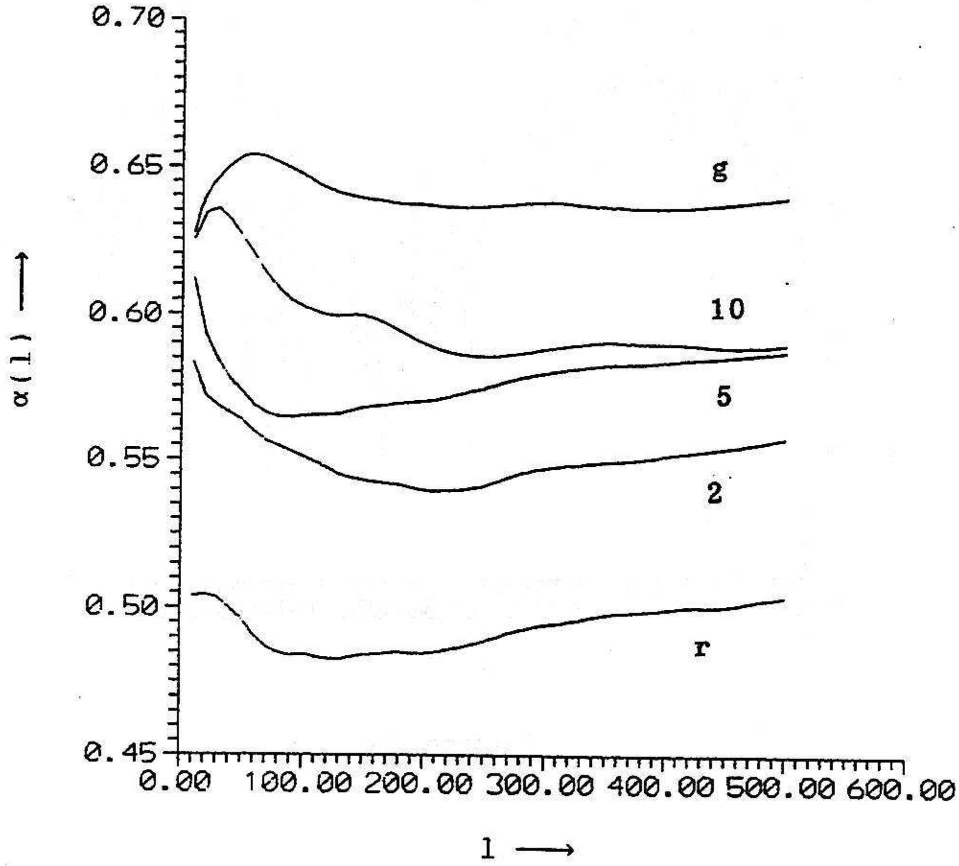
0.70
0.65
0.60
0.55
0.50
0.45
α(l) ⟶
0.00
100.00
200.00
300.00
400.00
500.00
600.00
l ⟶
g
10
5
2
r

Figure 2