*Indian J. Phys.* **73B** (2), 337-343 (1999)

**IJP B**
— an international journal

# Self-similarity and scaling exponent for DNA walk model in two and four dimensions

S Tarafdar*, P Nandy*, S Sahoo*, A Som*, J Chakrabarti* and A Nandy$

* Department of Physics, Jadavpur University, Jadavpur, Calcutta 700 032, India
* Department of Theoretical Physics, Indian Association for the Cultivation of Science, Jadavpur, Calcutta-700 032, India
$ Computer Division, Indian Institute of Chemical Biology, 4 Raja S C Mullick Road, Calcutta-700 032, India

**Abstract** . Recent investigations into characteristics of long DNA sequences have focused attention on the possible existence of fractal dimensions and long range correlations in such sequences Coupled with the search for patterns in base distribution, identification of coding and non-coding regions and the larger issue of indexation and classification of DNA sequences, these remain among the most challenging problems for molecular physics.

In this context, Xiao *et al* [1] and others, *e.g*, Jeffrey [2], have shown that self-similar patterns do exist in DNA sequences when mapped in specific manners, and that the fractal dimensions of introns and exons differ reflecting the differences in their structure and function In this paper, we consider the problem from slightly different perspective, analysing DNA sequences as a walk in a space of (a) two and (b) four dimensions as follows .

(a) The DNA sequence is mapped onto a two-dimensional metric space according to the prescription of Nandy [3], and we measure the length of the resulting DNA walk in different length scales. It is found that self-similarity exists and a scaling exponent can be defined which quantifies the "randomness" of the walk A marked persistence of the walk is observed for the intron segments. Systematics between different species is also noted

(b) In the second approach, we represent the DNA sequence as a directed walk in a four-dimensional metric where A, C, G, T represent the four coordinates and the four-dimensional length L(l) of the walk is calculated for different length scales l The difference between L(l) and the end-to-end distance Lo gives an idea about the correlation length of the sequence.

**Keywords** : DNA walk, graphical representation, scaling exponents

**PACS Nos.** : 87.10.+e, 87.15.-v

## 1. Introduction

The DNA sequence contains all relevant biological information of an organism in the form of a one-dimensional array of four bases : adenine (A), guanine (G), cytosine (C) and thymine (T). Identifying the presence of any pattern or any type of order in DNA sequences has been a long-standing but formidable problem, the elucidation of which is expected to facilitate comparative studies of DNA sequences and evolutionary signatures and consequences. Earlier work in this field have shown that long DNA sequences exhibit intriguing features of self-similarity or scale invariance [2, 4, 5]. Recently, Xiao *et al* [11] have done fractal studies of DNA sequences and reported significant differences in the fractal exponents of coding and non-coding regions and have claimed that this can be used as a signature to determine the nature of a raw sequence.

In this paper, we focus on techniques of fractal analysis and look for patterns or correlations by two approaches :

1. Calculating the scaling exponents in two-dimensional DNA walk model of DNA sequences [3] ;
2. Calculating the scaling exponent and identifying a characteristic length in a 4-dimensional directed DNA walk model. The question of long range correlations in DNA sequences that is a natural corollary of such an analysis is also addressed in this study.

In this analyses, we look for systematics in the scaling exponents for different genes, same genes for different species and for coding and non-coding regions.

## 2. Method

*Two dimensional DNA walk :*

A sequence can be mapped in 2-D by a "walk" where a step is taken in the negative x-direction for an A, a step in the positive y-direction for a C, a step in the positive x-direction for a G and a step in the negative y-direction for a T in the sequence. This generates a plot of the sequence reflecting the distribution of bases along the sequence as shown in the example in Figure 1 The length L of such a walk measured by joining points with different base intervals 'l' is found to obey a relation

$$L(l) \sim l^{D} \text{ for certain range of } l. \qquad (1)$$

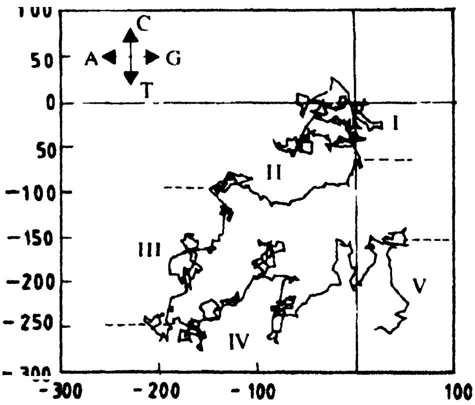

**Figure 1.** Example of a two-dimensional graphical representation of the rat skeletal myosin heavy chain gene (RNMHCG) mapped as stated in the text [7].

The value of D can be indicative of the presence or absence of long-range correlation. For D = 0.5 the walk is random with no correlation. This can be seen as follows : Let us take a sequence with $N_{tot}$ bases. We divide $N_{tot}$ into intervals of length l. The distance between two points separated by l is given by a power law

$$d(\mathrm{l}) \sim \mathrm{l}^{\alpha} \text{ for sufficiently large l}. \tag{2}$$

For a random walk we have the walk-known result, $\alpha = 0.5$. The number of intervals is $N_{tot}/\mathrm{l}$, so the total length of the 'walk' is

$$L(\mathrm{l}) \sim \mathrm{l}^{\alpha} N_{tot}/\mathrm{l}$$

$$\sim \mathrm{l}^{\alpha - 1} N_{tot} \, .$$

which is eq. (1) with $D = 1 - \alpha$. Thus, for the random walk where each step is independent of the previous step, $D = \alpha = 0.5$.

Departure of $\alpha$ towards a smaller or larger value indicates correlations. If the walk is persistent, *i.e.*, the probability of moving in the same direction as the previous step is higher than foi motion in other direction, then $\alpha > 0.5$, so $D < 0.5$. In the opposite case, *i.e.*, anti-persistence, the probability of continuing in the same direction is less compared to the probability of moving in the other directions. Here $\alpha < 0.5$ and $D > 0.5$. In both cases, the walker retains a memory of the previous step. This is correlation without bias (*i.e.* preference for moving in one particular direction).

It is to be noted that $D$ is different from the fractal dimension of the path representing the sequence. This would be obtained by measuring the track of the walker on different length scales to get a power law relation. Xiao *et al* [1] have done such a study.

*Four-dimensional DNA walk model :*

The previous approach shows interesting results reported elsewhere [6], but the mapping is dependent on the choice of co-ordinates and there are overlapping points. We can get a unique walk by mapping the sequence in four dimensions as a directed walk with $X_1(A)$, $X_2(G)$, $X_3(C)$ and $X_4(T)$ always increasing.

Our aim is to differentiate between walks – each starting at the origin $X_i = 0$ and ending at the same point $X_i(f)$ $i = 1, 2, 3, 4$ on the basis of the pattern of the walk, *i.e.*, the correlation, if any, in the definite sequence of the bases. Here we apply the prescription of Method 1 to

$$S(\mathrm{l}) = L(\mathrm{l}) - L_0 \, ,$$

where $$L_0 = \left[\Sigma X_i^2 (f)\right]^{1/2}, \tag{3}$$

thus eliminating the 'drift'.

We find that $S(\mathrm{l})$ *vs* l gives a very good power law fit with a characteristic exponent D' (Figure 2)

$$L - L_0 \sim \mathrm{l}^{-D'}$$

This work is still in progress. D' for several sequences are given in later.

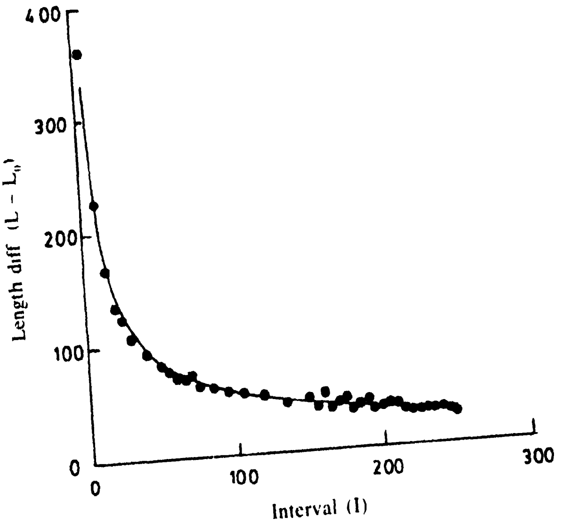

**Figure 2(a).** A typical plot of $L-L_0$ vs interval l for the total sequence of the chicken tubulin gene (GGTUB3B) with 2697 nucleotides including introns, exons and flanking regions

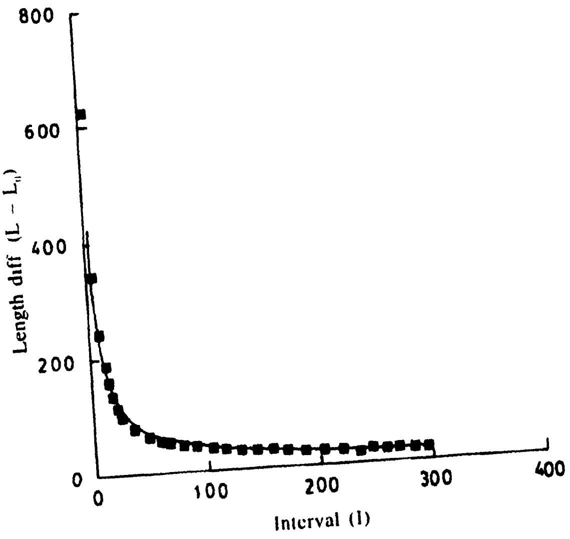

**Figure 2(b).** A typical plot of $L-L_0$ vs interval l for the coding sequence of the rat skeletal myosin heavy chain gene (RNMHCG) with 6034 nucleotides

*Characteristic Length :*

The 4-dimensional walk, Figure 3, reveals further features in the sequences. A sudden drop in the curve indicates the presence of a characteristic length above which the pattern appears more homogeneous. As an illustration, we show two walks, W1 and W2 on a two-dimensional map with only two types of bases A and C (Figure 3):

W1 ~ ACACACAC......................,

W2 ~ AAAACCCCAAAACCCC........ .

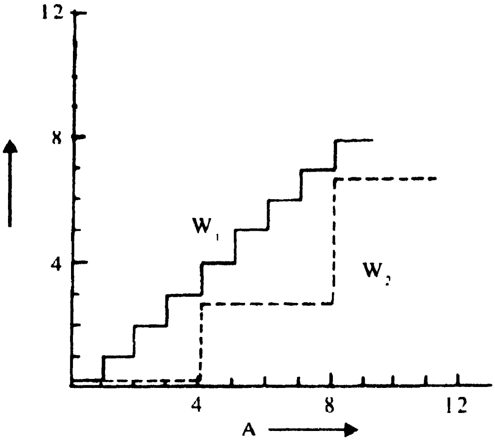

**Figure 3.** Two hypothetical walks to demonstrate the differences in L ( l ) values

For $W1$, $L(l)$ falls to $L_0$ as l increases from 1 to 2. But for $W2$, $L(l)$ has a constant value for $l = 1$ to $l = 4$ and drops to $L_0$ only when $l \sim 8$. So two walks having the four bases in equal numbers may be differentiated by defining a characteristic length $l_c$ where $S(l_c) = L(l_c) - L_0$ falls to some definite fraction of its initial value $S(1) = L(1) - L_0$. Assigning this fraction arbitrarily as 0.1, preliminary results indicate that $l_c \sim 100$ from the full sequence, but for the coding region along $l_c$ is remarkably lower, at around 50. For a meaningful comparison of course $L(n) = N_{tot}$ has to be equal in both cases.

Another interesting point is that $S(1)$ provides a measure of the bias, *i.e.* whether there is a preference for any particular base in the sequence or all four are equally represented. If A,C,G,T occur in equal numbers $L_0 = N_{tot}/2$, so $S(1) = N_{tot}/2$. A smaller value indicates an unequal distribution, the extreme case being all bases are of one kind, in which case $S(1) = 0$.

## 3. Results and discussion

Applying the scaling technique to the 2-dimensional graphical representations of different gene sequences we find the scaling exponents given in Table 1. The results have been averaged over different species for each type of gene in Table 1(a) and for each of the different kingdoms in Table 1(b). We note that the scaling exponents for the non-coding regions are always smaller than those for the coding regions, and both are smaller than the characteristic value for random walks, 0.5. By the definition given above, this would imply greater persistence in the case of non-coding regions, which arises from the fact that introns generally have larger repeats and duplications. We note in passing the obvious fact that sequences where the non-coding regions are significantly larger than the coding regions the values of the exponents for the total sequences including introns and flanking regions are greater than for the non-coding regions, but smaller than for the coding regions. The same differences between coding and non-coding regions are seen in the kingdomwise table also.

**Table 1(a).** Scaling exponents for different gene sequences

| Gene | All | cDNA | Non-coding |
|---|---|---|---|
| α-globin | 0.137 + 0.021 | 0.310 + 0.060 | 0.144 + 0.037 |
| β-globin | 0.259 + 0.038 | 0.425 + 0.039 | 0.248 + 0.060 |
| tubulin | 0.193 + 0.086 | 0.282 + 0.052 | 0.146 + 0.017 |
| histone H4 | 0.227 + 0.082 | 0.309 + 0.078 | |
| Heat Shock Protein | 0.264 | 0.376 | 0.220 |
| Myosin Heavy Chain (Invertebrates only) | 0.048 + 0.014 | 0.052 + 0.015 | |

**Table 1(b).** Scaling exponents for different kingdoms

| Kingdom | cDNA | Non-coding |
|---|---|---|
| Plants | 0.243 + 0.021 | |
| Avian | 0.329 + 0.069 | 0.163 + 0.027 |
| Amphibians | 0.412 + 0.002 | |
| Mammals | 0.354 + 0.060 | 0.157 + 0.065 |
| Invertebrates | 0.048 + 0.014 | 0.052 + 0.051 |
| Vertebrates | 0.353 + 0.063 | 0.161 + 0.054 |

In the case of the four-dimensional directed walks, the scaling exponent for the normal walk shows very little variations between the different genes (Table 2). However, when the effect of the directional walk is taken out by subtracting the gross length from the calculated length, the exponents show wide variations. However, contrary to the two-dimensional DNA walks, in this case the value of the exponent is almost always > 0.5.

**Table 2.** Scaling exponents for different genes in 4-D DNA walk model

| Gene | D' |
|---|---|
| α-globin | 0.649 |
| β-globin | 0.453 |
| tubulin | 0.598 |
| Heat Shock Protein | 0.568 |
| Myosin Heavy Chain | 0.652 |

In this case also, when $L - L_0$ is plotted against l, it shows a sharp drop from which, as remarked earlier, we define a correlation length $l_c$ (Figure 2). This characteristic length is found to be ~ 100 for the case of non-coding regions and ~ 50 for the coding regions.

## 4. Conclusion

Thus we find that the distribution of bases in a DNA sequence follows some sort of scale invariance leading to scaling exponents that are quite different from the case of random distributions. In fact, our observations from the two-dimensional graphical representation calculations show that while the distribution pattern in coding regions is close to random, the base distribution in non-coding regions contain a persistent pattern that results in low exponent values. The surprising result that the exponent for the coding regions in the four-dimensional model is greater than 0.5 can be understood from the differences in the methodology of plotting of points in the two models. Further work is in progress to understand these phenomena. The concept of characteristic length for coding and non-coding regions in the 4-D model is new and further work is being carried out in this area also.

## References

[1] Y Xiao, R Chen, R Shen, J Sun and J Xu *J Theor Biol* **175** 23 (1995)

[2] H J Jeffrey *Nucleic Acids Res* **18** 2163 (1990)

[3] A Nandy *Curr Sci* **66** 309 (1994)

[4] C-K Peng, S V Buldyrev, A L Goldberger, S Havlin, F Sciortino, M Simons and H E Stanley *Nature* (London) **356** 168 (1992)

[5] R Voss *Phys Rev Lett* **68** 3805 (1992)

[6] S Tarafdar, P Nandy, S Sahoo, A Som, J Chakrabarti, C Raychaudhury and A Nandy *National Symposium-cum-Workshop on Trends in Bioinformatics* (Bose Institute, Calcutta, March 24-27) (1998)

[7] A Nandy and P Nandy *Curr. Sci* **68** 75 (1995)