

## LONG-RANGE CORRELATIONS IN DNA SEQUENCES USING TWO-DIMENSIONAL DNA WALKS\*

Jin Chen<sup>d</sup>, Lin-xi Zhang<sup>a\*\*</sup> and De-lu Zhao<sup>b</sup>

<sup>a</sup>Department of Physics, Wenzhou Normal College, Wenzhou, 325027, China

<sup>b</sup>Polymer Physics Laboratory, Center of Molecular Science, Institute of Chemistry,  
Chinese Academy of Sciences, Beijing 100080, China

**Abstract** The characterization of long-range correlations and fractal properties of DNA sequences has proved to be a difficult though rewarding task mainly due to the mosaic character of DNA consisting of many patches of various lengths with different nucleotide constitutions. In this paper we investigate statistical correlations among different positions in DNA sequences using the two-dimensional DNA walk. The root-mean-square fluctuation  $F(l)$  is described by a power law. The autocorrelation function  $C(l)$ , which is used to measure the linear dependence and periodicity, exists a power law of  $C(l) \sim l^{-\mu}$ . We also calculate the mean-square distance  $\langle R^2(l) \rangle$  along the DNA chain, and it may be expressed as  $\langle R^2(l) \rangle \sim l^\gamma$  with  $2 > \gamma > 1$ . Our investigations can provide some insights into long-range correlations in DNA sequences.

**Keywords:** Autocorrelation function; DNA sequence; Long-range correlation.

### INTRODUCTION

The explosive accumulation of DNA sequence data in the last two decades has provided a rich source of raw material that hides valuable hints about the evolutionary mechanisms of the genome organization. Unveiling the patterns in those sequences has become an exciting challenge to the present generation of statistical physicists and information scientists. Recently there has been considerable interest in the finding of long-range (power-law) correlations in genomic DNA sequences<sup>[1]</sup>. Different techniques including mutual information functions<sup>[2]</sup>, autocorrelation functions<sup>[3]</sup>, power spectra<sup>[4]</sup>, “DNA walk” representation<sup>[5]</sup>, Zipf analysis<sup>[6]</sup> were used for statistical analysis of DNA sequences. For instance the method of “DNA walk” representation was widely used to paint the DNA polymer sequence clearly.

It is well known that DNA polymer sequence might be considered as a string of symbols (A, T, G, C) with a mosaic structure<sup>[7–9]</sup>. One of the main obstacles to long-range correlation analysis in DNA sequence is the mosaic structure of these sequences, which are well known to be formed of “patches” (“strand bias”) of different underlying composition. These patches appear as trends in the DNA walk landscapes and are likely to introduce some breaking of the scale invariance. In the past, many scientists had developed the one-dimensional DNA walk ideally. In order to apply numerical methods to a DNA sequence  $\{n_i\}$  consisting of the four nucleotides A, C, T and G, they generate a binary sequence  $\{u_i\}$  for each DNA sequence<sup>[8]</sup>. As we all know, the four types of nucleotides have different probabilities respectively and every nucleic acid has especial function. However, in the one-dimensional DNA walk, these methods may bury the unattached characteristics of four nucleotides. To improve the veracity of the DNA walk, we extend the one-dimensional DNA walk to the two-dimensional DNA walk. In the two-dimensional DNA walk, four directions mean the four types of nucleotides; therefore, we can

\* This work was financially support by the National Natural Science Foundation of China (Nos. 29874012, 20174036, 20274040), and Natural Science Foundation of Zhejiang Province (No. 10102).

\*\* Corresponding author: Lin-xi Zhang (章林溪), E-mail: lxzhang@hzncnc.com

Received September 15, 2003; Revised December 23, 2003; Accepted December 25, 2003

investigate long-range correlations in DNA sequences in more detail. Some different results can be obtained in our modified model.

## TWO-DIMENSIONAL DNA WALK

In order to study the scale-invariant long-range correlations of a DNA sequence, a common strategy to represent a given DNA sequence graphically consists of transforming it into a d-dimensional random walk, a so-called DNA walk, by associating a space direction to each nucleotide (A, T, G, C). Here, A, T, G and C represents adenine, thymine, guanine, and cytosine, respectively. In particular, a very popular representation is the one-dimensional walk. There are many alternative DNA walks and the most complete one that considers the four bases equivalently in base space is the three-dimensional DNA walk<sup>[10]</sup>, in the sense that its projections on appropriate axes or planes recover all possible one- and two-dimensional walks. Here we introduce a graphical representation of DNA sequence, which we term a two-dimensional DNA walk according to the one-dimensional DNA walk. This model may be more reasonable in investigating the statistical correlations in DNA sequences.

A “DNA walk” is initiated from the first nucleotide of the sequence and continued to the last nucleotide. In a DNA sequence  $\{n_i\}$ , we “define” the nucleotide G as the coordinate (1, 0); the nucleotide A as (-1, 0); the nucleotide C as (0, 1) and the nucleotide T as (0, -1) in two-dimensional Descartes coordinate system. To describe this method in more detail, supposing at first we are in the first nucleotide of one sequence, when we meet the nucleotide G, we walk a right step along the  $X$ -coordinate; when we meet the nucleotide A, we turn a step to the left along the  $X$ -coordinate again. Here the length of step is the unit 1. In the same way, we take the steps of up and down along the  $Y$ -coordinate respectively in the case of the nucleotide C and T. This representation preserves some basic symmetries of the DNA such as complementarity, reflection, substitution, and compatibility. For the case of an uncorrelated walk, the direction of each step is independent of the previous steps. For the case of a correlated random walk, the direction of each depends on the history (“memory”) of the walker. Finally we can get the curve of DNA sequence in the two-dimensional  $XY$ -coordinate. The DNA walk allows one to visualize directly the fluctuations of the four types nucleotide in DNA sequence. Visual observation of DNA walks suggests that there be a long-range correlation in DNA sequence.

## RESULTS AND DISCUSSIONS

### Scaling Concepts in DNA Walk

The DNA walk provides a graphical representation for each gene and permits the degree of correlation in the nucleotide sequence to be directly visualized. In the case of two-dimensional walk, the “net displacement” ( $\bar{y}$ ) of the walker after  $l$  steps is calculated to analyze the correlation, which is the sum of the unit steps  $u_i$ , and the expression is listed here:

$$\bar{y}(l) = \sum_{k=1}^l a(k)\bar{i} + \sum_{k=1}^l b(k)\bar{j} \quad (1)$$

Here  $a(k)$  and  $b(k)$  are the values of  $X$ -coordinate and  $Y$ -coordinate respectively after  $k$  steps. ( $l$  is a integer between  $l$  and  $L$ , and  $L$  is the length of the sequence).

As we all know that in a DNA chain sequence the number of the four types of nucleotides A, T, G and C is so large that it is difficult for us to get the rule on distribution of all these nucleotides clearly. Here a DNA chain is represented as a DNA walk chain. In order to investigate the scaling relation in DNA walk, we define the mean square distance  $\langle R^2(l) \rangle$

$$\langle R^2(l) \rangle = \frac{\sum_{l_0=1}^{L-l} R^2(l)}{L-l} \quad (2)$$

with

$$\bar{R}(l) = \bar{y}(l_0 + l) - \bar{y}(l_0) \quad (3)$$

( $l_0 = 1, 2, 3, \dots, L-l$ ).  $l_0$  is the distance along a DNA sequence.

The double logarithmic plots of the mean-square distance  $\langle R^2(l) \rangle$  as a function of the linear distance  $l$  along the DNA chain for *Escherichia coli* genomic DNA (ECO110 K, 111401 bp) and human Tcr-C-delta gene (HUMTCRADCY, 97630 bp) are shown in Fig. 1. In Fig. 1, we get nearly ideal linear curve, and the slope  $\gamma$  of this linear line is 1.530 and 1.670 for ECO110 K and HUMTCRADCY, respectively. The difference of the slope  $\gamma$  between the two DNA sequences is 0.140, and it is not too high. At the same time we also calculate the more DNA sequences including coding and non-coding sequences, and the results are given in the last column of Table 1. The minimum value  $\gamma$  lies in human dystrophin (HUMDYS, 13957 bp) and the value is 1.395, and the maximum value is 1.910 for dictyostelium discoideum myosin (DDIMYHC, 6680 bp). Therefore, we can get the scaling law as follows:

$$\langle R^2(l) \rangle \sim l^\gamma \quad (4)$$

with the value  $\gamma$  ranges from 1.395 to 1.910. Most of them are greater than 1.500, and this means that the DNA walk chain is different from either random-flight walk chain or self-avoiding walk chain. The slope  $\gamma$  of each DNA chain is different shows the fact that there exists a difference of DNA walk for different DNA sequence. However, for each DNA chain, there exists a good scaling relation from the fact that there is a good linear relationship between  $\ln \langle R^2(l) \rangle$  and  $\ln l$ . This result can help understand the property of DNA walk.

**Table 1.** Values of the scaling exponents  $H_t$  and  $\gamma$  for a set of different coding and non-coding sequences

| Sequence                               | Code       | Comments | Length | $H_t$ | $\gamma$ |
|----------------------------------------|------------|----------|--------|-------|----------|
| Non-coding                             |            |          |        |       |          |
| <i>Caenorhabditis Elegans</i> Myosin   | CELMYUNC   | Gene     | 9000   | 0.736 | 1.444    |
| Human Tcr-C-delta                      | HUMTCRADCY | Gene     | 97630  | 0.700 | 1.670    |
| Human $\beta$ -cardiac MHC             | HUMBMYH7   | Gene     | 28452  | 0.599 | 1.411    |
| <i>Drosophila Melanogaster</i> Myosin  | DROMHC     | Gene     | 22663  | 0.756 | 1.655    |
| Chicken embryonic myosin               | CHKMYHE    | Gene     | 31111  | 0.704 | 1.762    |
| Coding                                 |            |          |        |       |          |
| <i>Drosophila melanogaster</i> MHC     | DROMYONMA  | cDNA     | 6338   | 0.753 | 1.785    |
| Human $\beta$ -cardiac MHC             | HUMBMYH7CD | cDNA     | 6008   | 0.770 | 1.880    |
| Human dystrophin                       | HUMDYS     | cDNA     | 13957  | 0.694 | 1.395    |
| <i>Escherichia coli</i>                | ECO110K    | cDNA     | 111401 | 0.671 | 1.530    |
| <i>Dictyostelium discoideum</i> myosin | DDIMYHC    | Gene     | 6680   | 0.768 | 1.910    |

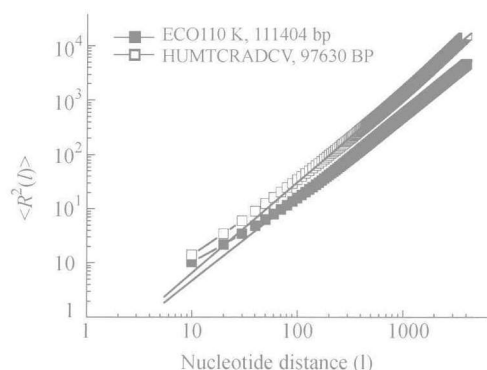


**Fig. 1** Double logarithmic plots of  $\langle R^2(l) \rangle$  versus  $l$  for ECO110 K (slope,  $\gamma = 1.530$ ); and HUMTCRADCY (slope,  $\gamma = 1.670$ )

### Long-range correlation of root mean square fluctuation $F(l)$

An important statistical quantity that characterizes the fluctuations of the average of the displacement of a quantity  $\bar{R}(l)$  defined by  $\bar{R}(l) = \bar{y}(l_0 + l) - \bar{y}(l_0)$  (see Eq. 3) is the root mean square fluctuation  $F(l)$  defined as:

$$F^2(l) = \overline{[\bar{R}(l) - \bar{\bar{R}}(l)]^2} = \overline{[\bar{R}(l)]^2} - [\bar{\bar{R}}(l)]^2 \quad (5)$$

We expect  $F(l)$  to increase with increasing  $l$  (here  $l$  is the distance along a DNA sequence). If there is no characteristic length, then fluctuations will be described by a power law

$$F(l) \sim l^{H_r} \quad (6)$$

The state is similar to the one-dimensional walk.  $H_r \neq 0.5$ . The case  $H_r = 0.5$  represents the absence of long-range correlations in the case of random and short-range correlated sequences. Any value different from 0.5 is the evidence of the existence of long-range correlation in the sequence.

The double logarithmic plots of the mean square fluctuation function  $F(l)$  as a function of the linear distance  $l$  along the DNA sequence for *Escherichia coli* genomic DNA (ECO110 K, 111401 bp) with the slope  $H_r = 0.671$ , and for Human Tcr-C-delta gene (HUMTCRADCV, 97630 bp) with the slope  $H_r = 0.700$  are shown in Fig. 2. Two good linear curves are found in Fig. 2. In Table 1, values of the scaling exponent  $H_r$  for a set of different coding and non-coding sequences are listed. We can get the conclusion that the data for all these DNA sequences are linear on double logarithmic plots and it confirms the theory  $F(l) \sim l^{H_r}$ . The minimum value  $H_r$  is 0.599 in human  $\beta$ -cardiac MHC (HUMBMVH7, 28452 bp) and the maximum value is 0.768 in dictyostelium discoideum myosin (DDIMYHC, 6680 bp). All the values of the curve slope are greater than 0.5. A least-squares fit produces a straight line with slope  $H_r$  substantially larger than 0.5, thus providing direct experimental evidence for the presence of long-range correlations. That means the long-range correlations obviously affect the DNA sequences, including coding and non-coding sequences.

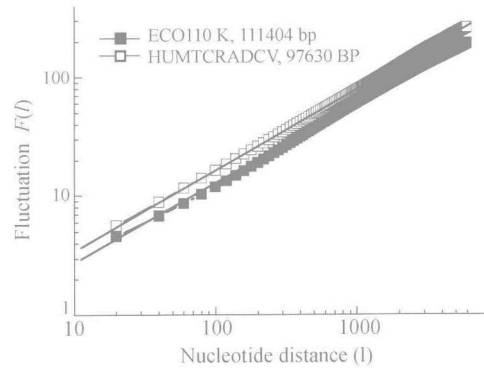


**Fig. 2** Double logarithmic plots of  $F(l)$  versus  $l$  for ECO110 K (slope,  $H_r = 0.671$ ), and HUMTCRADCV (slope,  $H_r = 0.700$ )

### The autocorrelation function $C(l)$

The autocorrelation function has been widely used in physics and signal theory as a measure of linear dependence and periodicity. The application of this measure to DNA-sequence analysis, although used previously, became “popular” in 1992 with the finding of power-law correlations in DNA sequences implying the presence of a high complexity and scale invariance in the heterogeneity of those sequences<sup>[1, 4, 11]</sup>. We define the autocorrelation function of the DNA sequence at distance  $l$  as

$$C(l) = \frac{1}{\sigma^2} l \frac{1}{L-l} \sum_{i=1}^{L-l} \tilde{y}_i \tilde{y}_{i+l} - \frac{1}{(L-l)^2} \sum_{i=1}^{L-l} \tilde{y}_i \sum_{i=1}^{L-l} \tilde{y}_{i+l} l \quad (7)$$

Here,  $\sigma^2$  is given as

$$\sigma^2 = \frac{1}{L-l} \sum_{i=1}^{L-l} (\bar{y}_{i+l} - \bar{y}_i)^2 - \left[ \frac{1}{L-l} \sum_{i=1}^{L-l} (\bar{y}_{i+l} - \bar{y}_i) \right]^2 \quad (8)$$

In fact, the autocorrelation function measures the deviation of  $\langle \bar{y}_i \bar{y}_{i+l} \rangle$  from  $\langle \bar{y}_i \rangle \langle \bar{y}_{i+l} \rangle$  (where the brackets denote the average along the sequence). If the deviation is zero, there are no linear correlations between the values of the sequence at position  $i$  ( $\bar{y}_i$ ) and the values at the distance  $l$  ( $\bar{y}_{i+l}$ ). The greater this deviation, the stronger the linear correlations between values separated a distance  $l$ .  $C(l) > 0$  means that the probability that  $\bar{y}_i$  and  $\bar{y}_{i+l}$  reach the same or similar values is higher than in a random sequence, whereas for  $C(l) < 0$  this probability is smaller than that expected in a random sequence. The definition of  $C(l)$  makes the implicit assumption of stationarity since it includes the variance of the whole sequence.

In Fig. 3, we plot the double logarithmic plots of autocorrelation function  $C(l)$  as a function of the nucleotide distance  $l$  using two-dimensional DNA walk for bacillus halodurans genomic DNA (AP001519, 303650 bp), *Escherichia coli* genomic DNA (ECO110 K, 111401 bp) and human chromosomes (NT\_011362, 2000000 bp). Here the curves are linear nicely with the slope of  $-1.455$ ,  $-1.355$ , and  $-1.655$ , respectively. We find that the solid line can be expressed in the form of

$$C(l) \sim l^{-\mu} \quad (9)$$

The reason why the autocorrelation function of  $C(l)|_{l=1}$  is greater than unit is that  $l \frac{1}{L-l} \sum_{i=1}^{L-l} \bar{y}_i \bar{y}_{i+l} - \frac{1}{(L-l)^2} \sum_{i=1}^{L-l} \bar{y}_i \sum_{i=1}^{L-l} \bar{y}_{i+l} \gg \sigma^2|_{l=1}$ . The value of  $C(l)|_{l=1}$  also increases with increasing the length of DNA sequences.

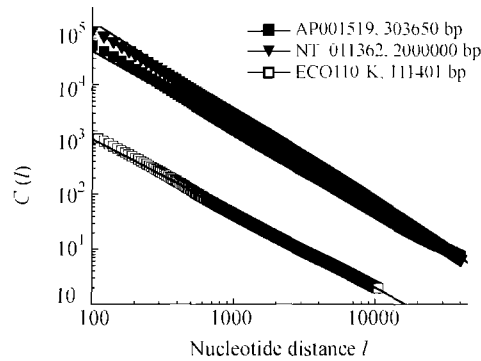


**Fig. 3** Double logarithmic plots of  $C(l)$  versus  $l$  for AP00151 (slope =  $-1.455$ ); ECO110 K (slope =  $-1.355$ ); and NT\_011362 (slope =  $-1.655$ )

### Summary

In order to use the standard techniques to study the correlations of DNA sequences, it is necessary first to eliminate the local biases in nucleotide composition, thus avoiding spurious effects due to the mosaic character of the sequence. In the two-dimensional DNA walk, four directions mean the four types of nucleotides; therefore, we can investigate long-range correlations in DNA sequences in more detail. The root-mean-square fluctuation  $F(l)$  can be described by a power law  $F(l) \sim l^{H_r}$  with slope  $H_r$  substantially larger than  $1/2$  for all DNA sequences, the autocorrelation function  $C(l)$  exists a power law  $C(l) \sim l^{-\mu}$  and the mean-square distance  $\langle R^2(l) \rangle$  along the DNA chain may be expressed as  $\langle R^2(l) \rangle \sim l^\gamma$  with  $2 > \gamma > 1$ . These investigations enable us to better understand the DNA sequences, and it seems more meaningful to predict and analyze the unrecognized DNA chains for their information content.

## REFERENCES

- 1 Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M. and Stanley, H.E., *Nature*, 1992, 356: 168
- 2 Azbel, M.Y., *Phys. Rev. Lett.*, 1995, 75: 168
- 3 Herzel, H. and Grobe, I., *Physica A.*, 1995, 216: 518
- 4 Voss, R.F., *Phys. Rev. Lett.*, 1992, 68: 3805; *Fractals*, 1994, 2: 1
- 5 Galvan, P.B., Carpena, P., Roldan, R.R. and Oliver, J.L., *Gene.*, 2002, 300: 105
- 6 Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havinm, S., Peng, C.K., Simons, M. and Stanley, H.E., *Phys. Rev. E.*, 1995, 52: 2939
- 7 Barbardi, G., Olofsson, B., Filipski, J., Zeral, M., Salinas, J. and Cuny, G., *Meunier-Rotival, Rodier, M., F. Sci.*, 1985, 228: 953
- 8 Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.K., Simon, M. and Stanley, H.E., *Phys. Rev. E.*, 1995, 51: 5084
- 9 Sun, T.T., Zhang, L.X., Chen, J. and Jiang, Z.T., *Chaos, Solitons, & Fractals*, 2004, 20: 1075
- 10 Luo, L., Lee, W., Jia, L., Ji, F. and Tsai, L., *Phys. Rev. E.*, 1998, 58: 861
- 11 Li, W. and Kaneko, K., *Europhys. Lett.*, 1992, 17: 555