Available online at www.sciencedirect.com
SCIENCE DIRECT
Physica A 369 (2006) 688–698
PHYSICA A
www.elsevier.com/locate/physa

A two-dimensional modified Lévy-walk model for the DNA sequences

Yuan-Yen Tai, Ping-Cheng Li*, Hsen-Che Tseng

Department of Physics, National Chung-Hsing University, 250 Guo-Kuang Road, Taichung 402, Taiwan, ROC

Received 25 October 2005; received in revised form 6 February 2006
Available online 22 March 2006

Abstract

A two-dimensional modified Lévy-walk model for the DNA sequences with only three parameters is proposed. This model is applied to simulate real DNA sequence data of the *Saccharomyces cerevisiae* genome. DNA sequences are converted into one- and two-dimensional random walk mappings for statistical study. Satisfactory statistical results are obtained when 16 chromosomes in the genome are all analyzed and compared to the model sequence data. Measurements on the root-mean-square deviations are in good agreement when real sequences and model sequences are compared. Lévy type statistics plays a major role in generating "long-range correlation" phenomena. Power values $\alpha = 0.64-0.68$ are obtained for the power law $F(l)\sim l^{\alpha}$ where $F(l)$ is the root-mean-square deviation at step size l. The same model is applied to sequences of some other species. Fair agreement is also obtained.

Keywords: DNA sequences; DNA-walk; Lévy statistics; Long-range correlation; Patchiness; Mapping; *Saccharomyces cerevisiae* genome

1. Introduction

As DNA sequence data becomes accessible through the Internet the study of statistical patterns in DNA sequences draw much attention to the scientific community [1–5]. An understanding of organization and evolution of life on the genome level may no longer be unattainable. Many authors analyzed DNA sequence data in a variety of ways. In 1992 C.-K. Peng of Boston University along with his colleagues published their well-known paper [1] and long-range correlations between the nucleotides were clearly shown for intron-containing genes and nontranscribed regulatory DNA sequences through the so-called 'DNA-walk' analysis. Despite the continuing debates of their biological interpretations the power law on a wide range of scales extending from tens to thousands of nucleotides appears to be enthralling. An immediate concern of whether the appearance of long-range power law correlations could arise from the mosaic organization of "patches" (excess of one type of nucleotides) [6,7] draw much attention from many researchers. In response to this issue C.-K. Peng et al. [8] made a thorough study of two classes of controls consisting of patchy nucleotide sequences generated by different algorithms (artificially generated sequences, one with and one without

*Corresponding author. Tel.: +886 4 22840427x714; fax: +886 4 22862534.
E-mail address: li@phys.nchu.edu.tw (P.-C. Li).

0378-4371/$ - see front matter
doi:10.1016/j.physa.2006.02.013

long-range power-law correlations) and conclude that patchiness itself is not sufficient to account for the observed long-range correlation properties in noncoding regions. In C.-K. Peng's papers [1,8] long-range correlations can be found in the intron-containing genes but not in the cDNA or intronless segments.

The patchiness structure revealed in the 'DNA-walk' analysis is itself an interesting feature. Its appearance is the issue of non-uniform concentration of different kinds of nucleotides and should bear some important biological meanings. Buldyrev et al. [9] has designed a generalized Lévy-walk model to generate a model sequence which is in many ways similar to the statistics obtained from the empirical sequence data. Their model contains two parameters. One is for the power law feature in the standard Lévy statistics and the other is to simulate randomness within some chosen "Lévy-walk" steps. Long-range correlations in noncoding DNA sequences and long sub-regions of biased walks within these correlated sequences are well accounted for in this model. While celebrating success is anticipated it is nevertheless a dichotomous model which reckons the whole DNA sequence as composed of two different kinds of nucleotides, viz. pyrimidines and purines. As we know for certain a DNA sequence contains four different kinds of nucleotides. A view of modeling these sequences in term of four directions in their "random-walk" motion would certainly be a more natural way of description. Abramson et al. [10] analyzed DNA sequences of the organism *Saccharomyces cerevisiae* (baker's yeast) in a two-dimensional mapping which produced a two-dimensional random-walk picture for study. A superposition of a Lévy-walk and white noise suffices to demonstrate the mean-square displacement dependence on the number of steps and a superdiffusive behavior is seen in these sequences.

Motivated by Guillermo's idea we look into a more general application of surveying sequences of various kinds as well as genomes of different species. If by a few parameters we are able to model sequences in a wide general data set it might give a possibility of putting things into categories. Hopefully some clues of biological information carried and hidden in these sequences may start to make their appearances.

In order to associate DNA sequences with a two-dimensional mapping we have to assign each nucleotide a direction either going along the x-axis in positive or negative direction or going along the y-axis in the same fashion. Totally there are 24 combinations but only three of them which we listed below as F1, F2 and F3 are truly independent mappings. Other combinations are just repetitions of these three forms with different orientations and do not carry further statistical information:

$$\begin{aligned}
&\mathrm{F1}: \mathrm{A}(\leftarrow), \mathrm{T}(\rightarrow), \mathrm{C}(\uparrow), \mathrm{G}(\downarrow),\\
&\mathrm{F2}: \mathrm{A}(\leftarrow), \mathrm{T}(\downarrow), \mathrm{C}(\uparrow), \mathrm{G}(\rightarrow),\\
&\mathrm{F3}: \mathrm{A}(\leftarrow), \mathrm{T}(\uparrow), \mathrm{C}(\rightarrow), \mathrm{G}(\downarrow).
\end{aligned}$$

A, T, C and G are, respectively, capitals of four different nucleotides (adenine, thymine, cytosine and guanine). Four different arrows denote directions along the x- or y-axis. DNA-walk graphic mappings of the chromosome I of *S. cerevisiae* are shown in Fig. 1. The overall diffusive direction in the first quadrant of the F1 mapping is the manifestation of the excess of (T, C) concentration over that of (A, G). The non-uniform concentration of the (A, T) set over the (C, G) set is more evident in the F2 and F3 mappings. Appling the definition of the mean-square deviation [1,8] as in the Guillermo's paper [10], we obtain from the F1 mapping exactly the same results as theirs. The mean-square deviation is defined as

$$F^2(l) = \langle \Delta \mathbf{r}^2 \rangle - \langle \Delta \mathbf{r} \rangle^2, \tag{1}$$

where $\Delta \mathbf{r}(l) = \mathbf{r}(l_0 + l) - \mathbf{r}(l_0)$ is the difference in position for the walker moves l steps downward while taking a initial position at step l_0. The averages are calculated through the entire track in the mapping. Typical results of the *S. cerevisiae* chromosome I are shown in Fig. 2. Superdiffusive nature is present in all F1, F2 and F3 mappings. Their slopes were all measured with values ranging from 0.68 to 0.70 which are clearly larger than 0.5, a value of a normal diffusive process.

2. Gauss–Lévy-walk model (GLWM)

In viewing these results an immediate idea is to further construct a model, based upon previous studies but somewhat different ways of modeling, which is able to simulate similar patterns as we see in the mappings and at the same time exhibit all kinds of statistical data at hand. Through inspections of various kinds of DNA