Complexities of human promoter sequences

Fangcui Zhao\textsuperscript{a}, Huijie Yang\textsuperscript{b,c,}\textsuperscript{*}, Binghong Wang\textsuperscript{d}

\textsuperscript{a}College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022, China
\textsuperscript{b}Department of Physics, National University of Singapore, Science Drive 2, Singapore 117543, Singapore
\textsuperscript{c}School of Management, University of Shanghai for Science and Technology, and Shanghai Institute for Systematic Science, Shanghai 200093, China
\textsuperscript{d}Department of Modern Physics, University of Science and Technology of China, Anhui Hefei 230026, China

Received 21 September 2006; received in revised form 28 March 2007; accepted 29 March 2007
Available online 1 April 2007

Abstract

By means of the diffusion entropy approach, we detect the scale-invariance characteristics embedded in the 4737 human promoter sequences. The exponent for the scale-invariance is in a wide range of \(0.3 \leq \delta_c \leq 0.9\), which centered at \(\delta_c = 0.66\). The distribution of the exponent can be separated into left and right branches with respect to the maximum. The left and right branches are asymmetric and can be fitted exactly with Gaussian form with different widths, respectively.

\(\copyright\) 2007 Elsevier Ltd. All rights reserved.

Keywords: Putative promoter sequences; Diffusion entropy; Complexity

1. Introduction

Understanding gene regulation is one of the most exciting topics in molecular genetics (Ohler and Niemann, 2001). Promoter sequences are crucial in gene regulation. The analysis of these regions is the first step towards complex models of regulatory networks.

A promoter is a combination of different regions with different functions (Werner, 1999; Pedersen et al., 1999; Zhang, 2002; Narang et al., 2005). Surrounding the transcription start site is the minimal sequence for initiating transcription, called core promoter. It interacts with RNA polymerase II and basal transcription factors. Few hundred base pairs upstream of the core promoter are the gene-specific regulatory elements, which are recognized by transcription factors to determine the efficiency and specificity of promoter activity. Far distant from the transcription start site there are enhancers and distal promoter elements which can considerably affect the rate of transcription. Multiple binding sites contribute to the functioning of a promoter, with their position and context of occurrence playing an important role. Large-scale studies show that repeats participate in the regulation of numerous human and mouse genes (Rosenberg and Jolicoeur, 1997). Hence, the promoter’s biological function is a cooperative process of different regions such as the core promoter, the gene-specific regulatory elements, the enhancers/silencers, the insulators, the CpG islands and so forth. But how they cooperate with each other is still a problem to be investigated carefully.

The structures of DNA sequences determine their biological functions (Buldyrev et al., 1995). Recent years witness an avalanche of finding nontrivial structure characteristics embedded in DNA sequences. Detailed works show that the noncoding sequences carry long-range correlations (Peng et al., 1992; Li and Kaneko, 1992; Yang et al., 2002). The size distributions of coding sequences and noncoding sequences obey Gaussian or exponential and power-law (Provata and Almirantis, 1997, 2000), respectively. Theoretical model-based simulations (Yang et al., 1994; Salerno, 1991; Lennholm and Homquist, 2003; Kalosakas et al., 2004) tell us that the parts of the promoters where the RNA transcription has started are more active than a random portion of the DNA. By means of the nonlinear modeling method it is found that along the
putative promoter regions of human sequences there are some segments much more predictable compared with other segments (Yang et al., 2006). All the evidences suggest that the nontrivial structure characteristics of a promoter determine its biological functions. The statistical properties of a promoter may shed light on the cooperative process of different regions.

Experimental knowledge of the precise 5' ends of cDNAs should facilitate the identification and characterization of regulatory sequence elements in proximal promoters (Trinklein et al., 2003). Using the oligocapping method, Suzuki et al. (2002) identify the transcriptional start sites from cDNA libraries enriched in full-length cDNA sequences. The identified transcriptional start sites are available at the Database, http://dbtss.hgc.jp/. Consequently, Leonardo et al. (2004) have used this data set and aligned the full-length cDNAs to the human genome, thereby extracting putative promoter regions (PPRs). Using the known transcriptional start sites from over 5700 different human full-length cDNAs, a set of 4737 distinct PPRs are extracted from the human genome. They have also counted eight-letter words within the PPRs, using z-scores and other related statistics to evaluate the over- and under-representations.

In this paper, by means of the concept of diffusion entropy (DE) we try to detect the scale-invariant characteristics in these putative promoter regions.

2. DE analysis

The DE method is firstly designed to capture the scale-invariance embedded in time series (Grigolini et al., 2001; Scafetta et al., 2001; Scafetta and Grigolini, 2002). To keep the description as self-contained as possible, we review briefly the procedures. We consider a PPR denoted with $Y = (y_1, y_2, \ldots, y_{3001})$, where $y_i$ is the element at the position $i$ and $y_i = A, T, C$ or $G$. Replacing $A, T$ and $C, G$ with $-1$ and $+1$, respectively, the original PPR is mapped to a time series $X = (x_1, x_2, \ldots, x_{3001})$. There is not a trend in this series, i.e., $X$ is stationary.

Connecting the starting and the end of $X$, we can obtain a set of delay-register vectors, which reads:

$$
\begin{align*}
T_1(t) &= (x_1, x_2, \ldots, x_t), \\
T_2(t) &= (x_2, x_3, \ldots, x_{t+1}), \\
&\vdots \\
T_{3001}(t) &= (x_{3001}, x_1, \ldots, x_{t-1}).
\end{align*}
$$

Regarding each vector as a trajectory of a particle in duration of $t$ time units, all the vectors can be described as a diffusion process of a system containing 3001 particles. The initial state of the system is

$$
\begin{pmatrix}
T_1(0) \\
T_2(0) \\
\vdots \\
T_{3001}(0)
\end{pmatrix} =
\begin{pmatrix}
0 \\
0 \\
\vdots \\
0
\end{pmatrix}.
$$

Accordingly, at each time step $t$ we can calculate displacements of all the particles. The probability distribution function (PDF) of the displacements can be approximated with $p(m, t) \sim K_m/3001$, where $m = -t, -t + 1, \ldots, t$ and $K_m$ is the number of the particles whose displacements are $m$. It can represent the state of the system at time $t$.

As a tenet of complexity theory (Bar-Yam, 1997; Mandelbrot, 1988), complexity is related with the concept of scaling invariance. For the constructed diffusion process, the scaling invariance is defined as

$$
p(m, t) \approx \frac{K_m}{3001} = \frac{1}{m} p\left(\frac{m}{m}\right),
$$

where $\delta$ is the scaling exponent and can be regarded as a quantitative description of the PPR’s complexity. If the elements in the PPR are positioned randomly, the resulting PDF obeys a Gaussian form and $\delta = 0.5$. Complexity of the PPR is expected to generate a departure from this ordinary condition, that is, $\delta \neq 0.5$.

The value of $\delta$ can tell us the pattern characteristics of a PPR. The departure from the ordinary condition can be described with a preferential effect. Let the element is $A, T$ (or $C, G$), the preferential probability for the following element’s being $A, T$ (or $C, G$) is $W_{\text{pre}}$. A positive preferential effect, i.e., $W_{\text{pre}} > 0.5$, leads to the value of $\delta$ larger than 0.5. While a negative preferential effect, i.e., $W_{\text{pre}} < 0.5$, can induce the value of $\delta$ smaller than 0.5. Hence, a large value of $\delta$ implies that $A, T$ or $C, G$ accumulate strongly in a scale-invariance way, respectively.

However, correct evaluation of the scaling exponent is a nontrivial problem. In literature, variance-based method is used to detect the scale-invariance. But the obtained Hurst exponent $H$ may be different from the real $\delta$, that is, generally we have $H \neq \delta$. And for some conditions, the variance is divergent, which leads the invalidation of the variance method at all. To overcome these shortages, the Shannon entropy for the diffusion can be used, which reads

$$
S(t) = -\sum_{m=-t}^{t} p(m, t) \ln p(m, t)
= -\sum_{m=-t}^{t} \frac{K_m}{3001} \ln \left(\frac{K_m}{3001}\right).
$$

This diffusion-based entropy is called DE. A simple computation leads the relation between the scaling invariance defined in Eq. (2) and the DE as

$$
S(t) = A + \delta \ln t,
$$

where $A = -\sum_{m=-t}^{t} \frac{K_m}{3001} \ln \left(\frac{K_m}{3001}\right)$.
where $A$ is a constant depends on the PDF. Detailed works show that DE is a reliable method to search the correct value of $\delta$, regardless the form of the PDF (Scafetta and West, 2003; Scafetta et al., 2002; Yang et al., 2004, 2005).

The complexity in the PDF can be catalogued into two levels (Pipek and Varga, 1992), the primary one due to the extension of the probability to all the possible displacements $m$, and the secondary one due to the internal structures. Consequently, we should consider also the corresponding shuffling sequences as comparison.

3. Results and discussions

The DEs for all the 4737 PPRs are calculated. As a typical example, Fig. 1 presents the DE results for the PPRs numbered 1, 1000, 2000 and 3000. In considerable wide regions of $t$, the curves of DE can be fitted almost exactly with the linear relation in Eq. (4).

For each PPR, there exists an interval, $t_0 - \Delta t$, in which the PDF behaves scale-invariance. Keeping simultaneously the standard deviation and the error of the scaling exponent for the fitting result in the range of $\leq 0.05$ and $\leq 0.03$, we can find the maximum intervals $\Delta t$ for all the PPRs. In the fitting procedure, the confidence level is set to be 95%. The distribution of $\Delta t$, as shown in Fig. 2, tells us that generally the scale-invariance can be found over two to three decades of the scale $t$. The concept of DE is based upon statistical theory, that is, $t_0$ should be large enough so that the statistical assumptions are valid. To cite an example, we consider a random series, whose elements obey a homogenous distribution in $[0, 1]$. Only the length of the delay-register vectors, $t$, in Eq. (1) is large enough, the corresponding PDF for the displacements, i.e., the summation value of each delay-register vector, approaches the Gaussian distribution. Consequently, $t_0$ is not a valuable parameter. The values of $t_0$ for different PPRs are not presented.

As shown in Fig. 3, the resulting scaling exponent $\delta$ distributes in a wide range of $[0.3, 0.9]$. The distribution can be separated into two branches with respect to the center $\delta_c = 0.66$. The two branches are asymmetric and can be fitted exactly with the Gaussian function, respectively. The widths and centers of the left and right branches are $(w_{\text{left}}, x_{\text{left}}) = (0.17, 0.67)$, $(w_{\text{right}}, x_{\text{right}}) = (0.10, 0.65)$. That is to say, the centers coincide with each other, $w_{\text{left}} \approx w_{\text{right}} \approx \delta_c = 0.66$. Comparatively, the right branch distributes in a significant narrow region.

![Fig. 1. Typical DE results. The results for the PPRs numbered 1, 1000, 2000 and 3000 are presented. In considerable wide regions of $t$, the curves of DE can be fitted almost exactly with the linear relation in Eq. (4).](image)

![Fig. 2. Distribution of the maximum interval $\Delta t$ in which one can find scale-invariant characteristics. Keeping the standard deviation of the fitting result in the range of $\leq 0.05$, we can find the maximum intervals $\Delta t$ for all the PPRs. The distribution tells us that generally the scale-invariance can be found over two to three decades of the scale $t$.](image)

![Fig. 3. The resulting scaling exponent $\delta$ distributes in a wide range of $[0.3, 0.9]$. The distribution can be separated into two branches with respect to the center $\delta_c = 0.66$.](image)
The centers coincide with each other, regarded as the departure from the ordinary condition, obey Gaussian form with different widths, respectively. A complexity index includes two asymmetric branches, which distribute in a wide range of $[0.3, 0.9]$. The distribution can be separated into two branches with respect to the center $\delta_c = 0.66$. The two branches are asymmetric and obey exactly the Gaussian function, respectively. The widths and centers of the left and right branches are $(w_{\text{left}}, x_{\text{left}}) = (0.17, 0.67)$, $(w_{\text{right}}, x_{\text{right}}) = (0.10, 0.65)$. The centers coincide with each other, $w_{\text{left}} \approx w_{\text{right}} \approx \delta_c = 0.66$. The right branch distributes in a significant narrow region.

In summary, by means of the DE method, we calculate the complexities of the 4737 PPRs. The distribution of the complexity index includes two asymmetric branches, which obey Gaussian form with different widths, respectively. A stochastic evolution model may provide us a comprehensive understand of these characteristics.

Acknowledgments

This work is funded by the National Natural Science Foundation of China under Grant Nos. 70571074, 10635040 and 70471033, by the National Basic Research Program of China (973 Program) under Grant No. 2006CB705500, by the President Funding of Chinese Academy of Science, and by the Specialized Research Fund for the Doctoral Program of Higher Education of China. One of the authors (H. Yang) would like to thank Prof. Y. Zhuo for stimulating discussions.

References


