Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 247 (2007) 645–649
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Complexities of human promoter sequences

Fangcui Zhao[a], Huijie Yang[b,c,*], Binghong Wang[d]

[a]College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022, China
[b]Department of Physics, National University of Singapore, Science Drive 2, Singapore 117543, Singapore
[c]School of Management, University of Shanghai for Science and Technology, and Shanghai Institute for Systematic Science, Shanghai 200093, China
[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, 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.

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

*Corresponding author. School of Management, University of Shanghai for Science and Technology, and Shanghai Institute for Systematic Science, Shanghai 200093, China.
E-mail addresses: yangzhaon@eyou.com (F. Zhao), huijieyangn@eyou.com (H. Yang).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.03.035

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. Each PPR consists nucleotides from −2000 to +1000 bp, relative to the corresponding transcriptional start site. 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_s is the element at the position s and $y_s=A,T,C$ 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,

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

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

$$p(m,t)\approx\frac{K_m}{3001}=\frac{1}{t^\delta}F\left(\frac{m}{t^\delta}\right), \tag{2}$$

where δ 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, $\delta\neq0.5$.

The value of δ 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_{pre}. A positive preferential effect, i.e., $W_{pre}>0.5$, leads to the value of δ larger than 0.5. While a negative preferential effect, i.e., $W_{pre}<0.5$, can induce the value of δ smaller than 0.5. Hence, a large value of δ 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 δ, 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

$$\begin{aligned} S(t)&=-\sum_{m=-t}^{t}p(m,t)\ln p(m,t)\\ &=-\sum_{m=-t}^{t}\frac{K_m}{3001}\ln\left(\frac{K_m}{3001}\right). \end{aligned} \tag{3}$$

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

$$S(t)=A+\delta\ln t, \tag{4}$$

where A is a constant depends on the PDF. Detailed works show that DE is a reliable method to search the correct value of δ, 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, $t_0 \sim t_0 + \Delta t$, 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 $\leqslant 0.05$ and $\leqslant 0.03$, we can find the maximum intervals Δt for all the PPRs. In the fitting procedure, the confidence level is set to be 95%. The distribution of Δt, 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 $\delta \pm 0.03$ 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 $(w^{left}, x_c^{left}) = (0.17, 0.67)$, $(w^{right}, x_c^{right}) = (0.10, 0.65)$. That is to say, the centers coincide with each other, $w^{left} \approx w^{right} \approx \delta_c = 0.66$. Comparatively, the right branch distributes in a significant narrow region.

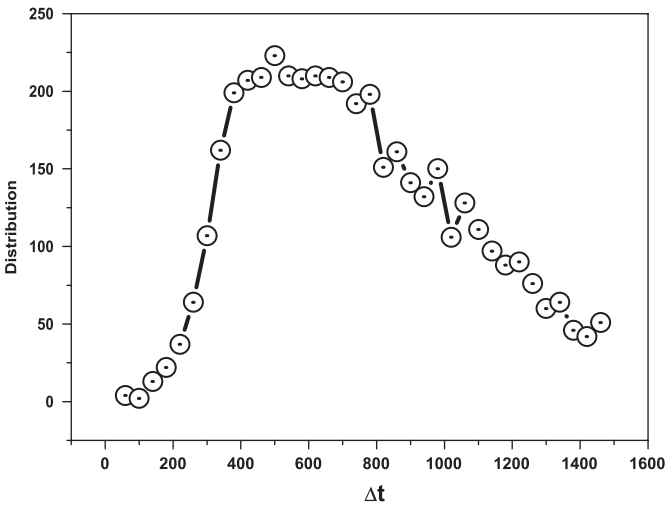

Fig. 2. Distribution of the maximum interval Δt in which one can find scale-invariant characteristics. Keeping the standard deviation of the fitting result in the range of $\leqslant 0.05$, we can find the maximum intervals Δt for all the PPRs. The distribution tells us that generally the scale-invariance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

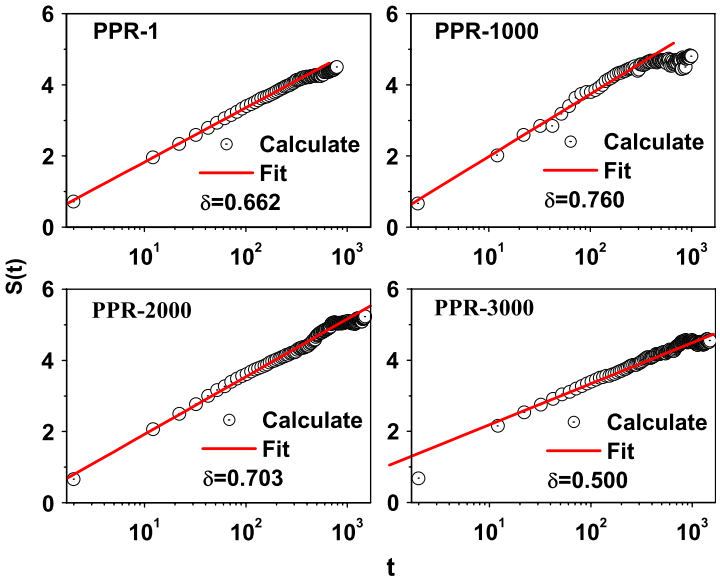

Fig. 1. (Color online) Typical DE results. The results for the PPRs numbered $1, 1000, 2000$ and 3000 are presented. In considerable wide regions of t, the curves of DE can be fitted almost exactly with the linear relation in Eq. (4).

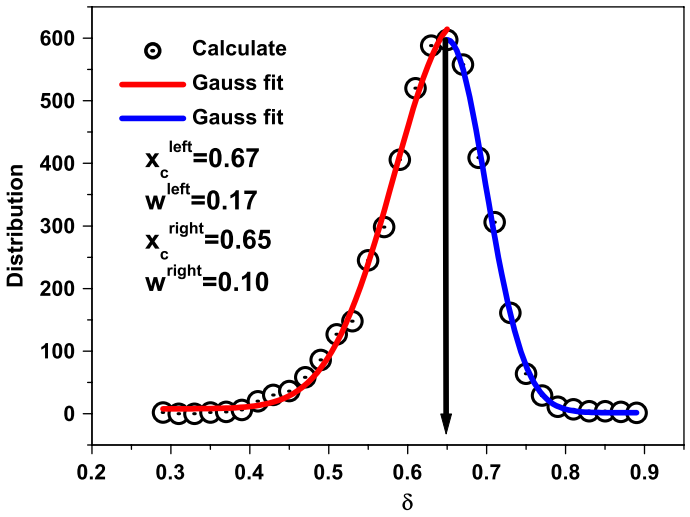

Fig. 3. (Color online) The complex index δ distributes in a wide range of $[0.3, 0.9]$. The distribution can be separated into two branches with respect to the center $\delta_c = 0.66$. The two branches are asymmetric and obey exactly the Gaussian function, respectively. The widths and centers of the left and right branches are $(w^{left}, x_c^{left}) = (0.17, 0.67)$, $(w^{right}, x_c^{right}) = (0.10, 0.65)$. The centers coincide with each other, $w^{left} \approx w^{right} \approx \delta_c = 0.66$. The right branch distributes in a significant narrow region.

The PPRs are shuffled also. For each PPR, the shuffling result is obtained by averaging over 10 shuffling samples. The scaling exponents are almost same, i.e., $\delta_{shuffling} = 0.5 \pm 0.03$. The detected scale-invariant characteristics are internal-structure-related.

How to understand the asymmetric characteristic of the distribution of the complexity index δ is an interesting problem. In literature, some statistical characteristics of DNA sequences are captured with evolution models, such as the long-range correlations and the over- and under-representation of strings and so on (Hsieh et al., 2003; Kloster, 2005; Messer et al., 2005). From the perspective of evolution, perhaps the distribution characteristics may favor a stochastic evolution model. The initial sequences have same complexity $\delta^{initial} = \delta_c = 0.66$. With the evolution processes the sequences diffuse along two directions, increasing complexity and decreasing complexity, i.e., the index δ increases and decreases, respectively. The diffusion coefficients for the two directions are significantly different, denoted with $D^{left} \neq D^{right}$. Based upon the widths of the two branches we can estimate that $D^{left}/D^{right} = \delta^{left}/\delta^{right} = 1.7$. It should be noted that, the complexity is regarded as the departure from the ordinary condition, $\delta = 0.5$. In the totally 4737 values of δ, only a small portion of them are less than 0.5. Accordingly, the PPRs may be catalogued into two classes, the PPRs with high complexity and the PPRs with low complexity. The former class evolves averagely with a slow speed while the later one with a high speed.

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

Bar-Yam, Y., 1997. Dynamics of Complex Systems. Addison-Wesley, Reading, MA.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.-K., Simons, M., Stanley, H.E., 1995. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys. Rev. E 51, 5084–5091.

Grigolini, P., Palatella, L., Raffaelli, G., 2001. Complex geometry, patterns, and scaling in nature and society. Fractals 9, 439–449.

Hsieh, L.-C., Luo, L., Ji, F., Lee, H.C., 2003. Minimal model for genome evolution and growth. Phys. Rev. Lett. 90, 018101.

Kalosakas, G., Rasmussen, K.O., Bishop, A.R., 2004. Sequence-specific thermal fluctuations identify start sites for DNA transcription. Europhys. Lett. 68, 127–133.

Kloster, M., 2005. Analysis of evolution through competitive selection. Phys. Rev. Lett. 95, 168701.

Lennholm, E., Homquist, M., 2003. Revisiting Salerno's sine-Gordon model of DNA: active regions and robustness. Physica D 177, 233–241.

Leonardo, M.-R., John, L.S., Gavin, C.K., David, L., 2004. Statistical analysis of over-represented words in human promoter sequences. Nucleic Acids Res. 32, 949–958 See also ⟨ftp://ftp.ncbi.nlm.nih.gov/pub/marino/published/hs_promoters/fasta/⟩.

Li, W., Kaneko, K., 1992. Long-range correlations and partial $1/f - \alpha$ spectrum in a noncoding DNA sequence. Europhys. Lett. 17, 655.

Mandelbrot, B.B., 1988. Fractal Geometry of Nature. W.H. Freeman, San Francisco, CA.

Messer, P.W., Arndt, P.F., Lassig, M., 2005. Solvable sequence evolution models and genomic correlations. Phys. Rev. Lett. 94, 138103.

Narang, V., Sung, W.-K., Mittal, A., 2005. Computational modeling of oligonucleotides positional densities for human promoter prediction. Artf. Intell. Med. 35, 107–119.

Ohler, U., Niemann, H., 2001. Identification and analysis of eukaryotic promoters: recent computational approaches. Trends Genet. 17, 56–60.

Pedersen, A.G., Baldi, P., Chauvin, Y., Brunak, S., 1999. The biology of eukaryotic promoter prediction—a review. Comput. Chem. 23, 191–207.

Peng, C.K., Buldyrev, S., Goldberger, A., Havlin, S., Sciortino, F., Simons, M., Stanley, H.E., 1992. Long-range correlations in nucleotide sequences. Nature 356, 168–171.

Pipek, J., Varga, I., 1992. Universal classification scheme for the spatial-localization properties of one-particle states in finite, d-dimensional systems. Phys. Rev. A 46, 3148–3163.

Provata, A., Almirantis, Y., 1997. Scaling properties of coding and non-coding DNA sequences. Physica A 247, 482.

Provata, A., Almirantis, Y., 2000. Fractal cantor patterns in the sequence structure of DNA. Fractals 8, 15–27.

Rosenberg, N., Jolicoeur, P., 1997. Retroviral pathogenesis. In: Coffin, J.M., et al. (Eds.), Retroviruses. Harbor Press, Cold Spring, pp. 475–586.

Salerno, M., 1991. Discrete model for DNA-promoter dynamics. Phys. Rev. A 44, 5292–5297.

Scafetta, N., Grigolini, P., 2002. Scaling detection in time series: diffusion entropy analysis. Phys. Rev. E 66, 036130.

Scafetta, N., West, B.J., 2003. Solar flare intermittency and the Earth's temperature anomalies. Phys. Rev. Lett. 90, 248701.

Scafetta, N., Hamilton, P., Grigolini, P., 2001. The thermodynamics of social processes: the teen birth phenomenon. Fractals 9, 193–208.

Scafetta, N., Latora, V., Grigolini, P., 2002. Levy scaling: the diffusion entropy analysis applied to DNA sequences. Phys. Rev. E 66, 031906.

Suzuki, Y., Yamashita, R., Nakai, K., Sugano, S., 2002. DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. Nucleic Acids Res. 30, 328–331.

Trinklein, N.D., Aldred, S.J., Saldanha, A.J., Myers, R.M., 2003. Identification and functional analysis of human transcriptional promoters. Genome Res. 13, 308–312.

Werner, T., 1999. Models for prediction and recognition of eukaryotic promoters. Mamm. Genome 10, 168–175.

Yang, H., Zhuo, Y., Wu, X., 1994. Investigation of thermal denaturation of DNA molecules based upon non-equilibrium transport approach. J. Phys. A 27, 6147–6156.

Yang, H., Zhao, F., Zhuo, Y., et al., 2002. Analysis of DNA chains by means of factorial moments. Phys. Lett. A 292, 349–356.

Yang, H., Zhao, F., Qi, L., Hu, B., 2004. Temporal series analysis approach to spectra of complex networks. Phys. Rev. E 69, 066104.

Yang, H., Zhao, F., Zhang, W., Li, Z., 2005. Diffusion entropy approach to complexity for a Hodgkin–Huxley neuron. Physica A 347, 704–710.

Yang, H., Zhao, F., Gu, J., Wang, B., 2006. Nonlinear modeling approach to human promoter sequences. J. Theor. Biol. 241, 765–773.

Zhang, M.Q., 2002. Computational methods for promoter recognition. In: Jiang, T., Xu, Y., Zhang, M.Q. (Eds.), Current Topics in Computational Molecular Biology. MIT Press, Cambridge, MA, pp. 249–268.