J Biol Phys (2016) 42:99–106
DOI 10.1007/s10867-015-9399-7

ORIGINAL PAPER

In search of coding and non-coding regions of DNA sequences based on balanced estimation of diffusion entropy

Jin Zhang[1,2] · Wenqing Zhang[1] · Huijie Yang[1]

Received: 28 March 2015 / Accepted: 30 July 2015 / Published online: 29 August 2015

Abstract Identification of coding regions in DNA sequences remains challenging. Various methods have been proposed, but these are limited by species-dependence and the need for adequate training sets. The elements in DNA coding regions are known to be distributed in a quasi-random way, while those in non-coding regions have typical similar structures. For short sequences, these statistical characteristics cannot be extracted correctly and cannot even be detected. This paper introduces a new way to solve the problem: balanced estimation of diffusion entropy (BEDE).

Keywords BEDE · Coding regions · Diffusion entropy · Non-coding regions · Self-similar structure · Time series

1 Introduction

In a eukaryotic DNA sequence, genes are usually subdivided into many segments called coding sequences (exons) and non-coding sequences (introns). Identifying coding regions in the analysis of DNA sequences is a major challenge in contemporary biology. Interdisciplinary research has contributed new methods for solving this problem, such as Markovian

✉ Jin Zhang
zdypaper@163.com

[1] Business School, University of Shanghai for Science and Technology, Shanghai 200093, China

[2] School of Information Science and Engineering, University of Jinan, Jinan 250022, China

approximations [1], correlation functions, and Fourier transform [2, 3]. However, these methods have two drawbacks. One is that they are species-dependent: to identify the coding regions in the DNA sequence of a species, it is necessary to construct a training set based on organism-specific data, which cannot be extended to other species. The other is the scale of this training set: only a sufficiently large training set can guarantee accuracy. Therefore, it is important to develop measures that are independent of species and training sets. Several novel methods have been proposed, such as entropy segmentation, the NM method, and mutual information function [4–6]. Unfortunately, these methods cannot determine both borders of each coding region, or cannot determine them precisely, so an effective technique is still urgently required.

One successful alternative is to construct all possible segments of a specified length from a stationary time series. When the length is treated as the time duration, each segment can be accepted as the trajectory of a particle starting from the original point. Then, the time series is mapped to an ensemble with the trajectories being realizations of stochastic motion. From the distribution function of the displacement, one can calculate the Shannon entropy, which Scafetta et al. called "diffusion entropy" (DE) [7]. DE is a powerful method for evaluating the scaling invariance embedded in time series in diverse fields, such as solar activity [8], the spectra of complex networks [9], physiological signals [10], and finance [11].

However, a time series of short length can induce large statistical fluctuations or bias in physical quantities, such as probability, moment, and entropy. The change of scaling from the short to long time region [12, 13] may be related to the large fluctuations. That is, the original DE method sometimes underestimates the value of the scaling exponent, or cannot even detect the scaling behavior due to the bias changing as the scale increases. To overcome this, the original form of the entropy can be replaced by a balanced estimator of DE (BEDE), which can evaluate the scale invariance in very short time series with considerable precision. In recent papers, BEDE was applied to detect scaling properties and structural breaks in stock price series on the Shanghai stock market [14], to evaluate scaling behaviors in heartbeat series for different sleep stages, and to assess stride time series for normal, fast, and slow walkers [15]. Here, we use BEDE (see Section 2.3) to find the coding and non-coding regions of DNA sequences; the results indicate that it reliably recognizes both borders.

With regard to the gene sequence of yeast, the Hurst exponents of the coding regions have a 0.5 error range and the nucleotides in them follow a stochastic distribution. In comparison, the Hurst exponents of non-coding regions are significantly larger than 0.5, i.e., they exhibit statistical features of long-range correlation and self-similar structure. We can obtain values of BEDE for several DNA segments by using a sliding window along the sequence. These values not only contribute to identifying coding and non-coding regions, but also help to determine the boundaries of these regions using differences in the statistical characteristics of both sides of the boundaries. The analysis correctly identified 15 of 19 recommended coding regions for an accuracy rate of 79%.

2 Materials and methods

2.1 DNA sequence of yeast

Yeast DNA sequence data (BK006948.2,1-48000bp) were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/), including 19 coding regions.

2.2 Diffusion entropy

Let us define a stationary time series $\{X_1, X_2, \cdots, X_{N-s+1}\}$ based on phase space construction for a segment of DNA sequence of length N [14]:

$$\begin{aligned} X_1 &= \{x_1, x_2, \cdots, x_s\} \\ X_2 &= \{x_2, x_3, \cdots, x_{s+1}\} \\ &\cdots\cdots \\ X_{N-s+1} &= \{x_{N-s+1}, x_{N-s+2}, \cdots, x_N\} \end{aligned}$$

where X_i represents the ith nucleotide, which is 1 (0) when the nucleotide of the ith position is C or G (A or T). Each vector X_i can be regarded as a case related to s particles. In other words, successive cases cover the entire DNA segment and reflect some statistical characteristics to an extent. The sum of displacements for each case is set by

$$Y_k(s) = \sum_{q=1}^{s} x_{k+q-1}, \quad k = 1, 2, \cdots, N-s+1 \tag{1}$$

After dividing the interval in which the displacements occur into $M(s)$ bins, $n(k, s)$ displacements occur in each bin, $k = 1, 2, \cdots, M(s)$. The displacement probability distribution function can be approximated by the following equation

$$p(j, s) \sim \hat{p}(j, s) = \frac{n(j, s)}{N-s+1}, \quad j = 1, 2, \cdots, M(s) \tag{2}$$

The consequent approximation of the Shannon entropy reads

$$S_{DE}(s) \sim S_{DE}^{naive}(s) = -\sum_{k=1}^{M(s)} \hat{p}(j, s) \ln[\hat{p}(j, s)]. \tag{3}$$

Provided that the stochastic process behaves in a self-similar way, we have

$$\hat{p}(j, s) \sim \frac{1}{s^{\delta}} F\{\frac{\min[Y(s)] + (j-0.5) \cdot \varepsilon(s)}{s^{\delta}}\} \quad j = 1, 2, \cdots, M(s) \tag{4}$$

Plugging (5) into (4) leads to

$$S_{DE}(s) = -\int_{-\infty}^{+\infty} F(y) \ln[F(y)] dy + \delta \ln(s) = A + \delta \ln(s) \tag{5}$$

where $A = -\int_{-\infty}^{+\infty} F(y) \ln[F(y)] dy$ is a constant. Therefore, the characteristics of a self-similar structure can be identified by calculating the DE and parameter δ can be determined. $\{X_1, X_2, \cdots, X_{N-s+1}\}$ exhibits long-range correlation if $0.5 < \delta < 1$ and behaves as a random walk if $\delta = 0.5$ [14, 16].

2.3 Balanced estimation of diffusion entropy

Unfortunately, replacing $p(j, s)$ with $\hat{p}(j, s)$ can lead to statistical and systematic error. Because $\hat{p}(j, s)$ is an unbiased estimate of $p(j, s)$, we define $\mu(j, s) = \frac{\hat{p}(j,s) - p(j,s)}{p(j,s)}$. After careful calculation, we have

$$S_{DE}(s) = S_{DE}^{naive}(s) + \frac{M(s)-1}{2(N-s+1)} + o(M(s)). \tag{6}$$

The error $\frac{M(s)-1}{2(N-s+1)}$ is negligible only when $N - s \to \infty$. Let us define $S_{DE}[p(j,s)] = -p(j,s)\ln[p(j,s)]$; then $S_{DE}(s) = \sum_{j=1}^{M(s)} S_{DE}[p(j,s)]$. Our goal is to find an acceptable estimation $\hat{S}_{DE}(s) = \sum_{j=1}^{M(s)} \hat{S}_{DE}[n(j,s)]$ for minimizing the systematic error $\Delta^2_{bias}(s) = (\langle \hat{S}_{DE}(s)\rangle - S_{DE}(s))^2$ and statistical error $\Delta^2_{stat}(s) = \langle(\hat{S}_{DE}(s) - \langle \hat{S}_{DE}(s)\rangle)^2\rangle$. To achieve this, we need to minimize the following function

$$\Delta^2(j,s) = \int_0^1 dp(j,s) \cdot w[p(j,s)] \cdot [\Delta^2_{bias}(j,s) + \Delta^2_{stat}(j,s)] \tag{7}$$

by means of $\frac{\partial \Delta^2(j,s)}{\partial \hat{S}_{DE}[n(j,s)]} = 0$, where

$$\Delta^2_{bias}(j,s) = (\langle \hat{S}_{DE}[n(j,s)]\rangle - S_{DE}[p(j,s)])^2 \tag{8}$$

$$\Delta^2_{stat}(j,s) = \langle(\hat{S}_{DE}[n(j,s)] - \langle \hat{S}_{DE}[n(j,s)]\rangle)^2\rangle \tag{9}$$

and $w[p(j,s)]$ is a weight function. Here, for convenience, we assume $w[p(j,s)] = 1$. After some complicated computations [14] we have

$$\hat{S}_{DE}(s) = \frac{1}{N-s+3} \sum_{j=1}^{M(s)} [N_j(s)+1] \cdot \sum_{k=N_j(s)+2}^{N-s+3} \frac{1}{k} \tag{10}$$

which is the balanced estimation of DE (BEDE).

3 Results and discussion

3.1 Identification of coding regions and non-coding regions

Recent research has demonstrated that long-range correlation and self-similar patterns exist in the non-coding regions of a gene and that the four bases A, C, G, and T are distributed randomly in the coding regions. Using this statistical difference, we can identify non-coding and coding regions.

A sliding window of length N is run along the DNA sequence. Each region that is covered by the window corresponds to one DNA segment of length N. First, we need to transform the segment into a stationary time series. Next, the DE in this region on different scales is calculated and parameter δ is determined. $0.5 < \delta < 1$ indicates that the region that is covered by the window belongs to the non-coding region class and δ=0.5 indicates that the region belongs to the coding region class (see Section 2.2). Note that the values of the DE are determined by BEDE (see Section 2.3). Figure 1 shows the difference in the statistical characteristics of the coding and non-coding regions. In this analysis, 15 of 19 coding regions were identified correctly, for a precision of 79%.

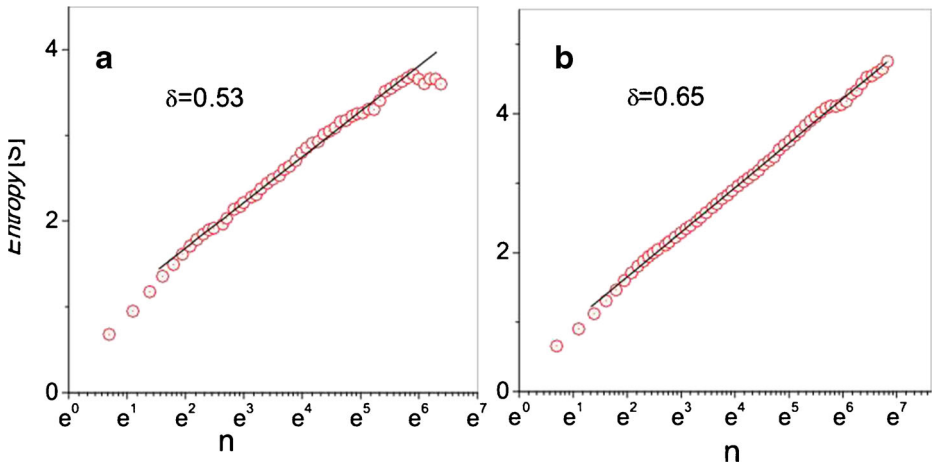

Fig. 1 Differences between coding and non-coding regions. **a** $\delta = 0.53$ corresponds to a coding region and **b** $\delta = 0.65$ corresponds to a non-coding region

3.2 Identification of coding region borders

This section describes how to identify the borders of the coding and non-coding regions. When the sliding window spans a coding region and a non-coding region, we have [17, 18]

$$\Delta P(t) = \sqrt{\frac{\sum\limits_{s} (P_{0s} - P_{ts})^2}{\sum\limits_{s} 1}} \tag{11}$$

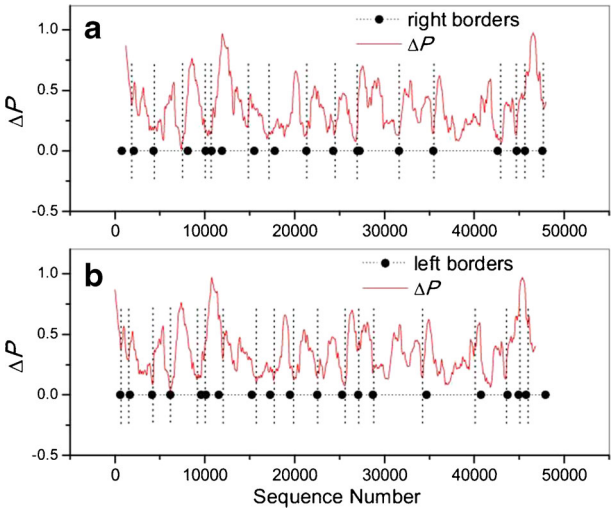

Fig. 2 The coding region 6175–7374 bp is used as the reference segment. *Black dots* represent true borders. **a** *right* and **b** *left* borders of the DNA sequence from yeast

where s is the length of X_i in time series $\{X_1, X_2, \cdots, X_{N-s+1}\}$(see Section 2.2), P_{0s} and P_{ts} are the values of the DE of a given reference region and the region covered by the tth sliding process, respectively. For example, we can choose a coding segment to construct the reference region. If the window enters a non-coding region from a coding region, $\Delta P(t)$ will increase rapidly. Conversely, $\Delta P(t)$ will decrease when the window enters a coding region from a non-coding region. Therefore, the bottoms of the valleys in the $t - \Delta P(t)$ curve should indicate the right borders of the coding segments [17]. For the left borders, we can reverse the DNA sequence and treat it in the same way.

For this study, the coding region [6175,8115] was used as the reference segment. The lengths of different sliding windows were set to 1200, 900, and 650 bp, respectively. For each window, the lengths of diverse cases were set to 10, 20, 30, 40, 50, and 60. This means that $\sum_s 1$ represents 6 and $\sum_s (P_{0s} - P_{ts})^2$ represents the sum of 6 squares in (11).

Figures 2a, 3a and 4a illustrate the results of searching for right borders and Figs. 2b to 4b illustrate the results of searching for left borders. They show the results of identifying the left and right borders of the coding regions using a sliding window of length N=1200 bp. We found that a large number of borders matched the bottoms of valleys.

Figure 3 shows the results of identifying the left and right borders of the coding regions using a sliding window of length N = 900 bp. We found that when using a smaller window, more borders matched the bottoms of valleys, compared with the results shown in Fig. 2.

Figure 4 shows the results for identifying the left and right borders of the coding regions using a sliding window of length N = 650 bp. When using a much smaller window, many more borders matched the bottoms of valleys compared with the results shown in Figs. 2 and 3.

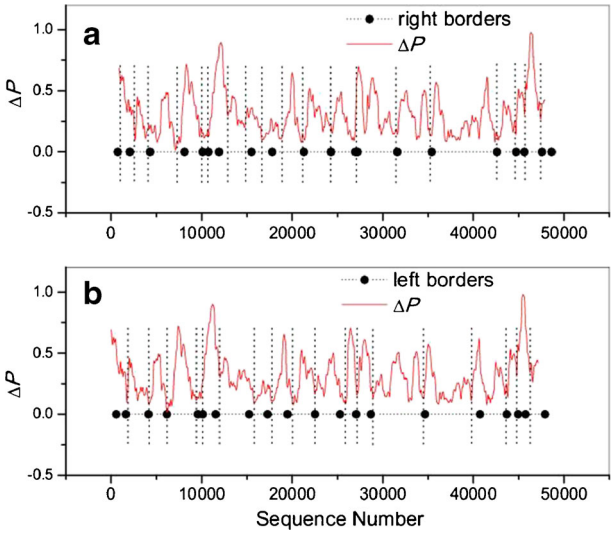

Fig. 3 The coding region 6175–7074 bp is used as the reference segment. *Black dots* represent true borders. **a** *right* and **b** *left* borders of the DNA sequence from yeast

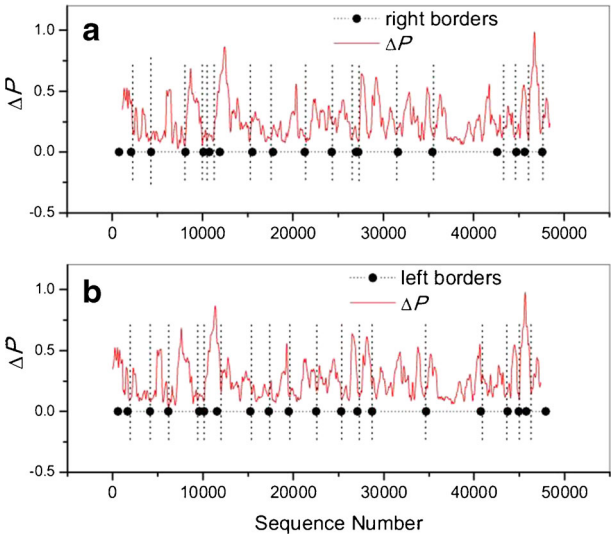

Fig. 4 The coding region 6175–6824 bp is used as the reference segment. *Black dots* represent true borders. **a** *right* and **b** *left* borders of the DNA sequence from yeast

4 Discussion

The identification of borders cannot be perfect for a variety of reasons, despite very satisfactory results (Figs. 2 to 4). That is, there is no consistent one-to-one match between each border and each valley bottom. The lengths of some coding or non-coding regions are too large and the nucleotides in these regions are distributed heterogeneously, which can lead to a few abnormal valley bottoms. Additionally, a small window can lead to much more precise results, as shown in Section 3.2, but it will also cause confusion because of the local microstructures in the $t - \Delta P(t)$ curve. A suitable strategy is to test the borders of the coding regions using windows of different sizes and to choose the best one.

5 Conclusions

The coding and non-coding regions behave in different ways in terms of long-range correlation and self-similar patterns. Using this statistical difference, we can take advantage of the self-similar structure parameter δ, which is deduced from BEDE to identify the coding and non-coding regions. To identify their borders, we choose a reference segment and calculate the entropy difference $\Delta P(t)$. The images of $t - \Delta P(t)$ exhibit strong dynamic fluctuations because of diverse trends in different types of regions. Numerical examples suggest that most of the valley bottoms in the $t - \Delta P(t)$ curve indicate the borders of coding or non-coding segments.

Conflict of interest The authors declare no conflict of interest.

References

1. Kotlar, D., Lavner, T.: Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. Genome Res. **13**(18), 1930–1937 (2003)
2. Lobzin, V.V., Chechetkin, V.R.: Order and correlations in genomic DNA sequences. The spectral approach. Physics-Uspekhi **43**, 55–78 (2000)
3. Anastassiou, D.: Frequency-domain analysis of biomolecular sequences. Bioinformatics **16**(12), 1073–1081 (2000)
4. Grosse, I., Herzel, H., Buldyrev, S.V., Stanley, H.E.: Species independence of mutual information in coding and noncoding DNA. Phys. Rev. E **61**(5), 5624–5629 (2000)
5. Bernaola-Galván, P., Grosse, I., Carpena, P., Oliver, J.L., Román-Roldán, R., Stanley, H.E.: Finding borders between coding and noncoding DNA regions by an entropic segmentation method. Phys. Rev. Lett. **85**(6), 1342–1345 (2000)
6. Barral, J.P., Hasmy, A., Jiménez, J., Marcano, A.: Nonlinear modeling technique for the analysis of DNA chains. Phys. Rev. E **61**(2), 1812–1815 (2000)
7. Scafetta, N., Hamilton, P., Grigolini, P.: The thermodynamics of social processes: the teen birth phenomenon. Fractals **9**(2), 193–208 (2001)
8. Grigolini, P., Leddon, D., Scafetta, N.: Diffusion entropy and waiting time statistics of hard-x-ray solar flares. Phys. Rev. E **65**(4), 046203 (2002)
9. Yang, H.J., Zhao, F.C., Qi, L.Y., Hu, B.L.: Temporal series analysis approach to spectra of complex networks. Phys. Rev. E **69**(6), 066104 (2004)
10. Yang, H.J., Zhao, F.C., Zhang, W., Li, Z.N.: Diffusion entropy approach to complexity for a Hodgkin–Huxley neuron. Physica A **347**, 704–710 (2005)
11. Cai, S.M., Zhou, P.L., Yang, H.J., Yang, C.X., Wang, B.H., Zhou, T.: Diffusion entropy analysis on the scaling behavior of financial markets. Physica A **367**, 337–344 (2006)
12. Scafetta, N., Latora, V., Grigolini, P.: Lévy scaling: the diffusion entropy analysis applied to DNA sequences. Phys. Rev. E **66**(3), 031906 (2002)
13. Allegrini, P., Bellazzini, J., Bramanti, G., et al.: Scaling breakdown: a signature of aging. Phys. Rev. E **66**(1), 015101 (2002)
14. Qi, J.C., Yang, H.J.: Hurst exponents for short time series. Phys. Rev. E **84**(6), 066114 (2011)
15. Zhang, W., Qiu, L., Xiao, Q., Yang, H.J., Zhang, Q., Wang, J.: Evaluation of scale invariance in physiological signals by means of balanced estimation of diffusion entropy. Phys. Rev. E **86**(5), 056107 (2012)
16. Stanley, H.E., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M.: Scaling features of noncoding DNA. Physica A **273**(1–2), 1–18 (1999)
17. Yang, H.J., Zhao, F.C., Zhuo, Y.Z., Wu X.Z.: Analysis of DNA chains by means of factorial moments. Phys. Lett. A **292**(6), 349–356 (2002)
18. García, P., Jiménez, J., Marcano, A., Molelro, F.: Local optimal metrics and nonlinear modeling of chaotic time series. Phys. Rev. Lett. **76**(9), 1449–1452 (1996)