Pattern Recognition 45 (2012) 947–955

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Exon prediction using empirical mode decomposition and Fourier transform of structural profiles of DNA sequences

Wei-Feng Zhang [a,*], Hong Yan [b]

[a] Department of Applied Mathematics, South China Agricultural University, 483 Wushan Road, Guangzhou 510642, China
[b] Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 1 November 2010
Received in revised form
9 August 2011
Accepted 14 August 2011
Available online 23 August 2011

Keywords:
DNA sequence analysis
Exon prediction
Discrete Fourier transform
Empirical mode decomposition
Structural features of DNA

ABSTRACT

Spectrum analysis approaches, such as the Fourier transform, wavelet transform and autoregressive model, have been successfully applied to solve the exon prediction problem due to their flexibility that requires no training data or prior knowledge. Detecting short exons is a difficult problem. The results achieved by the traditional methods are often unsatisfactory, because they cannot identify spectral patterns of short exons correctly. In this article, we propose an improved exon prediction method based on empirical mode decomposition and the Fourier transform. The proposed approach numerically represents the DNA sequences by their structural features, which can help to yield significant patterns that are rarely observed with the traditional methods. The structural profile is utilized to detect probable exons by examining the peaks of the local 1/3 frequency spectrum within a sliding window. The data in the window is firstly decomposed by empirical mode decomposition into a collection of intrinsic mode functions. Then the first intrinsic mode function is used to compute the local spectrum by fast Fourier transform. We compare our method with the traditional Fourier transform with binary representation method and the recently proposed paired spectral content method. Experiments on randomly selected Human genome dataset and the GENSCAN benchmark dataset illustrate that our method can enhance the signal-to-noise ratio of the analyzed sequences and improve the prediction accuracy of short exons.

1. Introduction

Recent high throughput sequencing technology has led to completion of more and more eukaryotic genome sequencing projects [1]. Efficient analysis of these sequence data, such as locating genes and exploring their functions, is receiving growing attention [2–4]. One of the key issues in the study of eukaryotic DNA sequences is the prediction of exons, also widely known as the protein coding regions identification problem. Unlike prokaryotes, in which the coding regions are not interrupted by intervening non-coding regions, eukaryotic gene is composed of isolated protein coding subregions (called exons) and non-coding regions such as introns or intergenic regions. Thus the prediction of exons in eukaryotes is far more difficult.

There has been a large body of literature looking into the possible prediction of exons [5–23], which can be roughly divided into two categories, model-based methods and non-model-based methods. Other kinds of classifications can be found in [3,4]. The model-based methods, which employ powerful machine learning techniques (such as neural networks and decision trees) or probabilistic learning models (such as Markov chain models and hidden Markov models), mainly involve MORGAN [7], GeneMark [9], GENSCAN [10,11,24], MZEF [18]. These methods are built upon large amounts of high-quality training data and often tend to be more precise than non-model-based methods. However, their performance depends heavily on the assumption that the sequenced organisms have the similar coding regions as the available training databases. This may not be true in many cases [3,12]. The non-model-based methods, which do not require prior organisms' genomic information or training dataset, are more applicable in the cases we wish to predict unknown exons in unlearned genes or organisms.

The two main ways of non-model-based methods are: the examination of various coding statistics [6,25] or the detection of the 1/3 frequency spectral pattern [2,12–17,19–23]. In this article, we focus on the latter. The protein coding sequence (CDSs) within genomic DNA sequences (named as exons in eukaryotic genes), usually exhibit a three-base periodicity property, which is absent in non-coding regions [6,12]. The three-base periodicity in exon region suggests that the Fourier magnitude spectrum at the frequency 1/3 should exhibit large values as peaks, while the

* Corresponding author. Tel.: +86 20 8528 0322.
E-mail addresses: zhangwf@scau.edu.cn (W.-F. Zhang), h.yan@cityu.edu.hk (H. Yan).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.08.016

non-coding regions do not exhibit such peaks. This characteristic, referred to as the 1/3 frequency spectral pattern, has been widely used in the prediction of exons by spectrum analysis methods.

These methods are under the same problem setting as that first converting the symbolic DNA sequence to numerical sequence, then applying certain spectrum analysis approach to examine the peaks of local 1/3 frequency spectrum within a sliding window. How to choose a proper numerical representation is critical in order to obtain good prediction performance. The most widely used numerical representation is the binary representation. It converts the DNA sequence into four binary signals, which respectively indicate the presence or absence of the chosen nucleotide A, C, G or T in the position index. Tiwari et al. used the Fourier transform with the binary representation of DNA sequence for exon prediction [12]. Kotlar et al. proposed an improved Fourier transform based method by using the information of phase angle distribution in coding regions [20]. However, their method is model-based, because in different organisms the coding regions have different phase angle distribution. Jiang et al. proved that the double curve representation of DNA sequence may expected to provide better performance than binary representation, and they used the double curve representation with Fourier transform for exon prediction [13]. Other numerical representation schemes have also been proposed, such as tetrahedron [26], Z-curve [27], paired numeric [21]. Comprehensive review of the numerical representation methods can be found in [21,22].

The choice of window size for spectrum analysis will directly influence the performance. To reduce this dependence, Choong et al. used a multi-scale spectrum analysis scheme with autoregressive (AR) model, which made use of multiple window sizes simultaneously [16]. Mena-Chalco et al. proposed a Gabor-wavelet based method to solve the window size selection problem. Recent progress has focused on the analysis of spectral properties of short exons [28]. Traditional spectrum analysis techniques, such as the Fourier transform and AR model, are based on the assumption that the analyzed signal is stationary. They do not work well for short exons, because the short exons by nature are non-stationary [14]. Due to the diversity in gene sequences and the changes inside sequences, the frequency components in different short exons change a lot, especially in the high-frequency components. Wavelet transform, which has the local time–frequency analysis property, can be used to analyze non-stationary signal. However, the wavelet transform is "non-adaptive", in the sense that its analysis results depend on the choice of the wavelet basis [29–31].

In this article, we propose a novel method, based on the empirical mode decomposition (EMD) and the Fourier transform with the structural profiles of DNA sequence, for short exon prediction. The empirical mode decomposition is a self-adaptive technique for spectrum analysis of non-stationary signal [14,29,32,33]. It can decompose a complicated signal into a collection of simple intrinsic mode functions (IMFs). Firstly the DNA sequences are numerically represented by the structural features, which have been proved to be very useful in eukaryotic core promoter prediction [34,35] and exon prediction [23]. We applied these structural profiles to exon prediction and show that some features can give better performance than traditional representation methods. Probable exons are detected by examining the peaks of local 1/3 frequency spectrum within a sliding window on the structural profiles. Slice in the window is firstly decomposed by EMD into IMFs. Then the first IMF is used to compute the frequency spectrum by fast Fourier transform (FFT). Experimental results demonstrate that our method can enhance the signal-to-noise ratio of analyzed sequences and improve the prediction accuracy of short exons.

2. Methods

2.1. Structural profiles of DNA sequence

A DNA sequence is made up of nucleotides, which can be distinguished by the four bases: adenine (A), cytosine (C), guanine (G) and thymine (T). Thus, a DNA sequence can be formally viewed as a symbol string, consisting of the four alphabet characters {A, C, G, T}. In order to employ spectrum analysis on a DNA sequence, firstly we need to convert its alphabetical representation into numerical form.

The most widely used numerical representation of DNA sequence is the binary representation [12,17]. It converts a DNA sequence into four binary sequences, which respectively denotes the occurrence of a chosen nucleotide A, C, G or T in the position index. For example, the binary representation of DNA sequence {CATTGCCAGT} is given by

$$\{x_A(n)\} = \{0,1,0,0,0,0,0,1,0,0\},$$

$$\{x_C(n)\} = \{1,0,0,0,0,1,1,0,0,0\},$$

$$\{x_G(n)\} = \{0,0,0,0,1,0,0,0,1,0\},$$

$$\{x_T(n)\} = \{0,0,1,1,0,0,0,0,0,1\},$$

where n represents the base index.

Unlike binary representation, which takes one base into consideration at a time, the double curve representation takes two bases at a time. By double curve representation, any DNA sequence can be converted to six numerical sequences, which are constructed based on the cumulative occurrence of a chosen nucleotide pair AT, AC, AG, TC, TG or CG. For example, the double curve representation of base pair AT is defined as

$$x_{AT}(n) = \sum_{i=1}^{n} u(i), \quad n = 1, \ldots, N, \tag{1}$$

where N is the length of the sequence and $u(n)$ is defined as

$$u(n) = \begin{cases} +1 & \text{for base A,} \\ -1 & \text{for base T,} \\ 0 & \text{for other bases.} \end{cases} \tag{2}$$

The other five signals $x_{AC}(n)$, $x_{AG}(n)$, $x_{TC}(n)$, $x_{TG}(n)$ and $x_{CG}(n)$ can be obtained in a similar way. The double curve representation of DNA sequence {CATTGCCAGT} is given by

$$\{x_{AT}(n)\} = \{0,1,0,-1,-1,-1,-1,0,0,-1\},$$

$$\{x_{AC}(n)\} = \{-1,0,0,0,0,-1,-2,-1,-1,-1\},$$

$$\{x_{AG}(n)\} = \{0,1,1,1,0,0,0,1,0,0\},$$

$$\{x_{TC}(n)\} = \{-1,-1,0,1,1,0,-1,-1,-1,0\},$$

$$\{x_{TG}(n)\} = \{0,0,1,2,1,1,1,1,0,1\},$$

$$\{x_{CG}(n)\} = \{1,1,1,1,0,1,2,2,1,1\}.$$

The results in [13,14,16] showed that the double curve representation is more informative than single base binary representation.

There are also other kinds of numerical representations, such as tetrahedron [26] and Z-curve [27], both of which convert a DNA sequence into three numerical series. Note that all the above numerical representations are based on only the nucleotide positions in sequence, which do not fully exploit the structural differences between protein coding and non-coding regions. Recently, Akhtar et al. proposed the paired numeric representation for exon prediction [21,22], which is based on the structural property that introns are rich in nucleotides A and T while exons are rich in nucleotides C and G. Their method assigns values of $+1$