Computers in Biology and Medicine 39 (2009) 392 – 395

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

An efficient sliding window strategy for accurate location of eukaryotic protein coding regions

Nini Rao*, Xu Lei, Jianxiu Guo, Hao Huang, Zhenglong Ren

School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China

ARTICLE INFO

Article history:
Received 7 November 2007
Accepted 28 January 2009

Keywords:
Protein coding region
Power spectrum analysis
Sliding window
Location

ABSTRACT

The sliding window is one of important factors that seriously affect the accuracy of coding region prediction and location for the methods based on power spectrum technique. It is very difficult to select the appropriate sliding step and the window length for different organisms. In this study, a novel sliding window strategy is proposed on the basis of power spectrum analysis for the accurate location of eukaryotic protein coding regions. The proposed sliding window strategy is very simple and the sliding step of window is changeable. Our tests show that the average location error for the novel method is 12 bases. Compared with the previous location error of 54 bases using the fixed sliding step, the novel sliding window strategy increased the location accuracy greatly. Further, the consumed CPU time to run the novel strategy is much shorter than the strategy of the fixed length sliding step. So, the computational complexity for the novel method is decreased greatly.

1. Introduction

Genome projects have given rise to an exponentially growing amount of genetic information. How to find out useful information from huge amounts of data through computational means are the problems that scientists focus on in current and future. The gene identification from DNA sequences of eukaryotic organisms is one of the most important and basic problems.

Many methods for gene identification of eukaryotic organisms based on distinctive features of protein coding sequences have been proposed. For example, neural net-based method [1–4], the method based on correlation function [5–7], base on power spectrum analysis [8,9] and so on. However, the comprehensive evaluation of various methods suggest that they cannot work equally well for all genes, and constant refinement is needed to evolve better methodologies.

The power spectrum analysis method is based on the feature of three-base periodicity in protein coding sequences. Fickett [10] verified that there existed a distinctive feature of protein coding regions of DNA sequences and the most prominent of these is a 1/3 periodicity, which has been shown to appear in coding sequences. Tsonis [11] and Voss [12] observed the signature of this periodicity through the Fourier analysis as a spectral peak. Furthermore, Tiwari [13] analyzed genomic sequences from different organisms, and confirmed that such periodicity is universal for protein coding sequences and is absent in genomic sequences which do not code for proteins.

In the most previous methods based on power spectrum analysis, the window is usually moved according to a fixed sliding step for acquiring the position information of protein coding regions. However, it is very difficult to select the sliding step since different organisms have different distribution of protein coding regions [13]. The sliding step of window, which is too long or too short, will seriously affect the accuracy of coding region location. Hence, we proposed a novel sliding window strategy to predict and locate eukaryotic protein coding regions. Using the novel strategy, the accuracy of predicting and locating the protein coding regions based on power spectrum analysis can be improved greatly, and meanwhile the running speed of the methods is reduced.

The remainder of this paper is organized as follows. In Section 2, we review the power spectrum analysis and describe the method of predicting and locating coding regions based on this technique with fixed sliding step. Furthermore, we describe the novel sliding window strategy and its application in predicting and locating coding regions. In Section 3, we present the results of our method for real DNA sequences. In Section 4, we analyze the effect of window length on prediction and location accuracy when the sliding step is fixed, the advantages and disadvantages of changeable sliding step proposed and finally conclude the work.

2. The method

When predicting and locating protein coding regions based on power spectrum analysis, DNA symbol string, $\{x_j, j=1, 2, \ldots, N\}$, where

* Corresponding author. Tel.: +86 028 83206489.
E-mail addresses: Raonn@uestc.edu.cn, Raonn@126.com (N. Rao).

0010-4825/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2009.01.010

Table 1
An example in which DNA symbol string was converted to four binary sequences.

DNA sequence	A	G	C	G	T	A	C	A	T	T	G	A	G	G	A	T	G	C	A
Apply X_A	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0	1
Apply X_T	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	1	0	0	0
Apply X_C	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
Apply X_G	0	1	0	1	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0

N is the length of DNA symbol string, x_j is one of the four symbols A, T, G and C, and denotes the occurrence of that particular nucleotide in position j, is firstly reflected into binary sequence. Here, the projection operator defined by Voss in 1992 was selected to realize this reflection [12], that is

$$U_\alpha(x_j) = \begin{cases} 1 & \text{if } x_j = \alpha \\ 0 & \text{other} \end{cases}$$

where $\alpha =$ A, T, G or C. Thus, any DNA sequence can be converted to four binary sequences, which can then be Fourier analyzed, as illustrated in Table 1.

According to digital signal processing theory, the total Fourier spectrum of the DNA sequence is the sum of these individual spectra, namely

$$S(f) = \sum_\alpha S_\alpha(f) = \sum_\alpha \frac{1}{N}\left|\sum_{j=1}^{N} U_\alpha(x_j)\exp\, 2\pi fj\right|^2 \quad (1)$$

where the discrete frequency $f = k/N$, $k = 1,2,\ldots,N/2$. $S_\alpha(f)$ is the partial spectrum corresponding to the symbol $\alpha =$ A,T,G or C.

The average of the total spectrum $\bar{S}$ can be calculated as follows:

$$\bar{S} = \frac{2}{N}\sum_{k=1}^{N/2} S(k/N) \quad (2)$$

There is a distinct peak at frequency $f = \frac{1}{3}$ in the Fourier spectrum of protein coding sequences, which reveals the characteristic periodicity of three. Non-protein coding sequences such as rRNA, intergenic spacers and introns, have a flat Fourier spectrum. In order to contrast the two types of spectra, the signal–to–noise radio of the peak at $f = \frac{1}{3}$ is defined as [13]

$$P = S(1/3)/\bar{S} \quad (3)$$

Here P is used as a discriminator between coding and non-coding sequences. Its values can be determined according to the statistical characteristics from a great amount of DNA sequences. The suggested P (baseline) value is 4 in [13].

2.1. Standard power spectrum analysis

Most existing methods based on power spectrum analysis mainly adopted the fixed sliding step strategies to move the window. Here we called them as standard power spectrum analysis.

Assume that the length of a sliding window is M. The method for predicting and locating coding sequences based on standard power spectrum analysis can be described as follows:

(1) Slide the window beginning from the first base position of DNA sequence by 3 base steps. If the length of DNA sequence that the last window corresponds to is less than M, extend its length to M by zero padding the sequence.
(2) Calculate the local spectrum of DNA sequence that each window corresponds to and its average value according to Eqs. (1) and (2), respectively.

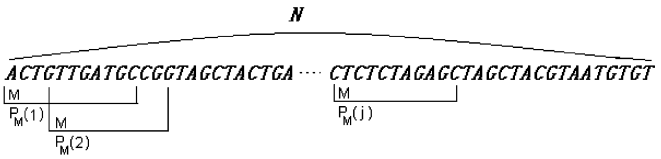

Fig. 1. The schematic diagram for predicting and locating coding sequences based on standard power spectrum analysis, where $M = 11$ and the sliding step is 3.

(3) According to Eq. (3), calculate the local signal-to-noise ratio $P_M(j)$, where j is the position of the center of window of length M and $j = 1,2,\ldots,\lfloor N/3 \rfloor$. Usually, those portions of the sequence with $P_M(j) > 4$ are identified as protein coding regions and the sequences with $P_M(j) < 4$ are identified as non-coding regions.

A schematic diagram that further interprets this method is shown in Fig. 1.

Any DNA sequence is scanned to a distance of M nucleotides up- and downstream of the above identified region so as to locate the initiation and termination codons, respectively, in any of the six possible reading frames.

2.2. The novel sliding window strategy

The location error is 54 bases on average using the standard power spectrum analysis [13]. In order to improve the location accuracy of the above method, we propose to slide the window according to the following strategy:

Step 1: Select a proper window with length of M bases using prior biological knowledge of organism. Calculate the local signal-to-noise ratio $P_M(j)$ (here j denotes the number of windows and $j = 1$) of DNA sequence that the first window corresponds to according to Eq. (3).

Step 2: Take an initial sliding step of K bases and set $K \leqslant M/2$ as an positive integer that satisfies $K = 2^n$, $n = 1,2,\ldots$.

Step 3: Slide the window one time in step of K beginning from the first base position of window $j(j = 1,2,\ldots,N-M)$ and calculate the local signal-to-noise ratio $P_M(j+1)$ corresponding to this window according to Eq. (3). Let $P_M(j)$ and $P_M(j+1)$ subtract P (baseline) value and their difference are denoted by $P'_M(j)$ and $P'_M(j+1)$.

Step 4: If the adjacent points $P'_M(j)$ and $P'_M(j+1)$ have different numerical symbols, go to step 5. Otherwise, let $P_M(j+i) = P_M(j)$ (here i denotes the number of bases and $i = 1,2,\ldots,K-1$), go to step 6.

Step 5: Set the sliding step $K = K/2$. If K is more than 1, slide the window in step of K beginning from the first base position of window j and calculate the local signal-to-noise ratio $P_M(j+1)$ corresponding to the novel window according to the Eq. (3). Let $P_M(j+1)$ subtract P (baseline) value and the difference is denoted by $P'_M(j+1)$. Go to step 4. Otherwise, go to step 6.

Step 6: $P_M(j) = P_M(j+1)$ and $j = j+1$. If $j > N-M$, stop iteration. Otherwise, go to step 2.

The discrimination between coding and non-coding sequences is same as that of standard power analysis. The proposed sliding window strategy is very simple and fast to run, and the sliding step of window is changeable.