Gene Prediction Based on One-Dimensional Image Segmentation

Fei-Yu Wang, Zi-Wei Zheng
College of Information Science and Engineering
Ningbo University
Ningbo, P. R. China
e-mail: sheldon.elec@gmail.com
ziwei_zheng@yahoo.com.cn

Wei-Hua Li
Faculty of Science
Ningbo University
Ningbo, P. R. China
e-mail: liweihua@nbu.edu.cn

***Abstract*—In the current study, using methods of signal processing to manage gene prediction has attracted great attention. At first, the voss mapping which can map the DNA alphabetic sequence into the numerical sequence and the 3-base periodicity of exon are introduced. Then a fixed-length sliding window approach and its feasibility are analyzed. It can be proved that when two exons are very close, gene prediction by only setting a threshold to the spectrum could not have good effect. To overcome this shortcoming, a new method based on one-dimensional image segmentation is proposed. Finally, simulation shows the short introns are culled commendably. Two evaluation indices are also introduced to demonstrate the effectiveness of this method.**

Keywords- gene prediction; image segmentation; voss mapping; 3-base periodicity; windowed fourier transform

I. INTRODUCTION

With a massive genomic data obtained after the accomplishment of the Human Genome Project, processing these data becomes more and more important. How to distinguish the coding sequence (i.e. exon) in a given DNA strand, also known as gene prediction, is an unsolved but the most basic and primary problem in bioinformatics.

One methodology for gene prediction is based on statistics [1]. However, statistical forecasting methods usually need plenty of DNA sequences (their code area have already been detected) as a training data set to determine the model parameters and thereby improve the accuracy of prediction. But if there is lack of information about gene, the accuracy of gene prediction would be significantly decreased. Therefore, discovering gene coding regions by methods of signal processing has attracted great attention [2] [3] in current study. Some researchers have proposed some algorithms to detect exons already. There are two most commonly used algorithms based on 3-base periodicity. One is a fixed-length sliding window approach [4] [5]. The other is a recognition method based on computing SNR (Signal-to-Noise Ratio) of the stepwise DNA segments of the target DNA sequences [6]. Due to the impact of the random noise in the DNA sequences, it is difficult to precisely locate the coding regions and set the two endpoints of the interval simply by setting a threshold of SNR to the spectrum curve.

In this paper, we use a series of measure of signal processing, such as FIR (Finite Impulse Response) [7] and image segmentation [8], to refine two endpoints of the coding regions. Two Evaluation Indices [9] [10] are introduced to measure the effectiveness and simulation results are given.

II. NUMERICAL MAPPING

In the study of DNA sequence, we should first mapping these alphabetic sequences which is a combination of four kind of nucleotides (*A, T, C, G*) to corresponding numerical sequences according to some certain rules. Then the digital signal processing could be made with these numerical sequences.

Let

$$I = \{A, T, G, C\} \tag{1}$$

Any DNA sequence whose length is N could be represented by:

$$S = \{S[n] \mid S[n] \in I,\ n = 0,1,2,\cdots,N-1\} \tag{2}$$

To any $b \in I$, let

$$u_b[n] = \begin{cases} 1, & S[n] = b \\ 0, & S[n] \neq b \end{cases},\ n = 0,1,2,\cdots,N-1 \tag{3}$$

which is so called Voss mapping [11] [12]. For example, supposing a given DNA fragment *S* = '*ATCGTACTG*', the four corresponding 0-1 sequences are generated respectively:

$$\begin{aligned} \{u_A[n]\} &: \{1,0,0,0,0,1,0,0,0\} \\ \{u_G[n]\} &: \{0,0,0,1,0,0,0,0,1\} \\ \{u_C[n]\} &: \{0,0,1,0,0,0,1,0,0\} \\ \{u_T[n]\} &: \{0,1,0,0,1,0,0,1,0\} \end{aligned} \tag{4}$$

Thus the four numerical sequences are called the binary indicator sequence of a DNA strand. They are converted into frequency signal by DFT (Discrete Fourier Transform) to study the characteristics of exon.

$$U_b[k] = \sum_{n=0}^{N-1} u_b[n] e^{-j\frac{2\pi nk}{N}},\ k = 0,1,\cdots,N-1 \tag{5}$$

Published by Atlantis Press, Paris, France.

Then four complex sequences can be obtained. Calculate the power spectrum of each complex sequences, and add up them to get the power spectrum of the DNA strand.

$$P[k]=|U_A[k]|^2+|U_T[k]|^2+|U_G[k]|^2+|U_C[k]|^2 \quad k=0,1,\cdots,N-1 \tag{6}$$

The power spectrums of the exon and intron in the same DNA strand usually exhibit different characteristics. At the frequency of $k=N/3$, it has a spectral peak in the power spectrum of exon, whereas the power spectrum of intron doesn't have such peak value. This statistical phenomenon is called 3-base periodicity [4] [13] [14].

III. Analysis of Algorithm

A. Sliding Window Approach and the Feasibility

The discovery of the 3-base periodicity of exon enables us to forecast and locate the coding regions in a DNA sequence that hasn't annotated before. Some researchers have proposed some algorithms to detect the exon already. There are two most commonly used algorithm based on 3-base periodicity. The more typical one is a fixed-length sliding window approach [4] [5].

We choose *M* (usually a multiple of 3, for example, *M*=99, 129, 255, 513, etc.) as the length of the window. To any *n* ($0 \le n \le N-1$), do DFT (7) to the indicator sequences from the segment (8) whose length is *M* and the middle point is *n*.

$$U_b[k]=\sum_{i=n-\frac{M-1}{2}}^{n+\frac{M-1}{2}} u_b[i]e^{-j\frac{2\pi ik}{M}},\ k=0,1,\cdots,M-1 \tag{7}$$

$$[n-\frac{M-1}{2},n+\frac{M-1}{2}] \tag{8}$$

Then all the values at the frequency of $M/3$ can be obtained as

$$P(n,\frac{M}{3})=\left|U_A[\frac{M}{3}]\right|^2+\left|U_T[\frac{M}{3}]\right|^2+\left|U_G[\frac{M}{3}]\right|^2+\left|U_C[\frac{M}{3}]\right|^2 \tag{9}$$

And then get the corresponding curve as a "spectrum", as in Fig. 1.

These red horizontal lines in the figure are the real intervals of exons in the DNA sequence. Obviously, the peak regions in the curve have a "Corresponding Relationship" with the intervals of coding regions. To determine these intervals, the common method is to set a proper threshold of SNR to the spectrum curve. Then those regions that higher than the threshold could be obtained as the coding regions. However, this method is inadequate because not all the points in these regions are the coding regions. There are some troughs in these regions, which, in theory, should not be considered as the coding regions because they don't obey the 3-base periodicity. Thus, in this paper, we proposed that eliminating those troughs and refining the two endpoints of the coding regions could be done based on first determine coding regions by the threshold.

To demonstrate the rationality and effectiveness of the fixed-length sliding window approach, the ideal periodic sequence is constructed. It is made up with ideal 3-Periodic subsequences and ideal 4-Periodic subsequences alternately. Ideal 3-Periodic subsequences are presumed as exons and ideal 4-Periodic subsequences are presumed as introns. As Fig. 2 shows, the spectrum of the sequence is obtained using the fixed-length sliding window approach.

As Fig. 2, the points at the steep slope are the boundary point of exon and intron. When two exons are very close, like the second one and the third one, we wouldn't detect the intron between them only by setting a threshold. It is also showed in Fig. 2 that the spectral value of the latter exons is very low comparing with the former ones. That is to say the length of the window would strongly affect the effectiveness of detecting the shorter exons.

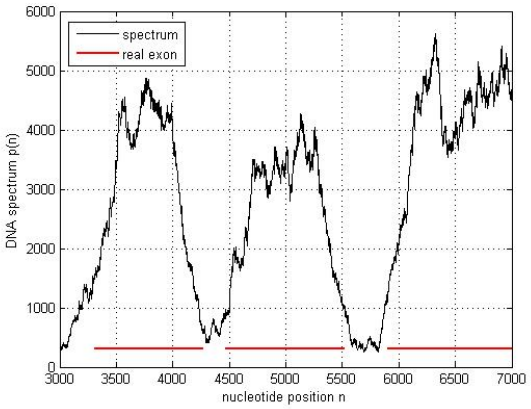

Figure 1. The spectrum using the fixed-length sliding window approach (homo mitochondrial gene, NC_012920_1)

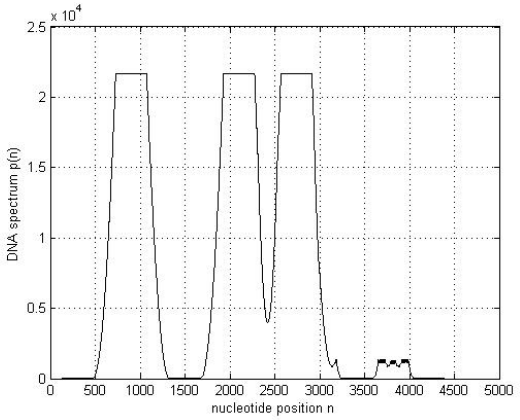

Figure 2. Spectrum of ideal sequence (the length of the window is 255)

B. Image Segmentation

With previous analysis, the key to predict the coding regions is to find the steep slope in the spectrum to determine the endpoints. Therefore the algorithm of image segmentation is used in this paper.

In the theory of image segmentation [8], the gradient of an image $h(x,y)$ is defined as the vector

$$\nabla h = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \frac{\partial h}{\partial x} \\ \frac{\partial h}{\partial y} \end{bmatrix} \tag{10}$$

This vector has the important geometrical property that it points in the direction of the greatest rate of change of *h* at location (x,y).

The magnitude of vector ∇h, denoted as $M(x,y)$, where

$$M(x,y) = mag(\nabla h) = [G_x^2 + G_y^2]^{\frac{1}{2}} \tag{11}$$

is the value of the rate of change in the direction of the gradient vector.

Obtaining the gradient of an image requires computing the partial derivatives $\partial h/\partial x$ and $\partial h/\partial y$ at every pixel location in the image. The simplest digital approximations to the partial derivatives in z_5 (as in Fig. 3) using masks of size 3×3 are called the Sobel operators, as (12), (13) and Fig. 3.

$$G_x = (z_7 + 2z_8 + z_9) - (z_1 + 2z_2 + z_3) \tag{12}$$

$$G_y = (z_3 + 2z_6 + z_9) - (z_1 + 2z_4 + z_7) \tag{13}$$

The problem in this paper is just one-dimensional image segmentation, so only the horizontal mask is needed.

C. Two Evaluation Indices

At the nucleotide level, the accuracy can be measured by comparing predicted and real exons. As in Fig. 4, *TP* (*the true positive*), *TN* (*the true negative*), *FN* (*the false negative*) and *FP* (*the false positive*) are used to represent the relationship between the predicted and real exons.

To measure the accuracy of the prediction, we often use *CC* (*Correlation Coefficient*) and *AC* (*Approximate Correlation*). They are obtained as (14) (15) and (16).

CC could measure the linearity between the real sequence and the predicted sequence, while *AC* represents the level of sensitivity and specificity. That is the difference between them in statistical sense [9] [10].

$$\begin{bmatrix} z_1 & z_2 & z_3 \\ z_4 & z_5 & z_6 \\ z_7 & z_8 & z_9 \end{bmatrix} \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ 1 & 0 & 1 \end{bmatrix}$$

Figure 3. A 3×3 region of an image and the Sobel operators

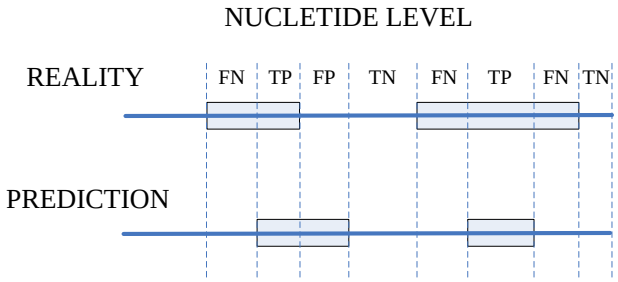

Figure 4. Comparison of reality and prediction on nucleotide level

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}} \tag{14}$$

$$ACP = \frac{1}{4}\left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right) \tag{15}$$

$$AC = (ACP - 0.5) \times 2 \tag{16}$$

IV. Implementation of Algorithm

Step 1: First gets the spectrum using sliding window approach from the unknown DNA sequence. Then filter the waveform by FIR to make it smoother and do the corresponding phase correction. After getting the new spectrum curve, set a threshold to it based on the length of the window. Reject the point below the threshold and then *N* initial intervals of exon are obtained, as (17).

$$U_i : [a_i, b_i], i = 1,2,3,\cdots,N \tag{17}$$

They are satisfies

$$0 < a_1 < b_1 < \cdots < a_i < b_i < a_{i+1} < b_{i+1} < \cdots < a_N < b_N < L \tag{18}$$

where *L* is the length of the unknown DNA sequence.

Step 2: Do the one-dimensional edge detection using Sobel operator to the waveform obtained in Step 1 and get a series of edge point, as in Fig. 5. Intuitively, these points show the greatest rate of change in the curve.

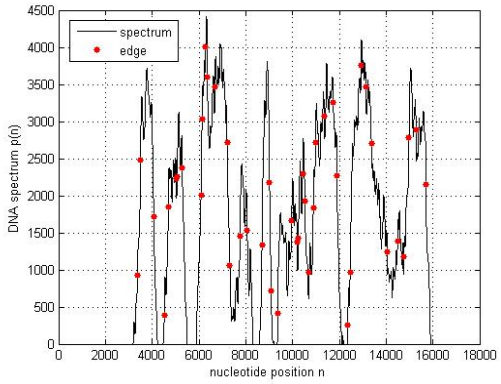

Figure 5. Edge detection to the spectrum (homo mitochondrial gene, NC_012920_1)

Step 3: To every interval

$$U_i, i = 1,2,3,\cdots,N \tag{19}$$

obtained in step 1, further optimization should be done. First identify these edge points at interval U_i are steep positive edge or negative edge and make them correspondingly labeled. Then get a series of subintervals as valley regions in U_i based on the matching relationship between the steep positive edge and negative edge. After culling these subintervals (as introns) we get *m* endpoints of exon in U_i, as (20).

$$x_{i,j},\ j \in [1,m] \tag{20}$$

And then obtain a series of coding regions

$$\begin{aligned}&[a_i, x_{i,1}],[x_{i,2}, x_{i,3}],\cdots,[x_{i,j-2}, x_{i,j-1}],\cdots,\\&[x_{i,j}, x_{i,j+1}],\cdots,[x_{i,m-2}, x_{i,m-1}],[x_{i,m}, b_i]\end{aligned} \tag{21}$$

Step 4: Refine the endpoints obtained in step 3. After Statistical Analysis on massive known DNA strand, the regions of intron always start with '*GT*' and end up with '*AG*'. So we could refine the endpoints of exon according to this characteristic. To every starting endpoint $x_{i,j}$, obtained in the previous step, get several points $x_{i,j,p}$ that satisfy

$$S(x_{i,j,p} - 2) = 'A', S(x_{i,j,p} - 1) = 'G' \tag{22}$$

in the neighborhoods of it. *S* is the map of position to the property of nucleotide. Similarly, to every terminal endpoint $x_{i,j+1}$, get several points $x_{i,j+1,q}$ that satisfy

$$S(x_{i,j+1,q} + 1) = 'G', S(x_{i,j+1,q} + 2) = 'T' \tag{23}$$

in the neighborhoods of $x_{i,j+1}$. Till then we get a new set of intervals

$$\{[x_{i,j,p}, x_{i,j+1,q}]\} \tag{24}$$

corresponding to

$$[x_{i,j}, x_{i,j+1}] \tag{25}$$

After calculating SNR in every interval in the set, we choose the biggest one and the corresponding interval

$$[x_{i,j,p_0}, x_{i,j+1,q_0}] \tag{26}$$

as the updated coding region

$$[x'_{i,j}, x'_{i,j+1}] \tag{27}$$

Finally we get the refined intervals of exon, as in Fig. 6 and Table 1. Intuitively, those valley regions are detected and culled as introns commendably.

Use the two evaluation indices, *AC* and *CC*, to measure the accuracy of this prediction. They are calculated as:

$$\begin{aligned}AC &= 0.753\\CC &= 0.752\end{aligned} \tag{28}$$

And to that only predicting by the threshold, the two indices are

$$\begin{aligned}AC &= 0.748\\CC &= 0.746\end{aligned} \tag{29}$$

The simulation indicates that the proposed method is useful for gene prediction and has better predictive results.

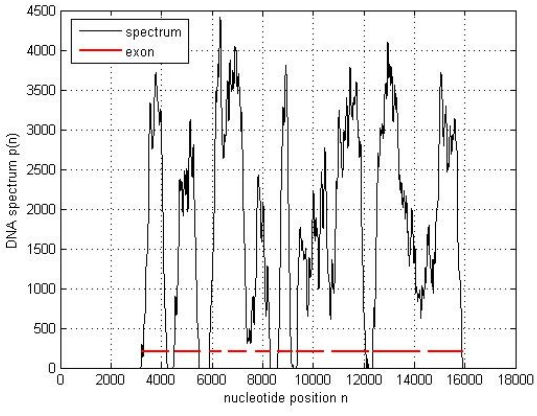

Figure 6. The final result of prediction (homo mitochondrial gene, NC_012920_1)

TABLE I. THE ENDPOINTS OF THE PREDICTED AND REAL EXONS

Endpoints (predicted)	3132..4083 4687..5399 6082..7058 7698..8165 8588..9177 9301..9961 10099..10413 11100..11899 12614..13812 14740..15919
Endpoints (real)	3307..4262 4470..5511 5904..7445 7586..8269 8366..8572 8527..9207 9207..9990 10059..10404 10470..10766 10760..12137 12337..14148 14149..14673 14747..15887

V. CONCLUSIONS

In this paper, a new method based on 3-base periodicity is proposed to manage gene prediction. First a fixed-length sliding window approach is introduced. However, when two exons are very close, prediction by only setting a threshold could not be very effective because it would have some valley regions in the predicted exon and those valley regions must be detected and culled as introns. To overcome this shortcoming, the one-dimensional image segmentation and a series of measure of signal processing are used and finally those short introns are culled commendably. Two evaluation indices are also introduced to demonstrate the effectiveness of this method. However, through the analysis of this method, the length of the window would strongly affect the effectiveness of detecting the shorter exons. So in the next step of work, we can consider the variable-length window. Meanwhile, some other properties that could be used to recognize the short exon may be taken into account.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation of China (No.60972063), the National Science and Technology Major Project (No.2011ZX03002-004-02), the Program for New Century Excellent Talents in University (No.NCET-08-0706), the Science Foundation of Zhejiang Province (No.R1110416), the Specialized Research Fund for the Doctoral Program of Higher Education (No.20113305110002), the Program for Technology Innovation Team of Ningbo Government (No.2011B81002), Zhejiang Scientific and Technical Key Innovation Team of New Generation Mobile Internet Client Software (2010R50009).

REFERENCES

[1] Chris Burge and Samuel Karlin, "Prediction of complete gene structures in human genomic DNA," J. Mol. Biol., vol.268, pp. 78–94, 1997.

[2] M. J. Berryman and A. Allison, "Review of signal processing in genetics," Fluctuation and Noise Letters, vol.5, no.4, pp. 13–35, 2005.

[3] Dimitris Anastassiou, "Genomic Signal Processing," IEEE Signal Processing Magazine, vol.18, no.4, pp. 8–20, 2001.

[4] Dimitris Anastassiou, "Frequency-domain analysis of biomolecular sequences," Bioinformatics, vol.16, pp. 1073–1081, 2000.

[5] Daniel Kotlar and Yizhar Lavner, "Gene prediction by spectral rotation measure a new method for identifying protein coding regions," Genome Research, vol.13, pp. 1930–1937, 2003.

[6] Changchuan Yin and Stephen S.-T. Yau, "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence," Journal of Theoretical Biology, vol.247, pp. 687–694, 2007.

[7] Sitanshu Sekhar Sahu and Ganapati Panda, "Identification of Protein-Coding Regions in DNA Sequences Using A Time-Frequency Filtering Approach," Genomics Proteomics & Bioinformatics, vol.9, pp. 45–55, 2011.

[8] Rafael C.Gonzalez and Richard E.Woods, "Digital Image Processing," 3rd ed., Prentice Hall, pp. 460-494, 2008.

[9] Zhuo Wang, Yazhu Chen and Yixue Li, "A Brief Review of Computational Gene Prediction Methods," Geno. Prot. Bioinfo., vol.2, no.4, pp. 216–221, 2004.

[10] M. Burset and R. Guigo, "Evaluation of Gene Structure Prediction Programs," Genomics, vol.34, pp. 353–367, 1996.

[11] S. D. Sharma, K. Shakya and S. N. Sharma, "Evaluation of DNA Mapping Schemes for Exon Detection," 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), pp. 71–74, 2011.

[12] Richard F. Voss, "Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences," Phy. Rev. Lett., vol.68, no.25, pp. 3805–3808, 1992.

[13] Jamal Tuqan and Ahmad Rushdi, "A DSP perspective to the period-3 detection problem," Proceedings of IEEE Workshop on Genomic Signal Processing and Statistics, pp. 53–54, 2006.

[14] Ming Yan, Zhe-Suai Lin and Chun-Ting Zhang, "A new Fourier transform approach for protein coding measure based on the format of the Z-curve," Bioinformatics, vol.14, no.8, pp. 685–690, 1998.