Bio-Medical Materials and Engineering 26 (2015) S1891–S1901
DOI 10.3233/BME-151492
IOS Press

The study of threshold determination of gene identification and its improvement algorithms

Chun Yan[a], Wei Liu[b,*], Yiwei Qiu[a] and Haitang Sun[a]
[a]*College of Mathematics and System Sciences, Shandong University of Science and Technology, Qingdao Shandong 266590, China*
[b]*College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao Shandong 266590, China*

Abstract. The problem of gene recognition based on the ratio of power spectrum, SNR, and Gabor transform and its implementation of the calculation were discussed. The optimal threshold could guarantee to identify the DNA sequences with the signal-to-noise ratio. It summarized three kinds of traditional ways to determine the threshold, and advanced the optimum entitled method showing the disparate degrees of highlight and the discrimination rate method of the exons or introns as far as possible to improve the rate of their accuracy. To evaluate different determination methods of threshold by using the calculation results of four kinds of DNA sequence. In order to ensure the analysis of DNA sequence more accurate, it adopted and improved gene identification method of Fourier transformation in a short time which is based on Gabor transformation. By using of the ergodic theory, the fixed percentage of the sequence length of exons in DNA has been improved to be the dynamic percentages which focus on different gene types. The exons of the DNA sequence which have been already discovered were identified by using the improved algorithm. With comparison of the results and the actual endpoint of exons, it confirmed that the improved algorithm can figure out the endpoint of the exons more accurate.

Keywords: The discrete Fourier transform, Gabor transform, gene identification, signal-to-noise ratio, threshold

1. Introduction

DNA is the carrier of biological genetic information. It uses genetic code to store information, and guides the synthesis of proteins. The accurate deliver of genetic information of protein could make the various life functions completely. Along with the successful completion of world human genome project, getting rich biological information from large amounts of DNA sequences through physical or mathematical methods has important theoretical significance and practical value in biology, medicine, pharmacy etc., and it is currently a hot research topic in the field of information biology. To identify the coding sequence of a given DNA sequence that namely exons, which is also called the gene prediction, is one of the most basic and important not solved completely problem in the field of

*Address for correspondence: Wei Liu, College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China, Tel.: +86 15063992767; Fax: +86 0532 86057638; E-mail: liuwei_doctor@yeah.net.

information biology.

Firstly, this essay will discuss the threshold determination of different species types of genes. To determine the threshold of genes types in different species, and to study the threshold determination method of each kind of representative gene sequence exons, and determine the threshold. And through the exploring the classification effectiveness of exons and non-exons, we can make an analysis on classified error in the process of classification recognition.

For DNA sequences with specific gene types, it has some experience and subjectivity to set the discriminant threshold of SNR R as $R_0 = 2$. The discriminant threshold we chose should according to different gene types. The traditional determination method of threshold includes weight method and optimization method [1-5]. But considering the different needs of exons and introns, there isn't a reasonable solution. This essay summarized the existing methods, and found a more reasonable and accurate method of threshold determination method to enhance the accuracy in gene recognition.

Secondly, this essay will study the realization of the gene identification algorithm. Our purpose is to detect and report all the genes coding sequences in DNA sequence which have not yet been annotated and complete (exons). There are usually two kinds of gene identification algorithms using SNR, one is the moving window method with fixed length [6, 7], and another is the recognition method of mobile signal-to-noise ratio curve [8-13]. At present, the gene recognition algorithms above can't get rid of the influence of random noise of DNA sequence, so it is difficult to determine the two endpoints of gene exons interval accurately.

We can consider the window method of the signal processing, use the Gaussian window function to deal with DNA indicated sequence, and determine the optimal threshold, thus determine the two endpoints of exons interval. And then to realize the gene identification algorithm, design and evaluate the accuracy of gene recognition algorithm, and further to forecast and analyze the six DNA sequence exonic range which are not annotated in gene6.

2. The illustration of model assumptions and symbols

2.1. Model assumptions

(1) The length of DNA sequences to be identified is a multiple of three, and intercept a fragment of length k_t (k_t is a multiple of 3), and deem k_t has a 3-cyclical. (2) No gene mutation and substitution exist. (3) The pulse signal of DNA sequence indicated is not stationary. (4) There are no influences among the bases far from each other.

2.2. Symbolic description

Let A denotes Adenine, G denotes Guanine, C denotes Cytosine, T denotes Thymine, I denotes the collection of four nucleotides, $\{S[n]\}$ denotes DNA sequence, $\{u_b[n]\}$ denotes indicate sequence of nucleotide B, $\{u_b[K]\}$ denotes the Fourier transform of indicate sequence of nucleotide B, $\{P[k]\}$ denotes power spectral sequence of DNA sequence, $\overline{E}$ denotes Average of power spectrum, L denotes DNA sequence length, M denotes the coefficient matrix of quadratic form, p_i denotes length of DNA fragment, V denotes the two order central matrix.

Table 1

Existing ways to determine threshold

Experience method	$R_0 = 2$			
weight method	weight		Threshold value R_0	Remarks
	w_1	w_2		
	$\frac{1}{2}$	$\frac{1}{2}$	$R_0 = \frac{E_1 + E_2}{2}$	—
	$\frac{E_1}{E_1 + E_2}$	$\frac{E_2}{E_1 + E_2}$	$R_0 = \frac{E_1}{E_1 + E_2} E_1 + \frac{E_2}{E_1 + E_2} E_2$	Mean value of exon E_1 Mean value of intron E_2
	$\frac{\sigma_1}{\sigma_1 + \sigma_2}$	$\frac{\sigma_2}{\sigma_1 + \sigma_2}$	$R_0 = \frac{\sigma_1}{\sigma_1 + \sigma_2} E_1 + \frac{\sigma_2}{\sigma_1 + \sigma_2}$	SD of exon σ_1 SD of intron σ_2
	$\frac{m_1}{m_1 + m_2}$	$\frac{m_2}{m_1 + m_2}$	$R_0 = \frac{m_1}{m_1 + m_2} E_1 + \frac{m_2}{m_1 + m_2} E_2$	Median of exon m_1 Median of intron m_2
Optimizati-on method	$\max f(x) = \sum_i \mathrm{sgn}(R_i^1 - R_0) + \sum_j \mathrm{sgn}(R_0 - R_j^2)$			$R_i^1 \in S_1$, S_1 is the SNR aggregation of exon; $R_j^1 \in S_2$, S_2 is the SNR aggregation of intron.

Note: SD is means standard deviation.

3. Threshold determination of different species types of gene

3.1. Traditional ways to determine the thresholds

Major ways to determine threshold can be summarized from references, such as experience method, weight method and optimization method. They are shown in Table 1.

The methods above are equivalent to estimate the exons and introns. In the optimization method, if the value of R_0 is small enough, exons can be misjudged as introns and the accuracy may be decreased, although the exons contained are as many as possible. In the similar way, if the value of R_0 is big enough, the misjudgment also happens frequently. If we need exons or introns controlled in the range which thresholds determined as many as possible, or the accuracy can be increased as much as possible, the results of the methods above may be less-than-desirable. So we give the improvement method as follows.

3.2. Modified ways to determine the threshold

3.2.1. Weighting optimization method

The aim of this method is to make exons or introns be contained in the sequence that have determined before as many as possible. The calculation equation is:

$$\max f(x) = w_1 \sum_i \mathrm{sgn}(R_i^1 - R_0) + w_2 \sum_j \mathrm{sgn}(R_0 - R_j^2) \tag{1}$$

In this equation, w_1 / w_2 means the importance of exons or introns. The smaller of w_1 / w_2, the more important exons are; the bigger of w_1 / w_2, the more important introns are.

3.2.2. Discriminant rate method of exons and introns

The aim of this method is to make the accurate rate of exons or introns in the discriminant result reach the highest level.

$$g_1(x) = \max \frac{\sum_i \mathrm{sgn}(R_i^1 - R_0)}{\sum_i \mathrm{sgn}(R_i^1 - R_0) + \sum_j \mathrm{sgn}(R_j^2 - R_0)},\ g_2(x) = \max \frac{\sum_j \mathrm{sgn}(R_j^2 - R_0)}{\sum_i \mathrm{sgn}(R_i^1 - R_0) + \sum_j \mathrm{sgn}(R_j^2 - R_0)} \tag{2}$$

If the accurate rate of exons needs to be high, equation $g_1(x)$ will be used. In the contrary, equation $g_2(x)$ will be used. The choice of different rate determined methods is based on the aim of the discrimination.

3.3. Evaluation index of threshold discriminant effect

There are three kinds of methods for determining the threshold above, but which kind of method is the best? Now suppose that the SNR classify threshold is R_0, that is to say $R \geq R_0$ is used as the discriminant rule for exons and $R \geq R_0$ is the discriminant rule for introns. Three kinds of index for evaluating the method of threshold discriminant, which were based on different types and different objective, were presented as follows.

(1) Discriminant rate of exons (or introns)

Most of the time, it is hoped that exons (or introns) in the DNA sequence can be determined as much as possible, even sometime introns are wrongly discriminated as exons or exons are wrongly discriminated as introns .

Discriminant rate of exons is $S_n = T_P /(T_P + F_N)$. Discriminant rate of introns is $S_p = T_N /(T_N + F_P)$. In the equation, T_P represents the number of exons that are discriminated correctly; T_N represents the number of introns that are discriminated correctly. F_N is the number of exons that are discriminated as introns by error, and F_P is the number of introns that are discriminated as exons by error.

(2) Total correct rate $A_c = \dfrac{S_n + S_p}{2}$

(3) The accuracy of exons in the discriminant result

Sometimes the total accuracy is not hoped to be large. In the contrary, the real content of exons in the discriminated ones is more valued. So the equation of accuracy of exons is defined as follows:

$$B_c = \frac{T_P}{(T_P + F_P)}$$

In this equation, the meanings of T_P and F_P are as same as above.

Table 2
Comparison chart of determining threshold methods

genes		Human beings	Yeast of liquor	Nematode's cosmidT24C4	Arabidopsis
Weight method	R_0^1	1.6156	3.9612	2.1781	1.3421
	A_c	0.7318	0.7858	0.7523	0.7011
	B_c	0.7657	0.7978	0.7352	0.7312
	S_n	0.7836	0.7931	0.7452	0.6892
Weighted average method with SD	R_0^1	2.295	6.5213	4.2518	1.7231
	A_c	0.6601	0.7018	0.6813	0.6621
	B_c	0.8073	0.8241	0.7693	0.7511
	S_n	0.7981	0.8013	0.7658	0.7541
Optimization method	R_0^1	1.1201	1.5621	1.1476	1.1135
	A_c	0.7823	0.9011	0.8312	0.7745
	B_c	0.8231	0.8567	0.8015	0.7553
	S_n	0.7731	0.7818	0.7743	0.7217
Weighted optimization method with SD	R_0^1	1.8762	1.8931	2.0176	1.2431
	A_c	0.7412	0.8123	0.7564	0.7634
	B_c	0.8169	0.8419	0.7884	0.7451
	S_n	0.7681	0.7519	0.8031	0.7018
Discriminant rate method of exons	R_0^1	0.8962	1.1213	1.1108	0.9852
	A_c	0.6891	0.7421	0.6985	0.7125
	B_c	0.8921	0.9314	0.8216	0.7947
	S_n	0.9012	0.9218	0.8321	0.8016
SNR threshold value is fixed as 2	R_0^1	2	2	2	2
	A_c	0.6768	0.8898	0.7269	0.6572
	B_c	0.6731	0.7076	0.653	0.6012
	S_n	0.6834	0.7215	0.7018	0.6547

3.4. Analysis and assess of the methods used for determining threshold

In order to assess the methods of determining threshold and compare them with the method of using a fixed threshold value about 2, we chose 4 representative groups of gene sequences based on the famous biological data sites: http://www.ncbi.nlm.nih.gov/guide/. The accuracy statistics can be shown in Table 2.

By analysis the date of Table 2, some can be discovered: Accuracy of weight method, weighted average method with SD and threshold value is fixed as 2 is less than the accuracy of optimization method. The percentage of real exons in discriminant rate method of exons is larger than other methods. Weighted optimization method with SD is the best to find introns when the introns are regarded as important. The threshold value in discrimination rate method of exons is the smallest one.

4. Improving of the gene recognition method

4.1. Disadvantages of the traditional method

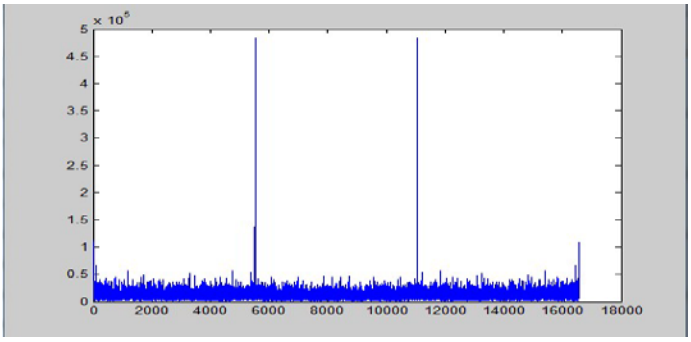

Fig. 1. The human mitochondrial genome sequence of DNA power spectrum.

Table 3

The table of exon actual position

Serial number	Exon position	Serial number	Exon position	Serial number	Exon position
1	3307-4262	6	8527-9207	11	12337-14148
2	4470-5511	7	9207-9990	12	14149-14673
3	5904-7445	8	10059-10404	13	14747-15887
4	7586-8269	9	10470-10766		
5	8366-8572	10	10760-12137		

The traditional Fourier transform method has a high time complexity, and because of the influence of DNA random series and so on, it's difficult to find the two endpoints of exon interval accurately. The power spectra of human mitochondrial genome DNA sequence (length is 16569bp) using the traditional Fourier transform method is shown in Figure 1.

The actual position of Exon in human mitochondrial genome DNA sequence can be seen in Table 3.

According to the Figure 1 and Table 3, we can obviously see that crest appears at the position of exons in this DNA sequence. But there is too much noise in its background, we cannot accurately figure out the exact endpoint on both sides of the crest. And it correctly identified only 2 of them in 13 exons. The correct rate is too low.

4.2. Improving of the gene recognition method

The short-time Fourier transform gene recognition method based on Gabor transform is given as follows. And it shows that the influence of background noise can be greatly reduced and it has a great accuracy through the data and figures.

4.2.1. Weighting optimization method

(1) Time-domain index

The average time in analysis window is $\bar{t} = 1/\|w(t)\|^2 \int_{\infty}^{\infty} t|w(t)|^2 dt$, which represents the "center of gravity" position of signal energy respected to time.

The duration of signal or the equivalent time width satisfies $(\Delta t)^2 = \sigma_t^2 = 1/\|w(t)\|^2 \int_{-\infty}^{\infty} (t-\bar{t})^2 |w(t)|^2 dt$.

if Δt is small, it means that the signal energy is concentrated around the average time. With the increasing time, signal energy will definitely decay at a higher speed, and the analysis window is a short-duration signal which has a better temporal resolution. On the contrast, if Δt is large, it means

the analysis window is a long duration signal, which has a poor temporal resolution.

(2) Localized index in frequency domain

The average frequency of the analysis window is $\overline{\omega} = 1/\left\|\hat{w}(t)\right\|^2 \int_{-\infty}^{\infty} \omega \left|\hat{w}(\omega)\right|^2 d\omega$, and $\hat{w}(\omega)$ is the continuous Fourier transform of the analysis window $w(t)$. Continuous width of the frequency domain or the signal bandwidth satisfies $(\Delta\omega)^2 = \int_{-\infty}^{\infty} (\omega - \overline{\omega}) \left|\hat{w}(\omega)\right|^2 d\omega$.

(3) Short-time Fourier transform

Using a window function on the signal to be measured, we analyzed the signal intercepted by the window once the window was panned, and this process is called short-time Fourier transform. This method can not only get the information about the frequency domain of the signal to be measured, but also observe the information of corresponding time domain. And because the shape of window remains unchanged during the transform process, it won't change the resolution of the time-frequency of the signal to be measured.

(4) Uncertainty principle

The limit of uncertainty principle is $\Delta x \cdot \Delta\omega \geq 1/2$, and the necessary and sufficient conditions for holding the quality is $W(x) = ce^{j\beta x} e^{-x^2/4a}$. Any window function of short-time Fourier transform will be subject to restrictions on uncertain principle.

(5) Gabor transform

The transform of using short-time Fourier transform with Gaussian-type window function is called Gabor transform. The Gaussian function is $g_a(x) = 1/2\sqrt{\pi a} e^{-x^2/4a}$. When $a > 0$, the expression of the time window and frequency window are:

$$\left\{ \begin{array}{l} t^* = \dfrac{1}{\|g\|_2^2} \int_R x \left|g(x)\right|^2 dx \\ \overline{\omega}^2 = \dfrac{1}{\|g\|_2^2} \int_R \overline{\omega} \left|g(\overline{\omega})\right|^2 d\overline{\omega} \end{array} \right\}.$$

4.2.2. Improving of the gene identification algorithm

(1) Determine the peak point

①Determine the center of time window and frequency window of Gaussian-type window function.

②For gene sequence $\{S\}$, in where a_i can be any value of parameter a, using the method of short-time Fourier transform can get a new sequence $\{S_i^{'}\}$.

③Superimposing all the new sequence $\{S_i^{'}\}$ obtained from all parameter a_i to get a new sequence $\{S^{'}\}$. The peak points of sequence $\{S^{'}\}$ are the exons of DNA fragment.

(2) Determine the endpoint of exons

① Get the sequence $\{RNS\}$ by sequencing the detected sequence $\{RN\}$ of DNA fragment in ascending order.

②Determine the threshold value. Use the corresponding element of $ss(ss = c \times n + 1)$ in the sequence $\{RNS\}$ as the threshold. And assign the elements less than the threshold in sequence $\{RN\}$ as 0.

③ The nonzero positions in $\{RN\}$ are the positions of exons. The zero positions are connection endpoints of exons.

(3) Improving of the model algorithm

The general value of c in $ss(ss = c \times n + 1)$ is 0.85, then we will improve it. Define $c_i \in (0,1)$, and there is only one threshold R_{i0} corresponded for $\forall c_i$.

4.2.3. Evaluation on improved gene identification algorithm

Use the improved algorithm to deal with the DNA sequence of human mitochondrial genome, and evaluate the algorithm with calculated results.

(1) Peak position.

A graph of the corresponding results for each base position is shown in Figure 2. The peak position is the position of exons.

(2) Fix the value of c_i

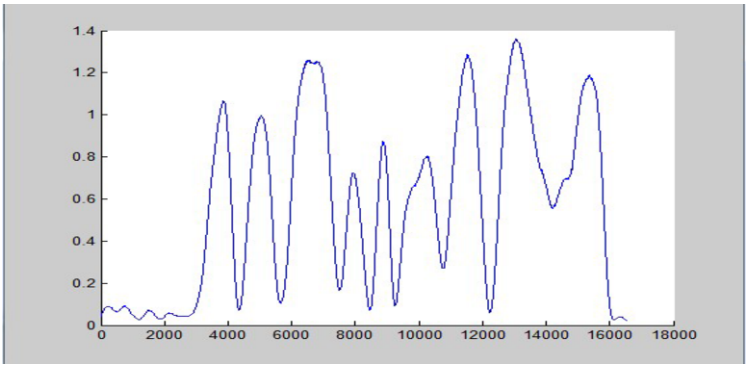

Fig. 2. The human mitochondrial genome sequence of DNA Gabor transformation curve.

Table 4

The table of human mitochondrial genome DNA sequence value of decision

Value i	value c_i	Number of successful predicted exons n_i	accuracy p_i
1	0.30	5	38.46%
2	0.35	5	38.46%
3	0.40	6	46.15%
4	0.45	6	46.15%
5	0.50	7	53.85%
6	0.55	9	69.23%
7	0.60	8	61.54%
8	0.65	7	53.85%
9	0.70	6	46.15%
10	0.75	6	46.15%
11	0.80	5	38.46%
12	0.85	4	30.77%
13	0.90	4	30.77%

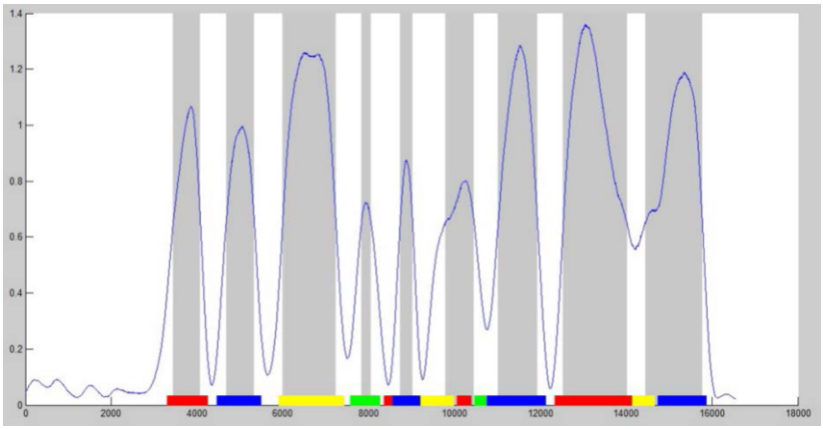

Fig. 3. Exon endpoint value and actual exon endpoint value comparison chart.

Table 5

The table of exon prediction interval

Serial number of predicted exons	Position of predicted exons	Serial number of predicted exons	Position of predicted exons
1	3445-4080	6	9786-10459
2	4861-5336	7	11020-11934
3	5996-7238	8	12536-14019
4	7833-8056	9	14446-14453
5	8734-9023	10	14455-15777

We can find that the optimal value c_i of human mitochondrial genome DNA sequence is 0.55 from Table 4.

(3) Fix the endpoint of exons

Exon endpoint value and actual exon endpoint value comparison chart is shown in Figure 3.

(4) Model Evaluation

By analyzing the Table 5, we can obviously see that 9 predicted exons (the ninth is deleted because of too short) are fully consistent in addition that some error occurs on the endpoint of exons comparing with these in Table 3. In the other word, we predicted 9 exons in 13 of them, the correct rate is 70%. It's better than the traditional Fourier transform method whose correct rate is 15%.

4.2.4. Gene prediction

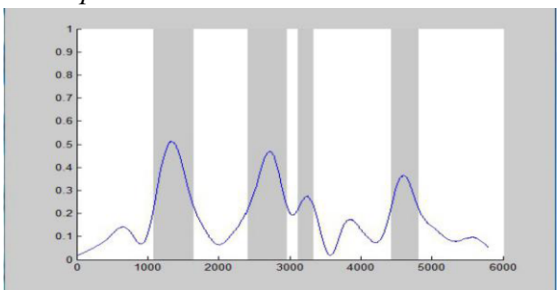

Fig. 4. Predicting outcomes of genes6-1DNA.

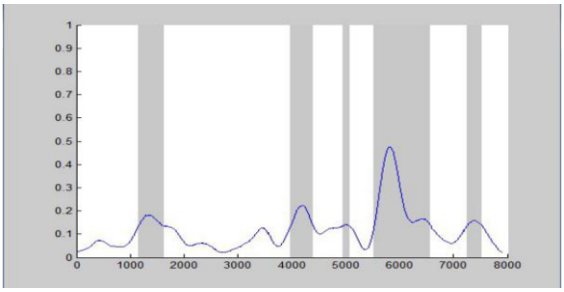

Fig. 5. Predicting outcomes of genes6-2DNA.

Then we will predict the exons of gene6 in the title database by the improved gene identification algorithm. set c_i as 0.7, and other value results of c_i in annex. The results are shown in Figures 4-9 and the exon prediction interval of DNA sequence in genes6 is shown in Table 6.

5. Conclusion

Based on the summary of three traditional threshold determination methods, this essay put forward weighted optimization method which can show different weight and discriminant rate method of exons and introns which can increase their accuracy. We used the results of four DNA sequences to evaluate different threshold determination methods. In order to analyze DNA sequence more accurately, this paper adopted and improved short-time Fourier transform method which based on Gabor transform for gene recognition. We used traversal thoughts to transform fixed percentage of exons in the DNA sequence length to dynamic percentage which aimed at different gene types. It used the improved

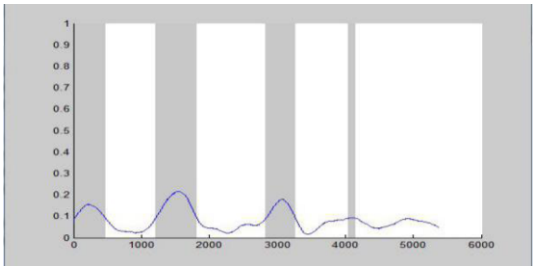

Fig. 6. Predicting outcomes of genes6-3DNA.

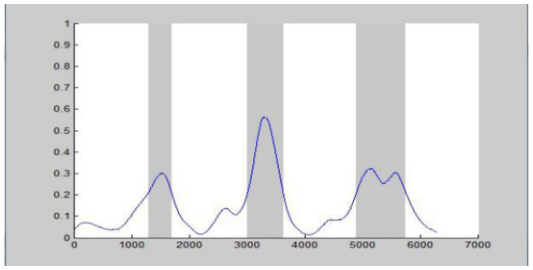

Fig. 7. Predicting outcomes of genes6-4DNA.

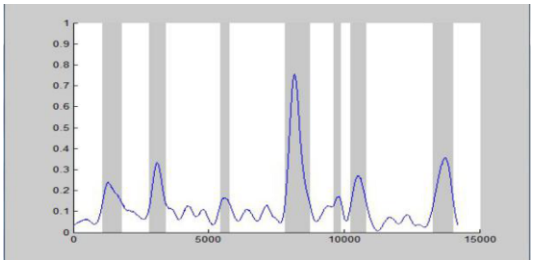

Fig. 8. Predicting outcomes of genes6-5DNA.

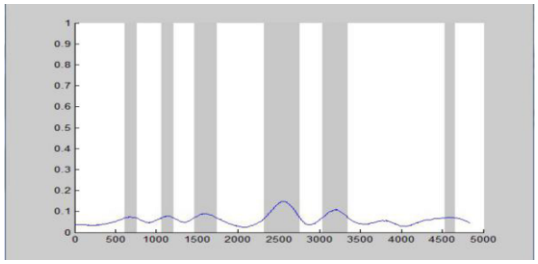

Fig. 9. Predicting outcomes of genes6-6DNA.

Table 6

Exon prediction interval of DNA sequence in genes6

Serial number	genes6-1 DNA	genes6-2 DNA	genes6-3 DNA	genes6-4 DNA	genes6-5 DNA	genes6-6 DNA
1	1085-1657	2415-2971	3123-3345	4426-4821	—	—
2	1153-1627	3978-4400	4941-5070	5529-6580	7254-7537	—
3	17-474	1206-1821	2833-3271	4049-4148	—	—
4	1291-1697	3009-3630	4987-5757	—	—	—
5	1073-1827	2933-3427	5442-5796	7816-8745	9631-9912	10250-10834
6	617-762	1067-1216	1470-1742	2327-2755	3030-3351	4528-4656

algorithm to identify the genes of the DNA sequence whose exons was known and compared the results with the actual endpoint of exons. It can confirm that the improved algorithm can judge out the

endpoint of the exons more accurately. We can imagine our method can be applied to detect novel genes which RNA-seq cannot capture and is general enough to be applied to other gene identification for the other genes.

Acknowledgments

This work was financially supported by the Project of National Natural Science Foundation of China (No. 61472228), the Project of Qingdao Applied Basic Research of Qingdao (special youth project) (No. 14-2-4-55-jch), Natural Science Foundation of Shandong province (No. ZR2014FM009) and the International Cooperation Training Project of Shandong Province Outstanding Young and Middle-aged Backbone Teachers.

References

[1] S.D. Sharma, K. Shakya and S.N. Sharma, Evaluation of DNA mapping schemes for exon detection, International Conference on Computer, Communication and Electrical Technology, Tamilnadu, 2011, pp. 18-19.
[2] C. Yin and S. Yau, A Fourier characteristic of coding sequences, origins and a non-Fourier approximation, Journal of Computational Biology **9** (2005), 1153-1165.
[3] J.F. Shao, X.H. Yan and W. Shao, 3-periodicity of DNA sequence signal, Journal of Nanjing University of Technology **34** (2012), 133-137.
[4] Z. Wang, Study on gene identification using signal processing methods, Master's thesis, Tianjin University, 2010.
[5] B.S. Ma, Y.S. Zhu and Y.Z. Chen, Gene identification using multiple statistical features, Computer Engineering and Applications **45** (2009), 118-119.
[6] D. Anastassiou, Frequency-domain analysis of biomolecular sequences, Bioinformatics **16** (2000), 1073-1081.
[7] D. Kotlar and Y. Lavner, Gene prediction by spectral rotation measure: A new method for identifying protein-coding regions, Genome Research **13** (2003), 1930-1937.
[8] C. Yin and S.T. Yau, Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence, Journal of Theoretical Biology **7** (2007), 687-694.
[9] Z.J. Liu, S.P. Feng and M.K. Pandey, Identification of expressed resistance gene analogs from peanut (arachis hypogaea L.) expressed sequence tags, Journal of Integrative Plant Biology **55** (2013), 453-461.
[10] H.B. An, H.C. Zheng and L. Zhang, Partial least squares based identification of Duchenne muscular dystrophy specific genes, Journal of Zhejiang University-SCIENCE B (Biomedicine and Biotechnology) **14** (2013), 973-982.
[11] J. Zhang, D. Xu, J. Nie, R. Han, Y. Zhai and Y. Shi, Comparative gene identification-58 (CGI-58) promotes autophagy as a putative lysophosphatidylglycerol acyltransferase, Proceedings of the National Academy of Science of the United States of America **110** (2013), 45-50.
[12] K.A. Zierler, D. Jaeger, N.M. Pollak, et al., Functional cardiac lipolysis in mice critically depends on comparative gene identification-58, The Journal of Biological Chemistry **288** (2013), 892-904.
[13] M.L. Stanly Paul, Amandeep Kaur, Ankit Geetea and M. Elizabeth Sobhia, Essential gene identification and drug target prioritization in Leishmania species, Molecular Biosystems **10** (2014), 1184-1195.