Journal of Biometrics & Biostatistics
ISSN: 2155-6180
Lin and Xiangfeng, J Biom Biostat 2016, 7:1
DOI: 10.4172/2155-6180.1000279
Research Article
Open Access

Using Mathematical Method to Solve Gene Identification Research

Dong Lin[1]* and Chu Xiangfeng[2]

[1]School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai, ROC
[2]College of Automotive Engineering University of Shanghai for Science and Technology, Shanghai, ROC

Abstract

For the identification of the gene sequences of the different types of biological "key" to construct the gene sequence screening model based on two -way clustering algorithm. First, the establishment of the FCM algorithm based on the primary model solution similar to clustering samples using two-way clustering algorithm optimized to filter out the "key" gene sequence. The problem of inaccurate forecasts for the experience of the threshold, the introduction of boots with a sampling algorithm based threshold model obtained cluster of clusters. Confidence level α = 0.05 under the highest confidence, in order to solve the species optimal threshold value selected. Checksum achieve the classification of genes coding interval 90% of the validity and accuracy of 88%, a 50% increase compared to the experience threshold algorithm. As for the random noise covering part of intron fluctuations, interfere with gene identification, the wavelet transform function is introduced into the DNA coding region prediction to filter the genes noise. Therefore, In order to solve drawbacks of coding region prediction imprecise, we establish a DNA sequence coding region prediction model based on wavelet transform. Using this model, the detection rate reached to 81%, 27% increase from the neural network method, the prediction accuracy reached to75%, 36% higher than the Fourier analysis.

Keywords: Gene identification; 3-cyclical; Mapping; Two-way clustering; Threshold, Bootlace sampling; Wavelet Transform

Introduction

With the human Genome project successfully completed, by using physical or mathematical method to obtain a wealth of biological information from a large number of DNA sequences have important theoretical significance and practical value in many aspects, such as: biology, medicine, pharmacy, etc. Gene Prediction is currently a hot research topic in the field of bioinformatics.

DNA is the carrier of genetic information, and its chemical name is Deoxyribonucleic acid, abbreviations for DNA. The DNA molecule is a long chain polymer, the DNA sequence consists of the four nucleotides adenine (A), guanine (G), cytosine (C), thymine (T). This nucleotide connected in a certain order. Wherein, the DNA fragment with the genetic information is called a gene. The DNA sequence fragments, some directly play a role in its own structure, and others involved in the regulation of the performance of the genetic message. As for the analysis of large and complex gene sequence, the traditional biology way to solve this problem is the experimental approach which based on the molecule, but its costly. In 1991 Nobel laureate W. Gilbert (Walter Gilbert, 1932 -; [the United States], first prepared by mixing DNA scientist) once pointed out: Now, based on the entire gene sequence will be known and electronic operational reside in the database, the starting point of the new biology research is the mode theory. Scientists from theoretical speculation starting, and then return to the exper iment to track or verify these theoretical assumptions.

Thus, in the study of gene prediction, signal processing and analysis methods to discover the gene coding sequences have also been extensive attention. Firstly, as for the DNA Sequence, we need to calculate the power spectra, the former specialists using the Discrete Fourier algorithm (DFT) to implement it. But, for the long DNA Sequence, the Discrete Fourier algorithm (DFT) need much time and could not afford it. Secondly, in regard to select the threshold, the experience threshold could solve most problems, but for some complex species, the older method brings inaccurate results. Here introduce bootlace-based sampling algorithm model to accomplish. Thirdly, for gene identification, random noise can interfere with the Fourier analysis, so we lead Wavelet Transform into it to filter the random noise. Therefore, the former statements indicate that we need for more investigation to continue the present study. Here, we use Fast Fourier Transform (FFT) to improve the operation efficiency, by doing this, the Operational efficiency enhance 200 times. Get help from two-way clustering algorithm model to select the represent genes. Use bootlace-based sampling algorithm model to select the proper threshold, the classification of genes coding interval 90% of the validity and accuracy of 88%, and increased by 50% compared to the experience threshold algorithm. As for the random noise covering part of intron fluctuations, interfere with gene identification, the wavelet transform function is introduced into the DNA coding region prediction to filter the genes noise. [1-3] Therefore, In order to solve drawbacks of coding region prediction imprecise, we establish a DNA sequence coding region prediction model based on wavelet transform. Using this model, the detection rate reached to 81%, 27% increase from the neural network method, the prediction accuracy reached to75%, 36% higher than the Fourier analysis.

Materials and Methods

Digital mapping and spectrum 3 – periodicity

As for the given DNA sequence, how to identify the coding sequence (exon), also known as gene prediction, the problem which not completely solved, but it is the most basic and most important problem in bioinformatics.

The problem of gene prediction method is based on statistics.

***Corresponding author:** Dong Lin, School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai 201620, Republic of China, Tel: 86-21-67791000; E-mail: knight0121@163.com

Received September 28, 2015; **Accepted** October 17, 2015; **Published** February 16, 2016

Citation: Lin D, Xiangfeng C (2016) Using Mathematical Method to Solve Gene Identification Research. J Biom Biostat 7: 279. doi:10.4172/2155-6180.1000279

Many international biological data on the site "gene identification algorithm". Such as the well-known data site which providing the gene recognition software such as GENSCAN (developed by Stanford University researchers, free use of gene prediction software), is mainly based on a hidden Markov chain (HMM) method. However, it is predicted 45,000 genes in the human genome, which is equivalent to twice the number of which is now widely recognized. In addition, statistical forecasting methods usually require the DNA sequence of the coding sequence information known as the training data set to determine the model parameters, thereby increasing the level of model predictions. But, in most case, we do not know much about the genetic information, gene recognition accuracy will be significantly decreased. Therefore, in the study of gene prediction, signal processing and analysis methods have given extensive attention in the way of discovering the gene coding sequences. In the study of the DNA sequence, at first, we should sequence the symbols of the four nucleotides A, T, G, C, based on certain rules mapped into the corresponding numeric sequence, order as on their digital processing.

I = {A,T,G,C}, Base Pair is bp for N any DNA sequence which is

$$S = \{S[n] \mid S[n] \in I, n = 0,1,2,\text{---}N-1\} \quad (1)$$

$$u_b[n] = \begin{cases} 1, & S[n] = b \\ 0, & S[n] \neq b \end{cases}, \; n = 0,1,2,\ldots N-1 \quad (2)$$

It is called *Voss* mapping, in this way creating the 0-1 sequence binary sequence

$\{u_b[n]\}$: $u_b[0], u_b[1], \ldots\ldots, u_b[\text{N-1}]$ (b$\in I$)。

For example, if some given *DNA* sequence tags is $S = ATCGTACTG$, the 0-1 sequence will be :

$\{u_A[n]\}$:{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};

So,the four DNA digital sequence is also called *DNA indicator Sequence*.

To study the DNA coding sequences (exon) characteristics, The sequence of instructions were discrete Fourier transform (DFT):

$$U_b[k] = \sum_{n=0}^{N-1} u_b[k] e^{-j\frac{2\pi nk}{N}}, \; k = 0,1,\ldots N-1 \quad (3)$$

This can get four length are N of the complex sequence $\{Ub[k]\}$, $b \in I$。Calculation every complex sequence $\{Ub[k]\}$ of square power spectrum, and additive, get the whole DNA sequence of power spectrum sequence $\{P[k]\}$:

$$P[k] = |U_A[k]|^2 + |U_T[k]|^2 + |U_G[k]|^2 + |U_C[k]|^2, \; k = 0,1,\ldots N-1 \quad (4)$$

For the same sequence of DNA, its exon and introns sequence fragment of power spectrum are usually show different characteristics (Figure 1).

Figure 1 Number BK006948.2 Yeast genes'DNA sequence of power spectrum (Because of symmetry, the actual given only half of the power spectrum). Figure 1 is genetically period exon (interval [81787,82920],length 1192 bp) corresponding power spectrum indicate sequence mapping, it has a 3 - cyclical; Figure 1 The following diagram is a gene on a period within the power spectrum of the intron (interval [96361,97551], length 1191 bp) indicates the sequence, it does not have a 3 - periodically.

It is can be seen: the exon sequence of the power spectrum curve at the frequency, with a larger Peak Value, whereas the intron has no similar peak. This statistical phenomenon is called the 3 -base Periodicity. As for the Long DNA sequences, the calculation of the power spectrum and signal-to-noise ratio, the overall amount of computation of the discrete Fourier transform (DFT) is large, it will affect the efficiency of gene identification algorithm design. Currently, DFT is an important transformation in signal analysis and processing. The drawback is the direct calculation of the DFT calculation amount is too large, and proportional to the square of the the transformation interval length of N, when N is large (greater than 210), the spectral analysis and real-time processing of the signal directly using the DFT algorithm is impractical.

Therefore, for the large calculation of directly use of the DFT, we introduce a optimization of fast algorithm, the DFT computation efficiency is improved by 1-2 orders of magnitude, creating the conditions for the digital signal processing technique is applied to real-time processing of the various signals. So, we optimized DFT algorithm speed, try to establish a Voss mapping analytical model, which based on optimized FFT algorithm to significantly improve the computing efficiency.

Gene sequence screening and threshold to determine

Set the total power of the DNA sequence spectrum of mean value for

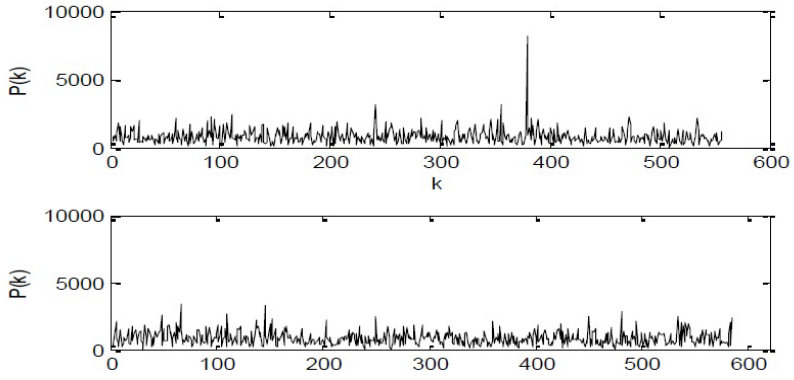


Figure 1: Number BK006948.2 Yeast genes'DNA sequence of power spectrum.

$$\bar{E} = \frac{\sum_{k=0}^{N-1} P[k]}{N} \quad (5)$$

And in a particular position of the DNA sequence, the place of power spectrum value, and the entire sequence of total power spectrum of the Ratio of average is called Signal Noise Ratio, "SNR".

$$R = \frac{P[\frac{N}{3}]}{\bar{E}} \quad (6)$$

The value of the size of the DNA sequence of the signal -to-noise ratio, represents both the relative height of Peak Value and reflect the coding or non-coding sequences 3 - periodically strength. SNR greater than a properly selected threshold value (for instance R0 = 2), is usually to meet the characteristics of the DNA sequence coding sequence fragments (exons), the intron is generally not having this property [4,5].

Two-way clustering algorithm model: On the type of DNA sequence of the particular gene, the judging threshold SNR is taken as 2 with a certain degree of subjectivity and empirical. As for different genetic types, the selected judging threshold may be different. Therefore, select a representative gene sequences is particularly important.

The selection of the general characteristics of gene sequences has a priori knowledge as a guide, that is known in the case of sample classification, the selection of characteristics contribute to the classification, there are some limitations for the selected gene sequences of different species characteristics. Due to lack of knowledge of the gene sequences of different species arranged characteristics, so we need a way to select represent genes; in the case of unsupervised represent the gene sequences of different species samples discriminate [6]. Therefore, the gene clustering function, related gene sequences classified by the expression pattern of similarity helps to study the sequence of genes of unknown function. We use the clustering algorithm model selected the representative gene sequence, that is, the direction of first characterized gene clustering, feature genes and then selected sample clustering of gene sequences. According to the representative entropy the size judgments gene clustering quality is good or bad, the introduction of the fluctuation coefficient selected to represent genes in the class. This algorithm is applied to gene expression data, the results show that the algorithm of this method, improving the accuracy of the gene selection under the lack of a priori knowledge [7].

Two-way clustering algorithm model solving steps:

(1) Network initialization: make sure SOM network initial neuron number, set the iterative times and Learning rate.

(2) The SOM gene clustering: genes as inputs, the similar expression patterns of genes classified as a class.

(3) The calculated fluctuation coefficient F: put F value calculation of the genes in each cluster respectively, singled out the gene of each cluster F value as the cluster on behalf of the Group.

(4) Calculate representative entropy H R: calculate average entropy of each gene cluster on behalf of the entropy HR and representatives of these clusters, calculate and Pick out the total characteristics of the genome represents entropy H S .

(5) Record HR and S R H of each neuron change: if the S R H is large and the HR is minimum, the neuron number is the best number of The SOM gene clustering, so do step 6. Otherwise, change the number of The SOM gene clustering (in case the number is 200), go back to step 2.

(6) The sample clustering: according to the characteristics of selected genetic makeup the new data sets of FCM clustering and get the sample parting results.

Bootlace-based sampling algorithm model: The value of the size of the DNA sequence of the signal-to-noise ratio, not only stand for Peak Value of the relative height, but also reflects the strength of the coding or non-coding sequences 3 - periodically. SNR greater than a properly selected threshold value R0 (such as R0 = 2), this is generally characteristics of the DNA sequence coding sequence fragments (exons) ,the intron is generally not having this property.

Voss mapping and Z-curve mapping described above, such as power spectrum analysis, regardless of the above analysis of the power spectrum, gene prediction accuracy is another important factor - the threshold. The setting of the threshold used to distinguish between a DNA sequence for the protein coding region and the non-coding region.

After performing a large number of simulation experiments, traditional experience threshold P = 4 does not apply to all biological, the experience threshold lack of versatility, and choose the same threshold for all kinds of different organisms lack of rationality, because different organisms have different gene structure characteristic of certain biological, if P = 4 election is too high, the predicted results with higher accuracy, and the detection rate is bound to decline, so the prediction accuracy is reduced. Conversely, as for some creature, select P = 4 may too low, a higher detection rate but low accuracy rate may occur, it also makes the prediction precision [8-10].

In addition to the factors of the different biological characteristics of its gene structure, the traditional experience threshold P = 4 does not have the versatility of other factors: power spectrum analysis of different window sizes, different power spectrum density calculating method will produce power spectral prediction change the amplitude of the curve, the apparent experience threshold will bring significant prediction error. Therefore, to solve the threshold selection problem in the power spectrum analysis method is crucial. Due to the empirical threshold will bring the prediction error, here select boots with a sampling-based algorithm, to establish Based boots with a sampling algorithm for the threshold model; use the Welch method of power spectral analysis method, predicted to the optimal threshold of distribution characteristics, and then use boots with a sampling algorithm, the optimal threshold within a certain confidence interval. Bootlace-based sampling algorithm to infer gene prediction threshold algorithm Flowchart follows (Figure 2).

Gene identification by Wavelet transform mod el

The purpose of identification of the gene is: to detect, to forecast not yet been annotated, the complete DNA sequence of the gene coding sequences (exons). The majority of gene identification algorithm results is not sufficient. For example, Fourier transform analysis of gene identification algorithm, due to the impact of the DNA sequence of the random noise, it is difficult to precisely determine the two endpoints of the interval exon [11-13].

Fast Fourier Transform algorithm model is the premise of Voss map using DFT algorithm to achieve the predicted sequence of the gene coding region; However, due to the random fluctuations of the Fourier technique, it will bring high -frequency noise, so it is difficult on the DNA sequence the coding region to achieve high -precision

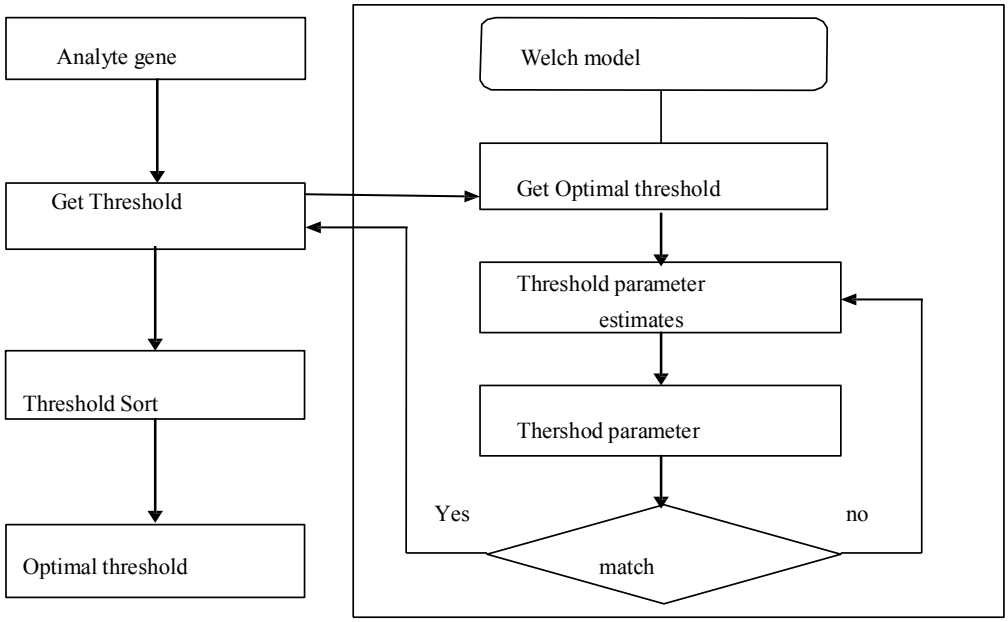


Figure 2: Welch algorithm to obtain the optimal threshold model.

prediction. Therefore, according to the Fourier analysis of the results, it is difficult for the DNA sequence coding region to achieve the high-precision prediction. So we base on Fourier transform, then rely on wavelet transform random fluctuations and separated from the useful signal, the wavelet transform can effectively remove high frequency noise caused by random fluctuations in the filter scale, so here we present a simple, quickly predict the DNA sequence of the coding region of the new method, and to improve prediction accuracy [14-17].

The wavelet transform is a new transformation analysis method; it is characterized by the transformation to fully highlight the characteristics of certain aspects of the problem, so the wavelet transform has been successfully applied in many fields. Wavelet function space satisfy the following conditions a function or signal$\psi(x)$

$$C_{\psi} = \int R' \frac{|\psi(\omega)|^2}{|\omega|} d\omega < \infty \quad (7)$$

As for every couple of real number (a,b), acquirability

$$\psi_{(a,b)}(x) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{x-a}{a}\right) \quad (8)$$

Wavelet transforms has the of the characteristics of resolution analysis, known as the analysis of signal microscope. Wavelet Transform In some filtering scale, can effectively remove the random fluctuations which caused by high -frequency noise. The predicting model based on wavelet transform DNA sequence coding region, Following model diagram as follows (Figure 3).

Results

Two-way clustering algorithm model screening gene sequence

As for specific gene sequences for several species selected gene sequence data, use two-way clustering model algorithm model, using Matlab2012a data processing, and then apply on FCM algorithm represent different biological as follows:

(1) AB304259.1Gene sequences representative of the (Saccharomyces cerevisia e ATP1a, ATP1b F1F0-ATP enzyme complex complete CD ATP1c gene) is located at [23867 bp, 26398 bp].

(2) AF100306.1 Gene sequence of (elegans cosmid T24C4, complete sequence) representative in [17856 bp, 19963 bp].

(3) CP002688.1 (Arabidopsis chromosome 5, SEQ 1 -21000) representative of the gene sequence located 10396 bp 12626 bp].

(4) NC_012920_1 (Human mitochondrial genome-wide) representative gene Sequences located 14568 bp 14862 bp].

Representative gene sequences of different species as follows (Table 1)

Optimal threshold based on the bootstrap sampling algorithm

The nucleotide sequence of Saccharomyces cerevisiae contains seven chromosomes (A-G) for the study of chromosome B sequence in the bio-data website, we find the serial number: CR3821314 of Saccharomyces cerevisiae. First, we perform optimal threshold inference method in Step 1 and Step 2 of the biological sequence to obtain the raw sample data, n = 10, experimental results are shown in the table below (Table 2).

As the optimum threshold value to obtain a 95% confidence interval (1.31, 1.65), the confidence level = 0.05. Take B = 1000 experiment from the optimal threshold distribution characteristics shown in Figure 4.

By implementing prediction method based on the power spectrum of the Welch method and set the prediction threshold of P = 1.48, the prediction accuracy is 0.90.

Similarly, using the same method can be achieved nematodes cosmid, Arabidopsis thaliana and Rodents, the prediction accuracy of the threshold value of human beings as well as on various genes are as follows (Table 3).

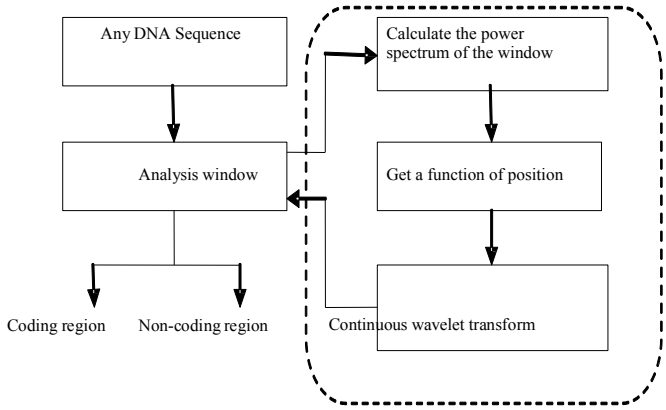


Figure 3: Coding region prediction model.

Name	Gene sequence position
Saccharomyces cerevisiae	[23867 bp, 26398 bp]
elegans cosmid	[17856 bp,19963 bp]
Arabidopsis chromosome	[10396 bp,12626 bp]
Human	[14568 bp,14862 bp]

Table 1: Representative gene sequences of different species.

bp	Optimal threshold	Specificity	Sensitivity	Prediction accuracy
Jan-00	P=1.5	0.94	0.87	0.91
6001-13500	P=2.0	0.76	0.87	0.82
13501-21000	P=1.1	0.42	1	0.71
21001-27300	P=1.6	0.94	0.88	0.91
27301-33900	P=1.1	0.55	0.89	0.72
33901-39300	P=1.5	0.86	0.87	0.87
39301-46200	P=1.8	0.97	0.94	0.96
46201-52500	P=1.6	0.95	0.76	0.86
52501-56400	P=1.4	0.91	0.94	0.93
56401-63300	P=1.2	0.93	0.87	0.9

Table 2: Saccharomyces cerevisiae original sample of experimental results obtained from the DNA sequence of the gene has been marked.

From the above analysis:

(1) Different threshold has an important impact on the prediction accuracy.

(2) The threshold selection is accurate or not directly determines the level of forecast accuracy.

(3) Thus, the gene of the spectrum or signal-to-noise ratio characteristics has great effectiveness on the classification of the coding and non -coding interval has intuitive.

Gene identification model based on wavelet transform

Firstly, from a famous website, one DNA sequence (M90075) of the selected ASYRVISP as experimental subjects. Known to the DNA sequence of the coding region, located at:522 -624,745-1041,1166-1334,1419-1584,1676-1915 and 2015-2113 respectively. Predict the coding region of the DNA sequence,using our method, the experimental results are shown in Figure 4

We can see from the Figure 5, Using Figure 5c corresponding to the filtering scale of a DNA sequence coding region of ASYRTlISP higher accuracy can be achieved in prediction and preliminary positioning. Using Figure 5c filtering scale, we predict that the coding region of The ASYRTlISP the gene DNA sequence (accession number), the results as shown in Table 4.

Discussions

From the research and experiments above, we believe that: Firstly, the computation speed of our method (FFT) is faster 200 times than the old method (DFT). Secondly, our model for select the threshold is much more exactly than the experience threshold. Thirdly, with the help of wavelet transform, our degree of accuracy for gen identification increased 27% from the neural network method.

For the first problem, the introduction of the fast Fourier transform to reduce the amount of computation, the mapping model based on optimized FFT algorithm; map the sequence of numbers under the power spectrum and the signal -to-noise ratio of the fast algorithm for mapping DNA sequence (when N = 1000, the optimal rate20,000% increase). For the second, with the help of Matlab2012a

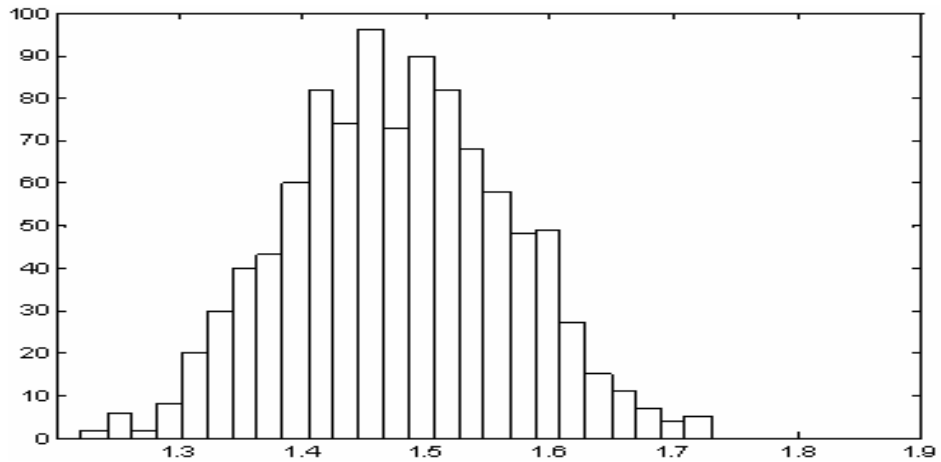


Figure 4: Optimal threshold distribution characteristics Figure.

name	Threshold	Prediction accuracy
human	1.03	0.78
nematodes cosmid	2.32	0.74
Arabidopsis thaliana	1.06	0.76
Rodents	1.16	0.81

Table 3: Several gene threshold prediction accuracy.

in bioinformatics toolbox, we ca n get the most representative of the several different biological gene sequence respectively, they are as follows (Figure 6).

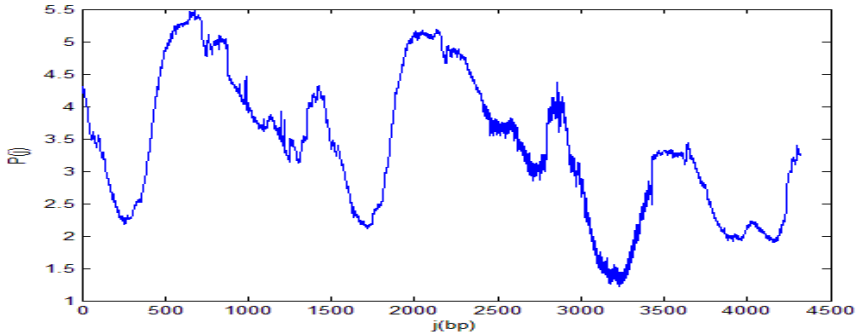
Figure 6 shows: the representative of the Saccharomyces cerevisiae gene sequence located at [23981 bp 26437 bp], with two-way clustering model algorithm selected the most representative gene sequence located at [23867 bp 26398 bp] ,essentially coincident the coincidence rate reached 98.6%; Therefore, the selection of the most representative gene sequence of two -way clustering model is more accurate. Similarly, several other creatures with the same method can be obtained:

(1) Elegans cosmid gene sequences representative located in [17926 bp, 19854 bp], the coincidence rate of 98.7%;

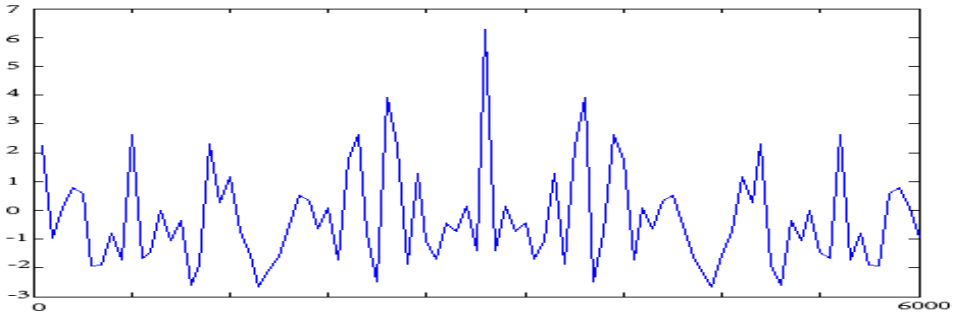
(2) Arabidopsis chromosome representative gene sequence is located [10368 bp, 12672 bp], the coincidence rate of 99.3%;

(3) The human mitochondrial genome gene sequences representative located [14302 bp, 14904 bp], the coincidence rate of 99.6%;

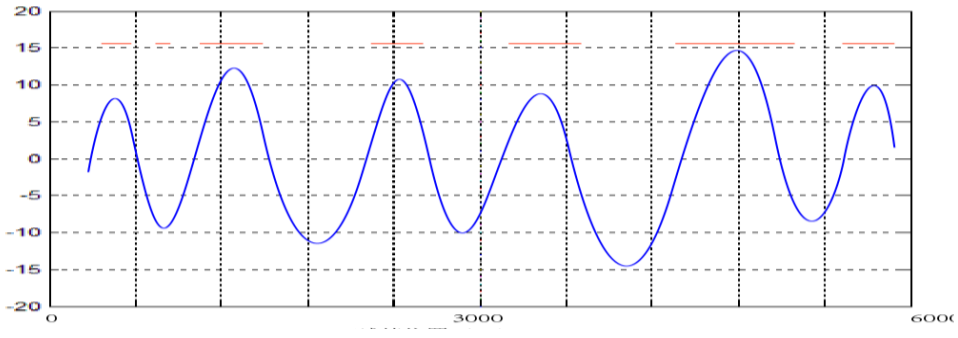
As for the last problem, For the identification of the gene sequences of th e different types of biological "key" to construct the gene sequence screening model based on two-way clustering algorithm. First, the establishment of the FCM algorithm based on the primary model solution similar to clustering samples using two -way clustering algorithm optimized to filter out the "key" gene sequence. The problem of inaccurate forecasts for the experience of the threshold,



(a)



(b)



(c) a=80

Figure 5: ASYRVISP a DNA sequence coding region of the experimental results.

Known coding sequence	The correct detection of the number of coding sequences	Erroneous detection of the number of coding sequences	The number of the coding sequence deletion	Detection n rates	Miss rate	Correct rate
6	6	1	0	1	0	0.86

Table 4: The predicted results of the coding region of accession number: M90075.

Figure 6: Representative of the Saccharomyces cerevisiae gene sequence diagram and Enlargement.

the introduction of boots with a sampling algorithm based threshold model obtained cluster of clusters. Confidence level α = 0.05 under the highest confidence, in order to solve the species optimal threshold value selected. Checksum achieve the classification of genes coding interval 90% of the validity and accuracy of 88%, a 50% increase compared to the experience threshold algorithm.

Therefore, the two-way clustering model selection of representative gene sequence is more accurate.

Therefore, in order to improve the prediction of the correct rate, we can set a

Threshold value C = 5, that is only to be considered greater than or equal to 5 peak. This can improve the accuracy of the prediction, of course, sometimes sacrificing detection rate price.

In the case of not set threshold and set threshold, the predicted sequence ASYRVISP DNA sequence (accession number: M90075) coding region of the predicted results shown in Figure 7, the prediction performance of the new method in both cases the following table (Figure 7) (Table 5 and 6).

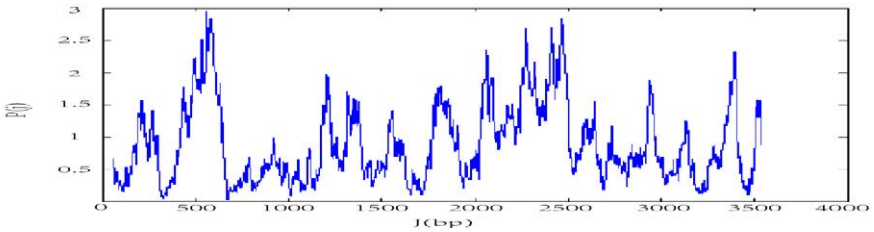
From Figure 7, we know that the DNA sequence has 6 of known coding region, and with the doors method can accurately predict 5 of them. Near 2000 bp coding region of the predicted four clutter interference. Therefore, if set threshold,it is possible to predict the 10 segments of the coding region. The prediction performance in both cases as shown in Tables 5 and 6. From the table we can see, setting a threshold can effectively improve the correct rate of forecasting methods to reduce the error rate of prediction. In order to further evaluate the effectiveness of the predicted DNA sequence of the coding region, the application of this method ASYRVISP entire data, after some strict standard screening and sorting out part of the data as follows (Table 7).

From the above table shows that:

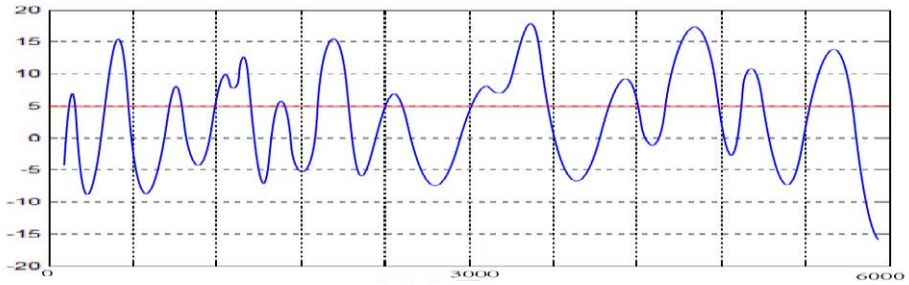
(1) The DNA sequence of the coding region based on wavelet transform prediction model, the predicted DNA sequence of the coding region of the detection rate was 81%.

(2) The DNA sequence coding region prediction model based on wavelet transform to predict the correct DNA sequence coding region was 75%.

As for the random noise covering part of intron fluctuations, interfere with gene identification, the wavelet transform function is introduced into the DNA coding region prediction to filter the genes noise. Therefore, In order to solve drawbacks of coding region prediction imprecise, we establish a DNA sequence coding region prediction model based on wavelet transform. Using this model, the detection rate reached to 81%, 27% increase from the neural network method, the prediction accuracy reached to75%, 36% higher than the Fourier analysis.



(a) the figure of PM (j)



(b) Filtering results of a=80

Figure 7: Prediction ASYRVISP sequence of the coding region of experimental results.

Known coding sequence	The correct detection of the number of coding sequences	Erroneous detection of the number of coding sequences	The number of the coding sequence of the deletion	detection rates	Miss rate	Correct rate
6	6	4	0	1	0	0.6

Table 5: Not set threshold M90075 sequence coding region of the predicted result table.

Known coding sequence	The correct detection of the number of coding sequences	Erroneous detection of the number of coding sequences	The number of the coding sequence of the deletion	detection rates	Miss rate	Correct rate
6	5	0	0	1	0	1

Table 6: Set the threshold M90075 sequence coding region of the predicted result table.

Known coding sequence	The number of probe to the coding region	The number of probe to the coding region（C≥5）	Erroneous detection of the number of the coding region.	detection rates	Correct rate
193	177	157	53	0.81	0.75

Table 7: Forecast ASYRVISP partial sequence results table.

References

1. Vinay KI, John GP (2008) Digital Signal Processing using MATLAB. Xi'a n Jiaotong University Press.
2. Yan M, Lin ZS, Zhang CT (1998) A new fourier transform approach for protein coding measure based on the format of the Z curve. Bioinformatics 14: 685-690.
3. Zhuo WA (2008) Brief review of computational gene prediction methods. Geno Prot Bioinfo pp: 217-219.
4. Arndt T (2008) Visual software tools for bioinformatics. Journal of Visual Languages & Computing 19: 291-301.
5. Gatto JG (2003) The changing face of bioinformatics. Drug Discov Today 8: 375-376.
6. Seth EM, Barbara HM, Robert AP (2011) Blurring the line between bioinformatics and patent analysis. World Patent In-formation 33: 257-259.
7. Nicolli A, Chiara F, Bortoletti I, Pasqualato F, Mongillo M, et al. (2011) [Release of metals from metal-on-metal hip prostheses]. G Ital Med Lav Ergon 33: 257-259.
8. Francesco M, Sushmita M (2009) Natural computing methods inbio informatics: A survey. Information Fusion 10: 211-216.
9. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78-94.
10. Altschul SF, Gish W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
11. (2002) Wag the dogma. Nat Genet 30: 343-344.
12. de Souza SJ, Camargo AA, Briones MR, Costa FF, Nagai MA, et al. (2000) Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. Proc Natl Acad Sci U S A 97: 12690-12693.
13. Anastassiou D1 (2000) Frequency-domain analysis of biomolecular sequences. Bioinformatics 16: 1073-1081.
14. Kotlar D, Lavner Y (2003) Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. Genome Res 13: 1930-1937.
15. Berryman MJ, Allison A (2005) Review of signal processing in genetics. Fluctuation and Noise Letters 5: 13-35.
16. Sharma SD, Shakya K, Sharma SN (2011) Evaluation of DNA Mapping Schemes for Exon Detection. International Conference on Computer, Communication and Electrical Technology- ICCCET.
17. Yin C, Yau SS (2007) Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J Theor Biol 247: 687-694.