Applied Mathematics & Information Sciences
An International Journal

Identification of 5'UTR Splicing Site Using Sequence and Structural Specificities Based on Combination Statistical Method with SVM

Lv Jun-Jie[1,2], *Wang Ke-Jun*[1], *Feng Wei-Xing*[1], *Wang Xin*[3] *and Xiong Xin-yan*[1]

[1]College of Automation, Harbin Engineering University, Harbin, China
[2]College of Bioinformatics Science and Technology, Harbin Medicine University, Harbin, China
[3]Cancer Research Centre, University of Cambridge, Cambridge U.K

Received: 9 Oct. 2012, Revised: 12 Nov. 2012, Accepted: 21 Nov. 2012
Published online: 1 Feb. 2013

Abstract: To identify untranslated regions (UTR) splice sites more accurately and efficiently, a method for the recognition of UTR splice sites using both splicing sequences and secondary structures of flank sequence information based on combination statistical method with support vector machine was proposed. The method consists of two stages: a statistical method is used in the first stage and a support vector machine (SVM) with polynomial kernel is used in the second stage. The statistical method serves as a pre-processing step for the SVM and takes UTR sequences as its input. It models the compositional features and dependencies of nucleotides in terms of probabilistic parameters around splice site regions. The probabilistic parameters are then fed into the SVM, which combines them nonlinearly to predict splice sites. Then the Mfold package in Vienna soft was used to predict the most stable secondary structure of flank sequences. The traditional four-letter alphabet was converted into eight-letter alphabet sequence. The sequence- structure combination strings were used for training models then recognized splice sites by the well trained models. Using the actual 5'UTR splice dataset of human gene tested the method; it shows a good performance for UTR splice sites recognition.

Keywords: Splice Sites; Untranslated Regions (UTR); Splice Sequence; Secondary Structure

1. Introduction

Gene untranslated region (UTR) will not be translated into protein, but it has an important role in regulating gene expression; many studies have shown [1, 2]: mutations and activity of UTR may be related to many diseases and even cancer. For example, globin gene in thalassemia patients, about $1/4$ of the nucleotide mutations in the intron 5' untranslated region or 3' untranslated region of conserved sequence [3], or directly interferes with the pre-mRNA normal splicing [4]. The 5' UTR length will affect the accuracy of the translation efficiency and start when the length between 17∼80nt [5], in vitro translation efficiency is proportional to its length change [6]. 5'UTR in base pairing formation of secondary structure will prevent the migration of 40S ribosomal subunits [7], and inhibition translation initiation [8]. 3 'UTR untranslated region has an important regulatory role of transcript stability, translation capabilities [9], and control of mRNA subcellular positioning to further understand the mechanisms of gene regulation [10], developing new therapies for genetic diseases [11, 12], and understanding the mechanism of cancer development and its treatment [13], it is necessary to make deep research into the mechanism and function of gene's UTR. Splice site recognition in UTR is the key question. Compared to the splice site recognition in coding region, one of the greatest difficulties in the UTR splice site recognition is: it cannot rely on the state transitions from coding to non-coding region. No matter how the intron removal from reading frame, for most of the methods to capture the conversion from non-coding region to the coding region is not difficult, which reduces the difficulty of identifying splice sites. Due to the lack of such a conversion in UTR, the accuracy of the method based on this conversion mechanism is greatly reduced in UTR splice site recognition. Moreover, the method using the DNA sequence similarity with the target protein cannot

* Corresponding author e-mail: lvjunjie525@126.com

be used to UTR splice site recognition (because UTR is not translated into protein).

Currently, only FIRSTEF [4] and NetUTR [5] can provide a comprehensive recognition of splice sites in gene UTR with relatively high accuracy, however, the recognition accuracy is not good. FIRSTEF is a powerful tool used to find the promoter and the first exon in the coding region and untranslated region, its recognition sensitivity of 86% and false positive rate of 17% for true splice site. The innovation of FIRSTEF is proposed existence conserved non-coding motif in untranslated region, for example: the CpG level near the transcription start point 500 bases. However, FIRSTEF can predict the first donor site, but cannot identify the location of the first acceptor sites, and more than 40% of the 5' untranslated region contains more than one exon, the data suggest that at least 9% of the 5' untranslated region contained second non-coding exon, at least 3% has the third non-coding exon, a special case, AF135187, contains four complete non-coding exon. FIRSTEF cannot to identify these acceptor sites. Eden and Brunak proposed NetUTR to identify splice sites in 5'UTR, It used neural networks to model splice sites, compared with FIRSTEF, and its recognition accuracy has been greatly improved. But for the strict data requirements, is not a very common method. Donor sites and acceptor sites in untranslated region are both located in the junction of intron and non-coding exon, to identify these splice sites is more difficult than traditional coding region.

To improve the accuracy of splice site recognition in UTR, we proposed a novel method, By analyzing the splicing sequences and secondary structures of flank sequence characteristics of donor sites and acceptor sites, donor sites in UTR identification based on the maximum correlation decomposition (MCD) with support vector machine model, acceptor sites UTR identification by the first order Markov model (MM1) method with support vector machine model were built respectively. Then the Mfold package in Vienna soft was used to predict the most stable secondary structure of flank sequences [6]. The predicted structures were converted to a string of two-symbol alphabet. With the combination of S and L symbols and four-letter nucleotide alphabet, each sequence was converted to an eight-letter alphabet sequence, the sequence-structure combination strings were used for training models, then recognized splice sites by the well trained models. Experimental results show that our proposed method is more effective than other existing methods.

2. Materials and Methods

2.1. Dataset

The data we need are all from ftp://ftp.ebi.ac.uk/pub/databases/UTR/data/, we extract 5'UTR sequence which contains at least one complete non-coding exons, and Comply with the GT-AG rule, removal the sequences Tags for alternative splicing site, in order to ensure the accuracy of the algorithm, we used BLAST algorithm to redundant processing for data sets, through similarity comparison, removal of excessive similarity UTR sequence. After the treatment, we get 380 UTR sequences, where the true donor sites and the true acceptor sites are all 453; 59276 false donor sites and 78721 false acceptor sites.

We set these data into training set and test set to train and test the model, the training set contains 300 UTR sequences, 359 true donor sites, 359 true acceptor sites, 49 163 false true donor sites, 65103 false acceptor sites; test set contains 80 UTR sequence, 94 true donor sites, 94 true acceptor sites, 10113 false true donor sites f, 13,618 false acceptor sites.

2.2. Secondary predicted

The Vienna package was used to predict the most stable fold for each flank sequence. The Mfold program in the package predicts the minimum free energy structure of a single sequence, based on the algorithm originally developed by Zuker and Stiegler [3]. The predicted structures were converted to a string of two-symbol alphabet (i.e. S, L) corresponding to whether each nucleotide is paired or unpaired, respectively. Then, with the combination of L and S symbols and four-letter nucleotide alphabet (i.e. A, T, C, G), each sequence was converted to an eight-letter alphabet sequence. The nucleotide sequences of splice sites (four-letter) and the sequence-structure combination strings (eight-letter) were used for UTR splice site recognition (see below).

2.3. Method description

The proposed method for UTR splice site recognition combined probability parameters with support vector machine consists of two main stages:

Overview of the proposed method as Fig.2.1:

The first stage used probabilistic model, and the second stage is a support vector machine (SVM) with polynomial kernel used the probability from the first stage as the input parameter. Probabilistic model serve as the SVM pre-processing stage to characterize the basic relationship and the sequences composition features in the vicinity of splice sites sequences in the form of probability parameters, then put these probabilities into the SVM with a polynomial kernel function for classification. Firstly, predicting the secondary structure for the input sequences combined the structural information with the sequence information; then building different probability models for donor sites and acceptor sites using different model in the pre-processing stage,

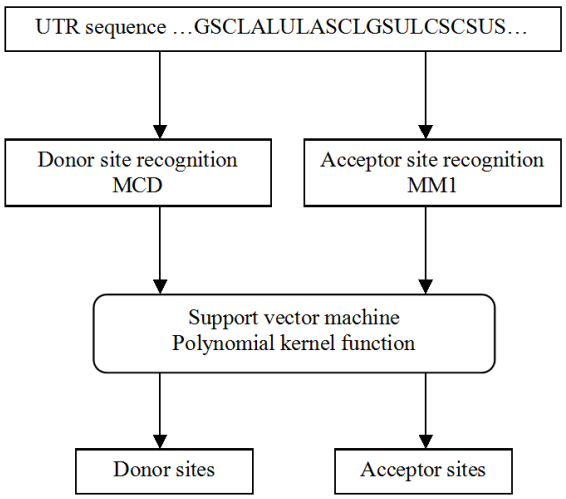

Figure 2.1 Schematic overview of the proposed method

respectively. In second stage the probability of the first phase serve as the input parameters and couple with support vector machines of polynomial kernel to identify the two kinds of sites.

2.4. Donor site recognition

Biological concept [10]: a striking similarity between the 5'UTR donor consensus patterns and the patterns in translated region donor sites, this similarity in local splice signal indicates that the splicing is constrained at the nucleotide level and both types of splice sites are complementary to the 5' end of the U1 snRNA. The only difference is an increased tendency for G and C at positions -6 to -10 and 7 to 9, respectively, at the 5'UTR donor sites. the six bases in front of the intron downstream and three bases at last of the non-coding exon upstream are still conserved, Considering this binding region as splice signal for building donor site signal model, we use the maximum correlation decomposition (MCD) method [5].

The MCD is to build a model; the model can capture the correlation between non-adjacent bases as the adjacent bases. It is to identify the bases with the greatest correlation with other locations in the sequence, divided the training data into two categories according to whether they contain this base. Repeat this process until each sub-class training data bases is less than a threshold value, and then, to build a low-order Markov model for each subclass, respectively. MCD captured correlation between non-adjacent bases through use conditional low-order Markov model to replace the unconditional low-order Markov model. To get a consensus sequence the MCD using the base of maximum probability of each position as a consistent base. According to the consensus sequence, define the variable C_i. When the bases at the sequence position i, the same as the position in the consensus sequence, $C_i = 1$ else $C_i = 0$; used χ inspection to verify the correlation between bases, and its implementation formula as following:

$$\chi^2(C_i, o_j) = \sum_{C_i=0}^{1} \sum_{o_j=A}^{T} \frac{(n^o_{C_i o_j} - n^e_{C_i o_j})^2}{n^e_{C_i o_j}} \tag{2.1}$$

Where $n^o_{C_i o_j}$ represents in class $C_i(C_i = 1/C_i = 0)$ data, the sequence observations number of the base O_j in position j, $n^e_{C_i o_j}$ is the expected number of the sequences. When assumptionO_j with C_i independent, $n^e_{C_i o_j}$ is:

$$n^e_{C_i o_j} = n_{C_i} n_{o_j} / N \tag{2.2}$$

Where n_{C_i} denotes the total number of class C_i, data, n_{o_j} represents all locations j at the base O_j is the number of sequences, N is the total number of all sequences. We set the interception probability $p = 0.001$, corresponding to the χ^2 with freedom 3, the statistics value is14.17. when the $\chi^2(C_i, o_j)$ values greater than 14.17, we believe that the independence assumption does not hold, i.e. j and i is correlation, taking the correlation between non-adjacent bases and the impact of the content of guanine G and cytosine C on splice site recognition, it appropriate to take larger sequence window. Here, taking -15 to +15 (splice sites for the origin) as the donor site recognition window, the output probability as a pre-process for donor site recognition, next, using the output probability as a support vector machine(SVM) input for classification.

The donor site recognition model shown in Fig.2.2.

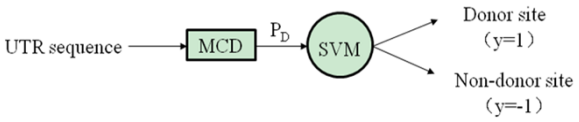

Figure 2.2 Donor site recognition model

Probabilistic model for donor site with MCD is calculated as follow:

$$p(o) = \prod_{j=1}^{N(d)} p_j(o) \prod_{i=1}^{L(d)} p^d_{i,o_i} \tag{2.3}$$

Where $N(d)$ is the number of the d-th PWM matrix decomposition, $p_j(o)$ is the selective probability of subsequence O in the j-th decomposition, $L(d)$ is the length of sequence in the d-th PWM matrix, p^d_{i,o_i} is the probability of base o_i at position i in the d-th PWM matrix.

The same method is used to build model for pseudo splice sites, the probability is calculated as $p'(o)$,for ease of calculation and improve the computing speed, math.log $p(o)$ and $p'(o)$ to and , then subtract the two logarithmic, according whether the subtraction result P_D is greater than 0 to evaluate the predicted sequence O is a real donor site sequence or not. P_D is calculated as follow:

$$P_D = \log p(o) - \log p'(o) \tag{2.4}$$

Then use the P_D which greater than 0 as the probabilistic output of donor recognition model, which will serve as the input of the second stage SVM. Set the output of SVM that$y = +1$ represent donor site, $y = -1$ for non-donor site. Prediction window's size is taken from -15 to +15.

2.5. Acceptor site recognition

Correlations in the 5'UTR acceptor splice sites were analyzed qualitatively and visualized using sequence logos made at the single nucleotide, dinucleotide and trinucleotide levels and compared with that of translated region splice sites [3]. In addition to the strong well known consensus at the acceptor site, the 5'UTR acceptor site pyrimidine tract typically extends through position -3 and gradually fades until position -26, where it stops. It has a weaker bias for cytosine at position -3 and slightly stronger bias at position -4 and 4 than that of coding region acceptor splice sites. The bias for thymine is stronger at several positions including -5, -6 and -12.

Through the analysis of acceptor sites, we can see that acceptor sites has a long conserved sequence, but the distance correlation between the base in sequence is not strong, first-order Markov model (MM1) could capture adjacent correlation between the location of bases, we get -30 to +10 around acceptor sites as identifiable information window, using the MM1 for pre-processing, and then apply the SVM for classification.

The acceptor site recognition model shown in Fig.2.3.

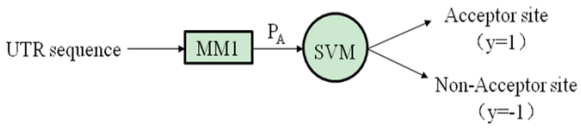

Figure 2.3 Acceptor site recognition model

Each nucleotide in a UTR sequence corresponds to a state in the Markov chain used, whose observed state variables are drawn from the alphabet:

$$S_{DNA} = \{AS, AL, CS, CL, GS, GL, TS, TL\} \tag{2.5}$$

We define an arbitrary sequence of length l (here $l = 40$):

$$l : \{s_1, s_2, \ldots, s_l\} \tag{2.6}$$

Where , $s_i \in \{AS, AL, CS, CL, GS, GL, TS, TL\}, \forall i \in \{1, 2, \ldots, l\}$ then the nucleotide s_i is a realization of the ith state variable of a Markov chain, and state transition is only allowed from state i to its adjacent state $i+1$. Hence, the model consists of states ordered in a series. It evolves from state s_i to s_{i+1} and emits symbols from the alphabet S_{DNA}, where each state is characterized by a position-specific probabilistic parameter. Assuming a Markov chain of order k, the likelihood of a sequence given the model is:

$$p(s_1, s_2, \ldots, s_l) = \prod_{i=1}^{l} p_i(s_i | s_{i-1}) \tag{2.7}$$

Where the Markovian probability $p_i(s_i) = p(s_i | s_{i-1}, s_{i-2}, \ldots, s_P i-k)$ denotes the conditional probability of a nucleotide at location i given the k predecessors. Such a model is characterized by a set of parameters: $\{p(s_i | s_{i-1}, s_{i-2}, \ldots, s_{i-k}) : s_i, s_{i-1}, s_{i-2}, \ldots, s_{i-k} \in S_{DNA}, I = 1, 2, \ldots, l\}$

MM1 is used to model a set of nucleotides in a UTR sequence. The Markovian parameters are expressed interms of position-specific first order conditional probabilities $(k = 1)$

$$p_i(s_i) = p(s_i | s_{i-1}) \tag{2.8}$$

The model is then characterized by the set of parameters: $\{p(s_i | s_{i-1}) : s_i, s_{i-1} \in S_{DNA}, i = 1, 2, \ldots, l\}$. It is shown that the likelihood of a sequence given a model M can be approximated by a polynomial of conditional probabilities:

$$p(s_1, s_2, \ldots, s_l) \approx p(s_1) \prod_{i=2}^{l} \sum_{j=1}^{i-1} b_{ij} p(s_i | s_{i-1}, \ldots, s_{i-j}) \tag{2.9}$$

Loi-Rajapakse has used this method with neural network to identify splice sites, and achieved good results [11]. Then we applied SVM with polynomial kernel to classify MM1 encoded splice site data. Based on the training, a SVM can classify splice site. The SVM is a canonical machine learning algorithm initially proposed by Vapnik. It uses a hypothetical space of linear functions in a high dimensional feature space trained with a learning algorithm based on optimization theory. The SVM description shown in Fig.2.4.

SVM classification is an optimization problem given by:

$$\begin{aligned} &Maximize\ f(a) \\ &= \textstyle\sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} a_i a_j y_i y_j K(x_i, x_j) \\ &s.t. \textstyle\sum_{i=1}^{l} a_i y_i = 0 \\ &0 \le a_i \le C, i = 1, 2, \ldots, l \end{aligned} \tag{2.10}$$

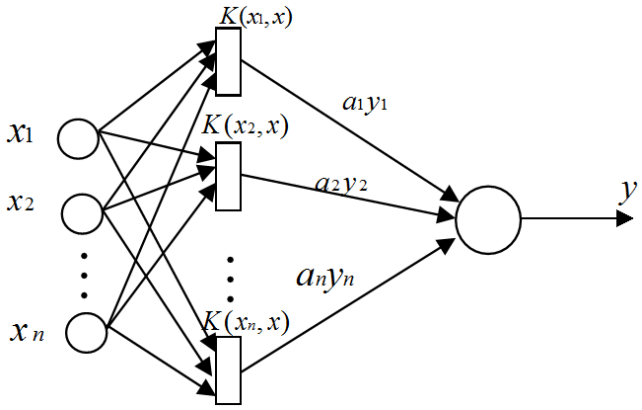

Figure 2.4 Description of the SVM

Where, l is the number of training examples, k is the kernel function, x is the input vectors, y is either -1 or +1 representing two different classes, is the variable to be optimized and C is a trade-off parameter for generalization performance. Each corresponds to one particular training example and after the training process, only a subgroup of will have non-zero values. This subgroup of and their corresponding training examples are called the support vectors. The class labels y in the two classifiers would then indicate true ($y = +1$) or false sites ($y = -1$) for splice site and non-splice site accordingly. Input x would always be a vector of MM1 probabilities.

Given a query UTR segment z, the trained SVM classifies based on the decision function:

$$o(z) = sign[\sum_{i \in v} a_i y_i K(x_i, z)] \tag{2.11}$$

Where v is the set of support vectors.

The kernel function in our classifiers is a second order polynomial :

$$K(x,z) = (< x \bullet z > +1)^2 \tag{2.12}$$

Where $< \bullet >$ indicates a dot product.

Expanding (2.8),we obtain:

$$K(x,z) = \sum_{(i,j)=(1,1)}^{n,n} (x_i, x_j)(z_i, z_j) + \sum_{i=1}^{n} (\sqrt{2}x_i)(\sqrt{(}2)z_i) + 1$$

Where n is the number of dimensions in vectors χ and z, and chi_i and z_i are the i-th element in vectors χ and z respectively. Substituting (2.13) into (2.11), the output $o(z)$ becomes a polynomial over z, with the polynomial constants determined by α and x of the set of support vectors. Since z is a vector of conditional probabilities of a sequence of length l:

$$z = [p(s_2|s_1), p(s_3|s_2), \ldots, p(s_l|s_{l-1})] \tag{2.13}$$

The output $o(z)$ in its polynomial form resembles equation (2.8).

The training sequences were aligned with respect to the consensus dinucleotides prior to stage one. The estimates of the MM1 are the ratios of the frequencies of each dinucleotide in each sequence position as shown in (2.15). Only the true splice site training sequences were used to create the Markov model. The desired output level is set to $+1$ or -1 depending on the true or false splice site class label.

$$\hat{p}_i(s_i) = \frac{\#(s^i_{i-k})}{\#(s^{i-1}_{i-k})} \tag{2.14}$$

3. Predictive accuracy measures

he classification performance is defined by the sensitivity (Sn), specificity (Sp), false positive ratio (FP %), and false negative ratio (FN %) of the model. The sensitivity, also known as true positive rate (TP %), is the percentage of correct prediction of true sites and specificity is the percentage of correct prediction of false sites. Specificity is the correct prediction of the false sites as defined below:

$$S_n = \frac{TP}{TP+FN} \tag{3.1}$$

$$Sp = \frac{TN}{TP+FP} \tag{3.2}$$

$$FP\% = \frac{FP}{FP+TN} \times 100\% \tag{3.3}$$

$$FN\% = \frac{FN}{TP+FN} \times 100\% \tag{3.4}$$

Where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives and FP is the number of false positives. Sp is proportion of predicted real sites that are actually real, while Sn is the proportion of real sites that have been correctly predicted as real. Since neither Sp nor Sn alone constitutes good measures of global accuracy, other measures are developed. The goal of our method is to get lower FP % when in higher Sn.

4. Results and discussion

The data we used for training and testing the proposed method are all download from ftp://ftp.ebi.ac.uk/pub/databases/UTR/data/, we got the results shown in TABLE I.

To evaluate the performance of the recognition algorithm, we comparative and analysis of the effect of the donor sites recognition model and acceptor sites recognition model with the existing corresponding recognition software, respectively. Compare the

Table 1 Splicing site recognition results

Splicing site	Sn (%)	Sp (%)
Donor site	87	83
	84	80
	75	73
	70	64
	60	58
Acceptor site	85	82
	80	75
	75	71
	70	62
	60	56

Table 2 Donor site recognition results comparison

Donor site	Sn (%)	Sp (%)
NetUTR	66	38
GeneSplicer	13	33
MCD-SVM	87	83

Table 3 Acceptor site recognition results comparison

Acceptor site	Sn (%)	Sp (%)
NetUTR	60	24
GeneSplicer	22	31
MM1-SVM	85	82

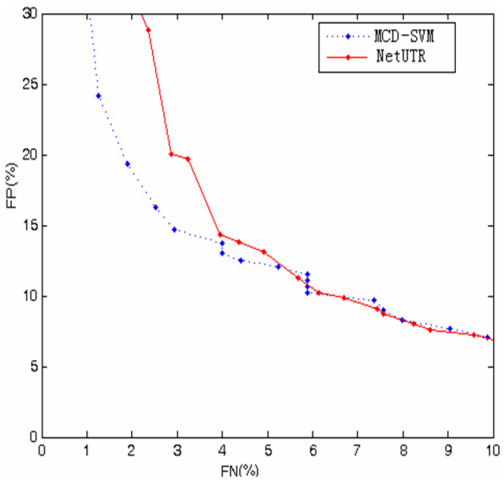

Figure 4.1 ROC curves for donor site identification

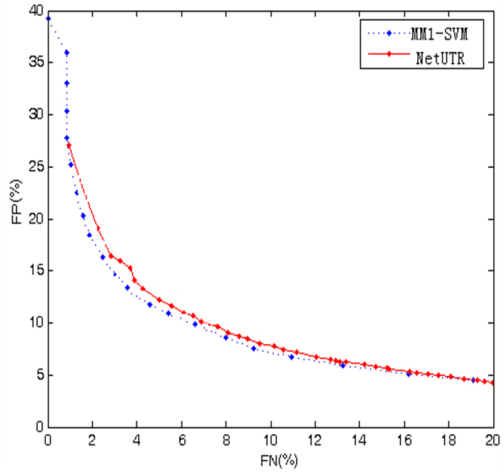

Figure 4.2 ROC curves for acceptor site identification

performance of donor sites and acceptor site recognition model with the existing recognition software, the results shown in TABLE II and TABLE III.

To further straightforward to see the prediction accuracies of these methods above, ROC curve (receiver operating curve) is used to evaluate model performance intuitively, the curve can be very manifest the performance. Fig.4.1. and Fig.4.2. show the ROC analysis of NetUTR, MCD-SVM, and MM1-SVM. When a ROC is created from the false positive rate FP% (the y axis) and the false negative rate FN% (the x axis) of a model, the ROC curve the closer to the origin $(0,0)$, the more accurate the model.

Through analyzing the results from TABLE I we can see that the performance of MCD-SVM is significantly higher than the other two on identifying donor sites, When evaluating the performance of the recognition method, higher sensitivity and specificity represent better result, in the experiment, the Sensitivity (Sn) of MCD-SVM is up to 87%, However, NetUTR and GeneSplicer method were 66% and 13%, respectively, the specificity(Sp) of MCD-SVM is 83%, NetUTR and GeneSplicer method were 38% and 33 %, under the indicator of specificity, MCD-SVM is much better than the other two; By comparing these methods, the Sensitivity (Sn) of NetUTR can be slightly lower than our MCD-SVM, but GeneSplicer is much lower than the other two, because GeneSplicer is design based on the characteristics of the coding region, the algorithm Used the characteristics of coding region, and UTR doesnt have these features, using it for UTR splice sites identified caused its recognition rate lower. As the results shown in TABLE II we can see that the performance of first-order Markov model combined with support vector machine method (MM1-SVM) is significantly higher than NetUTR and GeneSplicer for identifying acceptor sites, the Sensitivity (Sn) of MM1 -SVM is up to 85%, However, NetUTR and GeneSplicer are 60% and 22%, respectively, under the indicator of specificity, the superiority of MM1 -SVM is more obvious, the specificity (Sp) of MM1-SVM is up to 82%, while NetUTR and GeneSplicer were 24%

and 31%, respectively. After comparing the results from TABLE I and TABLE II we can also find that the recognition performance of acceptor sites is lower than that of donor sites in UTR, this phenomenon may be due to sequence conservation in the vicinity of acceptor sites is relatively poor than that of donor sites, its characteristics are not easy to extract. Through observing Fig.4.1. and Fig.4.2., we can see that the ROC curve of both MCD-SVM and MM1-SVM are all closer to the origin(0,0) than NetUTR, which indicating that the performance of our proposed method which combine probability and statistics with support vector machine is better than NetUTR for UTR splice site recognition.

5. Summary

In this paper we presented a new method for splice sites identification in eukaryotic gene untranslated coding regions (UTR). It based on splicing sequences and secondary structures of flank sequence information using incorporation statistical probability model with support vector machine, and in accordance with the fact that donor sites and acceptor sites have different statistical properties. Donor sites in UTR identification is based on the maximum correlation decomposition (MCD) with support vector machine model, and acceptor sites UTR identification using the first order Markov model (MM1) method with support vector machine model. Experimental results show that our proposed method is more effective than other existing methods. However, due to the understanding of splice sites in UTR is not deep, the establishment statistical models are only approximate, the biological information considered is not enough, and there is a theory recognition accuracy limit threshold of recognition for splice sites in UTR. To further improve the recognition performance, research on splice site prediction in UTR will focus on larger feature sets, since more biological information to achieve better results. Other future directions we would like to develop of more complex features that capture other nucleotide dependencies at the feature level.

Acknowledgement

This research was partially supported by China National Natural Science Foundation (61071174); China National 863 High-Tech Program (2008AA01Z148); Fundamental Research Funds for the Central Universities (HEUCFT1102); the Fundamental Research Funds for the Central Universities of China under Grant (HEUCF110424).

References

[1] M. J. Clemens and U. A. Bommer, Translational control: The cancer connection, Int J Biochem Cell Biol.31 (1999) 1-3.

[2] I. Korf, P. Flicek, and D. Duan, Integrating genomic homology into gene structure prediction, Bioinformatics. **17** (2001) 140-148.

[3] E. Eden and S. Brunak, Analysis and recognition of 5' UTR intron splice sites in human pre-mRNA, Nucleic Acids Res. **32** (2004) 1131-1142.

[4] M. Zuker, Mfold web server for nucleic acid folding and hybridization Prediction, Nucleic Acids Res. **31** (2003) 3406-3415.

[5] Y. L. Lai, J. R. Jiang, A Genetic Algorithm for Data Mule Path Planning in Wireless Sensor Networks, Appl. Math. Inf. Sci. **6** (2012) 53-59.

[6] M. Pertea, X. Y. Lin and S. L. Salzberg, GeneSplicer: a new computational method for splice site prediction, Nucleic Acids Res. **29** (2001) 1185-1190.

[7] E. Karapinar, I. M. Erhan and Y. Ulus, Fired Point Theorem for Cyclic Maps on Partial Metric Spaces, Appl. Math. Inf. Sci. **6** (2012) 239-244.

[8] D. C. Fischer, K. Noack, I. B. Runnebaum, et al. Expression of splicing factors in human ovarian cancer, Oncology reports. **11** (2004) 10851090.

[9] G. S. Wang and T. A. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. Nature reviews, Genetics. **8** (2007) 749-761.

[10] X. Wang, G. Wang, C. Shen, et al. Using RNase sequence specificity to refine the identification of RNA-protein binding regions, Appl. Math. Inf. Sci. **9** (2008) 1-17.

[11] B Jian, L. J. Zhou and Y. Yan. Analysis on complexity of neural networks using integer weights, Appl. Math. Inf. Sci. **6** (2012) 317-323.

[12] C. Berasain, S. Goni, J. Castillo, et al. Impairment of pre-mRNA splicing in liver disease: Mechanisms and consequences, World Journal of Gastroenterology. **16** (2010) 3091-3102.

[13] C. Zhang, M. A. Frias, A. Mele, et al. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls, Science. **329** (2010) 439-443.

Lvjunjie Lv received the Master degree in pattern recognition and intelligent system from Harbin Engineering University, Harbin, China, in 2010. Now study the PhD of pattern recognition and intelligent system in Harbin Engineering University from 2009. Her research interests include bioinformatics and biometric.