Recognition of Gene Acceptor Site Based on Multi-objective Optimization

Jing ZHAO[1,3,4], Yue-Min ZHU[2], Pei-Ming SONG[1], Qing FANG[1], and Jian-Hua LUO[1,3]*

[1] *School of Life Science & Technology, Shanghai Jiaotong University, Shanghai 200240, China;*
[2] *CREATIS, CNRS UMR 5515, INSA Lyon, 69621 Villeurbanne Cedex, France;*
[3] *Shanghai Center for Bioinformation and Technology, Shanghai 200235, China;*
[4] *Logistical Engineering University, Chongqing 400016, China*

Abstract A new method for predicting the gene acceptor site based on multi-objective optimization is introduced in this paper. The models for the acceptor, branch and distance between acceptor site and branch site were constructed according to the characteristics of the sequences from the exon-intron database and using common biological knowledge. The acceptor function, branch function and distance function were defined respectively, and the multi-objective optimization model was constructed to recognize the splice site. The test results show that the algorithm used in this study performs better than the SplicePredictor, which is one of the leading acceptor site detectors.

Key words multi-objective optimization; acceptor site; recognition

The process of cutting introns out of immature RNAs and stitching the exons together to form the final product is called RNA splicing. Most introns in nuclear mRNA precursors begin and end in the same way: exon/GU-intron-AG/exon. This is called the GU-AG rule [1]. The product of the nuclear mRNA precursor splicing, in which an adenosine nucleotide in the middle of the intron attacks the phosphodiester bond between the first exon and the beginning G of the intron, looks like a lariat. However, GU-AG motifs occur so frequently that a typical intron will contain several GUs and AGs within it. Therefore, using the GU-AG rule alone to predict the acceptor site will result in many false sites. The sequence characteristics around the donor and the acceptor site must be considered simultaneously.

Many methods have been developed for recognition of the acceptor site, such as the hidden Markov model, the BP neural network and the supported vector machine. Acceptor site recognition based on the hidden Markov model [2] considers the relativity between the nucleotides and the conservative sequences around the acceptor site. When applying the back propagation network to recognize the acceptor site [3,4], the learning factor and momentum factor affect the systematic error greatly. Raising the learning factor will increase the learning speed and decrease the error quickly, but will eventually make the neural network system unstable. Nowadays, the construction of the neural network is mainly based on experience and knowledge, especially when selecting the structure and parameters of the neural network. Additionally, the low learning speed of the neural network also limits its application in acceptor site prediction. During the prediction of the acceptor site, a high-level recognition of false sites and true sites at the same time is required as the number of false acceptor sites is greater than that of true sites in the intron. When employing the supported vector machine for acceptor site prediction [5], we can adjust the threshold to increase the recognition of one factor. But the increase is at the cost of a decrease of another factor.

In this paper, we introduce a novel algorithm to predict the acceptor site. By integrating the information relating to the acceptor site, branch site, distance between acceptor site and branch site, and base content, we can predict

Received: October 18, 2004 Accepted: April 19, 2005

This work was supported by the grants from the National High Technology Research and Development of China (No. 2002AA234021), the Major State Basic Research Development Program of China (No. 2002 CB 512800), the Science and Technology Committee of Shanghai (No. K02D 105501) and the Corporation Project of Shanghai Jiaotong University and Shanghai Second Medical University

Corresponding author: Tel, 86-21-54742963; E-mail, jhluo@sjtu.edu.cn

DOI: 10.1111/j.1745-7270.2005.00065.x

the acceptor site with multi-objective optimization. The results are compared with those obtained from the SplicePredictor [6].

Materials and Methods

Data collection

The exon-intron database (EID, available at http://www.meduohio.edu/bioinfo/eid/index.html) [7] was used for the analysis of the acceptor site. First, we downloaded 51,289 protein-coding gene sequences in which 287,209 exons were included. Second, we eliminated 17% of the sequences that were redundant and thus obtained 42,460 gene sequences. Third, by comparing the genome sequence with mRNA sequences, we constructed a subset with an approved acceptor site. This process resulted in a set of 11,242 genes that included 62,474 exons in total. Finally, we divided the data set of 11,242 genes into two sets randomly. One set, which included 70% of the genes of the data set, was the training set. The remaining 30% of the genes constituted the testing set.

Construction of signal model

The acceptor signal model is defined as the statistical character of the sequences in the intron 3′ end, while the acceptor site is defined as the first nucleotide in the exon. For every sequence in the training set, by extracting 7 nt from the 3′ end of the intron and the acceptor site following the intron, we obtain a sub-sequence of 8 nt. By computing the occurrence frequencies of the four bases A, T, G and C at the corresponding positions of the 8 nt sub-sequences, we obtained the statistical distribution of the 8 nt sub-sequences shown in **Table 1**, which is the acceptor signal model.

Branch site is located in the region of 18 nt to 40 nt upstream of the intron 3′ terminal and its consensus sequence can be written as $\text{Py}_{80}\text{NPy}_{87}\text{Pu}_{75}\text{A}_{100}\text{Py}_{95}$, where Py, Pu and N are pyridine, purine and any base, respectively, while the suffix indicates the frequency with which a base is found at that position. In this consensus sequence, base A is completely conservative and is called the branch site in our algorithm. The consensus sequence $\text{Py}_{80}\text{NPy}_{87}\text{Pu}_{75}\text{A}_{100}\text{Py}_{95}$ is used to extract the branch site region.

For every sequence in the training set, we first cut off 50 nt upstream of the acceptor site to construct a set consisting of 50 nt sub-sequences. According to the branch site analysis above, every 50 nt sub-sequence contains one branch site. Then we extracted the branch site by analyzing the 50 nt sub-sequences with the template $\text{Py}_{80}\text{NPy}_{87}\text{Pu}_{75}\text{A}_{100}\text{Py}_{95}$.

We compared the first 6 nt in the 3′ end of a 50 nt sub-sequence with the $\text{Py}_{80}\text{NPy}_{87}\text{Pu}_{75}\text{A}_{100}\text{Py}_{95}$ template and computed its score by multiplying the probabilities of the six nucleotides at the corresponding position in $\text{Py}_{80}\text{NPy}_{87}\text{Pu}_{75}\text{A}_{100}\text{Py}_{95}$. For example, suppose the first 6 nt in the 3′ end of one 50 nt sub-sequence is “CAGTCA”, then the score of this 6 nt sub-sequence is calculated as follows:

$$0.80\times0.5\times0.87\times0.75\times0\times0.05=0$$

The score calculated is the branch site probability of the 6th nucleotide of the 3′ end, which is “C” in this example. We repeated the process for every subsequent nucleotide in the 3′ end of the 50 nt sub-sequences. In this way, we obtained the branch site probability of every nucleotide except the last 5 nt of the 50 nt sub-sequence.

For every 50 nt sub-sequence, we took out the 6 nt with the highest branch site probability to construct a new data set. By computing the probability of each nucleotide at the corresponding position of the 6 nt sub-sequences in this new data set, we obtained the statistical distribution of the 6 nt sub-sequences shown in **Table 2**, which is the branch site model.

The distance between the branch site and acceptor site is defined as the number of nucleotides from the acceptor

Table 1 Acceptor signal model

	1	2	3	4	5	6	7	8
A	0.132120	0.126746	0.103878	0.242574	0.045736	0.984353	0.002442	0.229262
T	0.431535	0.436681	0.521663	0.242669	0.255678	0.002555	0.006658	0.110313
G	0.124229	0.104465	0.084434	0.269514	0.009245	0.004682	0.988862	0.522591
C	0.312116	0.332108	0.290022	0.245233	0.689341	0.008410	0.002038	0.137824

Table 2 Branch site model

	1	2	3	4	5	6
A	0.042652	0.152566	0.026258	0.324734	0.999989	0.002425
T	0.528883	0.339141	0.657152	0.134253	0.000011	0.498827
G	0.059365	0.154068	0.013853	0.326377	0	0.012468
C	0.369100	0.354225	0.302737	0.214636	0	0.486220

site to the branch site. By making use of the set of 50 nt sub-sequences constructed above and the positions with the highest branch site probability, we derived the distance between the acceptor site and branch site for every sequence. By carrying out a statistical analysis of all the distance data, as shown in **Fig. 1**, we found that the probability that the distance is below 9 nt was zero. Therefore, the branch site could not have been located very close to the intron. From **Fig. 1**, we can also see that the distance between the branch site and acceptor site is 9 to 48 nt.

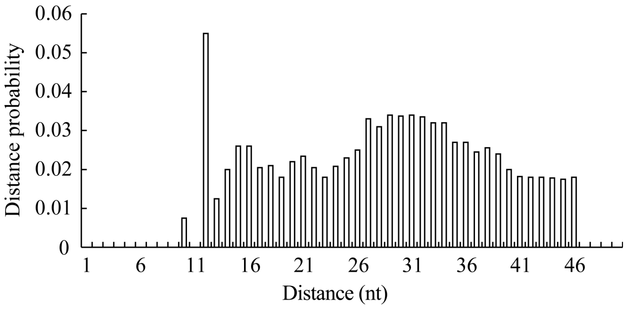

Fig. 1 Statistical distribution of the distance between the acceptor site and branch site

Algorithm of multi-objective optimization

For a sequence x with the length N, $W_8(i)$ is defined as its sub-sequence with the length 8 nt:

$$W_8(i)=\{w_{i1}w_{i2}...w_{i8}\}=\{x_{i-7}x_{i-6}...x_i\},\ i=8,9,\ldots,N \qquad (1)$$

The acceptor site signal function is defined as:

$$f_1(i)=\sum_{k=1}^{8}\ln(p_1[w_{ik}]),\ i=8,9,\ldots,N \qquad (2)$$

where k represents the kth nucleotide of the ith sub-sequence, and p_1 is the acceptor signal model, that is, the occurrence frequency of the corresponding base in **Table 1**. For example, suppose $W_8(i)$= "ACTATGCG". From **Table 1**, we obtain:

$$f_1(i)=ln0.132120+ln0.332108+ln0.521663+ln0.242574 +ln0.255678+ln0.004682+ln0.002038+ln0.522591$$

Since branch site j is located 9–48 nt upstream of the corresponding acceptor site i, j is constrained to $j=i-48, i-47,\ldots,i-9$. In addition, the branch site is the fifth base in the branch site template, so we have $j\geq5$. Denoting $i_0=\max\{5,i-48\}$, then j is constrained to $j=i_0,i_0+1,\ldots,i-9$. As for the acceptor site i, it must satisfy the condition of $i-9\geq5$. Thus we have $i=14,15,\ldots,N$. Considering that the acceptor site is rarely located at the 5′ or 3′ end of a sequence, we take i from 26 to $N-25$ in our algorithm.

Since the branch site is located at the 5th base of a 6 nt sequence, we define a sub-sequence $W_6(j)$ as follows:

$$W_6(j)=\{w_{j1}w_{j2}...w_{j6}\}=\{x_{j-4}x_{j-3}...x_{j+1}\}$$

Employing the branch site model, we define the signal function of branch site j corresponding to the acceptor site i as follows:

$$f_2^*(j)=\sum_{k=1}^{6}\ln[p_2(w_{jk})],\ j=i_0,i_0+1,\ldots,i-9$$

where p_2 is the branch site model; that is, the occurrence frequency of the corresponding base in **Table 2**. For the acceptor site i, we define its branch site signal function as the maximum of its signal functions at corresponding branch sites:

$$f_2(i)=\max\{f_2^*(j)\,|\,j=i_0,i_0+1,...,i-9\},\ i=26,27,\ldots,N-25 \qquad (3)$$

Let j^* be a branch site. Then, the branch-acceptor distance function is defined as:

$$f_3(i)=\ln[p_3(i-j^*)] \qquad (4)$$

where p_3 is the branch-acceptor distance model, that is, the distance probability in **Fig. 1**.

The GC content is an important signal of the splice site, so we define the GC content function as:

$$D_{\mathrm{GC}}(i) = \frac{N_{\mathrm{upGC}}(i) - N_{\mathrm{downGC}}(i)}{25} \tag{5}$$

in which, $N_{\mathrm{upGC}}(i)$ and $N_{\mathrm{downGC}}(i)$ are the number of GC within 25 nt upstream and downstream of the position *i*, respectively.

Model of multi-objective optimization

The correct recognition of the splice site *i* must satisfy the conditions of a strong acceptor site signal $f_1(i)$, a strong branch site signal $f_2(i)$ in a 10–45 nt region upstream of position *i* and an optimum branch-acceptor distance function $f_3(i)$. In addition, the downstream GC content of the position *i* must be obviously greater than the upstream GC content of the position *i*; that is, the value of the GC content function $D_{\mathrm{GC}}(i)$ should be a positive number. Therefore, the problem of predicting the splice site can be written as a multi-objective optimization:

$$\begin{cases} \max[f_1(i)] \\ \max[f_2(i)]; \\ \max[f_3(i)] \end{cases}$$

$$s.t.\ D_{\mathrm{GC}}(i)>C,\ i=26,27,\ldots,N-25 \tag{6}$$

where *C* is a given constant. Then the prediction of the splice site becomes a problem of finding a non-inferior solution in the solution space of the multi-objective optimization in **Equation (6)**. According to optimization theory, the multi-objective optimization shown in **Equation (6)** can be transformed into a single-objective optimization as follows:

$$V(i,\lambda) = l_1 f_1(i) + l_2 f_2(i) + l_3 f_3(i) + \lambda[D_{GC}(i) - C] \tag{7}$$

where λ is Lagrange's multiplier; l_1, l_2 and l_3 are the weight coefficients of $f_1(i)$, $f_2(i)$ and $f_3(i)$, respectively; and *C* is the threshold of the GC content function.

The algorithm for the multi-objective optimization shown in **Equation (6)** is illustrated in detail as follows:

Step 1: evaluate the *C* value. Statistical analysis shows that the GC content in exons near the acceptor is about 47.68% and that in introns is about 52.32%. For simplicity, we consider zero as the threshold.

Step 2: optimize respectively each of the objective functions in **Equation (6)**; that is, solve the following single-objective optimization problem:

$$\max[f_m(i)],\ m=1,2,3;$$

$$s.t.\ D_{\mathrm{GC}}(i)>C,\ 26\le i\le N-25$$

We obtain the optimum solutions i_m* and the reciprocal of Lagrange's multipliers γ_{m} (*m*=1,2,3).

Step 3: compute the weight coefficients by Method α; that is, solve the following system of linear equations and determine the value of l_1, l_2 and l_3:

$$\begin{cases} l_1 f_1(i_1^*) + l_2 f_2(i_1^*) + l_3 f_3(i_1^*) = \mathrm{d} \\ l_1 f_1(i_2^*) + l_2 f_2(i_2^*) + l_3 f_3(i_2^*) = \mathrm{d} \\ l_1 f_1(i_3^*) + l_2 f_2(i_3^*) + l_3 f_3(i_3^*) = \mathrm{d} \\ \qquad l_1 + l_2 + l_3 = 1 \end{cases}$$

Step 4: take the reciprocal of Lagrange's multiplier as:

$$\frac{1}{\lambda} = l_1\gamma_1 + l_2\gamma_2 + l_3\gamma_3$$

Step 5: optimize **Equation (7)**, and calculate the position *i* with the highest score. Then we finally obtain the optimum acceptor site.

Results and Discussion

In order to gauge the performance of the multi-objective optimization prediction method (MOPM), we compared the prediction result obtained using MOPM with that obtained by the SplicePredictor method (SPM) for the testing set we constructed. In our analysis, we used the Precision and Recall measures as measures of recognition performance [8]. These can be defined as:

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{and} \quad \text{Recall} = \frac{TP}{TP+FN}$$

where *TP*, *FP* and *FN* represent the number of true positives, false positives and false negatives, respectively. The results predicted by MOPM and SPM are shown in **Table 3**.

From **Table 3**, we can see that the Precision measure of MOPM reached 85.6%, which is better than that of SPM. The Recall measures for MOPM and SPM were 71.2% and 66.6%, respectively. The results show that the MOPM algorithm performs very well.

In order to improve the performance of our algorithm, we analyzed the sequences which were wrongly predicted.

Table 3 Comparison of MOPM and SPM

Measure	Precision	Recall
MOPM	85.6%	71.2%
SPM	74.5%	66.6%

There are two main causes of error. The first cause of error is that the false acceptor site has strong signals, especially the $f_1(i)$, and $D_{GC}(i)$ in our system of equations is too high. Modulation of the weight coefficients of these functions may improve the recognition accuracy. The second cause of error is that the true splice site is too weak to be recognized. Selecting the proper threshold of the GC content function also helps to improve the accuracy of recognition.

Conclusion

Based on the sequence characteristics from the exon-intron database and biological knowledge, we calculated the acceptor site function, the branch site function and the acceptor-branch distance function. We also constructed the multi-objective optimization model and used it to recognize the splice site. The results show that the multi-objective optimization algorithm performs well and is a novel and promising method for the prediction of the acceptor site.

References

1 Tong G eds. Gene and its Expression. Beijing: Science Press 1996

2 Xia H, Zhou Q, Li Y. Application of hidden Markov model in the recognition of splicing sites. J Tsinghua Univ (Sci & Tech) 2002, 42: 1214–1217

3 Ogura H, Agata H, Xie M, Odaka T, Furutani H. A study of learning splice site of DNA sequence by neural networks. Comput Biol Med 1997, 27: 67–75

4 Sun J, Xu J, Ling LG, Shen RQ, Chen RS. Predicting the splicing sites of mRNA by neural network. Acta Biophysica Sinica 1993, 9: 127–131

5 Wen F, Lu X, Sun ZR, Li YD. Splice sites prediction using support vector machine. Acta Biophysica Sinica 1999, 15: 733–739

6 Pertea M, Lin X, Salzberg SL. GeneSplicer: A new computational method for splice site prediction. Nucleic Acids Res 2001, 29: 1185–1190

7 Saxonov S, Daizadeh I, Fedorov A, Gilbert W. EID: The exon-intron database—an exhaustive database of protein-coding intron-containing genes. Nucleic Acids Res 2000, 28: 185–190

8 Saeys Y, Degroeve S, Aeyels D, Rouze P, van de Peer Y. Feature selection for splice site prediction: A new method using EDA-based feature ranking. BMC Bioinformatics 2004, 5: 64

Edited by
Yi-Xue LI