http://www.zju.edu.cn/jzus; E-mail: jzus@zju.edu.cn
ISSN 1009-3095 *Journal of Zhejiang University SCIENCE* V.4, No.5, P.573-577, Sep.-Oct., 2003

Splicing-site recognition of rice (*Oryza sativa* L.) DNA sequences by support vector machines*

PENG Si-hua(彭司华)†1, FAN Long-jiang(樊龙江)2, PENG Xiao-ning(彭小宁)3, ZHUANG Shu-lin(庄树林)4, DU Wei(杜维)1, CHEN Liang-biao(陈良标)†2

(1 *Department of Control Science and Engineering, College of Information Science and Engineering, Zhejiang University, Hangzhou 310027, China*)

(2 *Institute of Bioinformatics, Zhejiang University, Hangzhou 310029, China.*)

(3 *Verna and Mclean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas, TX 77030, USA*)

(4 *College of Science, Zhejiang University, Hangzhou 310027, China*)

E-mail: †1 pengsihua@zju.edu.cn; †2 liangbiao@zju.edu.cn

Received Sept.7, 2002; revision accepted Dec.8, 2002

Abstract: Motivation: It was found that high accuracy splicing-site recognition of rice (*Oryza sativa* L.) DNA sequence is especially difficult. We described a new method for the splicing-site recognition of rice DNA sequences. Method: Based on the intron in eukaryotic organisms conforming to the principle of GT-AG, we used support vector machines (SVM) to predict the splicing sites. By machine learning, we built a model and used it to test the effect of the test data set of true and pseudo splicing sites. Results: The prediction accuracy we obtained was 87.53% at the true 5' end splicing site and 87.37% at the true 3' end splicing sites. The results suggested that the SVM approach could achieve higher accuracy than the previous approaches.

Key words: Support vector machines, Machine learning, Intron, Splicing site, *Oryza sativa*.
Document code: A **CLC number:** Q756; TP181

INTRODUCTION

Correctly pinpointing splicing-sites in genomic DNA sequences is not an easy task, which is of great importance to the genome annotation and gene finding. Introns are generally divided into 3 classes, namely calss I, class II and common nucleus pre-mRNA. Intron of class I and II can go through the self-splicing processes while pre-mRNA's cannot. The conserved sequence around the 5' end splicing site of pre-mRNA is $^{C}_{A}$AAG|GTRAGT, and around the 3' end splicing site is Y_nYNYAG|G. This is the so-called GT-AG principle (Tong, 1998). Based on this principle, many methods have been proposed for splicing-site recognition, of which, the hidden Markov model (HMM) (Burge, 1997) and the neural networks (NN) (Ogura *et al.*, 1997; Sun *et al.*, 1993) approaches yielded relatively better results. Due to its limitations, such as difficulties in network design, and the training procedure to converge easily to some local minimum points, the NN approach has not been developed further. HMM model is so far the best method for predicting exons and introns, in which more factors are taken into account besides the GT-AG principle and achieved higher accuracy than others. The results are to some extent acceptable in some organisms such as *Arabidopsis thaliana*, but when the method was applied to rice genome, the recognition accuracy by those available softwares, such as GenScan+, FGeneSH, GeneMark, Glimmer, did not exceed 80% (Yu *et al.*, 2002). Novel algorithms are desirable to predict the introns in rice genome.

* Project partially supported by the Start-up Funding of Zhejiang University to Chen Liang-biao

A new general learning theory, support vector machines (SVM) (Vapnik, 2000), based on the statistics learning theory (SLT), was applied to improve the learning effect on a small set of samples. SLT and SVM are currently a highlight in the field of machine learning, and play important roles in various applications.

Pattern recognition, function simulation and estimation of the probability density are all learning problems based on some sample data. The important foundation of existing approaches is classical statistics, which require too many training data. When the sample data are not sufficient, the result is always not satisfactory. Statistical learning theory, proposed by Vapnik (2000), is a small sample statistical theory, extracts meaningful statistical results from a small quantity of sample data. SLT sets up a better framework for machine learning problem. And in this framework, a novel general algorithm, support vector machines (SVM) that can solve small sample data problems successfully, has been developed.

SVM technique has been used effectively in many disciplines. In the bioinformatics field, It has been applied in the prediction of protein secondary structure (Hua *et al*., 2001a), protein sub cellular localization (Hua *et al*., 2001b), and drug design (Burbidge *et al*., 2001).

In this paper, SVM algorithm is used to recognize the splicing sites of the rice genomic sequences, and analyzed the statistical model of the conserved sequences near the splicing sites.

ALGORITHM

1. The primal algorithm

Given training vectors $\boldsymbol{x}_i \in R^n$, $i = 1, \cdots, l$, in two classes, and a vector $\boldsymbol{y} \in R^l$ such that $y_i \in \{1, -1\}$, C-SVC (Cortes *et al*., 1995) solves the following primal problem:

$$\min_{\omega, b, \xi} \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + C\sum_{i=1}^{l}\xi_i \qquad (1)$$

$$y_i(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\phi}(x_i) + b) \geqslant 1 - \xi_i$$

$$\xi_i \geqslant 0, i = 1, \cdots, l.$$

Its dual problem is

$$\min_{\alpha} \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha} - \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha} \qquad (2)$$

$$0 \leqslant \alpha_i \leqslant C,$$

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0$$

Where $\boldsymbol{e}$ is the vector of all ones, $C > 0$ is the upper bound, $\boldsymbol{Q}$ is an l by l positive semi-definite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, and $K(x_i, x_j) \equiv \boldsymbol{\phi}(x_i)^{\mathrm{T}}\boldsymbol{\phi}(x_j)$ is the kernel. Here training vectors $\boldsymbol{x}_i$ are mapped into higher (maybe infinite) dimensional space by the function $\boldsymbol{\phi}$.

The decision function is

$$\mathrm{sign}\left(\sum_{i=1}^{l} y_i\alpha_i K(x_i, x) + b\right). \qquad (3)$$

To solve function Eq.(2) is not easy. Because Q_{ij} is in general not zero. We adopt a decomposition method proposed by Chang *et al*. (1997). The method is described below:

Step 1. Given a number $q \leqslant l$ as the size of the working set. Find α^1 as the initial solution. Set $k = 1$.

Step 2. If $\boldsymbol{\alpha}^k$ is an optimal of function Eq. (2), stop. Otherwise, find a working set $B \subset \{1, \cdots, l\}$ whose size is q. Define $N \equiv \{1, \cdots, l\} \setminus B$ and $\boldsymbol{\alpha}_B^k$ and $\boldsymbol{\alpha}_N^k$ to be sub-vector of $\boldsymbol{\alpha}^k$ corresponding to B and N, q respectively.

Step 3. Solve the following sub-problem with variable α_B:

$$\min_{\alpha} \frac{1}{2}\boldsymbol{\alpha}_B^{\mathrm{T}}\boldsymbol{Q}_{BB}\boldsymbol{\alpha}_B + (\boldsymbol{Q}_{BN}\boldsymbol{\alpha}_N^k)^{\mathrm{T}}\boldsymbol{\alpha}_B \qquad (4)$$

$$0 \leqslant (\boldsymbol{\alpha}_B)_t \leqslant C, t = 1, \cdots, q,$$

$$\boldsymbol{y}_B^{\mathrm{T}}\boldsymbol{\alpha}_B = \Delta - \boldsymbol{y}_N^{\mathrm{T}}\boldsymbol{\alpha}_N,$$

Where $\begin{bmatrix} \boldsymbol{Q}_{BB} & \boldsymbol{Q}_{BN} \\ \boldsymbol{Q}_{NB} & \boldsymbol{Q}_{NN} \end{bmatrix}$ is a permutation of the matrix $\boldsymbol{Q}$.

Step 4. Set $\boldsymbol{\alpha}_B^{k+1}$ to be the optimal solution of (4) and $\boldsymbol{\alpha}_N^{k+1} \equiv \boldsymbol{\alpha}_N^k$. Set $k \leftarrow k + 1$ and go to step 2.

2. Dealing with the unbalanced data

For normal DNA sequences, the data are unbalanced, namely, there are much more pseudo GT (AG) signals than the true GT (AG) signals in the sequence. We should take this property into account in our program. To solve this problem, we adopted an effective approach,

which was called penalty parameter method (PPM), and proposed by Osuna *et al*. (1997). Thus the function (1) can be converted into:

$$\min_{\omega, b, \xi} \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + C_{+}\sum_{y_i=1}\xi_i + C_{-}\sum_{y_i=-1}\xi_i \quad (5)$$

$$y_i(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\phi}(x_i) + b) \geqslant 1 - \xi_i$$

$$\xi_i \geqslant 0, i = 1, \cdots, l.$$

Its dual problem is

$$\min_{\alpha} \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha} - \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha} \quad (6)$$

$$0 \leqslant \alpha_i \leqslant C_{+}, \text{ if } y_i = 1$$

$$0 \leqslant \alpha_i \leqslant C_{-}, \text{ if } y_i = -1$$

3. Choice of the kernel function

In the SVM algorithm, it is very important to choose the right kernel function (Vapnik, 2000). Kernel function is the dot product of input vectors in the feature space. It reflects the distribution feature of the sample in the input space. For different length of sample data, we tested three kernel functions: polynomial function, radial basis function, and sigmoid function, and proved that the result was the most perfect when the fifth polynomial function was adopted as the kernel function. The formula of the kernel function is:

$$k(x, x_i) = [(x \cdot x_i) + r]^d \quad (7)$$

After repeated calculations, we chose the parameter $d = 5$, $r = 0.7$.

METHOD

1. Encoding the sequences

To treat the nucleotide sequence easily on the computer, nucleotide sequence data were encoded digitally. We adopted 4 bites binary encoding scheme described as follows.

A: 0001; C: 0010; G: 0100; T: 1000; others: 0000.

2. Choosing the sample-data length

On one hand, in our algorithm, we should consider the principle that the conserved sequence of an intron is $^{C}_{A}$AAG | GTRAGT around the 5' end splicing site, and Y_nNYAG | G around the 3' end splicing site. On the other hand, we also must take into account the fact that there are 18 – 40 nucleotides conserved in the branch target sequence in the upstream of the 3' end splicing site. To splice correctly, the segment of the target sequence is indispensable. The function of this branch target sequence is to distinguish the target site which will link with the 5' end splicing site and which is the closest to the 3' end splicing site. For yeast, a low-level eukaryotic organism, its target sequence is TACTAAC, which is highly conserved. For high-level eukaryotic organism, this sequence segment is not so conserved. But it has the identifiable sequence consensus. In other words, the positions of Purina and Pyridine are consistent. The sequence is Py80NPy87Pu75APy95 (Gao *et al*., 1999).

Based on upward consideration, the 3 segments of the sequence must be included when we choose the sample-data set length.

We chose the following lengths and their conserved sequences to train the algorithm:

L1 = 10, 15, 20, 30; L2 = 20, 30, 40; L3 = 20, 40, 60, 80; L4 = 10, 15, 20

After several rounds of test, we were able to identify the best lengths for L1, L2, L3, and L4.

DATA SET

A total of 218 rice genomic DNA entries containing at least one intron were extracted from GenBank based on two criteria: (1) those were translated into SWISS-PRO database entries or (2) have complete CDS (coding sequence) published in a certain journal. Generally, the features of these genes sequences are supported by mRNA data (i.e. experimentally derived transcripts). The 185 sequences are divided into two parts: training set which consists of 125 sequences containing a total of 1236 introns, and test set consisting of 93 sequences containing a total of 1007 introns.

The pseudo GT (AG) sequences were selected randomly from the entire sequence. By computation, we found that the probability of the occurrence of pseudo GT (AG) was over 100 times that of the true GT (AG) of introns. Al-

though we selected pseudo GT (AG) only 2 – 5 times more than the true GT (AG), the model we obtained after training could achieve about 95% prediction accuracy for pseudo GT (AG) data, and just 40% – 60% for the true GT (AG). After repeated calculation, we obtained comparatively better result when the ratio of the number of pseudo GT (AG) to the number of true GT (AG) was about 1.6. In this case, the model can predict test data at higher accuracy than either true GT (AG) data or pseudo GT (AG) data.

RESULTS

1. Recognition accuracy at the 5' site(GT site)

The accuracy for the 5' splicing site prediction at different sample data length was shown in Table 1, very high accuracy was obtained when the sample length was set to 15 for the L1 segment and 20 for the L2 segment(see Fig. 1 for the locations of L1 and L2, and the underlined number in Table 1).

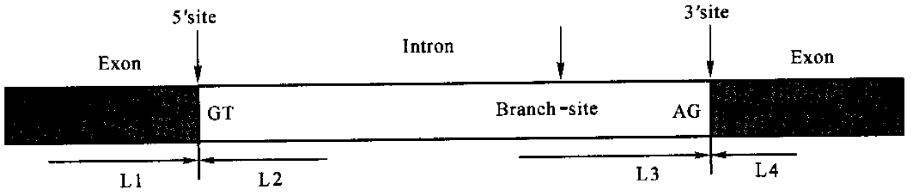

Fig.1 The sketch presentation of an Intron

The sequences consensus of splicing sites and the branch target sequence are shown. L1, L2, L3, L4 denote the locations and the lengths of the sequences we used as sample data to train the algorithm.

Table 1 Recognition accuracy (%) of GT datasets at the 5' sites

L1	L2 = 20		L2 = 30		L2 = 40	
	True GT	Pseudo GT	True GT	Pseudo GT	True GT	Pseudo GT
10	84.455	82.62	90.83	82.39	87.00	85.27
15	87.53	89.21	86.72	82.22	86.59	87.13
20	89.38	83.38	89.97	84.52	87.71	87.09
30	90.39	86.24	89.09	85.55	90.68	81.72

The best result is : L1 = 15, L2 = 20.

The recognition accuracy is:

87.53 % (to true GT), and 89.21 % (to Pseudo GT).

2. Recognition at the 3' site(AG site)

The accuracy of the 3' splicing site prediction at different sample data length was shown in Table 2, here the optimal results were obtained when the length of L3 was set to 40 nucleotides-long and L4 was set to 15 nucleotides-long(see the underlined number in Table 1 and Fig. 1 for the locations of L3 and L4).

Table 2 Recognition accuracy (%) on the 3' sites

L3	L4 = 10		L4 = 15		L4 = 20	
	True GT	Pseudo GT	True GT	Pseudo GT	True GT	Pseudo GT
20	83.88	82.17	83.52	82.29	82.76	81.51
40	84.53	82.69	87.37	85.23	85.08	82.42
60	82.08	81.31	82.76	78.24	84.20	82.98
80	83.76	78.52	83.09	79.67	85.12	80.23

The best result is : L3 = 40, L4 = 15.

The recognition accuracy is:

87.37 % (to true AG), and 85.23 % (to Pseudo AG).

DISCUSSION

Using SVM, we successfully designed an algorithm, and predicted the splicing sites in the rice genome sequences. A few conclusions can be drawn below.

(1) The selection of sample data length is a key factor for obtaining higher prediction accuracy. At the 5' site, if L1 is 15, and L2 is 20, the result turned out to be the best, suggesting that there were about 35 nucleotides meaningful for the correct recognition of the splicing at donor sites, 15 bases upstream to the GT and 20 bases following the GT. At the acceptor site, the best results were obtained when the L3 is 40, and L4 is 15, representing the importance of a total of 55 nucleotides in the correct recognition of the acceptor site by the mRNA splicing machinery. Although the branch site is not fixed, and the conservation of the sequence is at a low level, machine learning was able to find useful statisti-

cal features with high confidence. Obviously, these results are consistent with the molecular biology findings.

(2) The SVM algorithm, based on the GT (AG) principle, achieved relatively high recognition accuracy of the splicing sites, about 88% at the 5' sites and about 86% at the 3' sites. But this result is still far from perfect. This is because pre-mRNA contains much less information than intron II does. So the spliceosome, a huge complex of protein and RNA, is needed to guide the correct splicing. If we consider the information contained by those complex protein factors, the recognition accuracy may be improved greatly. Since the mechanism of the protein factors, which plays an important role in the splicing process, up to now, is unclear, there are many difficulties in the design of the machine-learning algorithm.

(3) Compared to other algorithms, SVM algorithm has an advantage. It can condense information in the training samples to provide a sparse representation using a very small number of samples, support vectors (SVs). In other words, these SVs could represent all the information of the training dataset. The results showed that the ratio of SVs to all training samples was only about 20%. This property has very attractive possibilities for much higher efficiency in treating a great deal of the dataset in bioinformatics field.

(4) However, if we use the model obtained from training rice sequences dataset to predict the introns in human DNA sequence, the results were far from satisfactory (data not shown). Therefore when using the SVM method, we should take the species specificity into account and train the model of different species separately.

(5) Compared with other methods, SVM algorithm obtained an accuracy of about 88% at the 5' sites and about 86% at the 3' sites, the mean value of the predicting accuracy is about 87% in the rice genomic DNA sequences. Reports shown that HMM algorithm could achieve 75% – 80% prediction accuracy in human DNA sequences (Burge, 1997), and lower accuracy in rice (Yu *et al*., 2002). In the NN algorithm, the predicting accuracy at splicing sites is only 64% for eukaryotic organism (SUN *et al*., 1993). Therefore, we think that the SVM algorithm is a more effective method for predicting the splicing sites in rice genome, and its usefulness in other eukaryotic genomes deserves further investigation.

ACKNOWLEDGMENT

We thank YE Xiu-Xiu and SONG Xiao-feng for helpful discussion on molecular biology and SVM technique.

References

Burge, C., 1997. Identification of Genes in Human Genomic DNA. Doctoral Thesis, Stanford University.

Burbidge, R., Trotter, M., Buxton, B. and Holden, S., 2001. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers and Chemistry*, **26**: 5 – 14.

Chang, C.C., Hsu, C.W. and Lin, C.J., 2000. The analysis of decomposition methods for support vector machines. *IEEE Trans. Neural Networks*, **11**(4): 1003 – 1008.

Cortes, C. and Vapnik, V., 1995. Support-Vector networks. *Machine learning*, **20**: 275 – 297.

Gao, J.R. and Ye, L.B., 1999. Molecular Biology. Wuhan University Press, Wuhan, p.135 – 138 (in Chinese).

Hua, S.J. and Sun, Z.R., 2001a. A Novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.*, **308**: 397 – 407.

Hua, S.J. and Sun, Z.R., 2001b. Support vector machine approach for protein sub cellular localization prediction. *Bioinformatics*, **17**(8): 721 – 728.

Ogura, H. and Hideyuki, Agata, 1997. A study of learning splicing site of DNA sequence by neural networks. *Comput. Biol. Med.*, **27**(1): 67 – 75.

Osuna, E., Freund, R. and Girosi, F., 1997. Support Vector Machines: Training and Applications. AI Memo 1602, Massachusetts Institute of Technology.

Sun, J., Xu, J. and Lin, L.J., 1993. Using neural networks to recognize the splicing sites of mRNA. *Transactions of Biophysical Sinica*, **9**(1): 127 – 131 (in Chinese).

Tong, K.Z., 1998. Gene and its Expression. Science Press, Beijing.

Yu, J., Hu, S.N. and Wang, J., 2002. A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. Indica). *Science*, **296**: 79 – 92.

Vapnik, V., 2000. The Nature of Statistical Learning Theory. Traslated by Zhang Yuegong, Tsinghua University Press, Beijing (in Chinese).