Application of Hidden Semi-Markov Model to 3′ Splice Sites Identification*

FENG Xiu-Cheng, QIAN Min-Ping**, DENG Ming-Hua, MA Xiao-Tu, YAN Xi-Ting

(*School of Mathematical Sciences, Peking University, Beijing* 100871, China;
Center for Theoretical Biology, Peking University, Beijing 100871, China)

Abstract In order to improve exon level sensitivity and specificity of recent gene-finding programs, strong "search by signal" components are needed to identify splice sites, translation start and other biological signal sites. A new model for the identification of 3′ splice sites (acceptors) using Hidden Semi-Markov Model (HSMM) was introduced. This model is proved to be particularly suitable for modeling the biological structure of acceptors. When tested in Burset/Guigo dataset, this new method demonstrated an improved accuracy compared with existing method. The success of this model gives a deep understanding of the structure of acceptors and the biological process of splicing.

Key words Hidden Semi-Markov Model, splicing, eukaryotic gene structure prediction, EM algorithm

New generation of gene-finding programs developed in recent years have substantial improvement compared with older ones. However, the goal of computational gene finding is still far away. The exon level sensitivity and specificity is still unsatisfactory. When tested in HMR195 dataset by Rogic, the program that showed highest ESn and ESp, HMMgene, has ESn and ESp as 0.76 and 0.77[1]. Even when the best two programs, Genscan and HMMgene, are combined, ESn and ESp is only a little higher than 0.80[2]. In order to improve exon level accuracy, gene-finding programs have to have stronger "search by signal" component (signal sensors)[2].

Currently, most successful gene finding programs, such as Genscan[3] and HMMgene[4], usually use Hidden Markov Model (HMM)[5,6] as a higher-level model that integrates many sub-models. The sub-models include coding statistics, transcription signals, translation signals and splicing signals. If the sub-models are all improved, we can expect a substantially improved overall performance. In our recent work, we specifically focus in improving the sub-model used in identification of 3′ splice sites.

Current programs have lower acceptor site accuracy than donor site accuracy[2]. The reason is that the signal of acceptor sites is not well conserved and the branch point is notoriously difficult to model. The branch point is not only less conserved and the position of the branch point varies in a range of about 20 bases[3,7].

However acceptor site has its special structure. The branch site lies 18 ~ 40 nucleotides upstream of the 3′ splice site[8]. The branch site in higher eukaryotes is not well conserved, but has a preference for purines or pyrimidines at each position[9] and has an A nucleotide as target base to form a 5′-2′ bond with the 5′ terminus at the end of the intron which is released in the first transesterification reaction[10]. The branch site is also the binding site of U2 snRNP when splicesome is transformed from E complex into A complex. Downstream of the branch site is a pyrimidine rich tract. The pyrimidine rich tract is bound by U2AF splicing factor. The binding of U2AF with the pyrimidine rich tract is needed for U2 snRNP to bind with branch site[11,12]. The region around the 3′ splice is also a binding site for other ribonucleoproteins such as U5 snRNP during C1 complex stage and this binding is needed by the second transesterification reaction[12].

To deal with the intrinsic structure of acceptors, Burge has developed a modified Weight Array Model (WAM) method in his well-known gene-finding program Genscan[3]. Specifically, bases −20 to +3 relative to the intron/exon junction, encompassing the pyrimidine-rich region and the acceptor splice site itself are modeled by a first-order WAM model. He introduced a "windowed WAM model" to model the branch site region [−40, −21][6].

To characterize the special structure of splice sites, we have introduced a novel model for identification of acceptors using Hidden Semi-Markov Model (HSMM)[5]. In this article, it is shown that this model is more suitable for the acceptor site than existing models. This model primarily aims to improve the accuracy by combining the information contained in the consensus around splice site, the polypyrimidine tract and the branch site upstream.

1 Method

We use Hidden Semi-Markov Model instead of

* This work was supported by grants from The National Natural Sciences Foundation of China (90208002, 10271008), State 863 High Technology R&D Project of China (2002AA234011) and The Special Funds for Major State Basic Research of China (2003CB715903).

** Corresponding author.
Tel: 86-10-62752525, E-mail: qianmp@math.pku.edu.cn

Received: November 10, 2003 Accepted: December 28, 2003

HMM because we think HMM does not characterize the acceptor structure well and HSMM is more preferable. As we will see in Figure 2, the distribution of branch point position is quite different from 1-shifted geometric distribution. However, if we use HMM as a higher-level model, the distribution of state duration will always be 1-shifted geometric, which is not the case.

HSMM model is used to combine sub-models characterizing branch site, pyrimidine-rich region and region around AG consensus. The states used in HSMM model are I (intron), B (branch site), P (polypyrimidine tract), A (consensus around AG) and E (exon). The state transitions between states are trivial according to their biological meaning. But the state durations are flexible (Figure 1).

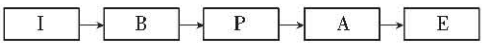

Fig. 1 State transitions

For the sub-models of the branch site region and the consensus around AG, a Linear Discriminant Function (LDF)[13] was used to combine Weight Matrix Model (WMM) and one-order WAM model. A homogenous two-order Markov Model was used to characterize the pyrimidine-rich region.

In training, as to the exact state sequence in the training set is unknown, EM Algorithm[14] is used to train the model. Random or arbitrary initial parameters are given. The result shows that the ultimate trained model is not sensitive to the initial parameter values. This suggests that the trained model is reliable. Figure 2 shows part of the trained parameter values, the distribution of branch point position. In testing Viterbi Algorithm[6] is used to find the optimal state sequence.

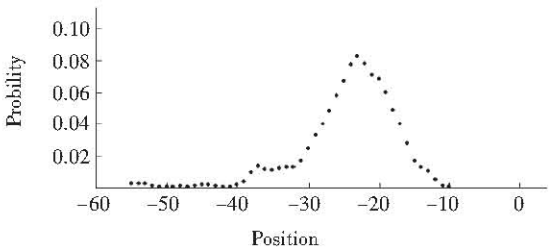

Fig. 2 Distribution of branch point position

2 Result

Our method was tested on Burset/Guigo set[15] of 570 vertebrate multi exon gene sequences using cross-validation. The 570 sequences were divided into four parts randomly. Three of them were chosen as training set and the other was used as test set. Then accuracy was averaged.

We first compared our method with two earlier models. One is the simplest, Weight Matrix Model (WMM). The second is modified WAM method used by Burge in Genscan. The result shows that the Genscan model is much better than WMM and HSMM is clearly better than the Genscan model (Table 1). Parameters of WMM were optimized to get best result. Parameters for Genscan splice site model were exactly the same provided by Genscan program.

WMM and modified WAM both use information in a rather short region. Consequently, when comparing with those two methods, we also take as input a short region around splice sites, which contains about 50 bp upstream and about 10 bp downstream. But a recent program GeneSplicer[16] takes into a large coding and noncoding area into consideration. It uses sequences of 80bp on either side of the splice sites. In order to compare with GeneSplicer, we produced another version of HSMM program, which use the same wide window as input. We call it HSMM with wide window (HSMM-W). From Table 1, we can see that our method is comparable with GeneSplicer and a little better.

Table 1 Comparison of various methods

	WMM		Modified WAM		HSMM		GeneSplicer		HSMM-W	
Sn	*Sq*	*Sp*	*Sq*	*Sp*	*Sq*	*Sp*	*Sq*	*Sp*	*Sq*	*Sp*
0.88	0.933	0.171	0.950	0.217	0.961	0.250	0.965	0.262	0.968	0.268
0.90	0.925	0.160	0.945	0.204	0.955	0.237	0.961	0.251	0.963	0.256
0.92	0.912	0.142	0.937	0.186	0.946	0.210	0.954	0.240	0.955	0.242
0.94	0.892	0.122	0.926	0.166	0.936	0.190	0.943	0.225	0.947	0.230
0.96	0.868	0.104	0.904	0.135	0.913	0.159	0.922	0.180	0.928	0.218

$Sn = TP/(TP + FN)$, $Sq = TN/(TN + FP)$, $Sp = TP/(TP + FP)$. *TP* is the number of acceptors that are predicted as acceptors. *FN* is the number of acceptors that are predicted as pseudo-acceptors. *TN* is the number of pseudo-acceptors that are predicted as pseudo-acceptors. *FP* is the number of pseudo-acceptors that are predicted as acceptors.

To give a profile of branch sites, we aligned the sequences identified as branch sites together and calculated the weight matrix. The result was given below (Table 2). From the result, we can see some clue that the vertebrate branch site is evolved from the original UACUAAC box in yeast.

Table 2 A weight matrix profile was calculated for the aligned branch sites that is identified by our method

	-6	-5	-4	-3	-2	-1	A site	1	2
A	0.236	0.224	0.264	0.115	0.126	0.305	1.000	0.091	0.130
T	0.289	0.289	0.329	0.202	0.692	0.144	0.000	0.349	0.354
C	0.240	0.306	0.236	0.549	0.149	0.248	0.000	0.456	0.335
G	0.233	0.179	0.169	0.132	0.031	0.301	0.000	0.102	0.179

To show some characteristics of the pyrimidine-rich region, we counted the frequencies of all sixteen dinucleotides that occur in the identified pyrimidine-rich regions. The result was given in Table 3. The most frequent four dinucleotides are all pyrimidine pairs and AG is the most unfrequent dinucleotide. The exceedingly low frequency of AG might be an evidence of the model suggested by Langford and Gallwitz. In 1983, they demonstrated the role of branch site might be telling the splicing machinery to splice to the first AG downstream[17].

Table 3 The frequencies of all sixteen dinucleotides in the identified pyrimidine-rich regions

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0.015	0.038	0.004	0.039	0.039	0.133	0.014	0.155	0.009	0.046	0.021	0.051	0.031	0.146	0.077	0.182

3 Discussion

We find that HSMM is especially suitable to identify a "hidden" structure which is composed of several relatively consensus motifs and the exact distance between the motifs is variable and unknown. Acceptor is such a kind of "hidden" structure. The method introduced here can also be used to identify other biological signals such as translation start and transcription start sites.

References

1 Rogic S, Mackworth A K, Ouellette F B. Evaluation of gene-finding programs on mammalian sequences. Genome Research, 2001, **11** (5): 817 ~ 832

2 Rogic S, Ouellette B F, Mackworth A K. Improving gene recognition accuracy by combining predictions from two gene-finding programs. Bioinformatics, 2002, **18** (8): 1034 ~ 1045

3 Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol, 1997, **268** (1): 78 ~ 94

4 Krogh A. Two methods for improving performance of an HMM and their application for gene finding. In: Gaasterland T, eds. Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology. USA: The AAAI Press, 1997. 179 ~ 186

5 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989, **77** (2): 257 ~ 286

6 杨文强, 钱敏平, Huang D W. 基于隐马氏模型对编码序列缺失与插入的检测. 生物化学与生物物理进展, 2002, 29 (1): 56 ~ 59
Yang W Q, Qian M P, Huang D W. Prog Biochem Biophys, 2002, **29** (1): 56 ~ 59

7 Burge C, Tuschl T, Sharp P A. Splicing of precursors to mRNAs by the spliceosomes. In: Gesteland R F, Cech T R, Atkins J F, eds. The RNA World. 2nd. New York: Cold Spring Harbor Laboratory Press, 1999. 525 ~ 560

8 Reed R, Maniatis T. A role for exon sequences and splice-site proximity in splice-site selection. Cell, 1986, **46** (5): 681 ~ 690

9 Zhuang Y A, Goldstein A M, Weiner A M. UACUAAC is the preferred branch site for mammalian mRNA splicing. Proc Nat Acad Sci USA, 1989, **86** (6): 2752 ~ 2756

10 Reed R, Maniatis T. Intron sequences involved in lariat formation during pre-mRNA splicing. Cell, 1985, **41** (1): 95 ~ 105

11 Parker R, Siliciano P G, Guthrie C. Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. Cell, 1987, **49** (2): 229 ~ 239

12 Burgess S, Couto J R, Guthrie C. A putative ATP binding protein influences the fidelity of branch point recognition in yeast splicing. Cell, 1990, **60** (5): 705 ~ 717

13 Solovyev V V. Finding genes by computer: Probabilistic and discriminative approaches. In: Jiang T, Smith T, Xu Y, Zhang M, eds. Current Topics in Computational Biology. USA: The MIT Press, 2002. 365 ~ 401

14 Dempster A P, Laird N M, Rubin D B. Maximum-likelihood from incomplete data via the em algorithm. J Royal Statist Soc Ser B, 1977, **39** (1): 1 ~ 38

15 Burset M, Guigo R. Evaluation of gene structure prediction programs. Genomics, 1996, **34** (3): 353 ~ 367

16 Pertea M, Lin X, Salzberg S L. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res, 2001, **29** (5): 1185 ~ 1190

17 Langford C J, Gallwitz D. Evidence for an intron-contained sequence required for the splicing of yeast RNA polymerase Ⅱ transcripts. Cell, 1983, **33** (2): 519 ~ 527

隐半马氏模型在3′剪接位点识别中的应用*

冯秀程 钱敏平** 邓明华 马小土 严熙婷

（北京大学数学科学学院，北京 100871；北京大学理论生物学中心，北京 100871）

摘要 新近的基因识别软件比先前的软件有着显著的提高，但是在外显子水平上的敏感性和特异性仍然不十分令人满意. 这是因为已有软件对于剪接位点，翻译起始等生物信号位点的识别还不够有效. 如果能够分别提高这些生物信号位点的识别效果，就能够提高整体的基因识别效率. 隐半马氏模型能够很好地刻画3′剪接位点（acceptor）的结构. 据此开发的一套对acceptor进行识别的算法在Burset/Guigo的数据集上经过检验，获得了比已有算法更好的识别率. 该模型的成功还使得我们对剪接点上游的分支位点和嘧啶富含区的概貌有了一定的认识，加深了人们对于acceptor的结构和剪接过程的理解.

关键词 隐半马氏模型，剪接，真核生物基因结构预测，EM算法

学科分类号 Q612

* 国家自然科学基金（90208002，10271008），国家高技术“863”计划资助项目（2002AA234011）和国家重点基础研究发展规划项目（973）（2003CB715903）资助.

** 通讯联系人.

Tel：010-62752525，E-mail：qianmp@math.pku.edu.cn

收稿日期：2003-11-10，接受日期：2003-12-28