一种新的真核基因剪接位点识别方法

王科俊[1],吕俊杰[1],冯伟兴[1],王 鑫[2],贺 波[1]
(1.哈尔滨工程大学自动化学院,黑龙江哈尔滨 150001;2.剑桥大学癌症分子研究中心,英国剑桥)

摘 要: 剪接位点识别是基因组分析的关键步骤.为提高真核基因剪接位点识别的精度,提出一种融合多种信息的方法.在采用序列信息与剪接位点信号信息的基础上,增加剪接调控元件信息,并引入结构信息,针对供体位点与受体位点的不同特点,为其建立不同的识别模型.实验结果表明:该方法对剪接位点的识别具有较好的效果,其识别精度可达 95%以上.

关键词: 位点识别; 信号信息; 序列信息; 调控元件; 结构信息
中图分类号: Q52 **文献标识码:** A **文章编号:** 0372-2112 (2011) 05-1210-04

A New Method for Identification of Eukaryotic Gene Splice Sites

WANG Ke-jun[1], LÜ Jun-jie[1], FENG Wei-xing[1], WANG Xin[2], HE Bo[1]
(1. *College of Automation, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*;
2. *Cancer Research Center, Cambridge University, Cambridge, England*)

Abstract: Splicing site recognition is the key step in the genome analysis. To improve the identification accuracy of eukaryotic gene splicing sites, a variety of information fusion recognition method of splicing sites is proposed. Based on the using sequence information and splicing site signal information, we increased splicing regulatory element information, and proposed the structure information. By analyzing the different characteristics of donor sites and acceptor sites, donor sites identification signal model, acceptor sites identification signal model, donor sites identification sequence model, acceptor sites identification sequence model were built respectively. Our results show that the accuracy of splice site recognition is greater than 95%, suggesting that the method has great potential to achieve a good performance for splice sites identification.

Key words: site identification; signal information; sequence information; regulatory element; structure information

1 引言

剪接位点的预测是真核基因组学的一个经典并且具有挑战性的生物信息学难题.现有的识别方法大多数只关注序列信息,忽视了结构信息以及剪接调节元件对剪接位点的影响[1~4],从而导致识别率较低.为提高剪接位点识别的精度,我们提出了一种采用多种信息的识别方法(G-SSS)[5],针对供体位点与受体位点剪接调控区域的不同特点,分别为其建立不同的信号模型,序列模型,并引入位点附近序列的二级结构信息,将结构信息与传统的序列信息相结合,用融合了结构信息的序列对建立的模型进行训练,最后用训练好的模型进行剪接位点的识别.实验证明,该方法使真核基因剪接位点的识别性能得到了提高.但是假阳率还较高,将虚假剪接位点识别为真实剪接位点的几率还很大,为进一步提高剪接位点识别方法的性能,我们在 G-SSS 的基础上进一步增加生物信息,提出一种 CS-Models 法,在对剪接序列建模时引入剪接调控元件信息,考虑稍远的剪接增强子、剪接沉默子的信息,用融合了信号信息,序列信息,结构信息,剪接调控元件信息的方法对真核基因剪接位点进行识别.经过实验验证与比较发现,CS-Models 法对真核基因的剪接位点识别取得了很好的识别效果.

2 材料与方法

方法的整体研究思路如图 1 整体研究框图 1 所示.

2.1 评价指标

常用敏感性 S_n,特异性 Sp,假阳率 $FP\%$.作为剪接位点识别的评价指标,定义如下:

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

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

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

收稿日期:2010-05-15;修回日期:2010-12-15
基金项目:国家自然科学基金(No.61071174);国家 863 高技术研究发展计划(No.2008AA01Z148)

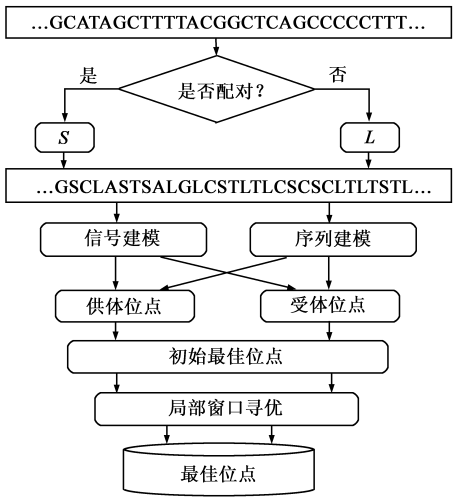

图1　整体研究框图

其中,*TP* 表示真阳,*TN* 表示真阴,*FP* 表示假阳,*FN* 表示假阴. S_n 表示所有实际真实的剪接位点中,被正确识别出来的比例;*FP*% 表示在所有的虚假剪接位点中,被识别为真实剪接位点的比例. 在评价方法的性能时,通常为:在取得较高的 S_n 同时得到较低的 *FP*% 值.

2.2　数据集

实验数据集 NN269[3],由 285 条人类基因序列组成,其中包含 1324 个真实供体位点,1324 个真实受体位点,4922 个虚假供体位点,5552 个虚假受体位点. 每个虚假的供体位点和受体位点也符合 AG-GT 规则. 将数据集分为训练集和测试集两个子集,训练集包含 1116 个真实受体位点,1116 个真实供体位点,4672 个虚假受体位点,4140 个虚假供体位点;测试集包含 208 个真实受体位点,208 个真实供体位点,881 个虚假受体位点,782 个虚假供体位点.

2.3　二级结构预测

已有生物信息学研究表明:剪接位点附近序列的二级结构会影响剪接位点的选择[6,7]. 但现有的识别方法中很少考虑结构因素的影响,在我们的方法中,引入位点附近序列的二级结构信息,应用 Vienna 软件包中 Mfold 程序包预测每个序列中核苷酸的二级结构[6]. 按序列中的每个核苷酸是否配对的原则,将结构信息转化为两字符的字母表{*S*,*L*},并将其与传统的 4 字符字母表{*A*,*G*,*C*,*T*}相结合,这样,每个序列就转化为用 8 字符字母表{(*AS*),(*AL*),(*GS*),(*GL*),(*CS*),(*CL*),(*TS*),(*TL*)}描述. 然后用这个 8 字符描述的序列对识别模型进行训练.

2.4　剪接序列建模

采用与 G-SSS 相同的方法,用隐马尔可夫模型(Hidden Markov Models,HMM)对剪接序列建模[8]. 建立的 HMM 包含两个随机过程,一个产生的输出为隐含的状态序列 $Q=(q_1q_2\cdots q_L)$,由(A,π)描述. 一个产生的输出为可观测到的符号序列 $O=(o_1o_2\cdots o_L)$,由 E 描述. HMM 各参数限定如下:

(1)给定状态集:$S=\{AS,TS,GS,CS,AL,TL,GL,CL\}$

观测符号集:$V=\{AS,TS,GS,CS,AL,TL,GL,CL\}$

(2) $\sum_{i=1}^{N}\pi_i=1$ $\sum_{j=1}^{N}a_{ij}=1$, $\sum_{k=1}^{M}e_{jk}=1$, $\pi_i,a_{ij},e_{jk}\in[0,1]$

(3) $p(q_t\mid q_1,q_2,\cdots,q_{t-1})=p(q_t\mid q_{t-1},q_{t-2})$

$p(o_t\mid q_1,q_2,\cdots,q_{t-1})=p(o_t\mid q_{t-1},q_{t-2})$

(4) $\forall\, t: a_{ij}=p(q_t=s_i\mid q_{t-1}=s_j)$

$e_{jk}=p(o_t=v_j\mid q_t=s_k)$

(5) $p(o_t\mid o_1,o_2,\cdots,o_{t-1},q_1,q_2,\cdots,q_t)=p(o_t\mid q_t)$

但是在序列的选取上,CS-Models 用与 G-SSS 不同的方法,我们考虑到剪接位点的选择是多种剪接调控元件综合作用的结果,将位点上下游稍远的剪接增强子,剪接抑制子考虑进来,选取较长的一段序列(上下游分别选取 300bp),由于核酸翻译的时候以三个连续的核苷酸作为一个密码,所以对供体位点与受体位点序列分别建立二阶 HMM:P_{dseq}和P_{aseq}.

2.5　剪接信号建模

剪接信号模型主要是对于位点附近的保守短片断进行建模. 由生物学知[9]:剪接因子 U1-snRNP 结合在 5′端剪接位点,这段结合区具有很强的保守性,将这段结合区作为供体位点信号序列,它由下游内含子最前面的 6 个碱基和上游外显子的最后 3 个碱基构成. 综合考虑模型的复杂程度和计算量,对供体位点信号建模采用最大相关分解(MDD)方法,供体位点信号模型的计算公式为:

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

其中,$N(d)$表示获得第 d 个 PWM 矩阵所进行分解的次数,$p_j(o)$表示子序列 O 在第 j 次分解中的选择概率,$L(d)$表示第 d 个 PWM 矩阵中序列的长度,p_{i,o_i}^{d}表示在第 d 个 PWM 矩阵中第 i 个位置出现碱基 o_i 的概率.

对虚假供体位点建立类似的模型,模型的计算概率为 $p'(o)$. 由于式(1)右端为乘法计算,计算复杂、速度慢,考虑到对数变换可将乘法运算变为加法运算,能够减小计算量,提高计算速度,将真实位点的概率计算公式与虚假位点的概率计算公式分别取对数 $\log p(o)$ 和 $\log p'(o)$,然后将两个对数做差,根据差值 P_D 是否大于 0,对测试序列 O 是否为真实的供体位点序列进行评价,P_D 的计算公式为:

$$P_D=\log p(o)-\log p'(o) \tag{2}$$

对于受体位点的建模,由于受体位点上游的分枝点附近存在着保守性,并且在分枝点和受体位点之间是富含嘧啶的区域[9].分枝点附近的保守性较弱,而且一般情况下它的位置并不清楚,所以把它与富含嘧啶的区域放在一起建模,同时考虑到训练的时间与效率,不同于 G-SSS,我们采用一阶马尔可夫模型描述这种相关性,其计算公式为:

$$p(o)=\prod_{i=2}^{L} p_{o_{i-1},o_i}^{i-1,i} \tag{3}$$

其中,$p_{o_{i-1},o_i}^{i-1,i}$表示在位置 $i-1$ 和 i 上分别出现碱基 o_{i-1}、和 o_i 的概率.对虚假受体位点建立类似的模型,虚假模型的计算概率为 $p'(o)$,然后分别取对数做差,根据差值 P_A 是否大于 0,对测试序列 O 是否为真实的受体位点序列进行评价,P_A 的计算公式为:

$$P_A=\log p(o)-\log p'(o) \tag{4}$$

2.6 识别模型

最终得到剪接位点识别判断公式为:

$$P_{dr}=P_{dseq}+P_D \tag{5}$$

$$P_{ar}=P_{aseq}+P_A \tag{6}$$

其中,P_{dr}表示供体位点识别概率,P_{ar}表示受体位点识别概率.为进一步降低识别的假阳性指标,采用类似 Brendel 等应用的局部寻优方法[2],在长度为 50bp 的窗口内首先确定 P 值大于设定阈值的剪接位点(文中取为 0.5,即取大于 50% 以上的),在这些剪接位点中选取 P 值最大的位点为真实剪接位点.

3 实验结果与分析

CS-Models 的识别性能如表 1 所示:

表 1 剪接位点的识别结果

位点	敏感性(%)	假阳率(%)
供体位点	95	5.1
	92	3.1
	80	1.6
	75	0.8
	60	0.7
受体位点	95	5.0
	92	3.0
	80	0.8
	75	0.7
	60	0.6

CS-Models 与较为经典的 GeneSplicer[4] 和 NNSplice[3] 以及 G-SSS 相比较,比较结果如表 2 所示:

由表 2 的比较结果可以看出,CS-Models 对供体位点识别性能优于其他三种方法.尤其高于 GeneSplicer 和 NNSplice 两种经典方法.通过比较还发现,*Sn* 越高,*FP*% 也随之越高,*Sn* 较低的情况下,*FP*% 也相应较低,这可能是因为虚假位点的大量存在,影响对真实位点的判断识别.对受体位点的识别性能,CS-Models 也优于其他三种方法.我们发现 *Sn* 值越低,CS-Models 的优越性较 GeneSplicer 和 NNSplice 越明显,但与 G-SSS 相当;比较表中供体位点识别结果和受体位点识别结果的数据还可以发现,受体位点的 *FP*% 指标总体上要小于供体位点的 *FP*% 值,这一现象可能是因为受体位点附近序列中保守碱基较多,构建的识别模型可以对其进行很好的描述.

表 2 剪接位点的识别结果比较

位点	敏感性(%)	假阳率(%)			
	—	GeneSplicer	NNSplice	G-SSS	CS-Models
供体位点	93	6.0	6.2	5.1	5.0
	90	3.3	3.5	3.2	3.2
	85	2.5	2.7	2.4	2.2
	75	2.1	2.2	1.9	1.8
	60	1.7	1.9	0.9	0.7
受体位点	91	5.6	5.8	4.6	4.6
	85	2.4	2.6	2.4	2.1
	80	1.4	1.3	0.8	0.8
	75	1.0	1.2	0.7	0.7
	60	0.8	0.7	0.6	0.6

对各方法性能的评价,更直观的为 ROC 曲线.以敏感性 *Sn* 为纵轴,特异性 *Sp* 为横轴绘制 ROC 曲线,曲线越靠近左侧边界,且越靠近 ROC 图的上侧边缘,模型的准确性越高,方法的识别性能越好.各方法的 ROC 曲线如图 2 供体位点识别 ROC 曲线和图 3 受体位点识别 ROC 曲线所示:

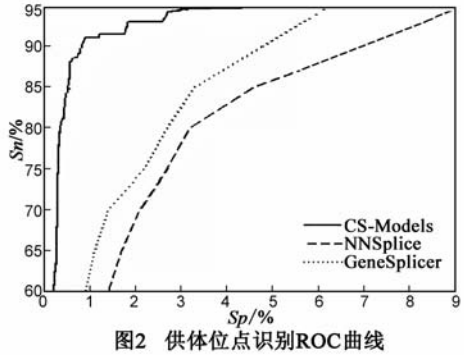

图2 供体位点识别ROC曲线

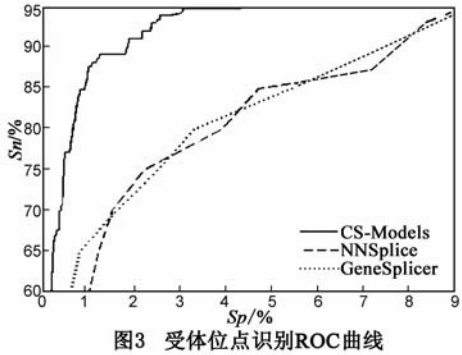

图3 受体位点识别ROC曲线

通过观察图 2 供体位点识别 ROC 曲线和图 3 受体位点识别 ROC 曲线可以看出,CS-Models 在对供体位点和受体位点识别时都要明显优于 GeneSplicer 和 NNSplice,在对供体位点识别时 GeneSplicer 的识别性能要优于 NNSplice,而在对受体位点识别时,两者的识别性能相当.

4 结论

将剪接调节元件信息、结构信息、剪接信号信息以及序列信息融合在一起对剪接位点进行识别,使得识别精度得到显著提高,但是由于对影响剪接位点的生物信息认识还不够全面,识别方法的精度还有待于进一步的提高.用系统的观点进行研究[10]将是未来的努力方向,接下来我们将寻求更为合理的结构信息及相关的生物信息,建立更符合生物学的识别模型,开发更为高效的人工智能识别方法.

参考文献

[1] Krogh A. Two methods for improving performance of a HMM and their application for gene finding [J]. Proc. Intell. Syst. Mol. Biol, 1997, 5(23): 179 – 186.

[2] Brendel V, Kleffe J. Prediction of locally optimal splice sites in plant pre-mRNA with application to gene identification in Arobidopsis thaliana fenormic DNA [J]. Nucleic Acids Res, 1998, 26(20): 4748 – 4757.

[3] Reese M G, Eeckman F H, Kulp D, et al. Improved splice site detection in genie [J]. Journal of Computational Biology, 1997, 4(3): 11 – 323.

[4] Pertea M, Lin X Y, Salzberg S L. GeneSplicer: a new computational method for splice site prediction [J]. Nucleic Acids Res, 2001, 29(5): 1185 – 1190.

[5] 吕俊杰.真核基因剪接位点识别方法研究[D].哈尔滨工程大学自动化学院,2010.41 – 49.
LÜ Jun-jie. Research on identification of eukaryotic gene splice sites[D]. Master's Thesis, College of Automation, Harbin Engineering University, 2010.41 – 49. (in Chinese)

[6] Hiller M, Zhang Z, Backofen R, et al. pre-mRNA secondary structure and splice site selection [J]. PLOS Genet, 2007, 3(1): 2147 – 2155.

[7] 闻芳,李衍达.基因表达调控与选择性剪接机制研究[J].电子学报,2001,29(12A):1735 – 1739.
Wen Fang, Li Yan-da. A bioinformatic analysis of alternatively spliced genes of human[J]. Acta Electronica Sinica, 2001, 29(12A): 1735 – 1739. (in Chinese)

[8] Krogh A. Using database matches with HMM Gene for automated gene detection in Drosophila [J]. Genome Res, 2000, 10(4): 523 – 528.

[9] Black D L. Mechanisms of alternative pre-messenger RNA splicing [J]. Annu Rev Biochem, 2003, 72(1): 291 – 336.

[10] 李衍达.以信息系统的观点了解基因组[J].电子学报,2001,29(12A):1731 – 1734.
Li Yan-da. Understanding genome from information system point of view[J]. Acta Electronica Sinica, 2001, 29(12A): 1731 – 1734. (in Chinese)

作者简介

王科俊　男,教授、博士生导师.1962 年 9 月出生于吉林省吉林市.中国人工智能学会理事,中国电子学会高级会员.1996 年在哈尔滨工程大学获工学博士学位.现为哈尔滨工程大学模式识别与智能系统学科带头人.发表论文 180 余篇,出版学术专著 3 部,国防教材 1 部,主审教材 2 部.主要研究方向:生物信息学、模糊混沌神经网络、模式识别.E-mail: wangkejun@hrbeu.edu.cn

吕俊杰　女,1982 年 1 月出生于黑龙江齐齐哈尔.哈尔滨工程大学模式识别与智能系统专业博士研究生,主要从事生物信息学方面的研究.E-mail: lvjunjie525@126.com