文章编号:1000-582X(2010)10-129-04

利用条件随机场实现 DNA 剪接位点的预测

杨王黎　许少华
(大庆石油学院 计算机与信息技术学院,黑龙江 大庆 163318)

摘　要:为解决传统的基因识别算法主要关注编码区的整体特性,而并不着重考虑个别位点的信息,因此难以准确地识别出剪接位点的缺点,提出了基于条件随机场的剪接位点预测方法,条件随机场能够更好的处理标记数据之间的依赖关系,并且能够避免数据标记偏置的问题。实验结果表明基于条件随机场的剪接位点预测方法是一种合适的方法,能够取得更好的效果。

关键词:剪接位点;条件随机场;基因编码区域
中图分类号:TP393　　**文献标志码**:A

Predicting the DNA splice sites with conditional random field

YANG Wang-Li, *XU Shao-hua*
(Daqing Petroleum Institute College of Computer Science and Engineering, Heilongjiang, Daqing 163318, P. R. China)

Abstract: Predicting the DNA splice sites has become the most attractive and important issue in the field of genome information research due to its great help to find gene coding area. Current gene recognition algorithms mainly consider the global features of the coding area, instead of the specific information of the splicing sites, which are usually unable to recognize the splicing sites accurately. A new method based on the conditional random fields (CRFs) is proposed for splice sites prediction. CRFs can capture long distance dependent relationships between labels and avoid label bias. Experimental results show that CRFs is suitable for splice sites prediction and can improve performance.

Key words: splice sites; conditional random field (CRF); gene coding area

随着人类基因组计划的完成,生物信息学的研究正在向后基因组时代迈进,人们更加关注基因组信息的结构分析、基因识别及功能预测的研究。其中剪接位点的预测可以对真核生物的剪接方式和 RNA 加工的研究提供大量有价值的信息,因此成为基因组信息研究中最引人关注的问题。如何提高剪接位点预测的精度是基因组信息研究中的一项重要内容。

在 DNA 剪接位点预测和分析方面,近年来出现了多种方法,比如贝叶斯网络[1-2]、支持向量机[3-5]、神经网络[6-7]、隐马尔可夫[8-11]及基于序列模式[12]等方法。这些方法大多是根据剪接位点的位置特性,结合外显子与内含子的统计特点而设计的。除隐马尔可夫方法和基于序列模式方法外,上述方法主要考虑到剪接位点附近存在的序列保守性,即存在所谓的 GT-A G 规则,没有考虑剪接位点附近序列各碱基之间的相关性,而这种相关性也是具有一定统计规律的。文献[10]的研究表明,利用碱基之间存在某种相关性能够更好地提取位点附近保守序列在边缘分布与条件分布(转移概率)上的统计特征。隐马尔可夫模型输出独立性假设,导致其特征

收稿日期:2010-04-02
基金项目:黑龙江省科技攻关资助项目(GZ07A103)
作者简介:杨王黎(1966-),女,大庆石油学院副教授,主要从事计算机软件研究,(Tel)13936811937;(E-mail)ywl008@126.com。

的选择仅能反映相邻碱基之间的相关性，而不能表示出非相邻碱基之间关系；基于序列模式的方法虽然在主题模式选取中能够考虑到碱基之间的相互关系，但需要进行主题模式的抽取。笔者使用条件随机场(CRF)[13]实现DNA剪接位点的预测。条件随机场不仅能够避免隐马尔可夫中特征选取受限的问题，而且能够解决最大熵隐马模型中局部的最优和标记偏见的问题。

1 条件随机场

条件随机场(CRF)是一种用于在给定输入结点值时计算指定输出结点值的条件概率的无向图模型(Lafferty et al.(2001))。用 X 表示一个值可以被观察的“输入”随机变量集合，Y 表示被模型预测的“输出”随机变量的集合，用 C 表示条件随机场中输入节点和输出节点所构成的无向图中的团的集合，$c\in C$ 上的势函数表示为 $\Phi_c(Y_c,X_c)$，CRFs将输出随机变量值的条件概率定义为与无向图中各个团的势函数(potential function)的乘积

$$P_\Lambda(Y\mid X)=\frac{1}{Z(X)}\prod_{c\in C}\Phi_c(Y_c,X_c),\qquad(1)$$

其中 $Z(X)=\sum_Y\prod_{c\in C}\Phi_c(Y_c,X_c)$ 是 X 的所有状态序列的归一化因子。根据特征集合$\{f_k\}$设定势函数为 $\Phi_c(Y_c,X_c)=\exp(\sum_k\lambda_k f_k(Y_c,X_c))$，以便$\{P_\Lambda\}$是1个指数族。一般情况下假定$\{f_k\}$是提前给定且固定的。模型的参数为 $\Lambda=\{\lambda_k\}$。使用一种线性链CRF，线性链CRFs假设在各个输出结点之间存在一阶马尔可夫独立性(二阶或更高阶的模型可类似扩展)。给定1个由式(1)定义的条件随机场模型，在已知输入数据序列 X 的情况下，最可能的输出序列可以由式(2)确定

$$Y=\underset{Y}{\operatorname{argmax}}P_\Lambda(Y\mid X)。\qquad(2)$$

建立CRFs模型要解决的2个关键问题是参数估计和特征选择。参数估计是从训练数据集学习每个特征的权重参数，即求解向量 $\boldsymbol{\Lambda}=\{\lambda_1,\lambda_2,\cdots,\lambda_K\}$ 的过程。模型训练过程如下：对于训练数据集 $D=\{X^{(i)},Y^{(i)}\}_{i-1}^N$，其中 $X^{(i)}=\{x_1^{(i)},x_2^{(i)},\cdots,x_T^{(i)}\}$ 是1个输入序列，$Y^{(i)}=\{y_1^{(i)},y_2^{(i)},\cdots,y_T^{(i)}\}$ 是相应的输出序列(也就是标记序列)。在训练集 $D=\{X^{(i)},Y^{(i)}\}_{i-1}^N$ 下的似然对数为

$$l(\boldsymbol{\Lambda})=\sum_{i=1}^N\log p(y^{(i)}\mid x^{(i)})。\qquad(3)$$

将式(1)带入式(3)，则有

$$l(\boldsymbol{\Lambda})=\sum_{i=1}^N\sum_{t=1}^T\sum_{k=1}^K\lambda_k f_k(y_t^{(i)},y_{t-1}^{(i)},x_t^{(i)})-\sum_{i=1}^N\log Z(x^{(i)})。\qquad(4)$$

为了避免多参数优化中过拟合情况，对数似然经常需要将参数作先验分布调整，采用高斯先验调整后

$$l(\boldsymbol{\Lambda})=\sum_{i=1}^N\sum_{t=1}^T\sum_{k=1}^K\lambda_k f_k(y_t^{(i)},y_{t-1}^{(i)},x_t^{(i)})-\sum_{i=1}^N\log Z(x^{(i)})-\sum_{k=1}^K\frac{\lambda_k^2}{2\sigma^2}。\qquad(5)$$

对式(5)求偏导有

$$\frac{\partial l}{\partial\lambda_k}=\sum_{i=1}^N\sum_{t=1}^T\lambda_k f_k(y_t^{(i)},y_{t-1}^{(i)},x_t^{(i)})-\sum_{i=1}^N\sum_{t=1}^T\sum_{y,y'}f_k(y,y',x_t^{(i)})p(y,y'\mid x^{(i)})-\sum_{k=1}^K\frac{\lambda_k}{\sigma^2},\qquad(6)$$

式中：第一项为特征 f_k 在经验分布下的期望值；第二项为特征 f_k 在模型 $\boldsymbol{\Lambda}$ 下的期望值。对于它们的计算，可采用动态规划高效实现。

2 实验

2.1 数据集

笔者选用的数据集来自人类剪接位点数据库HS3D[14](homo sapiens splice sites dataset)。HS3D是从Genebank Rel.123抽取的人类基因外显子、内含子和剪接位点序列的数据库，它是由Sannio-Benevento大学的计算机科学教授Salvatore Rampone建立的。目的是为了给基因的识别和特征标记的计算机方法实现提供标准的数据材料，以便用来训练并达到一定的位点预测精度。

在完整的GenBank(灵长类序列部分)Rel.123中有162 557个条目。HS3D取了包含完整编码序列CDS(complete coding sequence)的人类核酸DNA，并从中抽取了4 523个外显子和3 802个内含子。对这些外显子和内含子，HS3D还以包围剪接位点的140个碱基为窗口大小的形式抽取了3 799+3 799个供体和受体位点。这些序列数据中有65+74个序列不包含GT-AG规范的剪接点，有686+589个数据量不够(不能凑成140个碱基的窗口)的序列，有29+30个非AGCT碱基的序列，和218+226个冗余序列。丢弃掉这些序列后，共得到2 796+2 880个位点序列。

最后，通过在非剪接位点处搜寻规范GT-AG对找到271 937+329 374个伪剪接位点序列窗。其中，距离真实剪接位点+/−60个碱基的伪剪接位点被标记为最接近的。

在HS3D中，外显子、内含子、供体位点和受体位点的数据格式如图1所示。

该数据集的长度为140位，使用窗口截取其中的部分长度序列。对供体位点和受体位点各取其左右25个碱基组成样本序列。

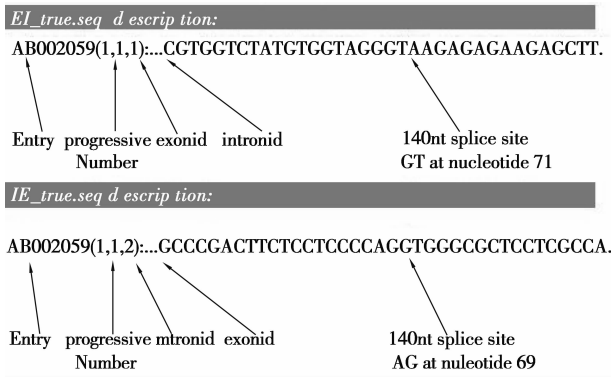

图 1　HS3D 数据集 DNA 序列格式

2.2　评价指标

对剪接位点的预测一般采用灵敏度（Sn，剪接位点预测准确率）和专一度（Sp，非剪接位点预测准确率）来评估模型的性能

$$Sn = \frac{TP}{TP + FN}; \tag{7}$$

$$Sp = \frac{TN}{TN + FP}, \tag{8}$$

其中：TP 为正确预测的剪接位点（供体位点或受体位点）数目；TN 为正确预测的非剪接位点数目；FN 为错误预测的剪接位点数目；FP 为错误预测的非剪接位点数目。

由于 $TN+FP$ 的数量要远远大于 $TP+FN$，基数较大，导致 Sp 的区分度不够，因此有文章提出了对于 Sp 的改进

$$Sp_2 = \frac{TP}{TP + FP}。 \tag{9}$$

对于位点预测的算法，既要求有较高的敏感性，也要求有较高的特异性。如果敏感性很高，但特异性比较低，则在实际应用中会产生高比率的假阳性；相反，如果特异性很高，而敏感性比较低，则会产生高比率的假阴性。需要对敏感性和特异性进行权衡，给出综合评价指标。常用的指标有[15-16]

$$F = \frac{2SnSp_2}{Sn + Sp_2}。 \tag{10}$$

$$q^9 = \begin{cases} \dfrac{TN - FP}{TN + FP}, \text{IF}: TP + FN = 0; \\ \dfrac{TN - FN}{TN + FN}, \text{IF}: TP + FP = 0; \\ 1 - \sqrt{2}\sqrt{\left(\dfrac{FN}{TP + FN}\right)^2 + \left(\dfrac{FP}{TP + FP}\right)^2}, \\ \text{IF}: TP + FN \neq 0, \text{AND}, TN + FP \neq 0。 \end{cases} \tag{11}$$

2.3　实验结果

实验中从 HS3D 数据库中分别随机抽取了供体位点和受体位点的真实样本数 500 个，虚假样本数 1 000 个。分别选取真实样本数 400 个，虚假样本数 800 个做为训练数据集，剩余的数据做为测试数据集，实验结果如表 1。

表 1　剪接位点预测结果

位点	Sn	Sp	Q_9	F
供体位点	0.967 9	0.997 3	0.967 3	0.349 5
受体位点	0.928 8	0.996 6	0.939 4	0.297 2

表 2 与表 3 中是研究方法与基于神经网络方法[7]、基于隐 Markov 模型的方法[8]及基于序列模式和 SVM 方法[12]等对剪接位点预测效果的对比，由对比结果可知，应用条件随机场进行剪接位点的预测的总体效果要好于其它方法。条件随机场是一种合适的剪接位点预测方法。

表 2　研究方法与基于序列模式特征和 SVM 方法、基于神经网络方法和 基于隐 Markov 模型的方法对供体剪接位点预测效果的对比

供体位点	Sn	Sp
研究方法	0.967 9	0.997 3
基于序列模式特征和 SVM 方法	0.9240	0.900 0
基于神经网络的方法	0.924 1	0.997 9
基于隐 Markov 模型的方法	0.951 8	0.926 0

表 3　研究方法与基于神经网络方法和基于隐 Markov 模型的方法对受体剪接位点预测效果的对比

受体位点	Sn	Sp
研究方法	0.928 8	0.996 6
基于神经网络的方法	0.873 4	0.996 8
基于隐 Markov 模型的方法	0.910 2	0.954 9

3　结　论

提出了一种基于条件随机场剪接位点预测方法，并对 HS3D 的部分数据集进行了实验，实验结果表明，基于条件随机场的方法能够较好的预测 DNA 序列中的剪接位点。但目前 CRFs 的参数训练和特征选择归纳还存在速度慢的问题，提升参数估计和特征选择归纳速度的算法将是今后的一个研究方向。

参考文献：

[1] DASH D, GOPALAKRISHNAN V. Modeling DNA splice regions by learning Bayesian networks [C]. CBMI Tech Report, 2001,11:33-40.

[2] 李骜,王涛,冯焕清,等.基于贝叶斯网络的 DNA 序列剪接位点预测[J].生物物理学报,2003,4:56-60.
LI AO, WANG TAO, FENG HUAN-QING, et al. Predicting splice junction site in DNA sequsence with bayesion network acta[J]. Biophysica Sinica, 2003, 04: 016.

[3] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines [J]. Machine Learning, 2000, 46(1-3): 389-422.

[4] SUN Y F, FAN X D, LI Y D. Identifying splicing sites in eukaryotic RNA: support vector machine approach [J]. Computers in Biology and Medicine, 2003, 33:17-29.

[5] 杨乌日吐,李前忠,刘利,等.用支持向量机预测人类基因 5′/3′选择性剪切位点[J]. 现代生物医学进展,2007,5:22-30.
YANG WU-RI-TU , LI QIAN-ZHONG, LIU LI, et al. Using support vector machine to predict alternative 5′/3′ splicing Sites of Human Genome [J]. Progress in Modern Biomedicine, 2007, 5: 22-30.

[6] 李银山,杨春燕,张伟.DNA 序列分类的神经网络方法[J].计算机仿真, 2003, 2:33-40.
LI YIN-SHAN, YANG CHUN-YAN, ZHANG WEI. The neural network method of classifications for DNA sequences[J]. computer simulation ,2003,2: 33-40.

[7] TONGLI C, QINKE P. Predicting the splice sites in DNA sequences using neural network based on complementary encoding method [C]//Proceedings of International Conference on Neural Networks. Beijing: IEEE Press, 2005,10(1): 473-476.

[8] 夏慧煜,周晴,李衍达.隐 Markov 模型在剪接位点识别中的应用[J].清华大学学报:自然科学版,2002,42(9):1214-1217.
XIA HUI-YU, ZHOU QING, LI YAN-DA. Application of hidden markov model in the recognition of splicing sites[J]. Journal of Tsinghua University: Science and Technology ,2002 ,42(9): 1214-1217.

[9] 曹胜玉,刘来福.隐马模型及其在基因识别中的应用[J].数学的实践与认识,2006(09):22-25.
CAO SHENG-YU, LIU LAI-FU. Hidden markov models and their applications in gene finding [J]. Mathematics Practice and Theory, 2006 36(9):55-61.

[10] 杨文强,邓明华,钱敏平.隐马尔可夫模型与剪切位点识别[C]//中国运筹学会第六届学术交流会.中国:[s.n],2006.

[11] 罗泽举,李艳会,宋丽红,等.基于隐马尔可夫模型的 DNA 序列识别[J]. 华南理工大学学报:自然科学版, 2007, 8: 26-29.
LUO ZE-JU, LI YAN-HUI, SONG LI-HONG, et al. Recognition of DNA sequences based on hidden markov models [J]. Journal of South China University of Technology: Natural Science Edition,2007,8: 26-29.

[12] 孙贺全,彭勤科,张全伟. 基于序列模式特征和 SVM 的剪切位点预测[J].计算机工程. 2009,35(5):180-183.
SUN HE-QUAN, PENG QIN-KE, ZHANG QUAN-WEI. Splice site prediction based on characteristics of sequence motif and support vector machine [J]. Computer Engineering, 2009 ,35(5) :80-183.

[13] LAFFERTY J, MCCALLUM A K, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//In Proceedings of 18th International Conference on Machine Learning. [s.l.]:IEEE,2001.

[14] POLLASTRO P, RAMPONE S. HS3D: homo sapiens splice site data set [J]. Nucleic Acids Research, Annual Database Issue ,2003,2:45-52.

[15] BURSET M, GUIGO R. Evaluation of gene structure prediction programs [J]. Genomics, 1996, 34: 353-367.

[16] ZHANG R. Evaluation of gene-finding algorithms by a content—balancing accuracy index[J]. J Biomol Struct Dyn, 2002,19:1045-1052.

（编辑 侯 湘）