ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 5,Sep. 2016,pp. 1010－1019
DOI:10.16337/j.1004-9037.2016.05.0019

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于序列模式挖掘的基因剪接位点

孙永山[1]　赵海峰[1]　汤振宇[1]　李　旦[2]　马　猛[1,2]　陈　荣[2]

(1.安徽大学计算机科学技术学院,合肥,230601;2.西奈山伊坎医学院遗传学与基因组学系,纽约,10029)

摘　要:剪接是基因表达过程中连接转录和翻译的中枢步骤,是一个高度调控的过程。剪接位点是基因剪接过程中的核心调控元件。本文通过挖掘剪接位点序列中蕴含的序列特征,提出了一个基于序列模式挖掘的基因剪接位点序列打分模型。通过该模型,实现对剪接位点序列信号强度的定量度量。实验结果表明,该模型可有效分类真假剪接位点序列,分类效果优于最大信息熵模型,模型具有良好的鲁棒性,并且可有效识别致病剪接位点序列突变。

关键词:剪接位点;序列模式;最大信息熵模型;致病突变

中图分类号:TP391　　**文献标志码**:A

Gene Splice Sites Based on Sequential Pattern Mining

Sun Yongshan[1], Zhao Haifeng[1], Tang Zhenyu[1], Li Dan[2], Ma Meng[1,2], Chen Rong[2]

(1. School of Computer Science and Technology, Anhui University, Hefei, 230601, China; 2. Department of Genetics and Genomic Science, Icahn School of Medicine at Mount Sinai, New York, 10029, USA)

Abstract: Gene splicing as a tightly regulated process, is a pivotal process between transcription and translation during gene expression. Splice sites are the kernel regulatory elements for gene splicing. Here, based on the sequential features minded from splice site sequences, we develop a score system for splice site sequences. Through this score system, splice site sequence can be measured quantitatively. The experimental results show that the canonical and pseudo splice site sequences can be discriminated effectively. Moreover, this model outperforms the maximum information entropy model with a great robustness, and the pathogenic splice site sequence mutations can be detected efficiently by the model.

Key words: splice sites; sequential pattern; maximum entropy model; pathogenic mutation

引　　言

人基因表达是一个高度调控的过程[1,2]。在人的基因序列中,编码片段通常被非编码片段隔开,其中编码片段称为外显子,非编码片段称为内含子。基因 DNA 序列经过转录得到前体 mRNA,从前体 mRNA 中剪除内含子,将外显子重新拼接,得到成熟 RNA 的过程,称为基因剪接[3-5]。基因剪接过程受多种剪接信号的调控,其中剪接位点是基因剪接过程中的核心调控元件。剪接位点通常指内含子两端

基金项目:国家自然科学基金(61300057, 81000321)资助项目;安徽省自然科学基金(1208085QF120, 1408085QF120, 1408085MKL94)资助项目;国家高技术研究发展计划("八六三"计划)子项目(2014AA015104)资助项目;安徽省教育厅重点(KJ2016A040)资助项目;教育部留学回国启动资金(教外司留【2014】1685)资助项目;2013 年安徽省留学人员择优资助项目。
收稿日期:2016-04-18;**修订日期**:2016-06-30

的保守二聚体。大约97%的人基因序列的内含子两端保守位点都是GT-AG。然而,人基因组中包含有非常多的GT-AG配对,大部分GT-AG并不定义内含子,仅有小于1%的GT-AG配对定义内含子,一个合理的猜测是GT-AG上下游序列片段包含有重要的特征信息,定义了GT-AG位点的生物功能。

多种计算模型已被用于剪接位点的识别。位置特异分值矩阵模型(Position specific scoring matrix, PSSM)最早被用于剪接位点建模[6];文献[7]应用支持向量机模型识别剪接位点;文献[8]结合剪接信号与调节元件信息设计了一个基于多层次支持向量机的剪接位点识别算法;文献[9]基于隐马尔可夫模型构建了剪接位点的分类器;文献[10]利用贝叶斯模型识别剪接位点[10];文献[11]提出了一种融合多种信息(调控元件信息、结构信息等)的方法提高了剪接位点识别精度;文献[12]利用最大信息熵模型研究剪接位点序列中的保守特征信息,并用于识别剪接位点;神经网络模型也被应用于剪接位点的分类识别[13]。但目前对剪接位点建模研究存在若干问题。例如,支持向量机、神经网络以及最大熵模型等都属于黑盒模型,在进行剪接位点分析时,只能给出预测结果,无法给出预测的依据。从生物学家的角度来看,一个好的预测模型不仅要有良好的预测精度,还要有透明的预测机构。PSSM模型假定单核苷酸位点间彼此独立,但这种假设不符合生物规律,这一点已为多项生物学研究所证实[14,15]。在具有特殊生物功能的碱基序列中,不同位置间实际上是相互关联的,而并非独立。PSSM模型固定了每个位置上的碱基频率,而没考虑到移动序列模式的存在,以下为若干个3′AG剪接位点序列:CTGTCTTC-CCTAGTGC;CCGTGCCCTGCAGGAG;CCTGCACACCCAGGGA;TGCCTCCCTGCAGAAG;CTGC-CCCTCACAGCCT;CATATGCCCACAGGGT;TGTTCTTTGCCAGGTT。可以看出,TGC在多个不同的位置上频繁出现,而PSSM方法无法建模移动模式TGC。

序列模式挖掘是数据挖掘领域中一个重要的研究问题[16,17]。数据挖掘中,顾客在超市购物的历史交易数据库中,顺序频繁出现的商品称为序列模式。假如将交易时间看作基因序列上的位置,商品为单核苷酸,则一位顾客的历史购买记录可以看作一个单核苷酸序列片段,一组序列可以看作一个顾客的历史交易数据库,其中包含的序列模式就可以看作这组序列中包含的保守性序列特征。基于足够的训练序列,移动模式可以被建模为序列模式,移动模式之外的位点可以用PSSM建模。由此,本文提出一个基于序列模式挖掘构建的剪接位点识别模型。

1 基于序列模式挖掘建模剪接位点序列

为了便于数学形式化描述,以长度为9的剪接位点序列为例子,给出相关概念说明。

(1)定义一个集合

$$I=\{i_{pn} \mid p\in\{1,2,3,4,5,6,7,8,9\} \quad n\in\{A,C,T,G\}\} \tag{1}$$

式中:i_{pn}为一个项目,表示9聚体的第p个位置出现了核苷酸n。集合I总共包含36个项目。例如i_{1A}表示单核苷酸A在序列的第1个位置出现。

(2)如果任何一个集合I包含k个项,则称该非空子集为k-项集(如$\{i_{1G},i_{3A}\}$为2-项集)。一个未知的9聚体q可以定义为满足如下约束的9项集,即有

$$q=\{i_{1n},i_{2n},i_{3n},i_{4n},i_{5n},i_{6n},i_{7n},i_{8n},i_{9n} \mid n\in\{A,C,G,T\}\} \tag{2}$$

例如:ACGTCGACT可以写成9项集为$\{i_{1A},i_{2C},i_{3G},i_{4T},i_{5C},i_{6G},i_{7A},i_{8C},i_{9T}\}$(或简写为$i_{1A}i_{2C}i_{3G}i_{4T}i_{5C}i_{6G}i_{7A}i_{8C}i_{9T}$)。

(3)将存在的真剪接位点和假剪接位点定义为4种类型:真5′端剪接位点(D_{C5SS})、假5′端剪接位点(D_{P5SS})、真3′端剪接位点(D_{C3SS})和假3′端剪接位点(D_{P3SS})。

(4)给定一个数据集D和其中的项集X,X的支持度定义为$\mathrm{Support}_D(X)$,表示数据集D中项集X的比例

$$\text{Support}_D(X)=\frac{C_D(X)}{|D|} \tag{3}$$

式中：$C_D(X)$ 为在 D 中含有 X 的所有项的项集数目；$|D|$ 为 D 中交易的总数量。

(5)对数据集 D，给定一个最小支持阈值 minsup，如果 $\text{Support}_D(\text{itemset}) \geqslant \text{minsup}$，则称该项集为频繁项集(Frequent item set，FIS)，如果项集中包含 k 个项目，则称该频繁项集为 k 项集或 k 项序列特征。

(6)从数据集 D 中得到的 FIS 指出了在序列中的某些特定位置上频繁出现的核苷酸，这些特定位置称为该序列特征的保守位置，其余的称为灵活位置。例如，频繁项集 $\{i_{1A}, i_{2A}, i_{3A}\}$ 确定了保守位置(1,2,3)和灵活位置(4,5,6,7,8,9)。其中，在 3 个固定位置的核苷酸分别为 A，其余 6 个灵活位置上核苷酸出现概率由满足该 FIS 的所有的序列统计得到。将灵活位置上的核苷酸出现概率矩阵与 FIS 结合，可获得扩展序列特征(Extended sequential feature，ESF)。例如：FIS$\{i_{1A}i_{2A}i_{3A}\}$ 的扩展序列特征 ESF 可描述为

AAA[A:23.5%，C:29.4%，G:41.2%，T:5.9%][A:35.9%，C:35.9%，G:23.5%，T:5.9%][A:35.9%，C:29.4%，G:17.6%，T:17.6%][A:23.5%，C:29.4%，G:41.2%，T:5.9%][A:35.9%，C:23.5%，G:35.9%，T:5.9%][A:35.9%，C:29.4%，G:17.6%，T:17.6%]。

对于每一个 ESF，定义 ESF 的支持度为对应 FIS 的支持度。

(7)给定扩展序列特征 ESF 和序列 q，则 q 满足 ESF 的概率计算如下

$$\text{Probability}_{\text{ESF}}(q)=\prod_{i\in\{\text{flexiblepositions}\}}\boldsymbol{PM}_{i,q_i} \tag{4}$$

式中：$\boldsymbol{PM}$ 为灵活位置上的核苷酸出现概率矩阵。则 q 属于数据类别 D 的概率可 用如下分值度量

$$\text{Score}_D(q)=\sum_{\text{esf}\in\{\text{ESF}_{D,q}\}}\text{Support}_D(ESF)\times\text{Probability}_{\text{ESF}}(q) \tag{5}$$

式中：$\text{ESF}_{D,q}$ 表示 q 满足的 ESF。

(8)对一个待预测的序列 q，分别根据真 5′端剪接位点 D_{C5SS} 和假 5′端剪接位点 D_{P5SS} 计算出对应分值：$\text{Score}_{D_{\text{C5SS}}}(q)$ 度量 q 属于 D_{C5SS} 的概率，$\text{Score}_{D_{\text{P5SS}}}(q)$ 度量 q 属于 D_{P5SS} 的概率。定义趋向值(Tendency ratio，TR)如下

$$\text{TR}(q)=\frac{\text{Score}_{D_{\text{C5SS}}}(q)-\text{Score}_{D_{\text{P5SS}}}(q)}{\text{Score}_{D_{\text{C5SS}}}(q)+\text{Score}_{D_{\text{P5SS}}}(q)} \tag{6}$$

趋向值取值范围为[−1，+1]，如果 TR 是个正小数，则提示 q 倾向表现为真 5′剪接位点，如果 TR 是个负数，则提示 q 倾向表现为假 5′剪接位点。对 3′端的定义类似于 5′端。图 1 为该模型分析流程图。

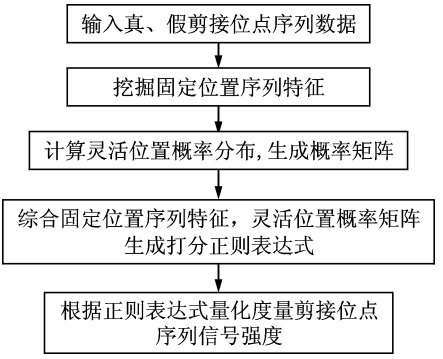


图 1 模型分析流程图

Fig. 1 Analysis flowchart for model

2 识别定义剪接位点的最佳上下游序列长度

本文从人基因组中紧邻剪接位点的上下游区域分别抽取内含子和外显子序列来构建真剪接位点序列，从内含子中抽取假剪接位点序列(详见 3.1 节)。为了确定定义剪接位点的最佳上下游序列长度，首先固定外显子序列长度为 3 bp，计算不同长度内含子序列对应的真、假 5′和 3′剪接位点序列趋向值。5′端内含子序列长度变化如表 1 所示，3′

表 1 5′端内含子长度变化表

Tab. 1 5′ Different lengths of intronic sequence bp

序列	外显子	内含子
L_7	3	4
L_8	3	5
L_9	3	6
L_10	3	7
L_11	3	8

端内含子序列长度变化如表 2 所示,5′端剪接位点序列趋向值分布盒图如图 2 所示,3′端剪接位点趋向值盒图如图 3 所示。

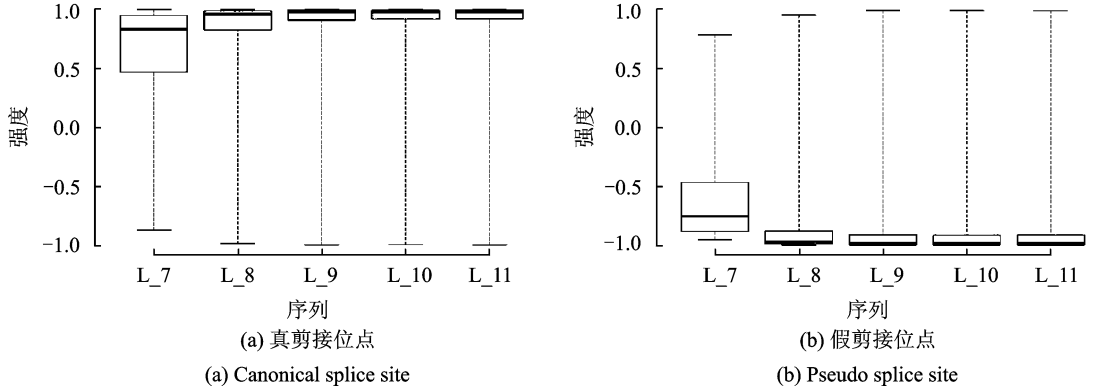


图 2 固定外显子序列(3 bp)对应的 5′端趋向值盒图

Fig. 2 Boxplot of tendency ratios for 5′ with 3 bp of exonic sequence

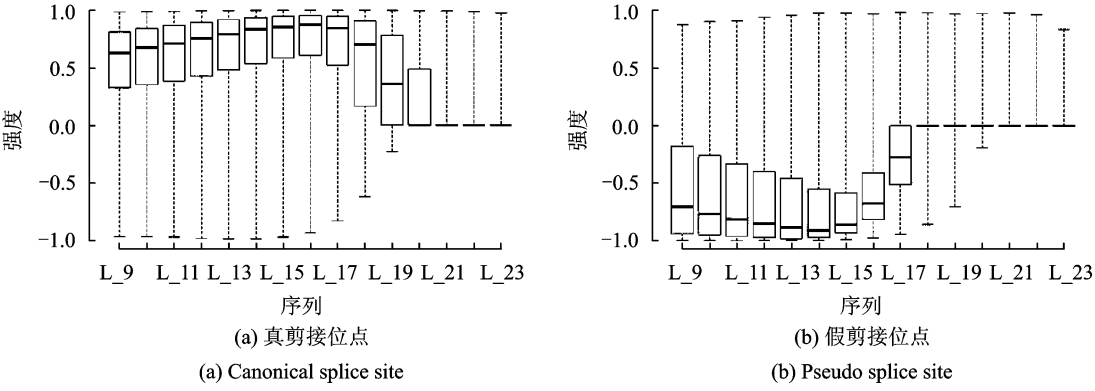


图 3 固定外显子序列(3 bp)对应的 3′端趋向值盒图

Fig. 3 Boxplot of tendency ratios for 3′with 3 bp of exonic sequence

在图 2 和表 1 中,真、假 5′端剪接位点序列在内含子序列长度为 6 时,对应趋向值差异最大,继续延长内含子序列,差异值没有显著改变,所以,对于真 5′剪接位点序列,长度 6 为最优内含子序列选择。在图 3 和表 2 中,真、假 3′端剪接位点序列在内含子序列长度为 12,13,14 时,对应趋向值差异显著,真 3′剪接位点序列在内含子序列长度取 13 时,对应趋向值最高,所以,对于真 3′剪接位点序列,选择 13 作为最优内含子序列长度。

表 2 3′端内含子长度变化表

Tab. 2 3′ different lengths of intronic sequence bp

序列	外显子	内含子	序列	外显子	内含子
L_9	3	6	L_17	3	14
L_10	3	7	L_18	3	15
L_11	3	8	L_19	3	16
L_12	3	9	L_20	3	17
L_13	3	10	L_21	3	18
L_14	3	11	L_16	3	19
L_15	3	12	L_23	3	20
L_16	3	13			

对 5′端剪接位点，固定内含子序列长度为 6 bp，对 3′端剪接位点，固定内含子长度为 13 bp，然后分别延长外显子序列长度，5′端外显子序列长度变化如表 3 所示，3′端外显子序列长度变化如表 4 所示。计算对应真、假 5′和 3′剪接位点序列趋向值，结果分别如图 4，5 所示。

表 3 5′端外显子长度变化表

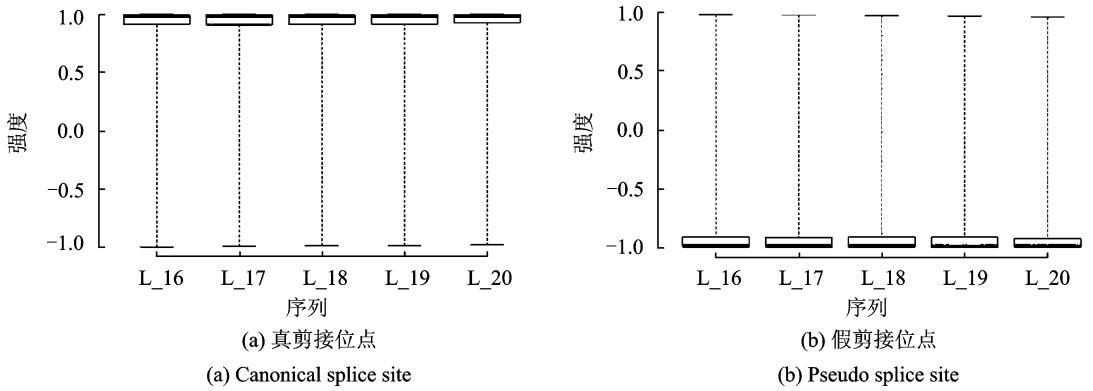
Tab. 3 5′ different lengths of exonic sequence bp

序列	外显子	内含子
L_9	3	6
L_10	4	6
L_11	5	6
L_12	6	6
L_13	7	6

表 4 3′端外显子长度变化表

Tab. 4 3′different lengths of exonic sequence bp

序列	外显子	内含子
L_16	3	13
L_17	4	13
L_18	5	13
L_19	6	13
L_20	7	13

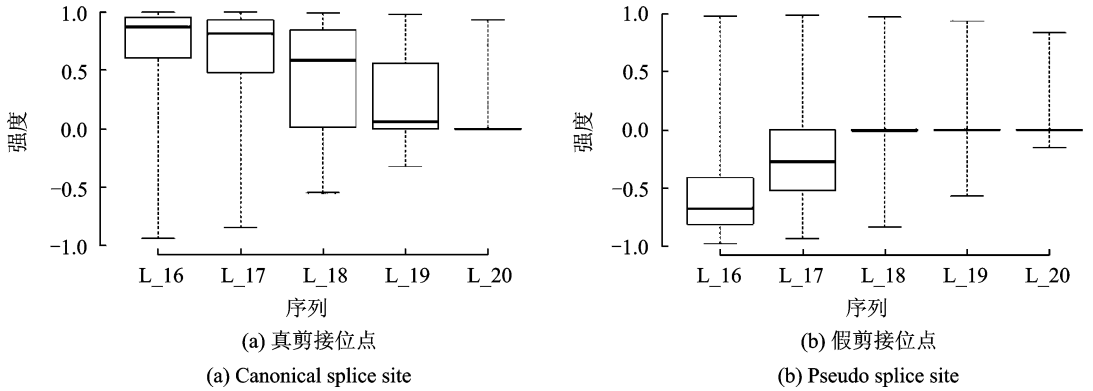


(a) 真剪接位点 (a) Canonical splice site

(b) 假剪接位点 (b) Pseudo splice site

图 4 固定内含子序列(6 bp)对应的 5′端趋向值盒图

Fig. 4 Boxplot of tendency ratios for 5′ with 6 bp of intronic sequence



(a) 真剪接位点 (a) Canonical splice site

(b) 假剪接位点 (b) Pseudo splice site

图 5 固定内含子序列(13 bp)对应的 3′端趋向值盒图

Fig. 5 Boxplot of tendency ratios for 3′ with 13 bp of intronic sequence

在图 4 和表 3 中，固定内含子序列长度为 6 bp，延长外显子序列长度，对应真、假 5′剪接位点序列趋向值差异没有显著改变，因此对于 5′剪接位点，选择 3 bp 作为最优外显子序列长度。在图 5 和表 4 中，固定内含子序列长度为 13 bp，延长外显子序列长度，对应真、假 3′端剪接位点序列趋向值差异显著缩小，因此对于 3′剪接位点，选择 3 bp 作为最优外显子序列长度。

根据以上实验结果,对 5′和 3′端剪接位点序列,最优外显子序列长度均为 3 bp,最优内含子序列长度分别为 6 bp 和 13 bp。

3 实验与结果

3.1 实验数据

本文主要分析剪接位点的邻近上下游序列片段,从中寻找定义保守位点 GT-AG 具有剪接位点功能的序列特征。从 UCSC 人基因组[18]中抽取真 5′和 3′剪接位点序列,剔除其中存在的重复剪接位点序列;为了避免其他基因剪接调控信号可能造成的干扰[19],剔除所有上下游发生可变剪接事件的剪接位点序列,人基因组中发生的所有可变剪接事件从 UCSC Genome Bioinformatics 的表格 KnownAlt 中获取[18];为了避免转录调控信号的影响[20],进一步剔除第 1 个和最后 1 个外显子的剪接位点序列。最后得到 2 343 条长度为 9 的真 5′剪接位点序列,和 69 081 条长度为 16 的真 3′剪接位点序列。从长度大于 2 000 bp 的内含子序列中间区域抽取含有保守位点 GT 或 AG 假 5′和 3′剪接位点序列,最后得到9 532 个长度为 9 的假 5′剪接位点序列,和 62 364 个长度为 16 的假 3′剪接位点序列。

3.2 真、假剪接位点

对真、假剪接位点序列,利用本文提出的序列模式挖掘模型,可首先分别计算出每一条序列的真、假剪接分值,然后进一步算出倾向值。真、假 5′和 3′端剪接位点倾向值柱状统计分布如图 6,7 所示。

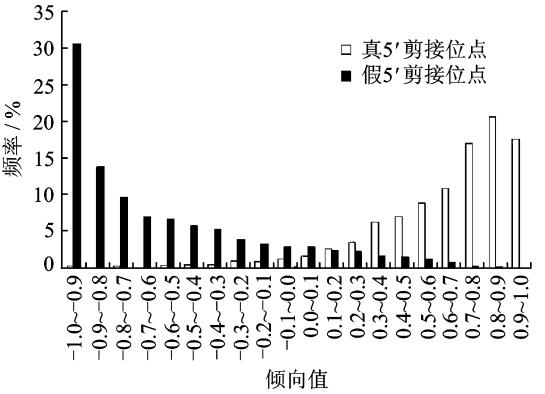


图 6 真、假 5′端剪接位点序列趋向值柱状统计分布图

Fig. 6 Histogram of tendency ratios for canonical and pseudo 5′ splice site sequences

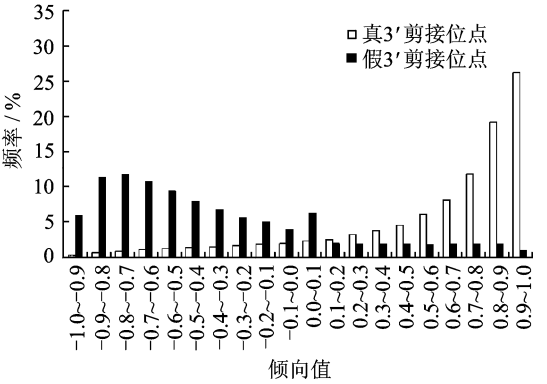


图 7 真、假 3′端剪接位点序列趋向值柱状统计分布图

Fig. 7 Histogram of tendency ratio for canonical and pseudo 3′ splice site sequences

由图 6,7 可知,超过 90%的假 5′剪接位点和超过 80%的假 3′剪接位点具有负趋向值;超过 90%的真 5′剪接位点和超过 85%的真 3′剪接位点具有正趋向值。趋向值可有效区分真、假剪接位点。

3.3 序列模式挖掘模型和最大信息熵模型的比较

最大信息熵模型(Maximum entropy model)是由 Dr. Christopher Burge 首次提出用于剪接位点序列建模的方法[12],多项研究均指出最大信息熵模型明显优于其他剪接位点建模模型[21,22]。对任一测试序列,最大信息熵模型可产生一个 MES 分值,分值越大,该序列为真剪接位点序列的可信度越大[12]。由于 MES 分值和序列模式挖掘模型产生的趋向值具有不同的值域,为了便于比较这两种模型,对两种分值进行 min-max 标准化处理。对真、假 5′和 3′剪接位点序列应用这两种模型,分布计算出每条序列的趋向值和 MES 分值,标准化处理后真、假序列分值分布如图 8 所示。

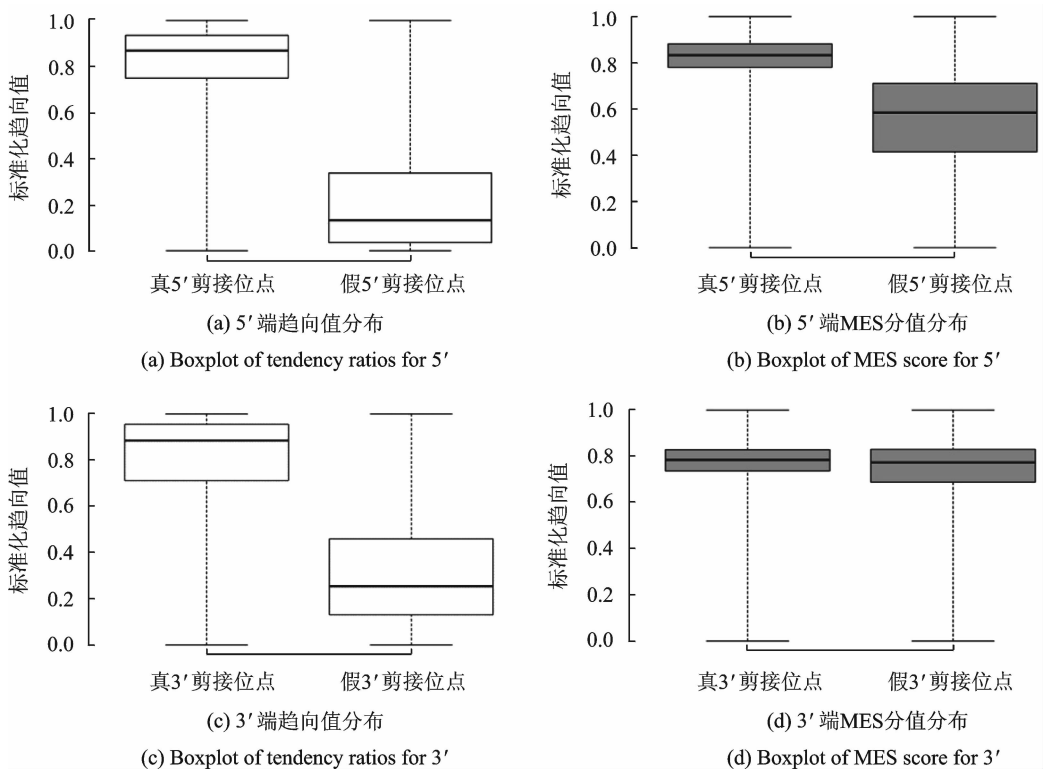


图 8 标准化处理后真、假剪接位点趋向值和 MES 分值分布盒图

Fig. 8 Boxplot of normalized tendency ratio and MES score for canonical and pseudo 5′ and 3′ splice site sequences

图 8(a,c)可有效区分真、假 5′和 3′剪接位点;图 8(b)基本可以区分真、假 5′剪接位点,但区分能力明显弱于标准化趋向值;图 8(d)中基本很难区分真、假 3′剪接位点。序列模式挖掘模型明显优于最大信息熵模型。

3.4 序列模式挖掘模型的鲁棒性

为了验证序列模式挖掘模型的健壮性,本文进一步选取模型产生的真剪接位点分值和趋向值作为特征向量,采用支持向量机模型作为分类器,使用留一法和分层十折交叉法进行计算验证。实验应用 LIBSVM[23] 实现支持向量机分类器,采用径向基函数作为核函数。下面以 5′剪接位点序列为例,说明如何实现留一法和分层十折交叉验证。

(1) 留一法:将真、假 5′端剪接位点序列混合,从混合数据集上每次抽取一个序列作为预测样本,其余作为训练样本,然后基于训练样本应用 LIBSVM 构建分类器,对测试样本进行类别预测;重复该过程,直到所有序列都被作为一次测试样本为止。

(2) 分层十折交叉法:将真、假 5′剪接位点混合数据分成 10 个子集,其中的 9 个子集,每个子集包含 234 条真 5′剪接位点序列和 953 条假剪接位点序列,第 10 个子集包含 237 条真 5′剪接位点序列和 955 条假 5′剪接位点序列;选择一个子集作为测试集,其余作为训练集应用 LIBSVM 构建分类器,对测试集样本进行类别预测;重复该过程,直到所有的子集均被用作一次测试集。

对真、假 5′和 3′剪接位点序列的留一法和分层十折交叉法验证实验结果见表 5 和表 6。对真、假剪接位点,两种验证方法得到的敏感度和特异度均大于 80%。两种方法对 5′剪接位点序列的预测精确度

均超过了 93%,对 3′剪接位点序列的预测精确度均超过了 84%。留一法和分层十折交叉法的验证实验结果表明序列模式挖掘模型具有良好的鲁棒性。

表 5 真、假 5′剪接位点序列的留一法和分层十折交叉法验证实验结果

Tab. 5 Experimental results of leave-one-out and cross ten-fold validation for canonical and pseudo 5′ splice site sequences %

方法	敏感度	特异度	精确度
留一法	80.28	87.41	93.83
分层十折交叉法	80.03	87.37	93.78

表 6 真、假 3′剪接位点序列的留一法和分层十折交叉法验证实验结果

Tab. 6 Experimental results of leave-one-out and cross ten-fold validation for canonical and pseudo 3′ splice site sequences %

方法	敏感度	特异度	精确度
留一法	84.42	86.39	84.76
分层十折交叉法	84.33	86.41	84.73

3.5 序列模式挖掘模型对致病剪接位点突变的识别

人类基因组如果在剪接位点序列发生了突变,可能会影响剪接位点的识别,甚至摧毁剪接位点,导致外显子跳跃而产生异常 mRNA 剪接体,促进疾病发生。为了验证序列模式挖掘模型是否可以有效识别致病剪接位点序列突变,本文从 DBASS 数据库[24-26]中分别收集了 294 个致病 5′剪接位点序列突变和 154 个 3′剪接位点序列突变,抽取对应的野生剪接位点序列和致病突变剪接位点序列,并在野生剪接位点序列上进行随机突变,产生对照组剪接位点序列。利用序列模式挖掘模型对野生组、对照组和致病突变组中的序列打分、计算趋向值,结果如图 9 所示。

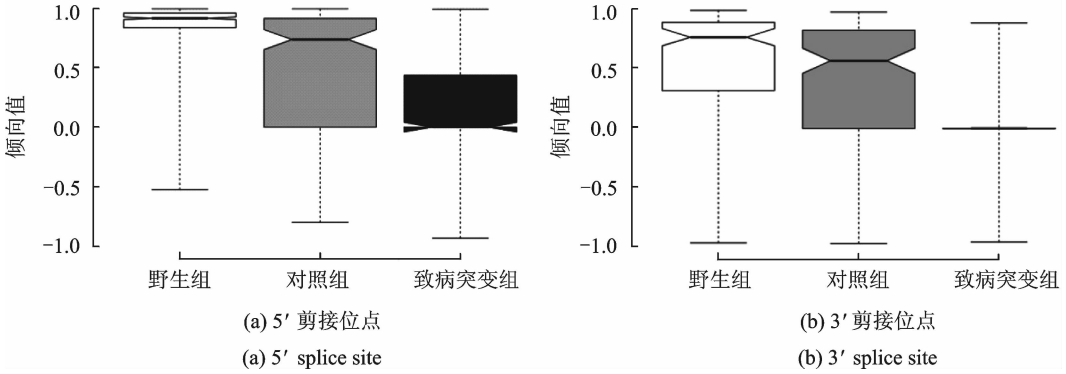


图 9 不同对照组的趋向值分布

Fig. 9 Boxplot of tendency ratio of variants from different control groups

在图 9(a)中,80%的野生组中 5′剪接位点序列的倾向值大于 0.8,超过 75%的致病突变组中 5′剪接位点序列倾向值小于 0.5,对照组倾向值分布介于两者之间;在图 9(b)中,超过 65%的野生组中 3′剪接位点序列的倾向值大于 0.5,超过 85%的致病突变组中 3′剪接位点序列的倾向值接近 0,对照组中倾向值分布介于两者之间。致病突变显著降低了剪接位点序列的倾向值,序列模式挖掘模型可有效识别致病剪接位点突变。

4 结束语

正确识别基因剪接位点是进一步探索基因选择性剪接机制的前提,也为治疗因剪接位点突变而导致的疾病提供依据和指导。本文应用序列模式挖掘对人基因剪接位点序列进行建模,构建出一套剪接位点序列的打分模型,对剪接位点信号进行强度量化和识别。本文提出的模型思想简单,具有良好的数学理论基础。实验结果表明,该模型可有效区分真假剪接位点,性能优于最大信息熵模型,具有良好的鲁棒性,并且该模型可有效地识别致病剪接位点序列突变。未来将进一步应用该模型分析恶性疾病病人基因组深度测序数据,探究疾病致病突变和致病基因。

参考文献:

[1] Gutierrez A M, Ongen H, Lappalainen T, et al. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing[J]. PLoS Genet, 2015, 11(1): 1-26.

[2] Ma M, Ru Y, Chuang L S, et al. Disease-associated variants in different categories of disease located in distinct regulatory elements[J]. BMC Genomics, 2015, 16(S8): 1-13.

[3] Wang Z, Burge C B. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code[J]. RNA, 2008, 14(5): 802-813.

[4] Wang Y, Ma M, Xiao X, et al. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules [J]. Nature Structural & Molecular Biology, 2012, 19(10): 1044-1052.

[5] Wang Y, Xiao X, Zhang J, et al. A complex network of factors with overlapping affinities represses splicing through intronic elements[J]. Nature Structural & Molecular Biology, 2013, 20(1): 36-45.

[6] Shapiro M B, Senapathy P. RNA splice junctions of different classes of eukaryotes:Sequence statistics and functional implications in gene expression[J]. Nucleic acids Research, 1987, 15(17): 7155-7174.

[7] Sonnenburg S, Schweikert G, Philips P, et al. Accurate splice site prediction using support vector machines[J]. BMC Bioinformatics, 2007, 8(S10): S7.

[8] 孙晓宗, 桑凌洁, 居理宁,等. 基于剪接信号和调节元件序列特征的剪接位点预测方法[J]. 科学通报, 2008, 53(19): 2298-2306.
Sun Xiaozong, Sang Lingjie, Ju Lining, et al. Predicting methods for splice sites based on splicing signals and regulatory element sequential features[J]. Science China, 2008, 53(19): 2298-2306.

[9] Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel[J]. Bioinformatics, 2003, 19 (S2): ii215-ii225.

[10] 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, 19(4): 431-436.
Li Ao, Wang Tao, Feng Huanqing, et al. Predicting splice junction site in DNA sequences with Bayesian network[J]. Acta Biophysica Sinica, 2003,19(4):431-436.

[11] 王科俊, 吕俊杰, 冯伟兴,等. 一种新的真核基因剪接位点识别方法[J]. 电子学报, 2011, 39(5): 1210-1213.
Wang Kejun, Lü Junjie, Feng Weixing, et al. A new method for identification of eukaryotic gene splice sites[J]. Acta Electronical Sinica, 2011, 39(5): 1210-1213.

[12] Yeo G, Burge C B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals[J]. Journal of Computational Biology, 2004, 11(2/3): 377-394.

[13] Reese M G, Eeckman F H, Kulp D, et al. Improved splice site detection in genie[J]. Journal of Computational Biology, 1997, 4(3): 311-323.

[14] Man T K, Stormo G D. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay[J]. Nucleic Acids Research, 2001, 29(12): 2471-2478.

[15] Roulet E, Busso S, Camargo A A, et al. High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites[J]. Nature Biotechnology, 2002, 20(8): 831-835.

[16] Shen W, Wang J, Han J. Sequential pattern mining:Frequent pattern mining[M]. Switzerland:Springer International Publishing,2014: 261-282.

[17] Mooney C H, Roddick J F. Sequential pattern mining—Approaches and algorithms[J]. ACM Computing Surveys (CSUR), 2013, 45(2): 1-19.

[18] Rosenbloom K R, Armstrong J, Barber G P, et al. The UCSC genome browser database: 2015 update[J]. Nucleic Acids Research, 2015, 43(D1): D670-D681.

[19] Matera A G, Wang Z. A day in the life of the spliceosome[J]. Nature Reviews Molecular Cell Biology, 2014, 15(2): 108-121.

[20] Bentley D L. Coupling mRNA processing with transcription in time and space[J]. Nature Reviews Genetics, 2014, 15(3): 163-175.

[21] Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome[J]. Nucleic Acids Research, 2014,42(22): 13534-13544.

[22] Hellen B. Splice site tools:A comparative analysis report [J]. National Genetics Reference Laboratory, 2009,35(12): 1-12.

[23] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.

[24] Královičová J, Christensen M B, Vorechovský I. Biased exon/intron distribution of cryptic and de novo 3′ splice sites[J]. Nucleic Acids Research, 2005, 33(15): 4882-4898.

[25] Vorechovsky I. Aberrant 3′ splice sites in human disease genes:Mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization[J]. Nucleic Acids Research, 2006, 34(16): 4630-4641.

[26] Buratti E, Chivers M, Královičová J, et al. Aberrant 5′ splice sites in human disease genes: Mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization[J]. Nucleic Acids Research, 2007, 35(13): 4250-4263.

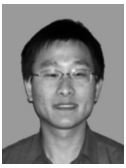
作者简介:

孙永山(1990-),男,硕士研究生,研究方向:数据挖掘、生物信息学,E-mail: sys_ahu@163.com。

赵海峰(1972-),男,副教授,研究方向:医学图像处理、模式识别。

汤振宇(1981-),男,讲师,研究方向:医学图像处理。

李 旦(1972-),男,副教授,研究方向:癌症基因组学。



马 猛(1978-),男,副教授,研究方向:数据挖掘、生物信息学。

陈 荣(1972-),男,副教授,研究方向:遗传学、生物信息学。