PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2019,46(5):496~503
www.pibb.ac.cn

基于统计差表与加权投票的高精度剪接位点预测*

曾 莹[1,2] 陈 渊[1] 袁哲明[1)**]

([1]) 湖南农业大学，湖南省农业大数据分析与决策工程技术研究中心，长沙 410128；[2)] 湖南农业大学东方科技学院，长沙 410128）

摘要 基于机器学习的高精度剪接位点识别是真核生物基因组注释的关键. 本文采用卡方测验确定序列窗口长度，构建卡方统计差表提取位置特征，并结合碱基二联体频次表征序列；针对剪接位点正负样本高度不均衡这一情形，构建10个正负样本均衡的支持向量机分类器，进行加权投票决策，有效解决了不平衡模式分类问题. HS^3D数据集上的独立测试结果显示，供体、受体位点预测准确率分别达到93.39%、90.46%，明显高于参比方法. 基于卡方统计差表的位置特征能有效表征DNA序列，在分子序列信号位点识别中具有应用前景.

关键词 剪接位点，位置特征，卡方统计差表，加权投票，支持向量机
中图分类号 Q51，Q61 **DOI**：10.16476/j.pibb.2018.0267

随着DNA测序技术的不断进步，基因组序列数据呈指数增长，迫切需要完成基因组序列注释，以深入理解基因的生物学功能. 基因识别是基因组注释的核心任务之一. 在大多数真核生物的基因结构中，编码区由外显子和内含子交替组成，而外显子与内含子间的边界即为剪接位点，其中，内含子的5'端被称为供体剪接位点，3'端被称为受体剪接位点. 如果能准确预测剪接位点，就能正确定位编码区，因此剪接位点预测是真核基因识别的关键环节. 目前真核基因剪接位点检测方法包括生物实验方法和计算机方法. 前者的结果准确，但成本高昂，无法大规模使用；后者成本低，但识别精度不如前者高. 因此，发展高精度的剪接位点计算机识别方法至关重要. 几乎99%的真核基因剪接位点都遵循“GT-AG”规则，即供体剪接位点为保守序列GT，受体剪接位点为保守序列AG[1]. 然而，这种较强的保守性并不能有效检测出剪接位点，因为更多的GT/AG存在于非剪接位点上. 因此，剪接位点预测可视作模式识别中的一个非平衡二分类问题，即将少量真实剪接位点（正样本）与大量满足“GT-AG”规则的虚假剪接位点（负样本）进行分类.

基于机器学习方法的剪接位点预测过程主要包括特征提取和分类器选择（或设计）. 提取的特征通常基于碱基位置信息[2-3]、序列组分信息[2-4]、相邻或非相邻核苷酸间的关联信息[4-6]、RNA二级结构信息[7]等. 常用的分类器有支持向量机（SVM）[8-12]、人工神经网络[13]、随机森林[14]等. 现有剪接位点识别方法虽然已取得了相对较高的识别精度，例如HS^3D数据集上报道的供体位点预测精度达到了90%以上，但由于基因序列中的GT/AG数量巨大，如人类基因组序列中约有1.87亿个GT/AG，即便总精度的细微提升也能极大增加真实剪接位点的检出数量，因此有必要进一步改进剪接位点预测方法.

本方法通过构建卡方统计差表来提取序列位置特征，并融入碱基二联体频次构成的序列组分特征，采用10个基于均衡子训练集的SVM分类器进行加权投票决策. 在相同的HS^3D数据集上，与其

* 国家自然科学基金（61701177），湖南省自然科学基金（2018JJ3225）和湖南省教育厅科学研究项目（17A096）资助.
** 通讯联系人.
Tel：0731-84613956，E-mail：zhmyuan@sina.com
收稿日期：2018-10-15，接受日期：2019-03-25

他方法的比较结果显示，本方法能获得更高的预测精度.

1　数据与方法

1.1　数据集

从HS^{3}D（homo sapiens splice sites data）数据集[15]中，抽取所有真实剪接位点序列（供体2 796条，受体2 880条）作为正样本，并随机抽取27 960/28 800条虚假供体/受体位点序列作为负样本. 从所有正样本中随机抽取1 957/2 016个供体/受体位点正样本用于训练，记作Tr-pos；余下的839/864个供体/受体位点正样本用于独立测试，记作Te-pos. 从所有负样本中随机无放回地抽取1 957/2 016个供体/受体位点负样本用于训练，重复10次，依次记作Tr-neg1、Tr-neg2、……、Tr-neg10，然后从剩余负样本中再随机抽取839/864个供体/受体位点负样本用于独立测试，记作Te-neg. 这样可得到10个正负样本均衡的子训练集，依次为Tr-pos：Tr-neg1、Tr-pos：Tr-neg2、……、Tr-pos：Tr-neg10，以及1个均衡独立测试集Te-pos：Te-neg. 所有样本序列的原始长度均为140 bp，且每个真实/虚假位点都满足“GT-AG”规则. 对供体位点序列，本文将其保守GT的位置设为00，上游区域位置分别标记为−1、−2、……、−70，下游区域位置则记为1、2、……、68；对受体位点序列，将保守AG的位置设为00，上游区域位置分别标记为−1、−2、……、−68，下游区域位置则标记为1、2、……、70.

1.2　窗口长度确定

基于训练集Tr-pos：Tr-neg1（序列长度为140 bp），对每个位置（除00位）构建一张2×4列联表，并计算对应的卡平方值. 卡平方值越高，说明该位置上的碱基在正负样本之间分布差异越大，则该位置越重要. 图1、2分别给出了供体、受体位点序列所有位置（除00位）对应的卡平方值. 可见，除受体−67、+22位外，其余位点的卡平方值均大于临界值$\chi^2(0.05,3)=7.81$，这表明除受体−67、+22位外，所有位置上的四种碱基分布在正负样本之间均差异显著. 进一步观察发现，供体位点序列−40、−3~+5、+7、+8、+10位的卡平方值高于其所有位置卡平方值的平均值，受体位点序列−21、−19~+1位的卡平方值高于其所有位置卡平方值的平均值. 考虑到窗口的连续性，我们最终确定供体位点序列的窗口长度为8 bp（即−3~+5位，不含00位），受体位点序列的窗口长度为20 bp（−19~+1位，不含00位）. 后文若无特别指出，则使用的供体/受体位点序列窗口长度均取8 bp/20 bp.

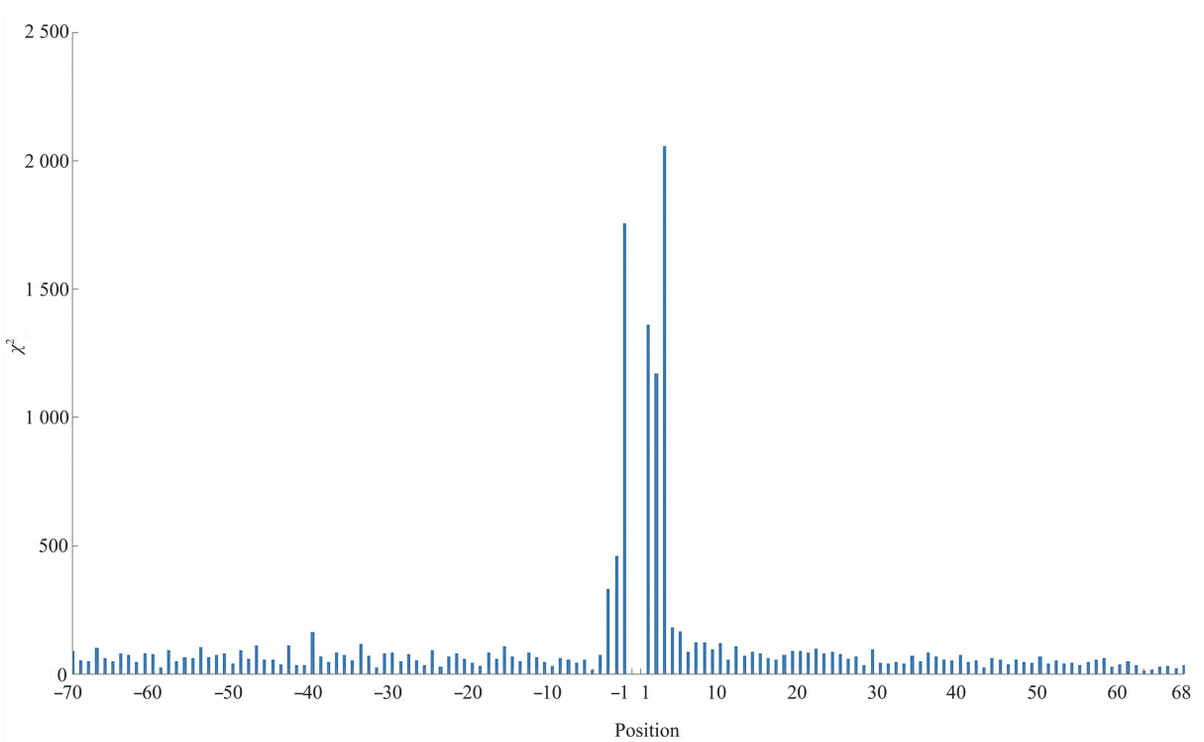

Fig. 1　Chi−square values for each position in donor splice site−containing sequences

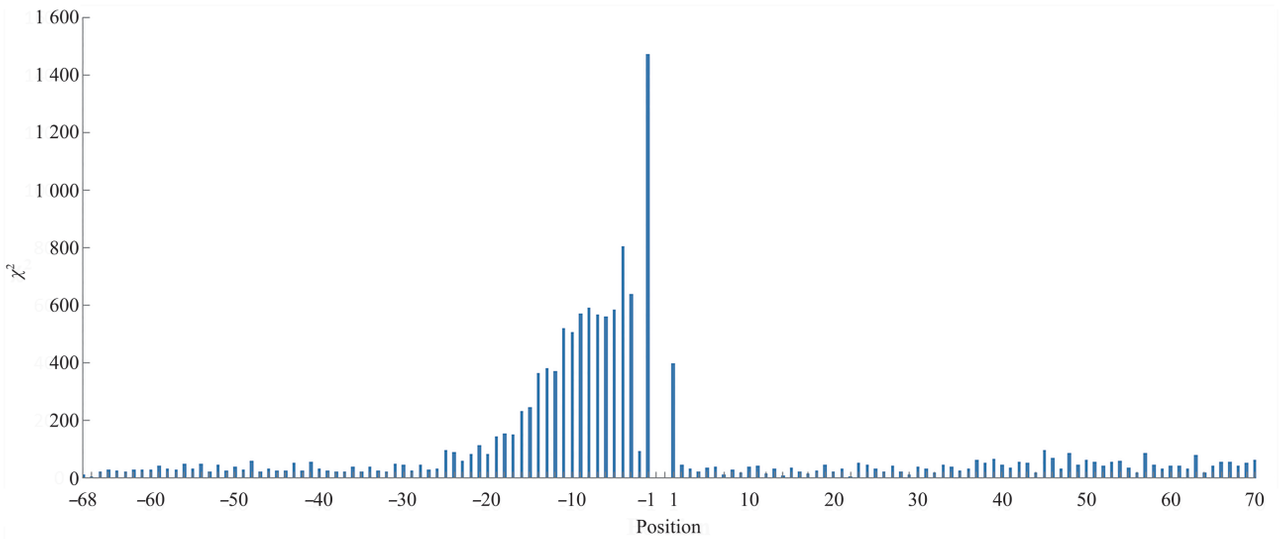

Fig. 2　Chi-square values for each position in acceptor splice site-containing sequences

1.3　特征提取

对每个8 bp/20 bp的供体/受体位点序列样本，我们提取了L（供体位点序列，L=8；受体位点序列，L=20）个位置特征，记作p_i（i=1，2，…，L）.具体过程描述如下：

在训练集中，分别统计四种碱基在第i个位置（i=1，2，…，L）正负样本中出现的频次，得到一张2×4列联表（表1）.

Table 1　Frequency distribution of four bases on the *i*th position

Sample	Base				Total
	A	T	C	G	
Positive	$f_{i,A}^{+}$	$f_{i,T}^{+}$	$f_{i,C}^{+}$	$f_{i,G}^{+}$	f_i^{+}
Negative	$f_{i,A}^{-}$	$f_{i,T}^{-}$	$f_{i,C}^{-}$	$f_{i,G}^{-}$	f_i^{-}
Total	$f_{i,A}$	$f_{i,T}$	$f_{i,C}$	$f_{i,G}$	N

表中，$f_{i,A}^{+}$、$f_{i,T}^{+}$、$f_{i,C}^{+}$、$f_{i,G}^{+}$分别表示碱基A、T、C和G在第i个位置正样本中出现的频次，$f_{i,A}^{-}$、$f_{i,T}^{-}$、$f_{i,C}^{-}$、$f_{i,G}^{-}$分别表示碱基A、T、C和G在第i个位置负样本中出现的频次，$f_{i,A}$、$f_{i,T}$、$f_{i,C}$、$f_{i,G}$分别表示碱基A、T、C和G在第i个位置所有样本中出现的频次，f_i^{+}、f_i^{-}分别为正、负样本的总数，N为所有样本的总数.位置i对应的卡平方值按下式计算：

$$\chi^2=\frac{N^2}{f_i^{+}\times f_i^{-}}\left[\sum_{j\in\{A,T,C,G\}}\frac{f_{i,j}^{+2}}{f_{i,j}}-\frac{f_i^{+2}}{N}\right] \quad (1)$$

若新增一个训练样本，其第i个位置为第j种碱基，先假设其为正样本，用$f_{i,j}^{+}+1$替换$f_{i,j}^{+}$，按式（1）算得一个卡平方值$\chi_{i,j}^{2+}$，再假设其为负样本，用$f_{i,j}^{-}+1$替换$f_{i,j}^{-}$，按式（1）算得一个卡平方值$\chi_{i,j}^{2-}$，则第i个位置为第j种碱基的卡方统计差表得分记为$\Delta\chi_{i,j}^{2}=\chi_{i,j}^{2+}-\chi_{i,j}^{2-}$，$j\in\{A,T,C,G\}$.由此，构建一张4×$L$（供体位点序列，$L$=8；受体位点序列，$L$=20）卡方统计差表，如表2所示.若序列样本的第$i$个位置出现第$j$种碱基，则其位置特征$p_i$（$i$=1，2，…，$L$）赋值为$\Delta\chi_{i,j}^{2}$.

Table 2　Chi-square statistical difference table

Base	Position				
	P_1	…	P_i	…	P_L
A	$\Delta\chi_{1,A}^{2}$	…	$\Delta\chi_{i,A}^{2}$	…	$\Delta\chi_{L,A}^{2}$
T	$\Delta\chi_{1,T}^{2}$	…	$\Delta\chi_{i,T}^{2}$	…	$\Delta\chi_{L,T}^{2}$
C	$\Delta\chi_{1,C}^{2}$	…	$\Delta\chi_{i,C}^{2}$	…	$\Delta\chi_{L,C}^{2}$
G	$\Delta\chi_{1,G}^{2}$	…	$\Delta\chi_{i,G}^{2}$	…	$\Delta\chi_{L,G}^{2}$

对每个8 bp/20 bp的供体/受体位点序列样本，我们还提取了16种碱基二联体在样本中的出现频次（记作f_{AA},f_{AT},f_{AC},f_{AG},f_{TA},f_{TT},f_{TC},f_{TG},f_{CA},f_{CT},f_{CC},f_{CG},f_{GA},f_{GT},f_{GC},f_{GG}），作为序列组分特征.以一个供体位点序列样本“TAA**GT**TCAAG”为例（不考虑00位上的GT），二联体AA在该样本中出现了2次，故$f_{AA}=2$，依此计算其他二联体的出现频次，最终得到f_{AA},f_{AT},f_{AC},f_{AG},f_{TA},f_{TT},f_{TC},f_{TG},f_{CA},f_{CT},f_{CC},f_{CG},f_{GA},f_{GT},f_{GC},f_{GG}的值依次为：2，0，0，1，1，0，1，0，1，0，0，0，0，0，0，0.

综上，对每个8 bp的供体位点序列样本，可用一个24维特征向量（8维位置特征+16维组分特征）来表征，对每个20 bp的受体位点序列样本，可用一个36维特征向量（20维位置特征+16维组分特征）来表征.

1.4　基于SVM和加权投票策略的分类决策

SVM是基于统计学习理论的一种机器学习方法.基于结构风险最小原则，SVM能够解决小样本、高维数、非线性、过拟合及局部最小等问题，且已成功应用于剪接位点预测.本文采用软件LIBSVM[16]实现SVM分类，其核函数固定为径向基核，参数c、g通过10折交叉测试搜索自动获取.

考虑到负样本（虚假剪接位点）数目远超过正样本（真实剪接位点），为有效解决不平衡模式分类问题，同时降低训练样本较大时SVM的时间复杂度，我们构建了10个均衡子训练集，即Tr-pos：Tr-neg1、Tr-pos：Tr-neg2、…、Tr-pos：Tr-neg10，构建方法详见本文第**1.1**节；并使用这10个子训练集分别建立10个SVM分类器.接下来，采用加权投票策略对1∶1独立测试集Te-pos：Te-neg进行分类决策，具体过程为：对第m个待测样本，设第k个SVM分类器判定其属于正类的概率为W_{mk}，则属于负类的概率为$1-W_{mk}$.若$\sum_{k=1}^{10}W_{mk}>\sum_{k=1}^{10}(1-W_{mk})$，则第$m$个样本判为正类，否则判为负类.这里，$W_{mk}$即为投票权重，由软件LIBSVM计算得到.

1.5　评价指标

采用敏感性（sensitivity，SN）、特异性（specificity，SP）、准确度（accuracy，ACC）、Matthew相关系数（MCC）来评估预测模型性能，这些指标定义如下：

$$SN=\frac{TP}{TP+FN} \tag{2}$$

$$SP=\frac{TN}{TN+FP} \tag{3}$$

$$ACC=\frac{TP+TN}{TP+FN+TN+FP} \tag{4}$$

$$MCC=\frac{TP\times TN-FN\times FP}{\sqrt{(TP+FN)\times(TP+FP)\times(TN+FP)\times(TN+FN)}} \tag{5}$$

其中，*TP*（ture positive）、*TN*（ture negative）、*FN*（false negative）、*FP*（false positive）分别表示正样本判对数、负样本判对数、正样本判错数、负样本判错数.

为了与参比算法进行比较，本文还使用了ROC曲线下的面积（area under ROC curve，AUC-ROC）及Q^9[17]作为综合评价指标，它们不受数据集类分布的影响，且已被广泛用于评价分类模型的性能.Q^9定义如下：

$$Q^9=\left(1+q^9\right)/2 \tag{6}$$

式中，$q^9=\begin{cases}(TN-FP)/(TN+FP), if\ TP+FN=0\\(TP-FN)/(TP+FN), if\ TN+FP=0\\1-\sqrt{2}\sqrt{\left[FN/(TP+FN)\right]^2+\left[FP/(TN+FP)\right]^2}, if\ TP+FN\neq 0\ and\ TN+FP\neq 0\end{cases}$

2　结　　果

2.1　加权投票结果

对独立测试集Te-pos：Te-neg，10个SVM分类器的加权投票结果见表3.为进行比较，我们还构建了一个1∶10非均衡训练集，即Tr-pos：（Tr-neg1+Tr-neg2+…+Tr-neg10），简记为Tr-pos：Tr-neg.比较得到：a.加权投票下的供体、受体位点独立预测ACC分别为93.39%、90.46%，较10个均衡子训练集的平均ACC（供体位点93.09%、受体位点89.88%）略有提高；b.基于1∶10非均衡训练集Tr-pos：Tr-neg的独立预测ACC（供体位点85.29%、受体位点83.99%）较加权投票明显下降，并且正样本识别率（供体SN=73.06%、受体SN=70.08%）远低于负样本识别率（供体SP=97.52%、受体SP=97.89%），这表明大量正样本被错判为负样本；c.基于多个均衡子训练集的加权投票策略能够有效解决不平衡模式分类问题.

Table 3　Independent test accuracy based on different training sets

Training set	Donor site				Acceptor site			
	SN	SP	ACC	MCC	SN	SP	ACC	MCC
Tr-pos：Tr-neg1	0.9511	0.8963	0.9237	0.8487	0.9132	0.8877	0.9005	0.8012
Tr-pos：Tr-neg2	0.9440	0.9118	0.9279	0.8562	0.9074	0.8981	0.9028	0.8056
Tr-pos：Tr-neg3	0.9440	0.9201	0.9321	0.8644	0.9190	0.8889	0.9040	0.8082
Tr-pos：Tr-neg4	0.9523	0.9190	0.9357	0.8718	0.9039	0.8819	0.8929	0.7861
Tr-pos：Tr-neg5	0.9523	0.9142	0.9333	0.8671	0.9144	0.8855	0.9000	0.8001
Tr-pos：Tr-neg6	0.9464	0.9166	0.9315	0.8633	0.9016	0.8935	0.8976	0.7952
Tr-pos：Tr-neg7	0.9440	0.9249	0.9345	0.8690	0.9097	0.8738	0.8918	0.7841
Tr-pos：Tr-neg8	0.9392	0.9154	0.9273	0.8548	0.9039	0.8785	0.8912	0.7827
Tr-pos：Tr-neg9	0.9428	0.9201	0.9315	0.8632	0.9201	0.8902	0.9052	0.8106
Tr-pos：Tr-neg10	0.9440	0.9190	0.9315	0.8632	0.9028	0.9005	0.9017	0.8032
Weighted voting	0.9499	0.9178	0.9339	0.8681	0.9144	0.8947	0.9046	0.8092
Tr-pos：Tr-neg	0.7306	0.9752	0.8529	0.7128	0.7008	0.9789	0.8399	0.7075

2.2　与其他算法的比较结果

在相同的HS^{3}D数据集中，分别与SVM-B[8]、MM1-SVM[9]、SAE[5]三种剪接位点识别算法比较（表4）. SVM-B和MM1-SVM是两种基于SVM分类器的经典剪接位点识别算法，它们使用的窗口长度为140 bp，其预测精度基于正负样本比例为1∶10（2 796/2 880个真实供体/受体位点，27 960/28 800个虚假供体/受体位点）的HS^{3}D数据集得到[18]. SAE是近年发展的一种基于短窗口（9 bp）的供体剪接位点识别新算法. 它通过构建碱基关联矩阵确定窗口长度，并定义绝对误差之和进行决策，虽然获得了较好的预测结果，但仅限于供体剪接位点预测. 表4给出的SAE算法预测精度是基于正负样本比例约为1∶5（2 796/15 000个真实/虚假供体位点）的HS^{3}D数据集得到[5]. 相同数据集中的比较结果表明，本算法的预测精度明显高于三种参比算法.

Table 4　Comparison with other algorithms

Algorithm	Donor site				Acceptor site		
	SN	*SP*	Q^9	*AUC*	*SN*	*SP*	Q^9
SVM-B	0.9406	0.9067	0.9212	--	0.9066	0.8797	0.8920
MM1-SVM	0.9256	0.9244	0.9247	--	0.8993	0.8869	0.8926
SAE	--	--	--	0.9450	--	--	--
The proposed	0.9499	0.9178	0.9319	0.9713	0.9144	0.8947	0.9040

3　讨　　论

3.1　基于卡方统计差表的位置特征的优点

在剪接位点预测中，常用的基于位置的序列表征方法有单碱基01编码、单碱基的统计特征多变量编码[19]. 在Tr-pos：Tr-neg1上，分别采用单碱基01编码、单碱基的统计特征多变量编码、卡方统计差表编码提取序列样本（供体8 bp/受体20 bp）的位置特征，然后采用5折交叉测试分别检验基于各种位置特征的预测精度（表5）. 结果表明，相较于单碱基01编码和单碱基的统计特征多变量编码，本文提出的卡方统计差表编码对应的位置特征维数最少，且预测精度最高. 采用单碱基01编码，不同位置上的同一碱基赋值相同，没有体现位置的差异性，而同一位置上的不同碱基编码没有体现碱基间差异程度. 例如，供体位点序列−1位上碱基A、C、G、T的含量分别为19.81%、5.69%、53.99%、20.51%，显然碱基A的含量与碱基T的含量相差很小，而碱基C的含量与碱基G的含量相差很大，若按01编码，则A-T与C-G间汉明距离都是2. 此外，单碱基01编码需用4维0/1特征表示每个位置，故产生的位置特征维数较高（4×L维，L为序列长度），且特征矩阵非常稀疏. 单碱基的统计特征多变量编码虽然考虑了碱基位置和含量的差异，但每个碱基仍需4个变量表示，特征维数也较高（4×L维，L为序列长度），特征矩阵同样非常稀

疏. 基于卡方统计差表编码的位置特征既能反映同一位置上不同碱基的差异，又能反映不同位置上同一碱基的差异，且具有特征维数少（L维，L为序列长度）、特征矩阵不稀疏等优点.

Table 5　5−fold cross accuracy based on the positional features by different coding

Coding	Donor site		Acceptor site	
	Positional feature dimension	ACC	Positional feature dimension	ACC
0/1 coding	32	0.9201	80	0.8798
Multivariate coding for statistical feature	32	0.9255	80	0.8941
Chi-square statistical difference table coding	8	0.9278	20	0.9001

3.2　补充碱基二联体频次的必要性

位置特征虽然能够有效表征序列，但对DNA序列中发生的碱基插入或缺失突变比较敏感，如下例所示：

突变序列是通过在原始序列的第4位上随机插入碱基A而产生的. 显然，突变序列的位置特征较原始序列发生了较大的变化，而各碱基二联体在序列中出现的频次改变较小. 因此，补充碱基二联体的出现频次，在一定程度上能够提高算法对于碱基插入或缺失突变的鲁棒性.

序列位置	1	2	3	4	5	6	7	8	9	10	11	12	13	14
原始序列	C	G	C	G	T	A	C	T	G	A	G	C	T	A
突变序列	C	G	C	**A**	G	T	A	C	T	G	A	G	C	T

3.3　采用8 bp/20 bp窗口长度的优势

在Tr-pos：Tr-neg1上，分别基于供体8 bp、受体20 bp窗口长度和供受体138 bp窗口长度提取位置特征和组分特征进行预测. 5折交叉测试结果（表6）显示，相较于原始138 bp窗口长度，8 bp/20 bp窗口长度下的预测精度更高，且总特征维数大幅减少. 这表明过长的窗口可能引入无关序列，进而降低预测准确率.

Table 6　5−fold cross accuracy based on different window sizes

Donor site			Acceptor site		
Window size（excluding GT at position 00）	Total feature dimension	ACC	Window size（excluding AG at position 00）	Total feature dimension	ACC
8 bp（−3~+5）	24	0.9331	20 bp（−19~+1）	36	0.9035
138 bp（−70~+68）	154	0.9297	138 bp（−68~+70）	154	0.9011

4　结　　论

剪接位点预测是基因识别的关键环节之一. 本文提出了一种基于卡方统计差表和SVM加权投票的剪接位点预测新方法. 实验结果表明：a. 在相同的HS^{3}D数据集上与其他算法相比，本算法能获得更高的预测精度；b. 提出的多个均衡子训练集加权投票策略能够有效解决不平衡模式分类问题；c. 基于卡方统计差表的位置特征具有维数少、特征矩阵不稀疏等优点，能有效表征DNA序列，在分子序列信号位点识别中具有良好的应用前景.

参　考　文　献

[1] Burset M, Seledtsov I A, Solovyev V V. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Research, 2000, **28**(21):4364-4375

[2] Degroeve S, Saeys Y, Baets B D, *et al*. SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics, 2005, **21**(8):1332-1338

[3] Li J L, Wang L F, Wang H Y, *et al*. High-accuracy splice site prediction based on sequence, component and position features. Genetics & Molecular Research, 2012, **11**(3):3432-3451

[4] 李琴, 张瑾, 骈聪等. 基于位置关联权重矩阵及序列组分的多样性增量识别剪接位点. 生物物理学报 2014, **30**(5): 391-400
Li Q, Zhang J, Pian C, *et al*. Acta Biophysica Sinica, 2014, **30**(5): 391-400

[5] Meher P, Sahu T, Rao A, *et al*. A statistical approach for 5' splice site prediction using short sequence motifs and without encoding sequence data. BMC Bioinformatics, 2014, **15**(1):1-14

[6] Zuo Y, Zhang P, Liu L, *et al*. Sequence-specific flexibility organization of splicing flanking sequence and prediction of splice sites in the human genome. Chromosome Research, 2014, **22**(3): 321-334

[7] Sun Y F, Fan X D, Li Y D. Identifying splicing sites in eukaryotic RNA: support vector machine approach. Computers in Biology &

Medicine, 2003, **33**(1):17-29

[8] Zhang Y, Chu C H, Chen Y, *et al*. Splice site prediction using support vector machines with a Bayes kernel. Expert Systems with Applications, 2006, **30**(1):73-81

[9] Baten A, Chang B, Halgamuge S K, *et al*. Splice site identification using probabilistic parameters and SVM classification. BMC Bioinformatics, 2006, 7(Suppl 5):S15

[10] Wei D, Zhang H L, Wei Y J. A novel splice site prediction method using support vector machine. Journal of Computational Information Systems, 2013, **20**(9): 8053-8060

[11] Garg D, Maji S. Hybrid approach using SVM and MM2 in splice site junction identification. Current Bioinformatics, 2014, **9**(1), doi:10.2174/1574893608999140109121721

[12] Goel N, Singh S, Aseri T C. An Improved method for splice site prediction in DNA sequences using support vector machines. Procedia Computer Science, 2015, **57**:358-367

[13] Nassa T, Singh S, Goel N. Splice site detection in DNA sequences using probabilistic neural network. International Journal of Computer Applications, 2014, **76**(4):1-4

[14] Meher P K, Sahu T K, Rao A R. Prediction of donor splice sites using random forest with a new sequence encoding approach. Biodata Mining, 2016, **9**:4

[15] Pollastro P, Rampone S. HS^3D, a dataset of homo sapiens splice regions, and its extraction procedure from a major public database. International Journal of Modern Physics C, 2002, **13**(8):1105-1117

[16] Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2007, **2**(3): 27

[17] Zhang C T, Zhang R. Evaluation of gene-finding algorithms by a content-balancing accuracy index. Journal of Biomolecular Structure & Dynamics, 2002, **19**(6):1045-1052

[18] Zhang Q, Peng Q, Zhang Q, *et al*. Splice sites prediction of Human genome using length-variable Markov model and feature selection. Expert Systems with Applications, 2010, **37**(4): 2771-2782

[19] 黄金艳, 李通化, 陈开 . 基于知识编码的剪切位点预测 . 同济大学学报(自然科学版), 2007, **35**(11):1548-1551
Huang J Y, Li T H, Chen K. Journal of Tongji University (natural science), 2007, **35**(11):1548-1551

High-accuracy Splice Site Prediction Based on Statistical Difference Table and Weighted Voting*

ZENG Ying[1,2], CHEN Yuan[1], YUAN Zhe-Ming[1]**

([1]*Hunan Engineering & Technology Research Center for Agricultural Big Data Analysis & Decision-making, Hunan Agricultural University, Changsha* 410128, China;

[2]*Orient Science &Technology College, Hunan Agricultural University, Changsha* 410128, China)

Abstract　High-accuracy splice site recognition based on machine learning is the key to eukaryotic genome annotation. In this paper, we used chi-square test to determine the window size of sequences, and constructed a chi-square statistical difference table to extract the positional features, and combined with the frequencies of dinucleotides to characterize sequences. For the problem that the positive and negative samples of splice sites are extremely imbalanced, 10 SVM classifiers based on the equal proportion of positive and negative samples were built for weighted voting, which effectively solved the imbalanced pattern classification problem. Independent testing results in HS^3D dataset showed that the prediction accuracy of donor and acceptor sites were 93.39% and 90.46% respectively, obviously higher than that of the compared methods. The positional features based on the chi-square statistical difference table can effectively characterize DNA sequences, and have application prospects in signal site recognition of molecular sequences.

Key words　splice site, positional features, chi-square statistical difference table, weighted voting, support vector machine (SVM)

DOI：10.16476/j.pibb.2018.0267

* This work was supported by grants from The National Natural Science Foundation of China（61701177）, Hunan Provincial Natural Science Foundation of China（2018JJ3225）and Scientific Research Project of Hunan Province Education Office（17A096）.

** Corresponding author.

Tel：+86-731-84613956，E-mail：zhmyuan@sina.com

Received：October 15，2018　Accepted：March 25，2019