第 35 卷 第 5 期 计 算 机 工 程 2009 年 3 月
Vol.35 No.5 Computer Engineering March 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)05—0180—03 文献标识码：A 中图分类号：TP18

# 基于序列模式特征和 SVM 的剪切位点预测

孙贺全，彭勤科，张全伟
(西安交通大学电信学院机械制造系统工程国家重点实验室，西安 710049)

**摘 要：**通过对 HS3D 数据集供点序列碱基的统计分析，利用供体位点邻域碱基出现规律构造模式(motif)作为 DNA 序列的属性。设置序列属性值将字符序列映射成数字向量，应用支撑向量机进行实验，实现对供体位点的预测分类。实验结果表明，与改进的 motif 得分模型方法相比，该文方法可有效去除数据中异常数据对分类的影响，将 DNA 字符序列变换到 motif 属性数字序列空间具有有效性和实用性。
**关键词：**序列模式；剪切位点；支撑向量机

# Splice Site Prediction Based on Characteristics of Sequence Motif and Support Vector Machine

**SUN He-quan, PENG Qin-ke, ZHANG Quan-wei**
(State Key Laboratory for Manufacturing Systems Engineering, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049)

**【Abstract】**Through statistic analysis on the donor site sequences in the dataset of HS3D, the rules that the bases appear in the adjacent sites around the splice sites are used for constructing motifs, which are then utilized as the attributes of the DNA sequences. And by setting the value of each attribute the literal sequences are transformed into numeric vectors, based on which a Support Vector Machine(SVM) model is constructed to predict splice sites. The experimental results indicate that compared with the improved motif scoring model, the proposed method has diminished the influence on the prediction generated by the abnormal data effectively and also shows that the new mapping method in virtue of motifs is practicable and effectual.
**【Key words】**sequence motif; splice site; Support Vector Machine(SVM)

## 1 概述

序列(motif)是指 DNA 或者蛋白质序列中保守的序列片断[1]。DNA 序列中的某些区域，如功能位点区域，由于对生物体的生存具有至关重要的意义，可能在进化的过程中更保守一些，并表现为序列中的模式[2]。

文献[3]提出从序列位点附近抽取序列，通过训练构造得分模型，进而选取得分阈值对位点序列进行分类的方法。该方法在应用于 HS3D(Homo Sapiens Splice Site Dataset)数据时受异常点的限制，不具备代表性的样本会严重影响模型的产生。

本文利用 HS3D 序列位点周围出现频繁的碱基构造 motif，通过建立学习集获得与构造的 motif 相似度在 80%以下的 motif 来共同数字化原序列。由数字化向量代表原序列输入支撑向量机进行位点预测实验，取得较好的分类效果。

## 2 本文方法

### 2.1 基本定义

(1) $\Sigma$ 字母表：所有位置上可能出现的字符的集合，对本文：$\Sigma$={A, C, G, T}；

(2) $S_k$ 序列：$S_k: x_1x_2\cdots x_{L_k}$，其中 $x_i \in \Sigma$，$i$=1, 2,…, $L_k$, $k$=1, 2,…, $n$; $|S_k|=L_k$ 代表序列 $S_k$ 的长度；$S_k(i)$代表序列 $S_k$ 第 $i$ 位上的字母，$1 \leqslant i \leqslant |S_k|$；

(3) $\Sigma^L$：长度为 $L$ 的序列 $S_k$ 组成的集合，$k$=1, 2,…, $n$，即 $\Sigma^L=\{S_k|k=1, 2, \cdots, n\}$；

(4) $S_k(i \rightarrow j)$：序列 $S_k$ 的子串，由 $S_k$ 中相继的字母所组成的序列，$i$ 和 $j$ 表示子串始末位，$1 \leqslant i<j \leqslant |S_k|$；

(5) $D(S_k, S_1)$，等长序列 $S_k$ 与 $S_1$ 的海明距离：$D(S_k, S_1)=\sum_{j=1}^{|S_k|} p(j)$，其中，$p(j)=\begin{cases}1 & \text{若 } S_k(j)\neq S_1(j)\\ 0 & \text{否则}\end{cases}$；

(6) $Sim$=1-$D(S_k, S_1)/|S_k|$：等长序列 $S_k$ 与 $S_1$ 的相似度。

### 2.2 模式抽取算法

**定义** (motif) 在给定的 $\Sigma^L=\{S_k|k=1, 2,\cdots, n\}$ 中，对序列 $S_k$，其长度为 $|S_k|$，$1 \leqslant k \leqslant n$，$n$ 为样本数目，寻找长度为 $L'$ 的子串，若满足：

(1)至少在 $K$ 个样本中出现；(2)每次出现最多有 $D$ 个不匹配碱基，则称其为序列的 ($L'$, $D$)-$K$ 模式。

本文的学习型 motif，取 $L'$=5，$D$=$K$=1，而对构造型 motif，保留全部不重复的构造结果。

正例序列位点邻位上碱基统计如图 1，分析：68→75 位

**基金项目：**国家自然科学基金资助项目(60774086, 60373107)
**作者简介：**孙贺全(1982－)，男，硕士研究生，主研方向：生物信息序列建模分析；彭勤科，教授、博士生导师；张全伟，博士研究生
**收稿日期：**2008-06-12 **E-mail：**sunhq2000@stu.xjtu.edu.cn

上碱基[A/C]-A-G-GT-[A/G]-A-G 依次出现概率较大(大于 0.3，[A/C]表示该位可取 2 种碱基 A 或 C)，而反例分析无明显类似特征。统计出的碱基用于产生构造型 motif。

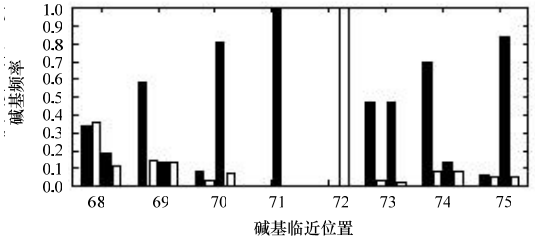


**图 1　正例位点邻域碱基频率(每位上依次为 ACGT)**

对给定训练集，模式抽取算法为：(1)构造包含 68→75 位上碱基的子串，分别是 $S1$=AAGGTAAG, $S2$=AAGGTGAG, $S3$=CAGGTAAG, $S4$=CAGGTGAG，重新设定其始末位标号为 1→8；

(2)用 5 位长滑动窗抽取(1)中 motif 的子串：$S_{i1}(1\to 5)$，$S_{i2}(2\to 6)$，$S_{i3}(3\to 7)$，$S_{i4}(4\to 8)$，$i$=1, 2, 3, 4 存放入 $Temp1^5$；

(3)将 $Temp1^5$ 中具有相同始末位的 motif 两两比对：若 $Sim$=1，保留 1 个于集合 $M1^5$ 中；否则全部保留，最终得到 8 个构造型 motif；

(4)取训练集中 $m$=200 个正例组成学习集，用 8 位长窗抽取其中所有序列的子串 $stu_j(68\to 75)$，$j$=1, 2, ⋯, $m$；用 5 位长滑动窗继续抽取子串用以提取学习型 motif，即得到 $stu_j(1\to 5)$，$stu_j(2\to 6)$，$stu_j(3\to 7)$，$stu_j(4\to 8)$，存放入 $Temp2^5$；

(5)将 $Temp2^5$ 中具有相同始末位的子串两两比对：若 $Sim\geqslant 80\%$，仅保留任一个于 $Temp2^5$；否则全部保留，直到该集合中任意 2 个子串的 $Sim$<80%；

(6)将子集 $Temp2^5$ 中子串与 $M1^5$ 中具有相同始末位的 motif 两两比对，若 $Sim\geqslant 80\%$，则从 $Temp2^5$ 中删除该子串：否则保留，最终得到 27 个学习型序列；

(7)将 $Temp2^5$ 与 $M1^5$ 合并，构成序列集 $M^5$，见图 2。

| 学习型 motif：始末位 | | | |
|---|---|---|---|
| 1→5 | 2→6 | 3→7 | 4→8 |
| 01.TATGT<br>02.TGCGT<br>03.CCTGT<br>04.GTTGT<br>05.TCAGT<br>06.CGAGT<br>07.AGTGT | 10.GCGTT<br>11.GAGTG<br>12.CTGTT<br>13.TCGTA<br>14.AAGTC<br>15.GGGTC | 18.TGTAC<br>19.CGTTG<br>20.AGTAG<br>21.AGTGT<br>22.CGTCA<br>23.AGTTC<br>24.CGTGC<br>25.GGTCC | 28.GTACT<br>29.GTTGA<br>30.GTTCC<br>31.GTCCA<br>32.GTCTC<br>33.GTAGC |
| 构造型 motif：始末位 | | | |
| 1→5 | 2→6 | 3→7 | 4→8 |
| 08.CAGGT<br>09.AAGGT | 16.AGGTA<br>17.AGGTG | 26.GGTAA<br>27.GGTGA | 34.GTAAG<br>35.GTGAG |
| Motif 总数＝9+8+10+8 | | | |

**图 2　序列集合 $M^5$**

此外，正例序列下游子串(73 位→140 位)碱基 G 出现概率高于 0.3，而反例序列无此特征，如图 3 和图 4。因此，经测试后再引入子串 GGG 属性。

将 GGG 属性与 motif 属性共同构成序列的属性向量，用于将每条序列表达成一个 37 维的向量:[ATGT, 2. TGCGT, ⋯, 35.GTGAG,GGG,LASS]，其中第 37 属性为类别标识属性 CLASS。

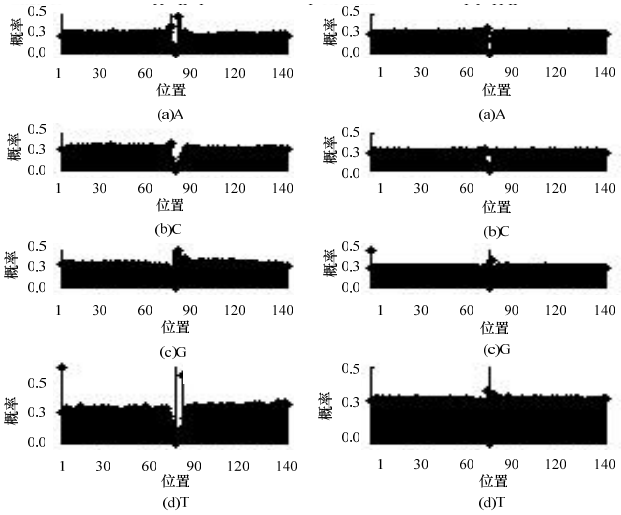


**图 3　正例序列碱基频率**




**图 4　反例序列碱基频率**

## 3　基于支撑向量机的位点预测算法

支撑向量机使数据集 $T$*中的 2 类数据通过特征空间的超平面得到线性划分：对一个未知序列 $z$，由下面的决策函数来预测为+1(正例)或-1(反例)：$pred(z)=t(wz+b)$，

其中，$t(x)=\begin{cases}+1, x\geqslant Td\\ -1, \text{else}\end{cases}$，$Td$ 是一个阈值。

本文利用 Lagrange 优化方法求解最优分类面，即在约束条件

$$\begin{cases}\sum_{i=1}^{n} y_i a_i = 0\\ c\geqslant a_i \geqslant 0,\quad i=1, 2, \cdots, n\end{cases}$$

对 $a_i$ 求解如下函数：

$W(a)=\sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j K(x_i, x_j)$ 的最大值，$a_i$ 为与每个样本对应的 Lagrange 乘子，其解中将只有一部分 $a_i$ 不为 0，对应的向量就是支撑向量，$K(x_i, x_j)$为选取的核函数。

测试后，本文采用多项式核函数：$K(x_i, x_j)=(\gamma x_i^T x_j + r)^d$，$\gamma$>0，其中，$d$ 为阶数，选取参数 $c\in\{1, 10, 20, \cdots, 500\}$，$c$ 为在实验过程中用来调整噪声对模型的影响惩罚因子。预测算法过程如下：

(1)8 位长窗口抽取 140 位长序列 $S_i$ 的子串 $S_i(68\to 75)$；

(2) 5 位长窗口抽取子串 $S_i(68\to 75)$ 的属性，每个子串得到 4 个属性子串；

(3)按对应始末位关系，分别计算(2)中属性子串与 $M^5$ 中 motif 的 $Sim$；若 $Sim\geqslant 80\%$，设置 $S_i$ 的 $j$.motif 属性值为 1，$j$=1,2,⋯,35，否则为 0，直到这一序列的所有 motif 属性值全部设置完毕，并最终组合成向量 $m_{i1\times 35}$；

(4)将子串 $S_i$ 所在序列下游，即 73 位→140 位上的子串 GGG 的出现次数进行统计：若该次数大于 2，则设置该序列 GGG 属性 $g_i$ 为 1，否则为 0；

(5)设置原序列的类别属性 $c_i\in\{1, -1\}$；

(6)组合序列数字化向量 $d_{i1\times37}=(m_{i1\times35}, g_i, c_i)$：

$Sam\_set^*=\{d_{i1\times37} \mid i=1, 2, \cdots, Seq\_Num\}$ ， $i=1, 2, \cdots, Seq\_Num$ 为样本序数；

(7)将 $Sam\_set^*$ 划分，组成为训练集、测试集：

$Train\_Set^*=\{dtr_x \mid x=1, 2, \cdots, Train\_Num\}$；

$Test\_Set^*=\{dte_x \mid x=1, 2, \cdots, Test\_Num\}$；

其中，$Train\_Num$ 和 $Test\_Num$ 分别为用于训练测试的样本数；

(8)用 $Train\_Set^*$ 输入支撑向量机进行训练，得到模型；

(9)用 $Test\_Set^*$ 对模型进行测试，分析结果。

## 4 实验结果分析

本文数据选自 HS3D 供点数据集[4]：每条序列长度 $|S_i|=140$，供体位点保守碱基为 $S_i(71 \rightarrow 72)$=GT，$1 \leqslant i \leqslant n$。

首先从数据集中任选正例 1 796 个，反例数∈ {2 000, 2 500, 3 000}，组合训练集 $Train\_set^*$ 建立模型；再取正例 1 000 个，反例 1 000 个，形成测试集 $Test\_Set^*$ 对各个模型进行测试。定义性能指标[5]：

正例测准率： $TP\_Rate=TP/(TP+FN)$

反例测准率： $TN\_rate=TN/(TN+FP)$

综合评价指标： $Q^9=(1+q^9)/2$

其中， $q^9=1-\sqrt{2\times((FN/(TP+FN))^2+(FP/(TN+FP))^2)}$。

经测试，本文得到评价指标如表 1 所示。其中，参数 2000-c050 表示，选取 2 000 个反例与 1 796 个正例混合构成训练集，训练时设定的核函数参数 $c$=50，其余类似。

**表 1 本文模型对 HS3D 数据的测试结果**

| 参数 | 样本个数 | *TP_Rate* | *TN_Rate* |
|---|---|---|---|
| 2000-c050 | 1 000 个正例<br>1 000 个反例 | 0.933 | 0.891 |
| 2500-c120 | | 0.931 | 0.901 |
| 2500-c250 | | 0.930 | 0.905 |
| 2500-c150 | | 0.928 | 0.903 |
| 2500-c100 | | 0.922 | 0.904 |
| 3000-c250 | | 0.916 | 0.911 |
| 3000-c100 | | 0.910 | 0.917 |

文献[6]对 MaishengYin 的 motif 方法进行改进，应用新的得分方法在一定程度上避免了异常数据对结果的影响，其建立的模型对 HS3D 数据集供点序列 1 000 个正例和 1 000 个反例测试结果如表 2 所示。

**表 2 改进后 Motif 模型对 HS3D 数据的测试结果**

| 类别 | 样本数 | 识别正确样本数 | 识别错误样本数 | *TP_rate/TN_rate* |
|---|---|---|---|---|
| 正例 | 1 000 | 831(*TP*) | 169(*FN*) | 0.831 |
| 反例 | 1 000 | 922(*TN*) | 78(*FP*) | 0.922 |

选取供体位点上下游不同长度($L_2/L_1$)的序列，正交编码后，应用神经网络 BP 算法建立模型，对 HS3D 数据集 500 个正例和 1 000 个反例进行测试，结果如表 3[6]($Q^9$是对原文的附加指标)。其中，50/50 表示选取的序列为位点左侧有 50 个字符，右侧有 50 个字符，其余类似。

本文从数据集中任意选取正例 2 295 个，反例数∈{2 000, 3 000}，组合训练集 $Train\_set^*$ 建立模型；取正例 500 个，反例 1 000 个，形成测试集 $Test\_Set^*$ 对各个模型进行测试，结果如表 4 所示。其参数表示含义同表 1。

表 1 与表 2 的指标值比较结果表明，模型对反例的识别率下降了 1%到 3%，但是对正例的识别率提高了 8%到 10%；表 3 与表 4 的比较结果表明，反例识别率的下降换取了几乎等量的正例识别率的提高，同时综合指标略微提高；以上均说明应用 motif 属性将 DNA 序列映射到高维数字空间进行分析的方法是可行的；另外，由于 motif 主要是在位点邻域取得的，结果也说明了剪切位点的信息主要集中在位点邻域之内。

**表 3 神经网络模型对 HS3D 数据的测试结果**

| 参数 | *TP_rate* | *TN_rate* | $Q^9$ |
|---|---|---|---|
| 5/30 | 0.892 | 0.935 | 0.906 0 |
| 10/30 | 0.896 | 0.935 | 0.906 6 |
| 20/30 | 0.886 | 0.939 | 0.909 8 |
| 30/30 | 0.890 | 0.936 | 0.906 9 |
| 30/20 | 0.886 | 0.928 | 0.897 0 |
| 30/10 | 0.876 | 0.937 | 0.905 8 |
| 30/5 | 0.898 | 0.930 | 0.901 2 |
| 50/50 | 0.893 | 0.950 | 0.924 0 |

**表 4 本文模型对 HS3D 数据的测试结果**

| 参数 | *TP_Rate* | *TN_Rate* | $Q^9$ |
|---|---|---|---|
| 2000-c250 | 0.924 | 0.900 | 0.911 2 |
| 3000-c060 | 0.914 | 0.903 | 0.908 3 |
| 3000-c110 | 0.920 | 0.903 | 0.911 1 |
| 3000-c140 | 0.920 | 0.902 | 0.910 6 |
| 3000-c200 | 0.922 | 0.903 | 0.912 0 |

## 5 结束语

本文方法与改进后的 Maisheng Yin 的 Motif 方法相比，应用于 HS3D 数据集时，通过对 motif 属性映射得到的数字序列应用支撑向量机，避免了得分模型阈值选取的问题，去除了噪声数据的影响；在保证反例预测效果变化不大的前提下使正例的预测效果由 83.1%提高到 91%以上，效果显著；与神经网络方法相比，反例识别率($TN\_rate$ =1-$FP\_Rate$)略有下降：由 93%左右到 90%左右；但正例识别率略有提升：由 89%左右到 91%左右，同时总的评价指标略有提升。

## 参考文献


[1] Yang Zhengrong. Decision Trees: a Novel Method for Decisive Template Selection-mining SARS-CoV Protease Cleavage Data Using Non-orthogonal[J]. Bioinformatics, 2005, 21: 2644-2650.

[2] 李冬冬, 王正志, 杜耀华, 等. DNA 序列中模式发现的一种快速算法[J]. 生物物理学报, 2005, 21(2): 121-129.

[3] Maisheng Y, Jason T L W. Algorithms for Splicing Junction Donor Recognition in Genomic DNA Sequences[C]//Proc. of IEEE International Joint Symposia on Intelligence and Systems. [S. l.]: IEEE Computer Society, 1998: 169-176.

[4] Salvatore R. HS3D, a Dataset of Homo Sapiens Splice Regions, and Its Extraction Procedure from a Major Public Database[J]. International Journal of Modern Physics C, 2002, 13(8): 1105-1117.

[5] Wren J D, William H H, Chandrasekaran S, et al. Markov Model Recognition and Classification of DNA/Protein Sequence Within LargeText Databases[J]. Bionformatics, 2005, 21(21): 4046- 4053.

[6] 雷 静, 阮晓刚. DNA 序列与剪接位点的关联性分析[J]. 北京工业大学学报, 2004, 30(3): 295-298.


编辑 金胡考