PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2009, 36(7): 863~871
www.pibb.ac.cn

基于序列和结构特征分析植物 TATA 和 TATA-less 启动子 *

左永春　李前忠 **
(内蒙古大学物理科学与技术学院，呼和浩特 010021)

摘要　分析启动子区域内调控元件是阐明基因转录起始机制的重要前提．利用从 PlanPromDB 数据库下载的植物 Pol-Ⅱ TATA 和 TATA-less 启动子数据，深入分析了两类启动子 GC 偏好、位点结构保守性、序列碱基组分、保守模体分布、TATA box 位点分布及关联位点保守性等特点，统计出两类植物启动子许多特有的序列组分和结构规律，这些规律对进一步揭示植物 Pol-Ⅱ启动子的转录调控机制有一定的帮助．通过构建能够同时考虑位点保守性和关联性的位点关联性权重矩阵扫描模型(PCWM)，利用相应打分函数(Score)对两类启动子进行区分，得到了较好结果，说明 PCWM 的预测性能要优于单碱基的位点权重矩阵(PWM).

关键词　植物 Pol-Ⅱ启动子，序列组成偏好特征，TATA 和 TATA-less 启动子，保守模体，位点关联权重矩阵(PCWM)
学科分类号　Q61　　**DOI:** 10.3724/SP.J.1206.2008.00727

基因转录起始是基因表达调控中最重要的一个环节，真核生物基因的精确转录是在一段被称为启动子的 DNA 序列上完成的．启动子承载着基因表达最核心的调控信息，它决定着基因的转录起始、转录频率及何时何地如何表达．因此，深入研究启动子的结构和功能是正确揭示基因转录调控机制和表达模式的关键，已经成为现代分子生物学研究的热点．认识启动子的传统方法是通过生物学实验方法来测定的，如基因报告检测、染色质免疫沉淀、Oligo-capping、CAGE、SAGE 等[1]，这些方法对认识启动子序列特征及它们的作用机制起到很大帮助，然而，这些实验的局限性很强，且费用较高．随着大规模基因组测序的完成，利用生物信息学方法分析启动子特征正逐渐走向成熟，与实验确定相比，利用生物信息学构建的软件可以针对大规模的数据进行统计研究，且具有成本低，时间短，结果可靠等优点，大大提高了对启动子研究的效率和速度.

真核生物有 3 种 RNA 聚合酶(RNAp)，每一种都有自己特定的启动子类型．Pol-Ⅰ主要负责转录 rRNA，Pol-Ⅲ主要负责转录 tDNA 和 5SrDNA，其启动子位于转录的 DNA 序列之内，称为下游启动子．Pol-Ⅱ负责编码蛋白质基因和部分 snRNA 基因的转录，其启动子类型最多，结构也最为复杂，是启动子中最重要且研究最多的一类[2]．随着这一类启动子数据的增加和研究的深入，Pol-Ⅱ启动子通常可分为含有 TATA box 的启动子(TATA promoter)和不含有 TATA box 的启动子(TATA-less promoter)两大类.

植物启动子是真核启动子中非常重要的一类[3]. 随着大量植物基因 mRNA 测序的完成，植物启动子的结构和特征已成为分子生物学家研究的热点之一．然而与哺乳动物(人类和小鼠)相比，植物转录起始区域调控方式的相关研究报告非常少．缺乏实验证实的启动子数据是制约植物启动子研究的原因之一．本文利用 PlantPromDB 数据库上给出的经实验证实的 Pol-Ⅱ类型的 TATA 启动子和 TATA-less 启动子数据[4]，通过对这两类启动子序列特征、模体特征及结构特征的统计分析，发现这

* 国家自然科学基金资助项目(30560039).
** 通讯联系人.
Tel: 0471-4992958, E-mail: qzli@imu.edu.cn
收稿日期：2008-10-23，接受日期：2008-11-25

两类植物启动子在序列组成及结构特点上均存在很大差异．另外，本文改进了原有的位点权重矩阵(PWM)，构建出位点关联性权重矩阵(PCWM)扫描算法对两类启动子进行区分预测，取得了较好的分类效果，说明转录起始区的调控元件大都存在很强的位点关联性．

1 数据及方法

1.1 数据集

本文研究的植物 Pol-Ⅱ启动子数据从 PlantPromDB 中获取[4]，该数据收集的植物 Pol-Ⅱ启动子都是经实验证实的，非冗余的，且有相应的转录起始位点(TSS)和翻译起始位点(TIS)信息．数据库共收集 305 条植物 Pol-Ⅱ启动子序列，其中 TATA 类型启动子序列 175 条，TATA-less 类型启动子 130 条．每条序列长度均为 251 bp，包含转录起始位点上游 200 bp 和下游 50 bp，其中 TSS 为 0 位点．另外，该数据库对每条启动子序列基因编号、所属物种、编码的蛋白质基因名称都进行了注释．

1.2 GC/AT-Skew

大多哺乳动物启动子在转录起始区域上游都存在一个 CpG 岛[5]，人类基因组中长度为 100～1 000 bp 且富含 CpG 二核苷酸的 CpG 岛总是处于未甲基化状态，并且与 56%的人类基因组编码基因相关，因此 CpG 岛往往是识别哺乳动物启动子的一个重要特征[6]．相反，在相应的植物近端启动子区域并未发现明显的 CpG 岛特征，而是存在很强的 GC-Skew，即碱基 C 的含量高于碱基 G 的含量[7]．基因表达水平越强，GC-Skew 越显著．统计研究表明，哺乳动物(人类)和果蝇基因启动子区域并未发现 GC-Skew 特征，这种碱基组分差异性特征应该是植物启动子特有的[8]．我们对 175 条 TATA 和 130 条 TATA-less 启动子位点 GC/AT-Skew 差异性进行对比统计．

1.3 结构特征分析

基因的转录起始主要是依靠 Protein-DNA 相互作用，即转录前复合物(PIC)识别 DNA 序列上的核心启动子元件进行转录起始．在转录起始中启动子区域内的 DNA 序列往往需要形成一定的局部空间结构，并在多种转录因子的辅助之下，才能被 RNA 聚合酶准确识别并与核心启动元件相结合，这使得启动子序列与其他序列相比具有更高的局部弯曲度和更低的双链稳定性[9]．目前，已有一些研究利用启动子区域的结构特征进行启动子识别的尝试，得到了较好的识别效果[10~12]，说明启动子区域具有某些特有的结构特征．最近 Goñi 等[13]运用分子动力学模拟方法，根据基因结构和同源保守性特征对人类基因启动子序列的 6 类结构特征进行统计预测，结果证明，在调节基因组表达上存在一套隐藏的物理编码，本文通过计算 TATA 启动子和 TATA-less 启动子相邻位点之间的 Twist、Tilt、Roll、Shift、Slide、Rise 等 6 类结构特征[13]，以便从物理结构特征上分析 TATA 启动子和 TATA-less 启动子各自的结构特征及两者在结构上差异性．这 6 种结构特征分别包括 3 种角度参数(Twist、Tilt、Roll)和 3 种距离参数(Shift、Slide、Rise)：Tilt、Roll、Twist 分别反映相邻碱基空间平面上下、前后、左右的夹角变化情况，Rise、Slide、Shift 分别反映相邻碱基空间相对位置上下、前后、左右的距离变化情况[14]，这 6 种参数通过定量地描述 DNA 序列在空间结构上的变化情况，使我们能够更深入地研究转录起始附近 DNA 序列的局部构象差异性．在 DNA 序列结构特征的研究当中，碱基最紧邻(二联体)特点研究的最为深入．由于受到碱基互补配对原则的约束，根据 Yanagi 等[15]的计算原则，碱基二联体独立的可能组合数目共有 10 类，它们分别是：AA(=TT)，AC(=GT)，AG(=CT)，AT，CA(=TG)，CC(=GG)，CG，GA(=TC)，GC，TA．6 类结构参数具体数值选取如表 1 所示[13]．

Table 1 The six physical feature parameters for different dinucleotides

	AA	AC	AG	AT	CA	CC	CG	GA	GC	TA
Twist	0.026	0.036	0.031	0.033	0.016	0.026	0.014	0.025	0.025	0.017
Tilt	0.038	0.038	0.037	0.036	0.025	0.042	0.026	0.038	0.036	0.018
Roll	0.02	0.023	0.019	0.022	0.017	0.019	0.016	0.02	0.026	0.016
Shift	1.69	1.32	1.46	1.03	1.07	1.43	1.08	1.32	1.20	0.72
Slide	2.26	3.03	2.03	3.83	1.78	1.65	2.00	1.93	2.61	1.20
Rise	7.65	8.93	7.08	9.07	6.38	8.04	6.23	8.56	9.53	6.23

1.4　转录因子结合区域模体差异比较

转录因子是转录起始过程中 RNA 聚合酶结合 DNA 序列时所需的辅助因子，转录因子结合位点(TFBS)是基因转录起始时转录因子识别的 DNA 区域[16]. 大部分转录因子结合位点都分布在启动子区域内，不同类型的启动子往往具有不同的转录因子结合位点，因此研究启动子序列内的转录因子结合位点的特征也是认识启动子类型及基因特点的一个十分重要的方面. 本文中我们利用 MEME 软件[17]搜索 TATA 启动子和 TATA-less 启动子区域内的保守转录因子结合位点，并利用 Webolog 软件对筛选出的转录因子结合位点进行保守性分析，进而更深刻地认识 TATA 启动子和 TATA-less 启动子在结合不同转录因子的特征差异.

1.5　*k*-mer 位点关联保守性分析

对于样本集的每个位点进行保守性分析，结合以前的工作及文献上的资料，我们定义 k-mer联体在具体位点的保守性 $M_k(l)$如下[18]：

$$M_k(l) = \sum_k \frac{(P(i,\ l)-1/4^k)^2}{1/4^k} \tag{1}$$

其中 k 表示联体碱基的长度，$P(i,\ l)$ 是第 i 个 k 联体在第 l 个位点出现的概率. 上式表示 $M_k(l)$随位点 l 的变化关系，$M_k(l)$的取值范围为$[0, (4^k-1)^2/4^k]$，它的意义可解释为样本集长度为 k 联体碱基片段在第 l 位点的保守程度. 可以看出，如果该位点保守性越强，则 $M_k(l)$的值越大，因此，我们可以利用 $M_k(l)$对某样本集进行位点关联保守性分析.

1.6　位点关联权重矩阵扫描算法(PCWM)

根据 Stormo[19]对位点权重矩阵(PWM)的阐述，位点权重矩阵是描述 DNA 序列上转录因子结合位点碱基分布的常用模型，最近 Li 和 Lin[20]构建出位点关联打分矩阵对大肠杆菌启动子进行了预测，得到很好的预测效果. 本文中我们依据文献[20]定义位点关联频率矩阵如下：

$$P_k(i,\ l) = \frac{f_{i,\,l} + s_i}{N_l + \sum_{k_l} s_i} \tag{2}$$

其中 $P_k(i,\ l)$是第 i 个 k 碱基联体在位点 l 出现的频率，$f_{i,\,l}$ 表示在 N_l 条序列中第 i 个 k 联体在位点 l 出现的次数，N_l 为 l 位点的序列总条数；S_i 为伪计数，为方便计算，我们取 $S_i = \sqrt{N_l}/4^k$.

定义相应的位点关联权重矩阵(PCWM)为：

$$W_k(i,\ l) = \ln[P_k(i,\ l)/P_k(i,\ 0)] \tag{3}$$

其中 $P_k(i,0)$是 k 联体的背景频率，本文中 $P_k(i,0)$从启动子随机序列集训练得到. 对于任意一条待判定的序列，用位点关联权重矩阵对该序列打分，打分函数表述为：

$$Score = \sum_{l=1}^{n} W_k(i,\ l) \tag{4}$$

$$Score(S,\ X^{\xi}) = \mathrm{Max}\{Score(S,\ X^P),\ Score(S,\ X^N)\} \tag{5}$$

这里 P 代表 TATA 启动子，N 代表 TATA-less 启动子. 给定任意一条序列，利用打分函数(公式 4)分别计算该序列 n 个位点在 TATA 启动子集和 TATA-less 启动子集中的打分. 这样每条序列都会得到 2 个分值，哪个分值高，就被判定为哪一类，见公式(5).

1.7　评价指标

启动子识别常用的评价指标有敏感性(S_n)、特异性(S_p)、平均预测率(AAc)和相关系数(Mcc)等.

敏感性(sensitivity): $S_n = TP/(TP+FN)$.

特异性(specificity): $S_p = TN/(TN+FP)$.

平均预测率(average accuracy): $AAc = (TP+TN)/(TP+FN+FP+TN)$.

相关系数(average accuracy): $Mcc = (TP+TN)/(TP+FN+FP+TN)$.

定义：TP 为 TATA 启动子被识别为 TATA 启动子的数目，FN 为 TATA 启动子被识别为 TATA-less 启动子的数目，TN 为 TATA-less 启动子被识别为 TATA-less 启动子的数目，FP 为 TATA-less 启动子被识别为 TATA 启动子的数目.

2　结果和讨论

2.1　碱基组分及 GC/AT−Skew 分析

2.1.1　碱基组分分析. 对植物 175 条 TATA 启动子和 TATA-less 启动子的 251 个位点分别进行单碱基和二联体频率统计，统计结果显示，TATA 和 TATA-less 启动子在单碱基及二联体组成上均具有相似的特征，在植物启动子区域富含 AT 碱基，A+T 含量高达 60%以上，这与人类启动子区域富含 GC 的特点正好相反. 从二联体含量我们还可以看出，排在前四位的二联体分别是：AA，TT，AT 和 TA. 另外，最近 Pandey 和 Krishnamachari[21]的研究报告指出，植物启动子区域的弯曲度与非启动子区域的弯曲度有着明显的不同，这是识别植物启动子的一个很好的特征.

2.1.2 GC/AT-Skew 分析．植物启动子的碱基含量与人类启动子存在很大的不同，并不存在哺乳动物启动子特有的 CpG 岛． Fujimori 等[8]发现高等植物启动子在其转录起始位点(TSS)附近存在很强的 GC-Skew．结合我们对植物 TATA 启动子特征分析的结果[22]，通过对 175 条 TATA 启动子和 130 条 TATA-less 启动子进行位点 GC-Skew (C > G) 和 AT-Skew(T > A)对比分析，我们发现，在 TATA 启动子近端启动子区域内存在很强的 GC-Skew (Value > 0)，而该偏好在 TATA-less 启动子中虽然存在但并不明显，而对于 AT-Skew，在两类植物启动子中的特征都不是十分明显，所以 GC-Skew 主要集中在植物 TATA 启动子中，这对进一步阐明植物 TATA Pol-Ⅱ启动子编码的基因转录起始机制会有一定帮助．

2.2　位点结构特征差异性分析

根据表 1 列出的二联体片段对应的参数表，实际的计算以步长为 1 bp 的窗口沿待测序列滑动，以碱基间隔作为位点，这样长度为 251 bp 的序列共有 250 间隔位点．分别计算 TATA 和 TATA-less 启动子 6 类结构特征，结果如图 1 所示．

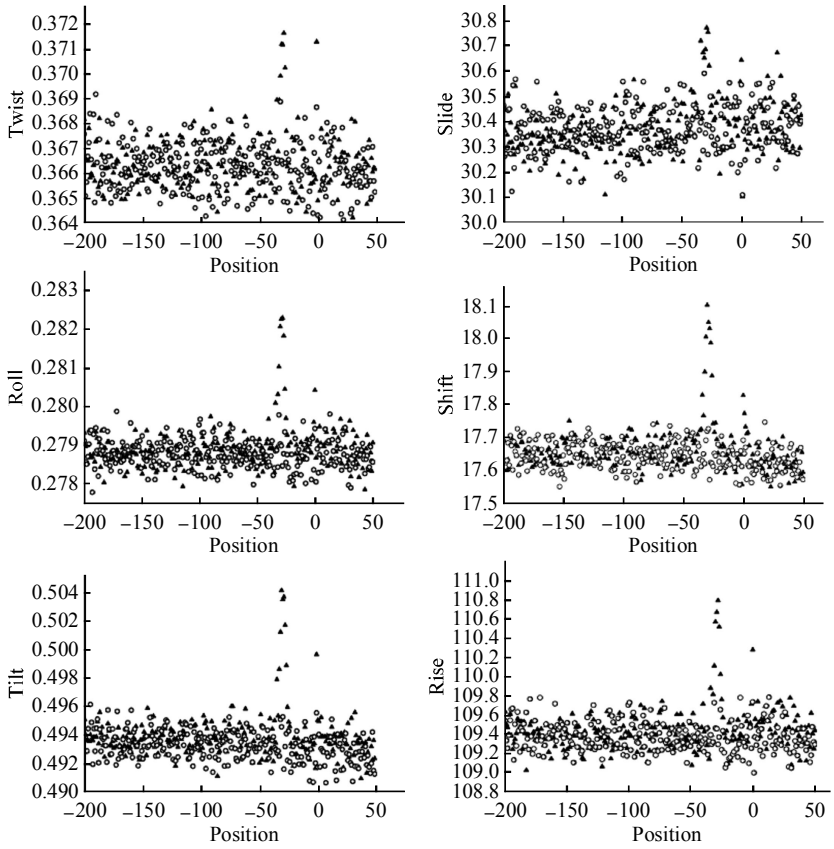


Fig. 1　The position conversation of six physical structure features

▲: The TATA promoter; ○: The TATA-less promoter.

通过对植物 175 条 TATA 启动子和 130 条 TATA-less 启动子的 6 种结构特征进行位点保守性分析表明，TATA 启动子在结构特征上总是十分保守的．在−30 bp 和 TSS 附近这 6 种结构特征保守性都很强(图 1)．相反，TATA-less 启动子并没有很强的保守区域．最新研究报告表明[23]，大部分宽峰启动子都缺少 TATA box．也就是说，大部分 TATA-less 启动子往往都具有多个转录起始位点，它们离散地分布在一个大约 100 bp 范围内的区域内． 另外，大部分组织特异性基因的启动子都是 TATA-less 类型的，它们往往没有特定的转录起始位点，这也可能使得 TATA-less 启动子并不存在特别保守的结构区域．

2.3　转录因子结合位点分布

2.3.1 TATA box 保守性及位点分布．TATA box 是 TATA 启动子的最重要的特征元件，是起始复

合物的主要装配点，TATA box 与 TFⅡD 的一个亚单位结合构成 TATA 结合蛋白(TBP)，决定着基因转录的精确起始．TATA box 与转录起始位点的距离(TSS)与下游转录物的特异性有很强的关联性[24]．最近 Ponjavic 等[24]对小鼠的 5′cDNA 统计发现，小鼠基因的 TATA-box 通常位于 TSS(0 bp)上游的 –32 bp～–29 bp 一个很小的范围内，其中–31 bp 和 –30 bp 2 个位点的跳跃选择就表现出很强的组织特异性．相反，如果 TATA box 位于 TSS 上游的 –28 bp 处，基因便无法正确转录起始．我们使用 MEME 软件对 175 条 TATA 启动子的 TATA box 进行搜寻分析，分析它们的位点保守性并统计它们与 TSS 位点距离，结果如图 2 和图 3 所示．

由图 2 可知，TATA box 的碱基位点保守性与哺乳动物 TATA 启动子基本一致，存在一个长约 9 bp 左右的保守区域，其正则表达式为 CTATA[AT]A[TA]A．观察图 3 的统计结果可看出，与小鼠的 TATA box 位置分布相比，TATA box 在植物启动子中分布的相对广泛一些．约 73.7% 的 TATA box 分布在距离转录起始位点 –35 bp～–29 bp范围内，约 16.5% 的 TATA box 分布在距离转录起始位点 –50 bp～–36 bp 范围内，另外有近 3.6% 的 TATA box 分布在距离转录起始位点(TSS)上游 60 bp 以外的范围内，极个别则分布在 TSS 的下游区域．植物启动子 TATA box 这种分布范围广的特点可能是导致植物多样化及多态性的原因之一．

2.3.2　保守模体分布差异性统计分析．为进一步分析 TATA 启动子和 TATA-less 启动子在转录因子

Fig. 2　The WebLogo images showing the plant TATA box

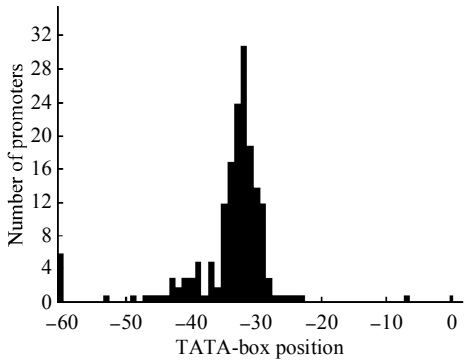


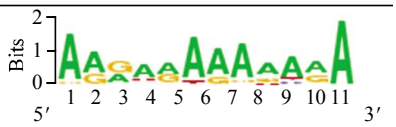
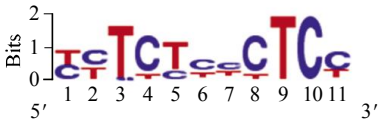
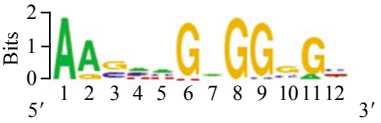
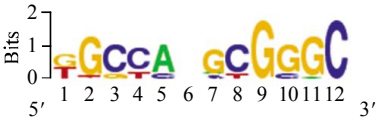
Fig. 3　The TATA-box located around TSS site

结合位点上的特征，我们使用 MEME 模体搜寻软件[17]及 Weblogo 位点保守性分析软件[25]，对这两类启动子转录起始位点附近近端区域内存在的保守模体搜寻统计，发现两类启动子在序列模体分布上存在很大的差异．我们将在两类启动子中搜寻到的保守模体按出现频数排列，将两类启动子 Top01～Top10 模体的正则表达式及它们的位点保守性特点分别列于表 2 和表 3 中．

Table 2　Regular expression of probable TFBSs discovered by MEME

TFBS	TATA	Frequency	TATA-less	Frequency
Top01	CTATA[AT]A[TA]A[CG][CA]	175	A[AG][GA]A[AG]A[AG]A[AG]A[AG]A	53
Top02	C[TA][TC][TC][TC]T[CT][TC][CT]T[TC]	79	[TC][CT]TC[TC][CT][CT]CTC[CT]	45
Top03	[AG]AA[GAC]A[AG][GA][AGC]A[AG]A[GA]	79	AA[GC][ACT][ACT]G[AG]GG[GC]G[CTA]	23
Top04	[CT]ACGTGG[CT][AT][TC][TC]C	29	[GC]G[AGT]GG[AT]GG	17
Top05	[TG][GC]C[AC]T[GC]CA[AT]GC[AT]	25	G[AC]CAC[GC]TGTC[ATG][CT]	12
Top06	G[CG][CT][CG][GA]C[CA]GGC[GC]G	13	[GT]GC[CT]A[CT]GCGGGC	11
Top07	[CTG]GG[AG]GA[GAT]GA[GT]GC	13	[CG][AG][CG]C[TG]TGGGCCC	6
Top08	[GC]TGC[CA][AG][CG]CCC[GT]G	9	[CG][AG]TGC[AG]CGTGCT	4
Top09	GACTTGACC[GA]TC	7	CCGCG[GA]CGCGA	3
Top10	GAGTC[TC]GGTA[TC]C	7	CGGACGGCTCGG	2

Table 3　The position conversation of different motifs discovered by MEME

TFBS	TATA	TATA-less
Top01		
Top02		
Top03		
Top04		
Top05		
Top06		
Top07		
Top08		
Top09		
Top10		

从表 3 中不难看出，对于 TATA 启动子，MEME 模体搜寻到的最强保守模体的正则表达式为 C**TATA**[**AT**]**A**[**TA**]**A**[**CG**][**CA**]，事实上这个保守模体正是 TATA 启动子序列最主要的转录因子结合位点——TATA box．对于 TATA-less 启动子，由于这类启动子调控的基因大多具有多个转录起始位点(TSS)，转录因子结合位点分布的位点特异性相对较弱，A[AG][GA]A[AG]A[AG]A[AG]A[AG]A 模体在 TATA-less 启动子中保守性最强．两类启动子内部保守模体分布差异性较大，说明两类启动子转录起始往往需要结合不同类型的转录因子进行转录起始．另外，TATA 启动子中也存在与 TATA-less 启动子相似的转录因子结合位点．例如，TATA 启动子的 Top02 与 TATA-less 启动子中的 Top02 就属于同种转录因子结合位点．具体模体的碱基位点保守性在表 3 中给出．总之，从两类启动子保守模体含量对比可以看出，两种启动子类型在序列模体组成上存在较大差异．

2.4　位点保守性分析及 PCWM 预测检验

利用 $M_1(l)$((1)式，k=1)对两类启动子进行位点保守性分析，单碱基位点保守性如图 3 所示，可以看出，TATA 启动子与 TATA-less 启动子除在–30 bp 的 TATA box 位置和转录起始位点处区别明显外，其他区域的单碱基位点保守性区别较小. 从图 4 中我们还发现，TATA-less 类型的启动子在单碱基分布上没有明显的保守位点，这使得对 TATA-less 启动子的转录调控研究更加困难，这类启动子引导的基因大多具有多个转录起始位点且组织特异性较强. 我们进一步利用 $M_k(l)$对两类启动子进行位点关联保守性分析(k = 2，…，6)，结果与 $M_1(l)$分析基本类似. 因此，本文只列出 $M_1(l)$和 $M_4(l)$的结果. 从图 5 中可以看出，当 k=4 时 TATA 启动子的位点保守性并未出现大的改变，而 TATA-less 启动子在其转录起始位点上游范围内则出现了几个相对比较保守的区域，例如，–190 bp、–150 bp、–105 bp 等位点. 这些位点的 4-mer 保守性要强于 TATA 启动子，TATA-less 类型的启动子在这些位点都可能结合相应的转录因子来进行转录起始，使得 TATA-less 启动子往往存在多个可变转录起始位点.

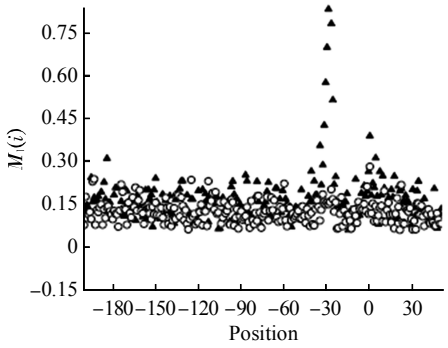


Fig. 4　The graph of $M_1(l)$ values at different position for plant promoters

▲: TATA; ○: TATA-less.

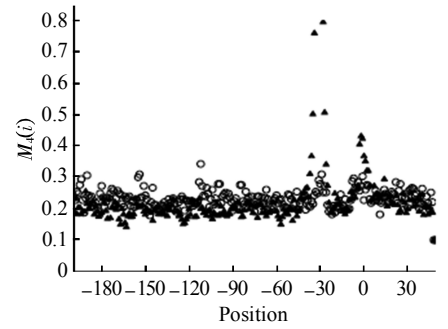


Fig. 5　The graph of $M_4(l)$ values at different position for plant promoters

▲: TATA; ○: TATA-less.

结合图 2，图 3 以及上述的分析结果，我们选取启动子序列–30 bp 附近区域内保守性较大的 9 个位点 (–35，–34，…，–27)构建 k-mer 碱基片段位点关联性打分矩阵(PCWM)，利用打分函数 $Score(S, X^{\xi})$(公式 5)对 TATA 启动子和 TATA-less 启动子做分类，采用 10-fold 交叉检验，计算平均值敏感性(S_n)，特异性(S_p)，平均预测率(AAc)及相互关联系数(Mcc)，结果列于表 4.

Table 4　The results of 10-fold cross validation test by k-mer PCWM

k-mer	Sn (%)	Sp (%)	AAc(%)	Mcc
k=1	63.53	83.90	71.67	0.47
k=2	68.24	90.20	77.33	0.58
k=3	72.12	93.27	77.00	0.59
k=4	**82.35**	**93.33**	**87.84**	**0.74**
k=5	36.47	99.17	63.67	0.44
k=6	50.00	99.23	71.33	0.54

分析表 4 可知，随着联体 k 的增加，预测特异性(S_p)呈单调递增趋势，而预测敏感性(S_n)，平均预测率(AAc)及相互关联系数(Mcc)则呈现先增后减的趋势. 当 k=4 时各个预测指标最优，ROC 曲线如图 6 所示. 当阈值(cutoff)等于–0.1836 时，平均预测敏感性为 82.35%，平均特异性为 93.33%，平均预测率也达到了 87%以上. 预测结果最好，充分说明了同时考虑碱基关联和位点关联的权重矩阵能够较好地进行分类预测并在一定程度上抑制预测算法的假阳率，从而保证了识别结果的可信性，优于单碱基位置权重矩阵(PWM)的分类性能.

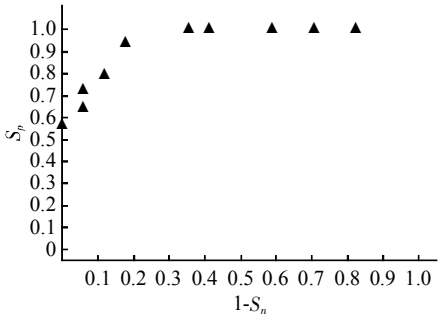


Fig. 6　ROC curves for promoter discrimination by using 4-mer PCWM

3　结　　论

植物启动子是真核启动子中非常重要的一类，是重要的顺式作用元件，是植物基因转录调控的中

心区域．由于植物启动子本身具有的多态性导致植物基因序列特异性弱，序列特征信号复杂的特点，目前针对植物启动子识别的相关文献相对较少，本文利用 PlantPromDB 数据库上给出的实验证实的 TATA 启动子和 TATA-less 启动子数据，从位点保守性，序列组成和结构特征，保守模体分布差异等多个方面入手，对两类植物启动子进行了深入对比研究，统计出植物 TATA 启动子和 TATA-less 启动子一些特有的特征规律．这些规律对更深层次地阐明植物 TATA 启动子和 TATA-less启动子在转录机制上的异同奠定了基础．另外，我们提出并构建了位点关联性权重矩阵(PCWM)扫描算法，对两类启动子进行分类预测并取得了优于单碱基位点权重矩阵(PWM)的分类效果，进一步提高了预测成功率，降低了假阳率，证实了基因组序列内部多数的特异性信号都具有位点关联的特征．相信这种模型对识别其他调控信号及进一步揭示基因转录调控机制会有一定帮助．

参 考 文 献

1 Akan P, Deloukas P. DNA sequence and structural properties as predictors of human and mouse promoters. Gene, 2008, **410**(1): 165～176

2 Allison L A, Moyle M, Shales M, *et al*. Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. Cell, 1985, **42**(2): 599～610

3 Rombauts S, Florquin K, Lescot M, *et al*. Computernal approches to identify promoters and *cis*-regulatory elements in plant genomes. Plant Physiology, 2003, **133**: 1162～1176

4 Shahmuradov I A, Gammerman A J, Hancock J M, *et al*. PlantProm: a database of plant promoter sequence. Nucleic Acids Res, 2003, **31**(1): 114～117

5 Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol, 1987, **196**(2): 261～282

6 Bajic V B, Tan S L, Sutuki Y T, *et al*. Promoter prediction analysis on the whole human genome. Nature Biotechnology, 2004, **22**(11): 1467～1473

7 Tatarinova T, Brover V, Troukhan M, *et al*. Skew in CG content near the transcription start site in Arabidopsis thaliana. Bioinformatics, 2003, **19**(Suppl 1): i313～i314

8 Fujimori S, Washio T, Tomita M. GC-compositional strand bias around transcription start sites in plants and fungi. BMC Genomics, 2005, **28**(6): 26

9 杜耀华, 王正志, 倪青山, 等. 一种基于特征筛选的原核生物启动子判别分析方法. 生物物理学报, 2006, **22**(1): 39～48
Du Y H, Wang Z Z, Ni Q S, *et al*. Acta Biophys Sin, 2006, **22**(1): 39～48

10 Abeel T, Saeys Y, Bonnet E, *et al*. Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res, 2008, **18**(2): 310～323

11 Uren P, Cameron-Jones M, Sale A, *et al*. Promoter prediction using physico chemical properties of DNA. The 2nd International Symposium on Computational Life Science, UK: Cambridge, 2006.

12 Florquin K, Saeys Y, Degroeve S, *et al*. Large-scale structural analysis of the core promoter in mammalian and plant genomes. Nucleic Acids Res, 2005, **33**(13): 4255～4264

13 Goñi J R, Fenollosa C, Pérez A, *et al*. Determining promoter location based on DNA structure first-principles calculations. Genome Biol, 2007, **8**(12): R263

14 Dickerson R E. Definitions and nomenclature of nucleic acid structure components. Nucleic Acids Res, 1989, **17**(5): 1797～1803

15 Yanagi K, Privé G G, Dickerson R E. Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. J Mol Biol, 1991, **217**(1): 201～214

16 Nikolov D B, Burley S K. RNA polymerase Ⅱ transcription initiation: a structural view. Proc Natl Acad Sci USA, 1997, **94**(1): 15～22

17 Bailey T L, Williams N, Misleh C, *et al*. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res, 2006, **34**(Web server issue): W369～W373

18 陈颖丽, 李前忠, 马克健. 大肠杆菌与酵母基因特定序列信息参量的研究. 生物物理学报, 2001, **17**(4): 676～684
Chen Y L, Li Q Z, Ma K J. Acta Biophys Sin, 2001, **17**(4): 676～684

19 Stormo G. DNA binding sites: representation and discovery. Bioinformatics, 2000, **16**(1): 16～23

20 Li Q Z, Lin H. The recognition of sigma 70 promoters in Escherichia K-12. J Theoretical Biology, 2006, **242**(1): 135～141

21 Pandey S S, Krishnamachari A. Computational analysis of plant RNA Pol-Ⅱ promoters. Biol Systems, 2006, **83**(1): 38～50

22 左永春, 李前忠, 杨 磊, 等. 基于 PCWM 扫描模型的植物启动子分析及识别. 内蒙古大学学报, 2008, **39**(5): 536～542
Zuo Y C, Li Q Z, Yang L, *et al*. Acta Scientiarum Naturalium Universitatis Neimongol, 2008, **39**(5): 536～542

23 Müller F, Demény M A, Tora L. New problems in RNA polymerase Ⅱ transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. J Biol Chem, 2007, **282**(20): 14685～14689

24 Ponjavic J, Lenhard B, Kai C, *et al*. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. Genome Biol, 2006, **7**(8): R78

25 Crooks G E, Hon G, Chandonia J M, *et al*. WebLogo: a sequence logo generator. Genome Res, 2004, **14**(6): 1188～1190

Analysis of Plant TATA and TATA-less Promoters by Using Sequence and Structure Features*

ZUO Yong-Chun, LI Qian-Zhong**

(*School of Physical Science and Technology, Inner Mongolia University, Hohhot* 010021, China)

Abstract Analysis of regular elements in promoter region is the base for elucidating the mechanism of gene transcription initiation. The TATA and the TATA-less promoters of plant RNA polymerase Ⅱ gene are chosen from the PlanPromDB. The GC bias, position structure conservation, nucleotide content and conservative motifs of sequences, position distribution of TATA box and conservation of correlation position are analyzed. Many specific regulars for the two types of promoters are found. These features can offer some help for revealing the transcription regulation of plant gene. A new prediction algorithm based on position-correlation weight matrix (PCWM) is proposed. The better discrimination results for two sort plant promoters are obtained by using score function. It is confirmed that the performance of position-correlation weight matrix (PCWM) is superior to single-base position weight matrix (PWM).

Key words plant pol-Ⅱ promoter, features of sequence content bias, TATA and TATA less promoter, conservative motifs, position-correlation weight matrix (PCWM)

DOI: 10.3724/SP.J.1206.2008.00727

---

*This work was supported by a grant from The National Natural Science Foundation of China(30560039).

**Corresponding author.

Tel: 86-471-4992958, E-mail: qzli@imu.edu.cn

Received: October 23, 2008 Accepted: November 25, 2008