·综述·

真核生物 RNA 聚合酶Ⅱ启动子的计算机预测

姚凤霞　张瑞芳　刘春宇　夏家辉　夏昆

摘要　启动子是基因组序列中靠近基因转录起始位点的区域，是影响基因表达的重要功能单位之一。除实验方法发现或验证序列的启动子外，现已经有多种启动子序列的计算机预测方法，如位点比重阵列(PWM)、隐马尔柯夫模型(HMM)、神经网络、低聚复合物和 CpG 岛等。本文综述了现有真核生物基因启动子序列的生物信息研究技术。

关键词　启动子；　转录起始位点；　生物信息学

随着人类基因组序列测定即将完成，当今面临的一个重要问题是如何正确地解读基因组的结构和功能。除寻找基因编码区外，基因表达中调控序列的认识将是一个更为艰巨的任务。调控序列包括位于基因编码序列上游的启动子(promoter)、正调控元件增强子（enhancer）以及负调控元件沉默子(silencer)等，其中启动子是 DNA 序列中 RNA 聚合酶Ⅱ的结合部位，也是启动转录的关键性的调控序列。对启动子区的认识，不仅有助于实验室分析研究，而且还可以为人类认识全基因组功能、基因表达调控机制以及人类疾病与启动子多态性或突变的关系提供很大的帮助。

1　真核生物启动子的结构和特征

真核生物基因表达调控主要包括染色体的结构、转录的起始、转录后加工等过程，但普遍认为在调控机制中转录的起始是一重要的环节。转录起始是前转录复合物（preinitiation complex, PIC）识别 DNA 序列上的核心启动子并启动转录的过程。PIC 除了 RNA 聚合酶Ⅱ(RNA polymerase Ⅱ, Pol Ⅱ)外，还包括一般转录因子（general transcription factors, GTFs），GTFs 通常包括 TFⅡA, TFⅡB, TFⅡD, TFⅡE, TFⅡF 和 TFⅡH 等[1]。核心启动子一般指 TATA 盒和起始子(initiator, Inr)。TATA 盒多位于转录起始位点（transcription start site, TSS）上游约 30bp 处，TATA 盒与 TFⅡD 的一个亚单位结合构成 TATA 结合蛋白（TATA binding protein, TBP）。TFⅡD 除了 TBP 亚单位外，还有许多 TBP 结合因子(TBP associated factors, TAFs)。Inr 位于转录起始位点附近，序列是 PyPyAN(TA)PyPy，其中 Py 代表嘧啶(C 或 T)，N 代表任意碱基[2]。有时两种核心启动子同时存在于一个基因的调控区中，有的没有任何一种，可能依靠下游启动子（downstream promoter element, DPE)来发挥结合作用(图 1)。

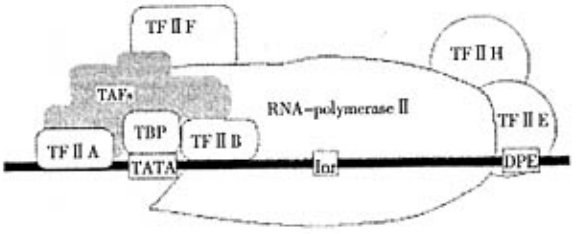

图 1　核心启动子结合有 RNA 聚合酶Ⅱ和一般转录因子[2]

另外，位于 TSS 上游约 70bp 处的上游调控元件 GC 盒，CCAAT 盒以及调控区的 CpG 岛都与启动子的认识有关。

2 真核启动子数据库（eukaryotic promoter database, EPD）

EPD[3]是由瑞士的实验癌症研究所(Swiss institute for experimental cancer research, ISREC）和生物

基金项目：863 计划重大专项资助项目(No.2002BA711A07-08)；国家自然科学基金资助项目(No.30100103；No.30270735)

作者单位：410078 长沙，中南大学医学遗传学国家重点实验室
通讯作者：姚凤霞(e-mail：fxyaocn@hotmail.com)

信息研究所(Swiss institute for bioinformatics,SIB)共同维护、注释的非冗余真核生物聚合酶Ⅱ启动子的数据库。该数据库所有的启动子均经过一系列的实验证实:如是否为真核 RNA 聚合酶Ⅱ启动子、是否在高等真核生物中有生物学活性、是否与数据库中的其他启动子有同源性等等。EPD 与其他的相关数据库也建立了相关链接,如 EMBL、SWISS-PROT、TRANSFAC 等[4]。在 2003 年 1 月发布的第 73 版中,EPD 将收集的启动子分为 6 大类:植物启动子、线虫启动子、拟南芥启动子、软体动物启动子、棘皮类动物启动子和脊椎动物启动子,共 2 997 个条目,其中脊椎动物中的人类启动子有 1 871 个,约占总数的 62%。该数据库是现有各种启动子预测方法的主要评价依据之一。

3 启动子预测

认识启动子的传统方法是实验研究和分析,近年来随着生物信息学(bioinformatics)这门新兴交叉学科的兴起和蓬勃发展,用各种方法进行计算机模拟启动子可以获得启动子的信息。与传统的实验研究方法相比,生物信息学分析具有节省人力和物力资源、且用较短时间可以预测大量的启动子序列等优点,成为现在启动子预测和研究的一个重要手段。以下是一些最常用的计算机预测启动子的方法。

3.1 位点比重阵列

位点比重阵列(position weight matrix,PWM),是一种特殊的多位点比对的方法。PWM 具体用在启动子预测中,是对特定的启动子元件,求每个位点出现每种碱基的比重,然后得到总的位点分值[1]。Bucher[5]构建了一个优化后 PWM,被广泛用于启动子预测方法中。一般认为 PWM 对转录因子结合位点提供的信息多于对启动子中的保守元件,当研究二者之间的联系时 PWM 是个好的选择[1]。以 PWM 为基础的预测软件有 Solovyev 和 Salamov 开发的 TSSG 和 TSSW[6],它们都是用线性判别函数来预测给定序列的可能功能域和定位,如启动子、转录因子结合位点等,其中 TSSG 中转录因子结合位点来源于 TFD[7],而 TSSW 则来源于 TRANSFAC[8]。TFD和 TRANSFAC 都是转录因子数据库。Zhang 实验室[9]的启动子预测软件 CorePromoter 则是用二次判别式分析(quadratic discriminant analysis,QDA),不仅考虑到 TSS 附近的元件,还把上游和下游的调控元件也考虑在这种方法中。

3.2 隐马尔柯夫模型

隐马尔柯夫模型(hidden Markov Model,HMM)是以概率为基础的动态统计学模型,被广泛用于序列分析中,如基因发现工具 GENSCAN[10]、蛋白质功能域和家族的多序列比对数据库 Pfam[11]等。

HMM 由两个相互关联的离散随机过程组成:观测序列 $X=X_1^T=(x_1,x_2,\cdots\cdots,x_T)$,以及隐藏在该观测序列背后的状态序列 $\Pi=\Pi_1^T=(\pi_1,\pi_2,\cdots\cdots,\pi_T)$,$X$ 和 Π 之间的关系见图 2。

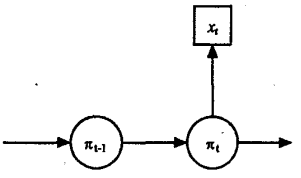

图2 X和Π之间的关系

注:正式地,一个基本的 HMM 可以用一个三元组 $M=(\Sigma,Q,\Theta)$ 来描述,其中:Σ 为字母表,含有 N 个观察值;Q 为状态值的有限集合,含有 N 个状态;$\Theta=(A,B)$为概率分布的集合,包括:状态转移概率 $A=\{a_{kl},k,l\in Q\}$,发散概率 $B=\{e_k(b),k\in Q$ 且 $b\in\Sigma\}$

利用 HMM 作启动子预测时,取 $\Sigma=\{A,T,C,G\}$,观测序列取为目的序列,而状态序列则取为各种调控元件(如启动子、增强子、转录因子结合位点等)。HMM 的使用分成"学习/训练"与"工作"两个阶段。在学习阶段,主要通过动态规划算法及 EM 算法逐步修正 Θ 中的各参数,目的是使得训练好的 HMM 对样本序列的适应性最佳。在工作阶段,则主要使用 Viterbi 算法,根据已知的 DNA 序列获得最可能的状态序列(即预测出各种调控元件序列)。

Audic 和 Claverie 开发的 Audic 是 HMM 模型结合 Bayes 定律通过对启动子与非启动子的转换概率比较后来确定启动子序列[13]。Ohler 和 Niemann 开发的 McPromoter 也是利用 HMM 模型并结合高斯分布曲线从序列的相似性和物理特性(如 B-DNA 双螺旋、核小体布局、DNA 结合度和 Z-DNA 稳定能量等等)两方面进行启动子的识别[14]。

3.3 神经网络原理

神经网络是模仿生物神经系统一类预测模型。基本的神经网络有 3 层结构:输入层、隐藏层以及输出层,每一层是由数量不等的神经元组成,各层神经元之间建立有完全的连接关系,信号可以通过这些连接从前一层传递到后继层,直到从输出层的神经元输出结果;在信号传递过程中,所经过的每个神经元要对信号作一个非线性的处理,而每个连接则要对这些信息作加权处理。具体对启动子预测

而言，一段未知 DNA 序列作为输入神经元，这里的 4 种碱基 A、C、G、T 由二元码表示，然后通过神经网络中某种算法得到结果并存储，这个结果会作为下一个神经网络的输入信号进行分析。最后输出的结果只有两个：启动子序列和非启动子序列。

与 HMM 相类似，神经网络也分为学习与工作两个阶段，学习阶段主要通过梯度下降等各种优化算法调整各神经元的处理函数及各连接权，在达到某种最优准则后，便可以将神经网络用于工作阶段了。

Knudsen[15]所开发的 Promoter 2.0 是用一个初始的人工神经网络扫描未知 DNA 序列，运用遗传算法（genetic algorithm）来进行优化和选择启动子。Reese 等[16]开发的 NNPP 是结合时间延迟神经网络（time-delay neural networks，TDNNs）和位点修剪而对真核和原核生物的启动子进行预测的。该软件是在对 TATA 盒和 Inr 认识的基础上建立的，时间延迟神经网络可以在不同序列的不同位置进行分析。NNPP 是先通过神经网络分析，用修剪程序剪切掉网络内部不可能成为启动子的序列，然后神经网络重新构建，这样周而复始，一直至满足最后条件为止。

3.4 六聚体（hexamer）和低聚复合物

启动子序列仅仅是一个基因的小部分序列，所以 Claverie 和 Bougueleret 把启动子序列与 DNA 中其他区域一并考虑从序列整体上预测启动子，从而产生了这种方法。Hutchinson 等[17]开发的 PromFind 是在启动子区、蛋白编码区和非编码区之间求 hexamer 频率，从而达到判定是否启动子序列的目的。Scherf 等[18]开发的软件 PromoterInspector 是根据 IUPAC（international union of pure and applied chemistry）标准把基因组序列分类，每一类是序列相似的低聚复合物集合，它们由通配符（用 N 表示，代表不确定碱基）来区别。如一个低聚复合物（AGC，GCA）有两个通配符，这个分类就是（AGCGCA，AGCNGCA，AGCNNGCA），通过对这些低聚复合物之间的序列分析然后得到启动子的序列信息。它的最大优点是可以降低假阳性率。Matthias 等通过对人 22 号染色体全基因组分析，预测出 465 个启动子中，其中有约 40%已得到了证实[19]。

3.5 CpG 岛

CpG 岛是最初在转录起始位点处发现的高 CG 含量的序列，约 50%的基因以及几乎所有的看家基因都在 5' 转录点存在 CpG 岛[20]。对 CpG 岛的判定有以下简单的规则[20]：①大于 200bp；②大于 50%的 G+C 含量，即 pG+pC>0.5；③该区域中至少有 0.6 的 CpG 岛频率，即 pCpG>0.6 × pG × pC。Hannenhalli 等[21]通过研究 CpG 岛与启动子的关系，发现结合 CpG 岛来预测启动子可以提高准确性，所以 CpG 岛是认识启动子的重要序列信号。

3.6 mRNA 与基因组序列比对

由于目前基因组序列比较完整，最近 Coleman 等[22]直接通过已知基因的最长 mRNA 序列在基因组中进行序列比对，提取 5' 端 500～700bp 基因组序列进行启动子活性的研究，结果表明现阶段约 75%的基因在这段区域中可以检测到启动子的活性。提示直接通过 cDNA 与 gDNA 比较就有较高的可能性确定有效的启动子区。但这种方法受限于已知基因的 5' 的完整性，不能区分 5' 非翻译区（5'-untranslated region，5'UTR）和启动子，同时对多启动子基因也无法鉴定。

3.7 综合多种方法

综上所述，在没有发现更有效的方法之前，可以综合以上各种机制预测启动子区，Liu 等[23]开发的 CONPRO 就是结合了 PWM、神经网络、以及低聚复合物等方法并经过优化后的软件。通过结合 GENSCAN，不仅对有完整全长的已知基因，而且对不完整的 EST 序列作了预测分析。结果表明约一半（37%～71%）的被分析序列中可得到启动子，其中有 85%～90%是真正的启动子。Bajic 等[24]开发的 DPF（dragon promoter finder）也是一种多原理综合分析软件，它是对一段未知序列依次进行五聚体筛选、PWM 位点分析、信号处理、人工神经网络（artificial neural network，ANN），最后得到预测结果。

Fickett 等[1]对 24 个启动子用现有的一些软件预测以达到评价各种软件的目的，得到如表 1 所示的结果。结果表明目前软件分析的可靠性还相当有限，假阳性率高达约 80%，而真阳性率很少超过 50%。表 2 列举出主要的启动子预测服务器地址。

表 1 各种软件对 24 个已知启动子的预测结果[1]

	TP	%TP	FP	%FP
Audic	5	21%	33	87%
Autogene	7	29%	51	88%
Promoter2.0(GeneID)	10	42%	51	84%
NNPP	13	54%	72	85%
PromFind	7	29%	29	81%
PromoterScan	3	13%	6	67%
TATA	6	25%	47	89%
TSSG	7	29%	25	78%
TSSW	10	42%	42	81%

注：TP：阳性数；%TP：阳性百分比；FP：假阳性数；%FP：假阳性百分比

表 2 启动子预测的服务器和软件

名称	网址
Audic	ftp://igs-server.cnrs-mrs.fr/pub/SELFID/
Autogene	ftp://ftp.bionet.nsc.ru/
CorePromoter	http://argon.cshl.org/genefinder/CPROMOTER/index.html
CONPRO	http://stl.bioinformatics.med.umich.edu/conpro/
DPF	http://sdmc.lit.org.sg/promoter/promoter1_3/DPFV13.html
McPromoter MM:II	http://genes.mit.edu/McPromoter.html
NNPP	http://www.fruitfly.org/seq_tools/promoter.html
PromFD	ftp://ftp.genetics.wustl.edu/pub/stormo/PromFD/
PromFind	http://www.rabbithutch.com/
PromoterInspector	http://www.genomatix.de/software_services/software/PromoterInspector/PromoterInspector.html
Promoter2.0(GeneID)	http://www.cbs.dtu.dk/services/Promoter/
PromoterScan	http://bimas.dcrt.nih.gov/molbio/proscan/
TSSG/TSSW	http://www.softberry.com/berry.phtml?topic=promoter

4 问题与展望

从 Fickett 等表 1 的评估可知目前启动子预测软件还存在明显不足。虽然 Scherf 等[18]开发的软件 PromoterInspector 假阳性大大降低了，但它属于商业性软件，免费分析的序列很有限。同时当前的软件也不同程度地存在假阴性现象。用计算机开发软件会遇到一些困难，如：①人类公共数据库中，只有极少数被实验证实的启动子，被收录到了 EPD 中，绝大多数基因的启动子仍然是未知的；②数据库中 cDNA 和 EST 簇经常是不完整序列，特别是 5' 端，故无法确定转录起始位点的确切位置，从而影响启动子的预测；③真核生物的启动子比原核生物的复杂，需要考虑多种因素才可以很好地预测；④即使部分启动子在 DNA 水平是客观存在的，可是通过体外实验却无法证实，可能是由于启动子不仅与它本身的位置以及结构有关，同时还与转录因子结合位点、DNA 的空间结构、染色体的功能区以及该基因的组织特异性表达等等有关。

在对真核生物启动子研究中，越来越多不同物种基因组序列的完成将为通过基因组的比较来发现基因调控序列提供重要基础。为了加速启动子的发现，新的大规模的实验技术和计算机算法还有待进一步开发，从而为研究基因的结构提供更加有效的工具。

参 考 文 献

1 Fickett JW, Hatzigeorgiou AG. Eukaryotic promoter recognition. Genome Res, 1997, 7 (9): 861-878.

2 Pedersen AG, Baldi P, Chauvin Y, et al. The biology of eukaryotic promoter prediction-a review. Comput Chem, 1999, 23 (6): 91-207.

3 http://www.epd.isb-sib.ch/

4 Praz V, Perier R, Bonnard C, et al. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucleic Acids Res, 2002, 30 (1): 322-324.

5 Bucher PJ. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. Mol Biol, 1990, 212 (4): 563-578.

6 http://www.softberry.com/berry.phtml?topic=promoter

7 http://www.ifti.org/

8 http://transfac.gbf.de/TRANSFAC/index.html

9 Zhang MQ. Identification of human gene core promoters in silico. Genome Res, 1998, 8 (3): 319-326.

10 http://genes.mit.edu/GENSCAN.html

11 http://www.sanger.ac.uk/Software/Pfam/I

12 Eddy SR. Hidden Markov models. Curr Opin Struct Biol, 1996, 6 (3): 361-365.

13 Audic S, Claverie JM. Detection of eukaryotic promoters using Markov transition matrices. Comput Chem, 1997, 21 (4): 223-227.

14 Ohler U, Liao GC, Niemann H, et al. Computational analysis of core promoters in the Drosophila genome. Genome Biol, 2002, 3 (12): research0087.

15 Knudsen S. Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics, 1999, 15 (5): 356-361.

16 http://www.fruitfly.org/seq-tools/promoter.html

17 Hutchinson GB. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. Comput Appl Biosci, 1996, 12 (5): 391-398.

18 Scherf M, Klingenhoff A, Werner T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol, 2000, 297 (3): 599-606.

19 Scherf M, Klingenhoff A, Frech K, et al. First pass annotation of promoters on human chromosome 22. Genome Res, 2001, 11 (3): 333-340.

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

21 Hannenhalli S, Levy S. Promoter prediction in the human genome. Bioinformatics, 2001, 17 (Supplement): S90-S96.

22 Coleman SL, Buckland PR, Hoogendoorn B, et al. Experimental analysis of the annotation of promoters in the public database. Hum Mol Genet, 2002, 11 (16): 1817-1821.

23 Liu R, States DJ. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. Genome Res, 2002, 12 (3): 462-469.

24 Bajic VB, Seah SH, Chong A, et al. Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics, 2002, 18 (1): 198-199.

(收稿日期：2004-02-27)
(本文编辑：王璐)