# 真核生物启动子的预测技术

孙吉贵 韩霄松 卢欣华 行 荣 仲 洋
(吉林大学计算机科学与技术学院 长春 130012)
(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

**摘 要** 启动子是基因表达过程中非常重要的调控序列,是影响基因能否转录的重要功能单位之一,真核生物的启动子预测已经成为生物信息学研究的热点。将结合人工神经网络、支持向量机、二次判别分析和位置权值矩阵技术,对国内外真核生物启动子的预测研究进行介绍,并在最后采用合适的技术应用到模拟真核生物基因表达过程的电子细胞模型 Analog-Cell 中。

**关键词** 启动子,人工神经网络,支持向量机,二次判别分析技术,位置权值矩阵,生物信息学

## Computational Prediction Technology of Eukaryotic Promoters

SUN Ji-gui HAN Xiao-song LU Xin-hua XING Rong ZHONG Yang
(College of Computer Science and Technology, Jilin University, Changchun 130012, China)
(Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University, Changchun 130012, China)

**Abstract** Promoter is a very important set of sequence in gene expression, and it has vital effect in gene transcription. Prediction of eukaryotic promoters has been a focus of bioinformatics. We will introduced the recently researches of eukaryotic promoters' prediction, which are using ANN, SVM, QDA and PWM. In the end, we will adopted proper technology to applying in E-Cell model Analog-Cell, which has simulated gene expression in eukaryotic cell.

**Keywords** Promoter, ANN, SVM, QDA, PWM, Bioinformatics

## 1 引言

随着人类基因组计划的胜利结束,人类得到的 DNA 序列数据大量增长,于是从 DNA 中提取有用的信息来解释生命现象,从而帮助人类更好地认识自己已成为一种迫切的要求。但是,在人类浩瀚的 DNA 序列中,仅仅有大约 3%的序列是编码蛋白质的序列,就是我们常说的基因。在基因表达成蛋白质的过程中,各种生化反应所用到的酶和调节因子等物质就必须精确地结合到待表达的基因序列上,来调控整个表达过程,尤其是转录过程。然而,在这数以万亿计的 DNA 序列中,酶和调节因子要精确地定位在待表达的基因的合适的结合位点上,无异于大海捞针。幸好生物漫长的进化过程形成了一套精密的基因表达体系,为每个基因提供了一些信号,即一些特异性的序列,专门用于结合起始转录需要的酶和调节因子。其中最重要的序列之一就是启动子序列。因此,启动子对于基因表达的作用是至关重要的,它控制了基因表达的第一步:mRNA 合成的起始。在现代分子生物学中,启动子的预测,尤其是真核生物启动子的预测一直是研究的热点之一。研究的开始阶段,主要采用生物实验中分析和鉴定启动子的方法进行预测,效率较低。然而,随着人类基因组计划的完成、人们先前启动子研究的积累以及生物信息学和计算机科学的发展,越来越多的科研人员利用生物信息学的方法通过计算机模拟和计算来预测启动子,这种方法成本低,所用的时间较短,并且得到的结果较为可靠[1]。当前国内外已经开发出很多有效的启动子预测软件。

## 2 启动子简介

启动子(Promoter)这个名称最早出现于 1964 年,用来描述原核生物乳糖操纵子中位于 3 个基因紧邻的上游区段。这个位点失活与否,影响着转录能否成功,就像下游基因的表达的开关一样。有研究表明,这个位点是负责转录操纵子的 RNA 聚合酶的结合位点。转录起始位点(Transcription Start Site, TSS)称为"+1"位,启动子序列一般位于转录起始位点上游 20～600 个核苷酸的位置。

而在真核生物中,启动子指的是对基因转录起始有重要作用的序列,不像原核生物那么保守,并且启动子的序列较多。这些序列有不同的功能,其中核心启动子(core promoter)负责与起始复合物结合,负责控制基因是否转录,包括 TATA 框和起始子;还有一些位于核心启动子上游的上游调控元件(UREs)序列。上游调控元件的有无,对于起始转录的效率至关重要,它们负责控制基因转录的频度,可以极大地提高核心启动子的低水平转录活性。比较常见的序列是 GC 框

到稿日期:2008-02-02 本文受国家自然科学基金(No. 607730097),国家教育部高等学校博士学科点专项科研基金(No. 20050183065)资助。
孙吉贵(1962－),男,博士,教授,博士生导师,计算机学会理事,主要研究领域为人工生命、约束程序;韩霄松(1984－),男,硕士研究生,主要研究领域为人工生命;卢欣华(1977－),女,博士研究生,助教,主要研究领域为人工生命;行 荣(1984－),女,硕士研究生,主要研究领域为人工生命、约束程序;仲 洋(1984－),女,硕士研究生,主要研究领域为计算智能、人工生命。

和 CCAAT 框[2]。真核生物的启动子结构见图 1。

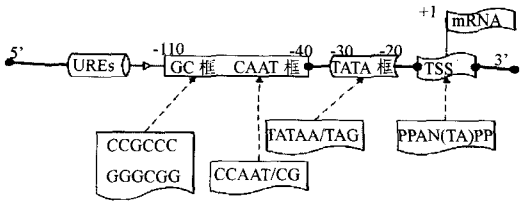


图 1　真核生物的启动子结构

对于真核生物转录起始很重要的另外一个结构是 CpG 岛，它是最初在转录起始位点处发现的高 CG 含量的序列，大约 50%的基因以及几乎所有持家基因都在 5'转录点存在 CpG 岛[3]。CpG 岛序列有以下 3 个特点：大于 200bp；大于 50%的 G+C 含量，即 pG+pC>0.5；该区域至少有 0.6 的 CpG 岛频率，即 pCpG>0.6×pG×pC。CpG 岛是预测启动子并提高预测准确性的重要序列[2]。

真核生物的 3 种 RNA 聚合酶识别不同的启动子序列，很多真核生物的核心启动子含有一个叫 TATA 框（TATA Box）的序列，位于转录起始位点上游大约 25～35 个核苷酸的位置。此序列比较保守，为 5'-TATA(A/T)A(A/T)-3'，负责与 TBP(TATA Box Binding Protein，TFIID)结合，从而吸引 RNA 聚合酶结合到 DNA 序列来起始转录。另外一个比较重要的核心启动子是起始子(Initiator，Inr)，序列是 PPAN(TA)PP，其中 P 代表嘧啶、N 代表任意碱基。起始子位于转录起始位点附近，可供 RNA 聚合酶 II 识别。无论 TATA 框存不存在，起始子对于启动子的强度和起始位点的选择都起很大作用。

总的来说，启动子是很重要的调控基因表达的 DNA 序列。如何从基因组中快速地找到启动子，已经成为生物信息学中一个非常重要的研究课题。本文主要是根据国内外研究成果所采用的技术来分类介绍启动子序列识别方法上的研究进展，并在最后采用合适的技术应用在已有的模拟真核生物基因表达过程的电子细胞模型 Analog-Cell 中。

## 3　启动子预测技术

当前的启动子识别技术根据研究目标可以分为共性启动子识别技术和特异性启动子识别技术。前者的目标是在基因组中找出基因的转录起始位点和核心启动子，后者指的是寻找一组特定基因的转录因子结合位点。目前国内的研究比较偏重于特异性启动子识别技术，而国外已经成型了几个基于共性启动子识别技术的系统。从这些成果所采用的技术看，主要分成 4 种：基于人工神经网络(ANN)技术、基于支持向量机(SVM)技术、基于二次判别分析(QDA)技术和基于位置权值矩阵(PWM)技术。以上 4 种技术都是将机器学习的方法运用到启动子预测的研究中，机器学习后用作分类器来判断待表达序列是否含有启动子序列。下面将基于这 4 种技术一一介绍识别启动子的技术。

### 3.1　基于人工神经网络的启动子预测技术

人工神经网络(Artificial Neural Networks，ANN)是指模拟人脑神经系统的结构和功能，运用大量的处理部件，由人工方式构造的网络系统。神经网络理论突破了传统的、线性处理的数字电子计算机的局限，是一个非线性动力学系统，并以分布式存储和并行协同处理为特色，虽然单个神经元的结构和功能极其简单有限，但是大量的神经元构成的网络系统所实现的行为却是极其丰富多彩的[4]。由于人工神经网络具有强大的模式识别能力，近年来也越来越多地被用到生物信息学，尤其是序列识别和预测的研究中。基本的人工神经网络有 3 层结构：输入层、隐藏层和输出层，每一层由若干神经元组成，各层神经元之间建立有连接，信号通过这些连接经过处理传到下一层，直到从输出层神经元输出结果。信号传递过程中，每个神经元都要对信号做非线性的处理，每个连接也通过连接上的权值对信号进行处理。人工神经网络分为学习和工作两个阶段，学习阶段主要通过各种学习算法调整各个神经元的阈值以及连接上的权值，即存储知识的过程；工作阶段就是通过学习完毕的神经网络进行工作，即知识的应用过程[4]。图 2 就是一个典型的人工神经网络结构。

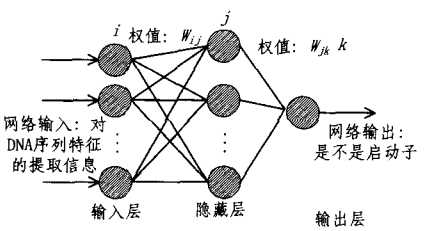


图 2　一个典型的人工神经网络结构

用到启动子预测中，人工神经网络的作用类似于一个分类器。一般是一个输出神经元确定是否含有启动子，其它的网络结构就要根据实际的算法来确定。首先是确定网络的输入，一般需要对 DNA 序列进行特征信息提取，因为原始 DNA 序列往往杂乱无章并且意义难于理解。序列特征信息是 DNA 原始序列所固有的，它可以有效地反映和描述序列的本质。然后根据提取的特征信息确定网络结构，经过正负样本的学习后，进行分类工作。对于学习的速度和效率根据采用学习算法的不同会有一定的差异。

应用到生物信息学的人工神经网络大多采用 BP 学习算法对网络进行训练。BP 算法是在多层前向网络中单向传播的算法，可以任意精度逼近任意连续函数，较为简单，实现方便。但是由于它是一个非线性优化问题，不可避免地存在着陷入局部极小问题以及收敛速度慢等问题。采用遗传算法(Genetic Algorithms，GA)对人工神经网络进行训练，可以很好地克服 BP 算法的诸多缺点[5]。遗传算法是一类通过模拟生物界自然选择和自然遗传机制的随机化搜索算法，由美国 J. Holand 教授于 1975 年首次提出。它利用某种编码技术作用于称为染色体的二进制数串，其基本思想是模拟由这些串组成的种群的进化过程，通过有组织地然而是随机地信息交换来重新组合那些具有更好适应性的串。遗传算法对求解问题的本身一无所知，它所需要的仅是对算法所产生的每个染色体用适应度函数进行评价，并根据适应性来选择染色体，使适应性好的染色体比适应性差的染色体有更多的繁殖机会[4]。遗传算法的基本操作有下面 4 个：编码（Coding）是将问题解的表示映射成遗传空间解的表示，即用字符串或位串构造染色体，其相反操作为解码；选择（Selection）操作是根据染色体的适应度，在群体中按一定概率选取可作为父本的染色体，选择的依据是适应度高的染色体被选中的概率大；交叉（Crossover）操作是按一定概率随机地选择染色体对，然

后对染色体对按一定概率随机地交换基因，以形成新的子染色体；变异（Mutation）操作是按一定概率随机地改变某个染色体的基因值[4]。算法的基本步骤如下：(1)编码后随即产生 $N$ 条染色体，形成初始种群。(2)通过计算每个染色体的适应度评价其优劣。(3)对当前种群进行选择、交叉、变异操作，产生下一代种群。(4)如果满足进化终止条件，达到一定的世代数或种群中最佳个体适应度达到一定水平，算法终止，输出最佳个体。否则返回第(2)步，进行新一代演算[4]。

Knudsen 等开发的 Promoter2.0[6] 就是采用了基于遗传算法的人工神经网络技术来预测脊椎动物 RNA 聚合酶 II 启动子序列的。Promoter2.0 用人工神经网络的方法确定 TATA 框、CAAT 框、加帽位点和 GC 框的位置和距离，4 个结构相同的人工神经网络分别识别以上 4 种元件。Promoter2.0 的神经网络结构如图 3 所示，只有一个输出神经元、一个隐层神经元和自定义固定窗口大小的序列输入神经元以及来自其他网络的影响信号的输入神经元。一个核苷酸由 4 个神经元确定，所以输入神经元共有(2+4×窗口大小)这么多个。每一轮学习需要对 4 个同样结构的网络用遗传算法进行学习：第一个网络进行学习找出这个序列中最活跃的位点(TATA 框)；第二个网络用另一组权值对原序列进行学习，并在输入中加入先前的活跃位点的影响，找出加帽位点；类似地再对第三、四个网络进行学习，分别找出 CAAT 框和 GC 框，一轮学习结束。然后按照遗传算法的步骤进行新一代的学习，如此反复。实验证明，在窗口大小为 6 时，经过大约 15000 代的学习，这个人工神经网络可以区分主要的启动子和非启动子，并且在分别采用其中 1 个、2 个、3 个和 4 个的网络的时候相关系数分别为 0.62，0.61，0.63，0.61[6]。

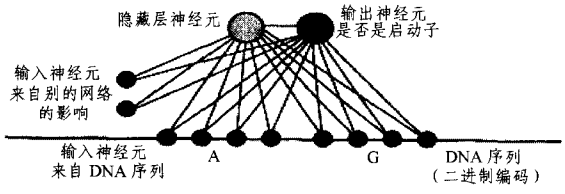


图 3　Promoter2.0 的网络结构

在国内，重庆大学的学者在启动子预测的研究中采用了和 Promoter2.0 相似的遗传算法和人工神经网络技术，所开发的系统对启动子序列能有效地进行识别，在训练集和测试集上的平均识别率分别高达 99%和 97%[5]。与 Promoter2.0 所不同的是，输入序列不是 DNA 序列，而是对 DNA 序列三元组经过处理后的数据。若直接输入三元组序列，64 个不同的三元组需要较多的隐层神经元来滤掉冗余信号，这样网络结构复杂并且学习速度慢。该系统采用三元组在每个区域的位置特异性频率分布进行打分，形成位置特异性权值矩阵。序列中每一位置的三元组在这个矩阵中的位置出现频率加和作为序列的特征参数，这样分别得到启动子、内含子和外显子的特征参数作为输入序列。系统所采用的人工神经网络结构为 3 个神经元的输入层、10 个神经元的隐层和 1 个神经元的输出层[5]。相比 Promoter2.0，由于采用了预先对 DNA 序列的加工，在每一代的学习中不用考虑上次学习的影响，学习速度更快，学习的代数更少，并具有很好的识别率。

### 3.2　基于支持向量机的启动子预测技术

支持向量机[7]（Support Vector Machines）是 Vapnik 等人在 1995 年提出的一类新型的建立在统计学理论之上的机器学习方法。支持向量机基于结构风险最小化原理，能够较好地解决小样本、高维数、非线性和局部最小点等问题，已经成为研究的热点之一[7]。由于其出色的学习性能，已经在人脸识别、文本分类、模式识别等领域得到了成功的应用。支持向量机的基本思想就是首先通过一个非线性映射把原本线性不可分的样本映射到高维特征空间，使之线性可分，再在这个高维特征空间里求解这些样本的最优分隔超平面来进行分类。因为许多生物问题的数据都是多维的，带有噪声并且输入的序列不一定等长，支持向量机的特点正好可以克服这些困难，在生物信息学应用方面具有得天独厚的优势。

支持向量机从线性可分情况下的最优分类面发展而来，应用于启动子预测的基本步骤也是对启动子序列正负样本进行学习，然后预测。首先对启动子序列正负样本进行特征提取，然后使用支持向量机把这些特征映射到高维空间中，使其线性可分。图 4 所示的实心点和空心点分别代表启动子和非启动子序列两类样本。图 4 表示一个高维空间，启动子和非启动子都在这个空间中，并且线性可分，H 为分类超平面，$H_1$，$H_2$ 分别代表各类中离 H 最近的样本且平行于 H 的面。如果这两类样本是线性可分的，机器学习的结果就是找到一个超平面，把样本分成两类。按照结构风险最小化原理，结果是一个最优的超平面，不仅能将两类样本分开，而且还要使分类间隔最大。分类间隔就是从分类超平面到两类样本中的最近的样本距离之和，如图中 $H_1$ 与 $H_2$ 之间距离，$H_1$ 与 $H_2$ 上确定了分类超平面的这些样本就是所谓的支持向量[8]。当样本集线性不可分的情况下，在确定超平面的条件中加入一个松弛项[8]，可得到与线性可分情况下一样的解。

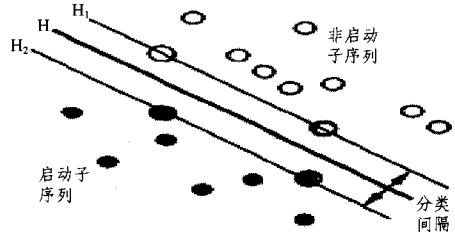


图 4　线性可分的分类超平面

在支持向量机中描述两个样本向量相似度的函数称为核函数(kernel)[8]。实际中难以线性可分的问题，分类样本可以通过选择适当的非线性变换映射到某个高维的特征空间，使得在目标高维空间这些样本线性可分，从而转换为线性可分问题[9]。对应到支持向量机技术中，在最优分类面中采用适当的核函数就可以实现某一非线性变换后的线性分类。有理论证明，支持向量机的支持向量集和核函数相关[10]。研究表明，非线性可分情况下，对于一个特定的核函数，给定的样本集中的任意一个样本都可能成为一个支持向量。这意味着在一个支持向量机下观察到的特征在其它支持向量机下(其它核函数)并不能保持。因此，对解决具体问题来说，选择合适的核函数是很重要的。常用的核函数主要有 4 种：线性核函数、多项式核函数、径向基核函数和二次神经网络核函数。

Rajeev Gangal 等开发的预测人类 RNA 聚合酶 II 的启动子工具 Prometheus[11] 就是采用了支持向量机技术。Prometheus 的识别精度可以达到 85%以上，并可以完全正确地识

别出人类 22 号染色体上所有的 20 个被体外实验验证的启动子。Prometheus 采用支持向量机技术，使用多项式核函数建立了两个模型：模型 1 输入包括多聚体的出现频率，CG 的百分比含量以及非线性时间序列描述符，模型 2 输入只包括多聚体的出现频率和 CG 的百分比含量。Prometheus 把一个 DNA 序列看成一个动态系统，把随时间连续观察，核苷酸的缺失、插入、变异产生的有顺序的观测值集合作为时间序列。支持向量机的机器学习过程采用 10-折交叉验证方法和独立的测试集，即所有数据的 80%用于训练，20%用于测试。然后训练集被分成 10 等份，其中 9 份用来训练，1 份用来测试。这样对这 10 份数据重复 10 次，最后用测试集进行测试。经过这样的学习，Prometheus 的敏感性达到 0.86，特异性达到 0.87，相关系数达到 0.74，是一个非常好的预测启动子工具[11]。

在国内，中国计量学院学者开发的预测人类核心启动子的系统[12]也采用了支持向量机的技术。在数据选取上，该系统采用了基于改进的位点权值矩阵(PWM)进行特征提取，然后筛选出含有特定启动子元件的作为正数据，再按上下文的比例选取负数据，最后划分成训练集和测试集。通过对比发现，该系统分别使用线性核函数、多项式核函数和径向基核函数时，都有较高的性能。在输入序列的编码上，该系统采用四维二进制码编码四种核苷酸，这种编码使每种核苷酸之间不具有线性相关性，适合支持向量机的特点。采用 SMO 训练算法的 LIBSVM 2.81 进行训练，在确定惩罚因子时采用 5-折交叉验证结合网格搜索法。实验证明，采用线性核函数，几种核心启动子识别的敏感性在 0.85 以上，特异性在 0.86 以上，相关系数在 0.77 以上；采用径向基核函数，几种核心启动子识别的敏感性在 0.84 以上，特异性在 0.86 以上，相关系数在 0.83 以上；采用多项式核函数，几种核心启动子识别的敏感性在 0.86 以上，特异性在 0.88 以上，相关系数在 0.80 以上[12]。特别地，含有 DPE(Downstream Promoter Element)的序列，用线性核函数效果比较好；含有 BRE(TFIIB Recognition Element)的序列，用多项式核函数效果比较好。综上可知，为达到理想效果，应该针对具体的问题采用不同的核函数。

### 3.3 基于二次判别分析的启动子预测技术

二次判别分析(Quadratic Discriminant Analysis)是一种经典的多元统计分析方法。根据它的原理建立的分类器形式简单有效，并且具有良好的扩展性。它可以通过特征筛选在备选特征中确定最优特征集，以提高模型的准确性和识别的正确率[13]。对于启动子预测来说，信息通过二次判别函数整合，给定启动子集合 $G_1$、非启动子集合 $G_2$ 以及由分类参数构成的 $r$ 维向量 $R(I_1, I_2, \cdots, I_r)$ 表示的待判别的序列 $S$，则判断 $S$ 类别的二次判别函数如下：

$$\xi=\log\frac{p0_1}{p0_2}-\frac{\delta_1-\delta_2}{2}-\frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|}$$

其中 $p0_i$ 是集合 $G_i(i=1,2)$ 的序列总数，即先验概率，$\delta_i=(R-\mu_i)^T\Sigma_i^{-1}(R-\mu_i)(i=1,2)$ 是 $R$ 和 $\mu_i$ 之间的马氏距离，$\mu_i$ 是训练集 $G_i$ 所有序列的 $r$ 维向量的平均，$\Sigma_i$ 是训练集 $G_i$ 的 $(r\times r)$协方差矩阵，$\Sigma_i$ 是矩阵 $\Sigma_i$ 的行列式值。

由 Ramana 等人开发的预测第一个外显子和启动子的软件 FirstEF[14]采用了二次判别分析技术，程序和 CpG 岛信息相结合，把序列分成 CpG 岛相关和不相关两类后再进行预测。FirstEF 搜索通过 5'-UTR 定位技术构建的第一外显子数据库，识别第一剪切点(first splicing donor site)，并结合 CpG 岛信息，确定启动子区，因此使用了 3 种不同的二次判别函数。这种方法使预测的敏感性和特异性都明显提高，下面仅对 FirstEF 的启动子预测过程进行介绍。FirstEF 首先扫描输入序列，确定是否现有潜在的剪切区，该过程通过一个二次判别函数给出可能性，再扫描确定是否与 CpG 岛相关。FirstEF 用一个 570bp(转录起始位点上游 500bp 和下游 70bp)的滑动窗口扫描候选的剪切区上游 1500bp。根据是否是 CpG 岛相关，通过两个不同的二次判别函数判断是不是启动子[14]。该程序预测含 CpG 岛的启动子的敏感性和特异性都高于 0.90，预测不含 CpG 岛的启动子的精确性相对略低。

国内由内蒙古大学学者开发的预测人类 RNA 聚合酶 II 启动子的系统同样采用了二次判别分析的技术，不同的是该系统采用基于多样性增量(Increment of Diversity，ID)的二次判别分析方法(IDQD)。多样性增量是衡量具有相同尺度特征条件下样本相似性的方法，最早在内含子剪接位点的预测问题中提出，并获得了成功应用。IDQD 的中心思想是信息经 ID 处理后用 QD 进行整合，最后按贝叶斯后验概率排序，寻找最佳分类点。由于信息的维数很高，这也是把分类特征向高维空间投影的一种识别算法[15]。多样性增量增加了除序列本身外的其他特征，即把特征向量扩展到了高维空间。该系统用到的多样性增量信息参数包括反映序列组分的特征、反映序列位点保守性的特征和反映 GC 含量变化的特征，另外包括两个非多样性增量的信息参数：反映序列具有某种读码框的非均匀指数和正链反链的高频数的 56 个模体序列频数。该系统预测结果的敏感性、特异性和关联系数分别达到了 0.93，0.92 和 0.85[15]。

### 3.4 基于位置权值矩阵的启动子预测技术

位置权值矩阵(Position Weight Matrix，PWM)用位置加权矩阵来表示一组已知位点样本，然后用该矩阵搜寻未知的位点，利用该矩阵给转录因子结合位点的各位置上每种可能出现的核苷酸都分配一个权值。对于一个特定的序列，将各位置的相应分数加和后给出该序列作为潜在位点的得分。通过与一个事先确定好的阈值进行比较，来判断该特定序列是否为位点。在某些情况下，这些分数与转录因子的结合近似成正比。1990 年 Bucher 最早提出用 PWM 的方法预测启动子[16]，该方法的原理是基于 Berg 和 Von Hippel(1987)提出来的关于调控蛋白(转录因子)选择 DNA 结合位点的理论[17]：为了解释在 DNA 序列中许多位点所具有的简并性，Berg 和 Von Hippel 提出特异性结合位点的选择是要满足一定功能性限制的，例如结合的亲和力或者活性必须在一定的允许范围内，所有能够满足这个要求的序列都可能成为特异性转录因子结合位点。这个理论已经通过实验手段检验，成为 PWM 方法的理论基础。PWM 方法大致思路如下：首先获得一组已知的转录因子结合位点集合，将该集合作为训练集，然后计算得到该组集合的位置加权矩阵，再利用优化程序，确定结合位点，判定阈值。最后对于任意 Lbp 长的 DNA 输入序列，利用得出的位置加权矩阵评估该段序列与训练集合中的转录因子结合位点的相似性。这个 DNA 序列的总得分为序列中每一位置碱基的得分之和作为其总得分，任何得

分超过特定阈值的序列就认为是一个转录因子结合位点。

PWM 技术用于启动子预测已经有十多年时间。十几年间，PWM 经过了多次优化，EPD 上也更新了多次 TATA 框的 PWM。其中比较有影响力的是俄亥俄州立大学的 Ilya 和 Gershenzon 等人在 2005 年对 GC 框的 PWM 的改进[18]。Gershenzon 修改了 Bucher 在 1990 年的矩阵中权值的公式，定义了一个平滑函数，代替了原来公式中的平滑系数，并且在优化阈值的时候使用相关系数代替原来的局部过表达参数来优化阈值。算法首先输入原来的 PWM，和在 TRANSFAC 中得到的 GC 框结合因子 SP1 的经过实验验证的结合位点序列，以及一个初始的敏感度，然后用新定义的权值计算公式来优化这个 PWM，并利用相关系数优化阈值和 GC 框的长度。使用这种方法，使 GC 框的敏感度由原来的 0.68 提高到了 0.73[18]。Gershenzon 等人根据这些优化，设计了启动子分类工具 Promoter Classifier[18]，这个工具已被广泛用到启动子分析预测工作中来。

国内，清华大学的学者用改进的 PWM 技术扫描肝脏特异基因的启动子序列[19]。该方法采用 3 阶 Markov 过程来描述背景集合的核酸分布情况，通过统计背景集合中 4 种核酸的 3 阶相关性，得到 3 阶 Markov 概率转移矩阵，生成新的启动子序列。有研究表明，考虑核酸序列之间相关性情况下，采用 Markov 随机过程来描述这一背景集合的自然特性[19]。构建 PWM 的方法和传统的 Bucher 方法类似，得到矩阵后，采用阈值累加法对整个序列转录因子 PWM 扫描结果进行评分，从而更全面地评价所选序列的结合性能。将转录因子在完全随机序列（指一条各个位置上 4 种碱基等概率分布的序列）上的结合性能作为噪声的评估标准。假定扫描结果大于噪声标准 10 倍的结合位点是有意义的，并以此标准筛选 PWM 扫描结果，将有意义的结果累加求和。这样克服了最大值方法导致一些有意义的结合位点被忽略的缺点，又能有效地抑制噪声位点的干扰，从而可以比较准确地衡量一条序列上转录因子结合位点的整体结合情况。而采用支持向量机法进行分类，经过交叉检验，阈值累加法得到的分类准确率为 93.33%[19]。

## 4 Analog-Cell 中的启动子预测技术

Analog-Cell[20,21] 作为模拟细胞基因表达过程的电子细胞，是我们开发研制的目前国内第一个虚拟真核细胞模型。它以图形化的方式成功模拟了 DNA 序列转录得到 mRNA，mRNA 翻译得到相应多肽链这一基因表达过程。该多肽链是各种酶或其它调控因子的前体蛋白质，它经过加工将会最终影响整个细胞的活动，并在转录和翻译期间实现了部分基因表达调控机制。研究显示，基因的表达调控多是在转录水平上进行，而转录水平上的基因调控很大程度上依赖于基因启动子区域上的具有特定调控作用的转录因子对 DNA 序列的识别和结合情况。可见，启动子区域上的转录因子结合位点的分布对于真核基因的表达调控起着十分重要的作用。因此，在 Analog-Cell 中模拟启动子识别的功能可以更逼真地描述细胞的基因表达过程。

综合比较以上 4 种启动子识别技术，其中基于 PWM 的方法实现简单，效率较高，因此在 Analog-Cell 中，我们采用 PWM 的技术识别核心启动子元件、TATA 框。首先，我们把待表达的序列用训练好的 PWM 进行扫描。如果有超过阈值的位点，我们就认为这段序列有启动子，可以进行转录。否则，就认为没有，不能进行转录。我们采用的 TATA 框的 PWM 是基于 Bucher 方法训练得到的矩阵，如表 1 所示。

表 1 TATA 框的 PWM[16]

| Pos | W(A) | W(C) | W(G) | W(T) | Cons |
|---|---|---|---|---|---|
| −3 | −1.02 | −0.28 | 0.00 | −1.68 | g/c |
| −2 | −3.05 | −2.06 | −2.74 | 0.00 | t |
| −1 | 0.00 | −5.22 | −4.28 | −2.28 | A/t |
| 0 | −4.61 | −3.49 | −4.61 | 0.00 | T |
| 1 | 0.00 | −5.17 | −3.77 | −2.34 | A/t |
| 2 | 0.00 | −4.63 | −4.73 | −0.54 | A/T |
| 3 | 0.00 | −4.12 | −2.65 | −3.65 | A |
| 4 | 0.00 | −3.74 | −1.50 | −0.37 | a/t |
| 5 | −0.01 | −1.13 | 0.00 | −1.40 | g/a |
| 6 | −0.94 | −0.05 | 0.00 | −0.97 | g/c |
| 7 | −0.54 | 0.00 | −0.09 | −1.40 | c/g |
| 8 | −0.48 | −0.05 | 0.00 | −0.82 | g/c |
| 9 | −0.48 | −0.11 | 0.00 | −0.66 | g/c |
| 10 | −0.74 | −0.28 | 0.00 | −0.54 | g/c |
| 11 | −0.62 | −0.40 | 0.00 | −0.61 | g/c |
| 位点 0 常在 TSS 上游−20 到−36 范围内 | | | | | |
| 优化后阈值：−8.16 (97%) | | | | | |

我们知道，转录是从转录起始位点开始的，识别出启动子后的工作就是确定转录起始位点的位置，这其实是另一个研究领域的问题，但是核心启动子 INR 正好在转录起始位点附近，我们在 Analog-Cell 中就可以利用 INR 的 PWM 来预测序列的转录起始位点，我们认为 INR 的保守序列的第三位是转录起始位点，INR 的 PWM 如表 2 所示。

表 2 INR 的 PWM[16]

| Pos | W(A) | W(C) | W(G) | W(T) | Cons |
|---|---|---|---|---|---|
| −2 | −1.41 | −1.61 | −0.75 | 0.00 | t/g |
| −1 | −5.26 | 0.00 | −5.26 | −5.26 | C |
| 0 | 0.00 | −5.21 | −5.21 | −2.74 | A/t |
| 1 | −1.51 | −0.41 | 0.00 | −0.29 | g/t/c |
| 2 | −0.65 | −0.45 | −4.56 | 0.00 | T/c/a |
| 3 | −0.55 | 0.00 | −0.86 | −0.36 | c/t |
| 4 | −0.91 | −0.29 | −0.38 | 0.00 | t/c/g |
| 5 | −0.82 | −0.18 | −0.65 | 0.00 | t/c |
| 位点 0 常在 TSS 上或下游+5 范围内 | | | | | |
| 优化后阈值：−3.75 (81.4%) | | | | | |

**结束语** 经过以上介绍，了解了基于人工神经网络、支持向量机、二次判别分析和位置权值矩阵的 4 种启动子预测技术，并结合国内外实际的 8 个例子看到了这些技术的效果。可以说，这 4 种技术都可以达到比较好的效果。其中，人工神经网络技术应用得比较早，技术比较成熟，关于这种技术的资源比较丰富，也易于理解，与遗传算法相结合可以更好地应用于生物信息学方面。支持向量机的研究比较晚，但是由于它克服了传统机器学习的种种缺点，并且在增加维数的基础上不增加计算量，已经在各个领域获得了成功的应用，相信支持向量机仍然具有很大潜力[10]。二次判别分析来源于贝叶斯理论，有很强的数学背景，虽然难于理解，但是效果很好，内蒙古大学学者的系统结果已证明好于 Rajeev Gangal 的 Prometheus[15]。位置权值矩阵的方法实现简单，应用范围广，有着浓厚的统计学特征，而且有研究发现 PWM 用在某种特异启

（下转第 33 页）

工程的研究内容划分成6个层次，每个层次增加了新的侧重领域，如表2所示。研究的方向按照由上到下的层次逐步增量地扩大，研究目标的粒度由教育级逐渐过渡到产品最终达到教育软件战略级的发展方向。该结构说明在不同的层次方向，研究的范围侧重和应用各有不同。

**结束语** 教育软件工程是一种非常重要的新的交叉领域。它使得软件工程和教育成为一个协同体，用教育和软件的视角来分析和处理宏观教育软件行业和微观教育软件关系单元和产品的题目，对于教育软件产品决策和教育软件产业分析具有重要的现实意义。教育软件工程对于未来教育软件产业将产生深远影响。

## 参 考 文 献


[1] 陈莉. 我国中小学教育软件资源建设的现状分析与建议. 中国电化教育，2002(3)：46-49

[2] 郑永柏. 中国教育软件发展的过去、现在和未来. 中国远程教育，2001(4)：61-62

[3] 李克东，费玉珍. 美国中小学生远距离教育和计算机辅助教学考察报告. 中国电化教育，1997(1)：9-13

[4] 李怀璋，李明树. 基于ISO9000和CMM的软件过程改进及其实践工程. 软件学报，2002，13(增刊)：113-119

[5] 艾伦，王陆. 教育与信息技术的交叉理科与工科培养模式的结合—教育技术学系教育软件工程方向的建设思路. 电化教育研究，2003(7)：23-24

[6] 李昭智，卞蔡，于长云，等. 时代呼唤教育软件工程学. 宁夏大学学报：自然科学版，1996，17(增刊)：3-8

[7] 袁宏伟. 谈教育软件的开发-在第四届全国计算机应用联合学术会议上的交流报告. 计算机系统与应用，1997(11)：54-55

[8] 王万良，王陆，杨卉. 教育软件工程专业人才培养模式的探讨. 电化教育研究，2006(2)：39-42

[9] 胡礼和. 现代教育技术学. 武汉：湖北科学技术出版社，2000

[10] 李昭智，于长云. 教育软件开发的基本原理与准则的探讨. 微小型计算机开发与应用，1996(6)：7-10

[11] 师书恩. 计算机辅助教育. 北京：北京师范大学出版社，1995

[12] 李运林，李克东. 电化教育导论. 北京：高等教育出版社，1986

[13] 李克东，费玉珍. 美国中小学生远距离教育和计算机辅助教学考察报告. 中国电化教育，1997(1)：9-13

[14] 何克抗. 计算机辅助教学研究与发展. 北京：高等教育出版社，1996

[15] 师书恩. 计算机辅助教育基本原理. 电子工业出版社，1995

[16] 桑新民. 当代信息技术在传统文化—教育基础中引发的革命. 教育研究，1997(5)：18-23

[17] 方海光. 软件工程经济研究和发展趋势. 计算机工程，2006(18)：3-6

[18] Boehm B W，Sullivan K J. Software Economics：a Roadmap // Proceedings of the ICSE22. 2000：7-11


(上接第9页)

动子预测上具有较高的效率[19]。

但是目前启动子的预测尤其是真核启动子的预测由于以下原因还是一个巨大的挑战：收录的已知体外实验验证的启动子序列太少，很多基因的启动子没有被发现；收录的基因序列的上游序列一般是不完整的；真核生物的启动子比原核生物的复杂得多；尽管有些启动子在DNA上是客观存在的，但是无法用体外实验证明，因为启动子的功能还和其他很多因素有关。这个挑战无疑会带来许多新的惊奇。本文介绍的4种技术尽管在一定领域内预测启动子都能得到较好的结果，但没有一个通用的模型去预测所有物种、所有基因的启动子，那么不妨就先从预测某类基因的启动子序列开始，慢慢地积累各类启动子的知识。随着生命科学和计算机科学进步，可以预期对真核生物启动子预测工作肯定会有所突破，新的生物信息学工具的研发也会有助于该问题的解答，最终这个挑战会迎刃而解。

## 参 考 文 献


[1] 姚凤霞，张瑞芳，刘春宇，等. 真核生物RNA聚合酶II启动子的计算机预测[J]. 国外医学遗传学分册，2005，8(1)：6-9

[2] P. C. 特纳，A. G. 麦克伦南，A. D. 贝茨，等. 分子生物学[M]. 第二版. 刘进元，李文君，王薛林，等译. 北京：科学出版社，2000

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

[4] 周春光，梁艳春. 计算智能[M]. 长春：吉林大学出版社，2005

[5] 熊清，王远强，李志良. 基于遗传神经网络的启动子识别系统[J]. 生物医学工程学杂志，2006，23(4)：730-733

[6] Knudsen S. Promoter2. 0：for the recognition of PolII promoter sequences [J]. Bioinformaties，1999，15(5)：356-361

[7] Vapnik V N. The nature of statistical learning [M]. New York：Springer-Verlag，1995

[8] 祁亨年. 支持向量机及其应用研究综述[J]. 计算机工程，2004，30(10)：6-9

[9] Zhang Ling. The Relationship Between Kernel Functions Based SVM and Three-layer Feed Forward Neural Networks[J]. Chinese J. Computer，2002，25(7)：1-5

[10] Zhang Ling，Zhang Bo. Relationship Between Support Vector Set and Kernel Functions in SVM [J]. J. Comput. Sci. &-Technol.，2002，17(5)：549

[11] Gangal R，Sharma P. Human polII promoter prediction：time series descriptors and machine learning [J]. Nucleic Acids Research，2005，33(4)：1332-1336

[12] 徐文韬，叶子弘，俞晓平. 基于支持向量机(SVMs)的人类核心启动子的识别[J]. 安徽农学通报，2006，12(13)：64-66

[13] 杜耀，王正，倪青，等. 一种基于特征筛选的原核生物启动子判别分析方法[J]. 生物物理学报，2006，22(1)：39-48

[14] Ramana V D，Gross I，Zhang M Q. Computational identification of promoters and first exons in the human genome [J]. Nat Genet，2001，29(4)：412-417

[15] 吕军，罗辽复. 人类polII启动子的识别[J]. 生物化学与生物物理进展，2005，32(12)：1185-1191

[16] Bucher P. Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 5023 unrelated promoter sequences[J]. J. Mol. Biol.，1990，212：563-578

[17] Berg O G，von Hippel P H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters[J]. J. Mol. Biol.，1987，193：723-750

[18] Gershenzon N I，Stormol G D，Ioshikhes I P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites [J]. Nucleic Acids Research，2005，33(7)：2290-2301

[19] 荀靓，张利，甄一松，等. 基于PWM扫描算法的启动子区域统计分析[J]. 清华大学学报：自然科学版，2006，46(7)：1267-1270

[20] Lu Xinhua，Sun Jigui，Ren Ying，et al. The Regulation of Gene Expression in E-Cell[C] // Proceedings of Third International Conference on Natural Computation (ICNC 2007). Volume 3，Los Alamitos：IEEE Computer Society Press，2007

[21] 卢欣华，孙吉贵. Analog-Cell：一种新的电子细胞图形模型[J]. 电子学报，2007，35(1)：49-53