基于"快速投票"算法的 HMM/SVM 混合识别模型及应用*)

罗泽举[1] 朱思铭[2]

(重庆工商大学计算机科学与信息工程学院 重庆 400067)[1] (中山大学数学与计算科学学院 广州 510275)[2]

摘 要 提出一种基于隐马尔可夫模型(HMM)和支持向量机(SVM)的双层过滤识别系统。根据隐马尔可夫模型训练中不同结构的序列其 L 值分布范围不同的特点,对传统多类"投票模型"进行改进,提出一种"快速投票"算法。先用 HMM 对人类内含子和外显子进行识别,同时,对于 L 值区域有重叠造成识别率较低的部分,再用支持向量机进行第二次识别过滤。这一模型克服了传统用单一 HMM 识别方法的不足,实现了 HMM 和 SVM 的优势互补。实验表明,用 HMM/SVM 进行两类识别,其平均识别率达到了 90%,进行多类识别,平均识别率达到了 91.5%。

关键词 HMM/SVM 模型,"快速投票"方法,内含子和启动子识别

An HMM/SVM Mixed Recognition Model Based on "Fast Voting" Algorithm and Application

LUO Ze-Ju[1] ZHU Si-Ming[2]

(School of Computer Science and Information Engineering, Chongqing Technology and Business University, Chongqing 400067)[1]

(School of Mathematics Computational Science, Sun Yat-Sen University, Guangzhou 510275)[2]

Abstract Propose a kind of HMM/SVM double layer filter recognition system. According to the characteristic that the L value is different while the sequence structure is different in the HMM training, improve the traditional "voting model", put forward a "Fast Voting" algorithm. First, use HMM models to recognize human intron and exon, meanwhile, for the part that L value range has part overlapped, use SVM for the second recognition. This model overcomes the deficiency of the traditional single HMM recognition method, realizes the mutual supplement with each other's advantages of HMM and SVM. Experiment indicates that carrying on recognition of two kinds of questions with HMM/SVM, the average rate of accuracy is up to 90%, carrying on multiple classifying, the average rate of accuracy is up to 91.5%.

Keywords HMM/SVM models, "Fast voting" algorithm, Intron and promoter recognition

1 引言

外显子是基因上用于编码一个完整蛋白质的特定 DNA 序列,在断裂基因及其初级转录产物上出现,并表达为成熟 RNA 核酸序列;内含子是基因上的非编码序列,在断裂基因及其初级转录产物上出现,而在剪接过程中被除去的核酸序列,不参与蛋白质的编码组成。启动子是位于结构基因 5 端上游的一段 DNA 序列,能够指导全酶(holoenzyme)同模板正确结合,活化 RNA 聚合酶,启动基因转录。鉴别和区分内含子,外显子及启动子序列是后基因组时代生命科学的重要课题。了解 RNA 剪接方式的不同,排除蛋白质编码过程中的噪声,对于了解基因的功能结构,探索生命的起源从而最终解码生命都具有重要意义。

隐马尔可夫模型和支持向量机均基于统计学习理论,它们都是当前机器学习的重要模型,在实际中都有广泛的应用,目前在语音识别[1,2]、图像识别[3,4]等领域都有很好的实验结果;支持向量机基于严格的统计学习理论,在分类和识别中有自己突出的优势;而隐马尔可夫模型由于能很好地模拟生物的进化过程,使它在生命科学特别是生物信息领域独受欢迎。但是它们有各自不同的算法特点。

本文利用隐马尔可夫模型(Hidden Markov Models, HMM)训练中 L 值分布的特殊性,对传统 HMM 分类识别中的"投票模型"进行改进,利用"快速投票"算法,将内含子和外显子区分;同时,对于 L 值区域有交叉、识别率较低的部分,再用支持向量机(support vector machine, SVM)进行进一步过滤识别,这一方法克服了单一分类方法的不足,达到了良好的识别效果。

2 隐马尔可夫模型

2.1 HMM 模型

隐马尔可夫模型由五元组 $\lambda=(S,\Sigma,A,B,\pi)$ 构成,其中 $S=\{S_1,S_2,\cdots,S_N\}$ 为状态集合,$\Sigma=\{O_1,O_2,\cdots,O_M\}$ 是观察符号或观察向量的集合,$A=(a_{ij})$ 为状态转移概率矩阵,其元素 a_{ij} 表示从状态 S_i 转移到状态 S_j 的转移概率,记为 $a_{ij}=P(q_{t+1}=S_j|q_t=S_i)$,$1\leqslant i,j\leqslant N$;$B=(b_j(k))$ 表示在状态 S_j 时产生观察符号 $v_k\in\Sigma$ 的离散概率值(v_k 为离散符号)或连续概率密度(v_k 是连续的观察矢量)矩阵:$b_j(k)=P(v_k|q_t=S_j)$,$1\leqslant j\leqslant N$,$1\leqslant k\leqslant M$;$\pi=(\pi_j)$ 是初始状态分布矩阵,其中:$\pi_j=P(q_1=S_j)$,$1\leqslant j\leqslant N$. s. $t\pi_j\geqslant 0$,$\sum\pi_j=1$。隐马尔可夫模型中最重要的是模型参数是 A,B,π,一般将模型简记为 $\lambda=(A,B,\pi)$。当我们用 HMM 模型进行识别的时候,首先用 EM 算法[5]对参数 A,B,π 进行重估,然后利用新参数来计算产生当前序列概率的对数值(log likelihood, L 值),再根据 L 值来识别对象所属的类。

*)国家自然科学基金资助项目(No. 10371135)。罗泽举 博士,主要研究方向:机器学习与模式识别,生物信息学。朱思铭 教授,博士生导师,主要研究方向:应用数学、常微分方程、计算机应用。

2.2 向前算法的改进[6]

由模型 λ 产生序列 $O_1^* O_2^* \cdots O_k^*$ 的概率是：

$$P(O_1^* O_2^* \cdots O_k^* | \lambda) = \sum_{allpath} \pi_1 b_1(O_1^*) a_{12} b_2(O_2^*) \cdots a_{k-1k} b_k(O_k^*) \quad (2\text{-}1)$$

产生序列 $O_1^* O_2^* \cdots O_k^*$ 所需计算量是 $O(kN^k)$，若 $N=10$，观察序列长度是 $k=100$，则 10^{100} 级的计算量计算机是根本吃不消的！为此必须对算法进行改进，定义向前变量 $\alpha_t(i)$ 如下：

$$\alpha_t(i) = P(O_1^* O_2^* \cdots O_t^*, q_t = S_i^* | \lambda) \quad (2\text{-}2)$$

故关于评估问题 $P(O^* | \lambda)$ 算法的可以改进为：

①初始化：$\alpha_1(i) = \pi_i b_i(O_1^*)$，$1 \leqslant i \leqslant N$， (2-3)

②迭代向前：$\alpha_{t+1}(i) = (\sum_{i=1}^{N} \alpha_t(i) a_{ij}) b_j(O_{t+1}^*)$，

$1 \leqslant t \leqslant k-1$，$1 \leqslant j \leqslant N$ (2-4)

③终止：$P(O_1^* O_2^* \cdots O_k^* | \lambda) = \sum_{i=1}^{N} \alpha_k(i)$， (2-5)

由此可知，改进后的算法，其运算量减少为 $O(kN^2)$，比起改进前的 $O(kN^k)$，其减少的量级是指数级的。

设当前模型为 $\lambda=(A,B,\pi)$，重估后的模型为 $\lambda^* = (A^*, B^*, \pi^*)$。Baum 等证明[5]只要：$\max_{\lambda^*} Q(\lambda, \lambda^*) = \sum_{Q} P(Q|O,\lambda) \log\{P(O,Q|\lambda^*)\}$，则有 $P(O|\lambda^*) \geqslant P(O|\lambda)$，且算法收敛到一个局部最大值。其中 $O = O_1^* O_2^* \cdots O_n^*$ 为观察序，$Q = q_1 q_2 \cdots q_n$ 为状态序列，表明由重估后模型生成观察序列 O 的概率比用当前模型生成观察序列 O 的概率要大，从而计算的 L 值更加可信。

由于不同的 DNA 家簇有着不同的结构，因而有不同的 L 值范围，再根据 L 值范围区分内含子和外显子；而且可以根据 L 值的多个范围可以同时识别多类 DNA 序列，这是隐马尔可夫模型独具的特点。SVM 利用函数只将数据分为两类，而不是 HMM 算法中的根据 L 值的多个范围可以同时划分多类数据，这是两类算法显著的不同点。

3 HMM 的“快速投票”算法

传统 HMM 进行多类识别的方法是采用所谓“投票原则”，算法是事先根据每一类单独训练出一组模型参数，即若有 N 类数据，就先训练出 N 组模型参数，再用各自的测试集进行“投票表决”；算法要进行 N 次训练，N^2 次投票，共 $N+N^2$ 次，复杂度为 $O(N^2)$，我们可以把这类投票模型叫做“多票决定制”，因为每一类在决定分类时都要投 N 次票后才能决定所属类。

本文我们将传统投票模型进行改进，使得每一类在决定分类时只投“一次票”便可决定它的所属类别，我们把它叫做“快速投票”算法。其算法步骤是：

Step1：只训练出一类 DNA 序列；

Step2：其它类的数据一部分作为训练数据用 Step1 的模型计算 L 值，得出各类不同的 L 值范围；

Step3：其它类的测试数据再用 Step1 的模型计算 L 值，看 L 值落在什么范围(投票)；

Step4：根据 Step3 识别和分类；

可见我们的模型训练 $1+N$ 次，投票 N 次，共 $2N+1$ 次，复杂度减少为 $O(N)$。

4 HMM/SVM 双层过滤系统

根据上节改进的“快速投票”算法，为了对内含子和外显子进行识别，我们建立如图 1 所示的双层过滤系统。

用双层过滤系统进行序列识别的步骤是：

Step1：用 EM 算法训练出内含子 DNA 序列的 HMM 模型；

Step2：用内含子、启动子序列的一部分数据作为训练数据由 Step1 的模型计算 L 值，得出各类不同的 L 值范围。

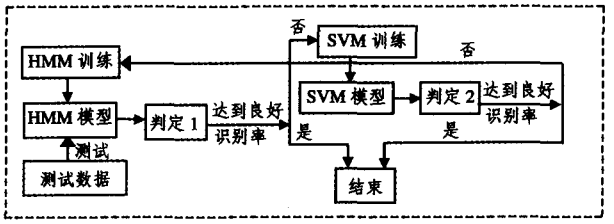

图 1 HMM/SVM 双层过滤模型

Step3：再用内含子、启动子序列的测试数据用 Step1 的模型计算 L 值，看 L 值落在什么范围(投票)；

Step4：根据 Step3 用 HMM 模型进行第一次识别和分类；

Step5：如果 Step4 的识别率达到要求，则转入 Step10，否则转入 Step6；

Step6：用训练数据进行 SVM 训练；

Step7：再用训练好的 SVM 模型对 Step4 中错分的数据进行第二次识别和分类；

Step8：如果 Step7 的识别率达到要求，则转入 Step10；否则转入 Step9；

Step9：用第二次仍然错分的数据加上第一次的训练集作为新的训练集，转入步骤 Step1；

Step10：停止迭代；

DNA 中外显子参与蛋白质编码，而内含子不参与，它们有不同的序列结构，从而决定 HMM 可以训练出不同的 L 值范围，达到有效的识别；然而由于 DNA 中碱基间隔相连，使得部分数据可能重叠，造成用 HMM 模型训练时产生 L 值的交叉，影响到识别率。但由于我们采用了两层过滤，将 HMM 的 L 值有交叉而 HMM 不能识别的部分再用 SVM 的最优分类超平面训练分隔，进行第二次数据过滤，就可以将识别率显著提高。

值得注意的是，如果经过多个循环的训练仍然未达到识别率的要求，则适当选取迭代次数停止迭代。

5 实验结果

5.1 DNA 序列的编码

DNA 序列都是以 A、G、T、C 进行横排的序列，而我们进行 HMM 训练需要的是特征向量，如何提取 DNA 序列的特征来装配向量是目前人们普遍关心的问题。但是什么是一个 DNA 序列的特征？S. Kundsen 等研究过根据序列 Motif 集作为编码[7]，也有人以基于序列相似性比对作为编码规则的[8]，但是这些编码模式并没有考虑序列本身的排列信息，转

移信息;不同于图像识别、语音识别,在提取 DNA 序列特征方面,目前还没有一个公认的描述序列特征的标准。

我们认为,在如此长的 DNA 序列中(例如人类核基因组 DNA 约有 3×10^9 个碱基对(bp)),有关序列的信息是直接隐藏在序列本身中的。由于原序列最能说明它自己的结构,有关排列信息、转移信息、碱基含量或以碱基位置等等,因此本文直接以序列本身作为装配对象, A、G、T、C 四个碱基分别用数字 1、2、3、4 表示,例如一个序列为:TAAGAGGTTGCT,可以用向量

X=(3,1,1,2,1,2,2,3,3,2,4,3)

来表示。而且这样表示正好和隐马尔可夫过程一致,表示刚好有四个观察符号的离散时间序列,通过隐马尔可夫训练,将其序列的信息反映在 L 值中,再根据其 L 值的分布状况来确定序列所属的类,从而来识别 DNA 序列。

5.2 建立识别内含子和外显子的两类识别模型

从美国俄亥俄州立大学生物信息学、蛋白质和遗传学的(bioinformatics and proteomics/genomics)BPG 内含子和外显子数据库(http://www.meduohio.edu/bioinfo/eid/index.html)下载大鼠内含子和外显子序列,该数据库含有完整的编码蛋白质的基因,基于 GenBank,版本为 built 2。从新加坡国立大学生物信息中心(NUS Bioinformatics Centre)的 Xpro 数据库下载真核生物(人类)的编码蛋白质 14 号染色体上的内含子和外显子序列(http://origin.bic.nus.edu.sg/),格式有 FASTA 和 XML 两种,文件基于 GenBank,版本 139.0,数据库共收集内含子和外显子序列 166555 个(2006-3-3 前)。我们从中各取 1000 条作为数据集,另外,也从 BPG 数据库下载几种收集极少的基因序列 Anadara trapezia β 球蛋白基因,α 毒素南极光基因,虾红素基因,伊蚊白基因的内含子和外显子序列各 5 条作为数据集。

由于有的基因序列非常长,我们取其中的 230 bps 作统一的长度,以方便可比性和训练,以人类内含子 800 条作为训练集,采用 200 个循环的训练,训练出 HMM 模型参数;采用"快速投票"算法,用余下的人类内含子和外显子序列各 200 条,大鼠内含子和外显子序列各 100 条作为测试集,另外,少数据基因序列 Anadara trapezia β 球蛋白基因,α 毒素南极光基因,虾红素基因,伊蚊白基因的内含子和外显子序列各 5 条也作为测试集。我们的算法可以叙述为"一类训练,多类测试"。

不同类的 DNA 内含子和外显子序列有不同的 L 值范围,这相当于定义了一个函数,将不同集合(类)的元素映射到不同的区间中去。利用这个显著特性,我们再用测试集进行测试,得到表 1 所示的识别结果。

从表 1 可以看出,当以 HMM 进行测试时,有的类识别率并不理想,例如只有 80%;但是当再一次进入 SVM 识别器进行二次识别后,就将识别率提高了平均 2 个百分点以上,说明 HMM/SVM 混合识别模型比起传统方法是非常有效的。

表 1 各类内含子和外显子序列的正确识别率

DNA 类型	测试数	正确识别率			
		第一次识别(HMM)(传统)		第二次识别(HMM+SVM)(新型)	
		intron	exon	inton	exon
Human	200	90%	89%	93%	92%
Mouse	200	88%	85%	90%	88%
Anadara trapezia beta globin gene	5	80%	60%	100%	80%
A. australis gene for alpha toxin	5	80%	80%	100%	100%
A. astacus astacin gene	5	100%	80%	100%	100%
Aedes aegypti white gene	5	80%	60%	80%	80%

注:SVM 分类器采用径向基核函数。

5.3 建立多类 DNA 序列分类模型

利用隐马尔可夫训练独有的多个 L 值范围进行多个 DNA 序列家簇的分类。从瑞士实验癌研究组织(The Swiss Institute for Experimental Cancer Research, ISREC)生物信息组的真核生物非沉余启动子数据库(The Eukaryotic Promoter Database, EPD)下载启动子序列数据,该数据库是基于欧洲分子生物实验室(European Molecular Biology Laboratory, EMBL),目前共收集有多个物种的 4809 个启动子序列(http://www.epd.isb-sib.ch/),(2006-3-3 前)其中脊椎动物共有 2540 个(包括人类启动子有 1871 个),植物质粒 198 个,另外,还有线虫(Nematode)、软体动物(Mollusc)、棘皮类动物(Echinoderm)等启动子序列。以人类启动子 100 个序列,选择植物质粒 100 个、线虫(Nematode)18 个序列、棘皮类动物(Echinoderm)20 个序列分别作为测试集,长度仍统一取 230 bps,以训练好的 HMM 模型进行 L 值测试和经过二层识别过滤,得到表 2 所示的分类结果,平均识别率达到了 91.5%。

HMM 迭代规定只要算法收敛就停止迭代,虽然我们选择了 200 个循环,但实际上平均只运行 73 步就收敛了,说明算法的收敛速度是很快的。图 2 为三条训练序列的收敛曲线图。

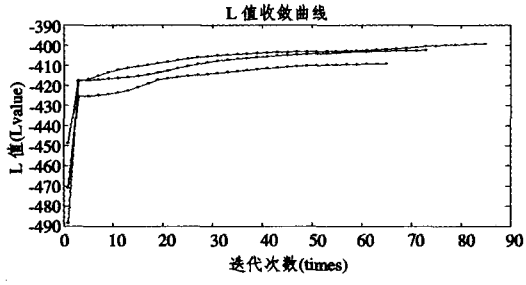

图 2 训练 L 值收敛曲线

从实验中我们惊奇地发现,棘皮类动物的启动子(Echinoderm promoters)第一次用 HMM 根本无法识别(表 2),因为它的 L 值范围完全包含在人类内含子训练的 L 值范围之内,有

(−420.2,−425.6)⊏(−402.3,−426.8).

可见,对于哪些 L 值重叠很多的序列,HMM 分类就将失效,但由于有 SVM 作第二层分类,可以将失效的部分挽救回来,这是 HMM/SVM 双层模型比单独用一种 HMM 的传统识别方法的优势。

(下转第 227 页)

参 考 文 献

1 边肇祺,张学工,等编著.模式识别[M](第二版).清华大学出版社,1999

2 Boser B Z, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers[A]. In: Proceedings of the 5th Annual ACM. Workshop on Computational Learning Theory[C]. Pittsburgh, PA, July ACM Press, 1992. 144～152

3 Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995

4 Schölkopf B, Smola A, Muller K R. Kernel principal component analysis[A]. In: W. Gerstner, ed. Artificial Neural Networks - ICANN'97[C], Berlin, 1997. 583～588

5 Mika S, Ratsch G, Weston J, Scholkopf B, Muller K R. Fisher discriminant analysis with kernels[A]. In: IEEE Neural Networks for Signal Processing Workshop[C], 1999. 41～48

6 Bach F, Jordan M I. Kernel independent component analysis[A]. Technical Report CSD-01-1166[R]. Computer Science Division, University of California, Berkeley, 2001

7 Bennett K P, Bredensteiner E J. Duality and geometry in SVM classifiers[A]. In: P. Langley, ed. Proceedings of the 17th International Conference on Machine Learning[C], San Francisco, California, Morgan Kaufmann, 2000. 57～64

8 Keerthi S S, Shevade S K, Bhattacharyya C, Murthy K R K. A fast iterative nearest point algorithm for support vector machine classifier design. Neural Networks[J]. IEEE Transactions on. 2000, 11(1): 124～136

9 邓乃扬,田英杰著.数据据挖掘中的新方法—支持向量机[M].科学出版社,2004

10 Mallat S. Mutiresolution Approximation and Wavelet Orthonormal Bases of L2(R)[J]. Trans. Amer. Math. Soc., 1989, 315: 69～87

11 Mallat S. A Theory for Multiresolution Signal Decomposition: the Wavelet Representation[J]. IEEE Trans. PAMI, 1989, 11(5): 674～693

12 Meyer Y. Ondelettes et Opérateurs[M]. Paris: Hermann Press, 1990

13 徐长发,李国宽著.实用小波方法[M].华中科技大学出版社,2004

14 Daubechies I. Orthonormal Bases of Compactly Supported Wavelets[J], Comm. Pure and Appl. Math, 1988, 41: 909～996

15 海金著,叶世伟,史忠植译.神经网络原理(原书第2版)[M].机械工业出版社,2004

16 王日爽编著.泛函分析与优化理论[M].北京航空航天大学出版社,2003

17 Kinderlehrer D. 变分不等方程及其应用[M].郭友中等译,科学出版社,1991

(上接第217页)

由表2看到HMM/SVM模型的提高的识别效果是显著的。但有的序列,例如Nematode promoters经过二次识别后虽然比第一次大幅度提高了识别效果,但仍然只有88.8%。笔者认为,有三种可能的原因,一是所训练的序列太少,导致L值范围的偏离,影响了识别效果;另一个是本身序列数据的原因,也许是提交的数据库序列存在测序上的误差;三是我们的模型还需进一步改进。

表2 双层识别模型进行DNA多家簇分类

DNA类	L值范围	测试数	第一次识别(传统)(HMM)		第二次识别(新型)(HMM+SVM)	
			正确识别数	识别率	正确识别数	识别率
Human intron	−402.3～−426.8	200	180	90%	186	93%
Man promoters	−433.3～−480.8	100	87	87%	91	91%
Plant promoters	−285.6～−305.9	100	88	88%	91	91%
Echinoderm promoters	−420.2～−425.6	20	0(无法识别)	0%	17	85%
Nematode promoters	−448.1～−453.7	18	15	83.3%	16	88.8%

近年来,SVM和HMMS都是模型识别的主流分支,SVM有它独有的小样本处理能力及基于核函数特征映射使之得到迅速发展;而HMMS由于能对基因序列符号进行编解码,其插入、删除、匹配正好模拟了基因的进化过程,使它在生命科学领域备受欢迎。国际上已经建立的著名的基于隐马尔可夫模型蛋白质家族识别数据库(基于多重序列比对)Pfam[http://pfam.wustl.edu/]就是一例。

讨论 本文利用隐马尔可夫模型训练中不同结构的序列其L值分布有不同范围的特点,对传统分类"投票模型"进行改进,提出一种HMM/SVM混合识别模型。先用HMM进行识别,同时,对于L值区域有重叠造成识别率较低的部分,我们再用支持向量机进行第二次过滤识别。该方法克服了传统单一HMM分类方法的不足,提高了识别效果。虽然隐马尔可夫模型也有许多不足,例如HMM只是线性模型、独立性假设(HMM假设观察符号之间是相互独立的,但它们之间可能相互依赖)、分类域值选取问题(过紧和过松的域值选取,会导致样本的少分或错分)、小样本(有的数据库仅有少数几条序列)训练带来的L值误差,这些缺点结合SVM可以得到弥补,因为SVM有极强的小样本处理能力,SVM欠缺的多分类能力又恰恰是HMM所具有的,所以HMM和SVM结合过滤,达到了优势互补。

模型参数的选取问题、如何科学地提取DNA序列的特征以及如何进一步优化模型仍是将来要继续研究的工作。

参 考 文 献

1 Vaseghi S V, Milner B P. Noise compensation methods for hidden Markov model speech recognition in adverse environments. Speech and Audio Processing, IEEETransactions on, 1997, 5(1): 11～21

2 Gordan M, Kotropoulos C, Pitas I. Application of support vector machines classifiers to visual speech recognition. Image Processing. In: 2002. Proceedings. 2002 International Conference on, June 2002, 3: 129～132

3 Lefevre S, Bouton E, Brouard T, Vincent N. A new way to use hidden Markov models for object tracking in video sequences. Image Processing, 2003. ICIP 2003. In: Proceedings. 2003 International Conference on Volume 3, Sept. 2003, 2(III-1): 17～20

4 Fu Y, Shen R, Lu H. Watermarking scheme based on support vector machine for colour images. Electronics Letters, 2004, 40(16): 986～987

5 Baum L E, Sell G R. Growth functions for transformations on manifolds. Pac. J. Math., 1968, 27(2): 211～227

6 Baum L E, Petrie T. Satistical inference for probabilistic functions of finite state Markov chains. Annmath. Stat., 1996, 37: 1554～1563

7 Kundsen S. Promoter 2.0: for the Recognition of Poll Promoter Sequences. Bioinformatics, 1999, 15: 356～361

8 Kasabov N, Pang S. TRANSDUCTIVE SUPPORT VECTOR MCHINES AND APPLICATIONS IN BIOINFORMATICS FOR PROMOTER RECOGNITION. In: IEEE int. Conf. Neural Networks & Signal Processing. Nanjing, China, December, 2003. 14～17