人类 pol Ⅱ 启动子的识别 *

吕 军[1)] 罗辽复**

(内蒙古大学物理学系，呼和浩特 010021)

摘要 依据基因启动子区和非启动子区碱基分布的特征，应用基于多样性增量的二次判别分析 (IDQD)，对人类 pol Ⅱ 启动子进行识别，识别精度达到 90%以上的水平，优于其他已发表的 (包括 SVM 分类器等) 识别算法. 使用 IDQD 算法也能对转录起始位点 (TSS) 进行较准确的预测，10-fold 交叉检验结果的敏感性和特异性分别为 86%和 91%. 这些结果表明 IDQD 是一个有效的分类器.

关键词 启动子，多样性增量，二次判别函数，转录起始位点

学科分类号 Q61

真核基因的识别问题一直是生物信息学的一个重要内容，基因启动子区的识别是完整基因结构识别中的重要一环[1, 2]. 人类启动子的识别是生物医学研究的基本需要，是构建基因调节网络的一个核心问题. 负责 mRNA 转录的 RNAP Ⅱ (pol Ⅱ) 启动子是数量最多且最重要的一类. 真核 pol Ⅱ 启动子的识别和预测，已有很多方法、软件及综述，如文献[3~19]及软件神经网络启动子预测器 (NNPP，http://www.fruitfly.org/seq_tools/promoter.html) [17]， Soft Berry (http://www.softberry.com)[13]， Promoter Scan (http://bimas.dcrt.nih.gov/molbio/proscan/) [18]， Dragon Promoter Finder (http://research..i2r.a-star.edu.sg/promoter)[6]， Promoter2.0 Prediction Server (http://www.cbs.dtu.dk/services/Promoter/)[5]等.

但识别效率都不能令人满意，所列 5 种软件对人类 pol Ⅱ 启动子预测的相关系数皆在 0.35 以下. 如果预测涉及转录起始位点 (TSS) 的定位，情况更差.

真核生物的 pol Ⅱ 启动子区含有丰富的转录因子结合位点(transcription factor binding sites，TFBS)，启动子序列基本上是由这些短序列组合而成，主要在 TSS 上游 1 kb 的范围内. 在 TSS 附近−60 bp 到 +40 bp 是核心启动子区，它对于精确转录是必须的最小单元[20]. 在启动子序列中普遍存在两个位置相对稳定的保守区，一个是 TSS 上游−30 bp 附近的 TATA 框，其序列为 TATA(A/T)A(A/T)，是起始复合物(preinitiation complex)的主要装配点. 另一个是转录起始点处的保守序列称为起始子(initiator，Inr)，其共有序列为 YYAN(T/A)YY(下划线处是 TSS). 当然，有些启动子没有 TATA 框，有些没有 Inr，或者二者都没有. 除了上述两个相对保守区之外，启动子区还包含大量的位置相对不固定的正向和反向模体序列(regulatory motifs)[21]. 以上这些序列特征是利用信息论方法识别启动子所必须考虑的主要信息来源.

从方法上看，有神经网络，隐马尔可夫链，支持向量机(SVM)等. 最近的文献[14]介绍了一种 Prometheus方法，使用 Tsallis 熵结合 SVM 对人类 pol Ⅱ 启动子进行识别，用 100 个人类 pol Ⅱ 启动子和 100 个 Intron 序列与现在通行的 5 个启动子预测软件进行了预测能力比较，证明 Prometheus 具有最高的识别能力，相关系数达 0.74. 看来，设计新的预测方法对于成功识别是当前研究工作的一个关键. 本文将采用基于多样性增量 (increment of diversity，ID) [22] 的二次判别分析 (quadratic discriminant analysis，QD)[23, 24] 方法(称为 IDQD)，该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[25]. IDQD 的中心思想是信息经 ID 处理后用 QD 进行整合，最后按 Bayes 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. 本文

*国家自然科学基金资助项目(90403010)和内蒙古自治区自然科学基金资助项目(200308020102).

** 通讯联系人.

1) 内蒙古工业大学物理系, 呼和浩特 010051.

Tel: 0471-4992676, E-mail: lfluo@mail.imu.edu.cn

收稿日期：2005-06-13， 接受日期：2005-07-28

应用此方法，用于真核启动子预测，结果表明，在使用与文献[14]相同的数据库以及相同口径的训练集和测试集的条件下，本文的算法识别精度明显优于 Prometheus 算法.

转录起始位点的精确定位是一个更大的难题. 文献[4] 用包含 24 个新确认的转录起始点的 18 个序列测试当时的所有程序，最多找出了一半的转录起始位点，假阳性率约为每千个碱基中一个. 在这个预测中，凡在实验 TSS 的 5′端 200 bp，3′端 100 bp 的范围内都算正确 (文献[16]的 TSS 预测也用此标准). 用我们的方法对 EPD[26]中标注的 TSS 进行了识别，识别结果优于现有算法.

1 数据和方法

1.1 数据集

我们使用的数据有三部分：第一部分人类启动子数据下载自真核启动子数据库第 82 版(The Eukaryotic Promoter Database Release 82， EPD，http://www.epd.isb-sib.ch/index.html)[26]. 从 EPD 下载了总数为 1 871，长度为 300 bp (TSS 上游 250 bp，下游 50 bp)的人类启动子序列，手工剔除了在此区间包含“N”的序列 26 个，剩余 1 845 个启动子序列，随机选取 1 000 个序列作为训练正集,与训练集独立的 800 个作为检验正集. 第二部分非启动子序列选自人类 22 号染色体的 CDS 和 Intron，该数据下载自 http://www.sanger.ac.uk/HGP/Chr22/，其中 CDS 和 Intron 分别随机选取长度为 300 bp 的 500 和 1 000 共 1 500 个序列作为训练负集，再随机选取与训练集独立的 Intron 序列 1 000 个作为检验负集. 第三部分取自文献[27]，包含人类 22 号染色体的经实验验证的 20 个启动子序列，其中 NCBI 登录号为“AB016655”的启动子序列因 mRNA 注释信息不明确，在本文中没有采用. 其余 19 个启动子序列作为另一个检验集.

1.2 序列结构的统计分析

对训练集长度为 300 bp 的 1 000 个启动子序列和 1 500 个非启动子序列的结构进行统计分析，结果表明，CDS 区 4 种碱基的分布明显呈现出三周期特征，且 GC 含量要稍高于 AT 含量，Intron 区 4 种碱基几乎等概率分布，十分接近随机分布，整个非启动子序列区域中的碱基分布起伏很小. 启动子区 4 种碱基的分布明显不同于非启动子区，整体 GC 含量明显高于 AT 含量，且随着向 TSS 的靠近 GC 含量逐渐增加. 在−30 ~ −20 bp 区间碱基分布有明显的起伏，显示该区域含有 TATA 框，在 +1 位点附近碱基分布也有明显的变化，显示该区域为 Inr 区(图略). 由此，序列的组分特征是识别启动子区与非启动子区的基本信息，特定位点的序列保守性特征是重要的识别信息. 为了更加明显地显示 GC 含量随位点的变化，我们又做了 GC 含量分布图(图 1)，这里可以明显地看出启动子区与非启动子区 GC 含量分布的不同，这是另一个重要的识别信息.

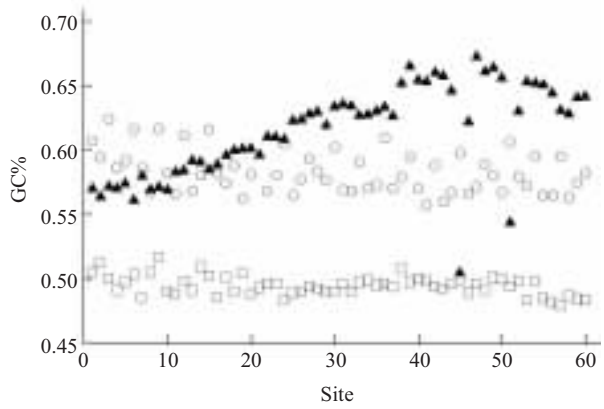

Fig. 1 G+C content distribution of promoter sequence and non-promoter sequence

The G+C content distributions for 1 000 promoter sequence (denoted as ▲), 500 CDS's (denoted as ○) and 1 000 introns (denoted as □) are given. Sequence length is taken to be 300 and divided into 60 intervals (each interval contains 5 sites). The percent of bases G and C in each interval is plotted.

1.3 方法

1.3.1 信息的多样性增量(ID)处理. 将样品的特征分为若干组. 一般说来，一组特征不是用样品本身就能表示清楚，而必须通过和大量标准样品(称为标准源)的比较来确定. 也就是说，由样品特征的多样性分布($D(X)$)和标准源特征的多样性分布($D(S)$)的比较来确定. 如特征具有 s 维，样品中第 i 个特征(例如某种碱基在某一位点上的数量)用整数 n_i 表示，n_i 为样品的信息参数 (i=1，…，s)；标准源中该特征用整数 m_i 表示，m_i 为标准源的信息参数 (i=1，…，s). 样品和标准源的多样性分别为：

$$D(X)=D(n_1,\ n_2,\ \cdots,\ n_s)=N\log_2 N-\sum_{i=1}^{s} n_i\log_2 n_i \qquad (1)$$

$$D(S)=D(m_1,\ m_2,\ \cdots,\ m_s)=M\log_2 M-\sum_{i=1}^{s} m_i\log_2 m_s \qquad (2)$$

$$\left(N=\sum_{i=1}^{s} n_i,\ \ M=\sum_{i=1}^{s} m_i\right)$$

两者的总多样性量 $D(X+S)$可由源$\{n_1+m_1,\ n_2+m_2,\ \cdots,\ n_s+m_s\}$类似地定义，而多样性增量 ID 定义为：

$$\mathrm{ID}(X,\ S)=D(X+S)-D(X)-D(S) \tag{3}$$

ID表征了样品 X 和标准源信息参数分布的差异性，它提供了样品 X 特征的数量表示. 可以证明，ID 值在区间$[0,\ D(N,\ M)]$中变化，

$$D(N,\ M)=(N+M)\log_2(N+M)-N\log_2 N-M\log_2 M \tag{4}$$

在应用中为使正集和负集能有效区分开，选择样品的特征时，要注意正集样品 X 的 ID$(X,\ S)$和负集样品 X 的 ID$(X,\ S)$取值尽量不同 (当 X 在标准源以外). 至于这个特征的具体选取，它可以是表征正集样品的量，也可以是表征负集样品的量，或者正集和负集都用它表征. 由于 s 可能很大 (信息参数的维数可能很高)，这个方法自然地包含了把分类特征扩展向高维空间的运算.

1.3.2 信息的二次判别函数(QD)整合. 上面考虑了一组特征的处理，实际问题中有 r 组特征，即有 r 个 ID，(当一部分特征可用简单数量表示，不需引入 ID，本节的理论形式同样适用)，须用二次判别函数把它们整合起来. 给定两个训练集合 G_1(启动子)和 G_2(非启动子)，对于任一待判别的序列 S，如果此序列由分类参数构成的 r 维向量 $\boldsymbol{R}(I_1,\ I_2,\ \cdots,\ I_r)$ 所表示，则判别其归属的二次判别函数由下面的(5)式给出.

$$\xi=\log\frac{p0_1}{p0_2}-\frac{\delta_1-\delta_2}{2}-\frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|} \tag{5}$$

其中 $p0_i$ 是集合 G_i (i=1，2，分别代表正集和负集)的序列总数，即先验概率，$\delta_i=(\boldsymbol{R}-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\boldsymbol{R}-\boldsymbol{\mu}_i)$ (i=1，2)是 $\boldsymbol{R}$ 与 $\boldsymbol{\mu}_i$ 之间的马氏距离，$\boldsymbol{\mu}_i$ 是训练集 G_i 所有序列的 r 维向量的平均，Σ_i 是训练集 G_i 的 $(r\times r)$ 协方差矩阵，$|\Sigma_i|$ 是矩阵 Σ_i 的行列式值. (5)式由 Bayes 理论导出，这里 ξ 是正负集后验概率比的自然对数. 如果特征 (信息参数) 选得合适，正负集可在 ξ 空间的 0 附近分得很开. 不过，由于正负集样本数的有限性，二者差异还可能很大，并且都可能对正态分布有偏离 (正态分布是导出(5)的前提条件)，两个集合在 ξ 空间的分界点不一定是 0，最佳分界值要由经验确定. 以上理论可从 2 分类推广到多分类，在 n- 分类情形，只须把(5)式中的 ξ 理解为 n 类中某 2 类的后验概率比的自然对数即可.

2 启动子识别

2.1 对启动子序列的识别

2.1.1 参数定义. 考虑到启动子和非启动子序列的结构特征，我们选取了如表 1 给出的信息参数，作为 IDQD 分类的信息源. 其中，

a. 反映序列组分特征的参数. 我们分别选取 −250 bp：−49 bp 的 6-mer 频次，定义 4 096 维向量 $\boldsymbol{X}_1$；−50 bp：−1 bp 的 6-mer 频次，定义 4 096 维向量 $\boldsymbol{X}_2$；0 bp：+49 bp 的 6-mer 频次，定义 4 096 维向量 $\boldsymbol{X}_3$.

b. 反映序列位点保守性的特征参数. 我们选取 −29 bp：−22 bp 的固定位点的 3 碱基频次，定义 $C_8^3\times64$ 维向量 $\boldsymbol{X}_4$；选取−1 bp：+3 bp 的固定位点的 3 碱基频次，$C_5^3\times64$ 维向量 $\boldsymbol{X}_5$.

c. 反映 GC 含量变化的特征参数. 我们选取 −250 bp：+49 bp 的每 5 bp 间隔的 GC 含量，定义 60维向量 $\boldsymbol{X}_6$.

d. 另外，定义 2 个非 ID 信息参数. 一个是反映序列具有某种读码框的非均匀指数 HI[28]，第二个是正链和反链的高频数的 56 个模体序列频数[21]，它们直接作为二次判别分析向量的 2 个元素. HI 的定义为[28]，

$$HI=\sum_{l=1}^{3}\sum_{\alpha=1}^{4}\frac{(N_\alpha^{\,l}-\frac{N^lN_\alpha}{N})^2}{\frac{N^lN_\alpha}{N}} \tag{6}$$

其中 N_α (α=A，C，G，T) 表示序列 4 种碱基数，$N=\sum_\alpha N_\alpha$，N^l (l=1，2，3)为 3 个子序列的长度，$N^l=\frac{N}{3}$，$N_\alpha^{\,l}$ 为第 l 个子序列中第 α 种碱基数.

$X_{1\sim6}$ 构成任意序列 X 的 6 个多样性源，相应地，由训练正集和负集的全部序列 S 也可建立 6 个多样性源，分别称为正集和负集标准源. 这样，我们可以依据(1)~(3)式去计算由任一序列 X 与正集 G_1 标准源和负集 G_2 标准源之间的 12 个多样性增量 $\mathrm{ID}_{1\sim12}$，作为序列 X 的分类参数，加上 HI 和模体频次共构成序列 X 的 14 维二次判别向量. 当 $X\in G_i$ (i=1，2)时，求得训练集序列的分类参数后，对集内全部序列求平均，便得到平均向量 $\boldsymbol{\mu}_i$ (i=1，2) 和协方差矩阵 Σ_i (i=1，2). 对于任一待识别序列 X，依据(5)式和 ξ 的阈值 ξ_0 就可以给出该序列的分类判别，当 $\xi>\xi_0$，序列 X 识别为真(启动子)，否则为假(非启动子).

Table 1 Definition of classification parameters (12 ID and 2 non-ID parameters)

Parameter	ID type	Source of information	Increment of diversity
I_1	$ID_1(4^6)$	6-mer frequency in −250 bp : − 49 bp	ID between X_1 and positive set
I_2	$ID_2(4^6)$	6-mer frequency in −250 bp : − 49 bp	ID between X_1 and negative set
I_3	$ID_3(4^6)$	6-mer frequency in −50 bp : −1 bp	ID between X_2 and positive set
I_4	$ID_4(4^6)$	6-mer frequency in −50 bp : −1 bp	ID between X_2 and negative set
I_5	$ID_5(4^6)$	6-mer frequency in 0 bp : +49 bp	ID between X_3 and positive set
I_6	$ID_6(4^6)$	6-mer frequency in 0 bp : +49 bp	ID between X_3 and negative set
I_7	$ID_7(C_8^3\times64)$	tri-nucleotide freq in −29 bp : −22 bp	ID between X_4 and positive set
I_8	$ID_8(C_8^3\times64)$	tri-nucleotide freq in −29 bp : −22 bp	ID between X_4 and negative set
I_9	$ID_9(C_5^3\times64)$	tri-nucleotide freq in −1 bp : +3 bp	ID between X_5 and positive set
I_{10}	$ID_{10}(C_5^3\times64)$	tri-nucleotide freq in −1 bp : +3 bp	ID between X_5 and negative set
I_{11}	$ID_{11}(60)$	G+C in 5 bp interval in −250 bp : +49 bp	ID between X_6 and positive set
I_{12}	$ID_{12}(60)$	G+C in 5 bp interval in −250 bp : +49 bp	ID between X_6 and negative set
I_{13}	—	non-homogeneous index *HI* in −250 bp : +49 bp	—
I_{14}	—	frequency of first 56 motifs in −250 bp : −1 bp	—

The second column gives the ID type of the parameter, the number in the bracket behind which denotes its dimension. The third column indicates the increment of diversity of X with positive or negative set. I_{13} and I_{14} are non-ID parameters.

2.1.2 结果

采用与文献[14]相同口径的训练集及检验集，即训练正集包含 1 000 个启动子序列，训练负集包含 1 500 个非启动子序列 (500 个 CDS 序列和 1 000 个 Intron 序列)，检验集包括 800 个启动子序列，1 000 个 Intron 序列，19 个经实验验证的人类 22 号染色体的启动子序列，检验集独立于训练集.使用 IDQD 识别算法，在阈值 $\xi_0 = 0$ 时，结果见表 2.

Table 2 The result of promoter recognition by use of IDQD

Predicted sequence	Seq number	True positive(TP)	False positive(FP)	True negative(TN)	False negative(FN)
Promoter	800	741 (707)	0 (0)	0 (0)	59 (93)
Intron	1 000	0 (0)	60 (97)	940 (903)	0 (0)
Human chromosome	19 (20)	17 (20)	0 (0)	0 (0)	2 (0)
22 experimentally verified promoters[1)]					

Numbers in bracket are results give by Prometheus. [1)]Twenty promoters of human chromosome 22 can be found in [27].

文献[14]又随机挑选 100 个启动子序列和 100 个 Intron 序列，和现在通行的 5 个启动子预测软件进行了预测能力比较，我们也参照此做法进行预测比较. 定义敏感性指标 S_n 和特异性指标 S_p 及相关系数 CC 如下：

$$S_n=[TP/(TP+FN)]\times100\%$$

$$S_p=[TN/(TN+FP)]\times100\%$$

$$CC=\frac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \quad (7)$$

比较结果列在表 3 中. 可以看出，本文的 IDQD 算法优于现有的其他识别算法. 另外我们还随机选取了 100 个 CDS 序列进行了预测，结果为 $TP = 0$，$TN = 97$，$FP = 3$，$FN = 0$，$Sp = 97\%$，$CC = 0.90$. 这里 IDQD 算法采用了 10 折交叉检验.

另外，为了进一步检验 IDQD 算法的预测能力，我们在保持训练正集不变的情况下，在训练负集中又加入了 1 500 条 Intergenic 序列(Intergenic 序列取自人 22 号染色体)，重新按照前述的方案进行

了启动子序列识别. 随机选取了 100 个 Promoter 序列和 100 个 Intergenic 序列，采用 10 折交叉检验，在 $\xi_0 = 0$ 时进行预测，结果为 $TP = 89$，$TN = 95$，$FP = 11$，$FN = 5$，$Sn = 89\%$，$Sp = 95\%$，$CC = 0.84$，在最佳阈值 $\xi_0 = -1$ 时进行预测，结果为 $TP = 92$，$TN = 94$，$FP = 6$，$FN = 8$，$Sn = 92\%$，$Sp = 94\%$，$CC = 0.86$ (比较表 3).

Table 3 Comparison of the promoter prediction by use of different computational methods

Algorithm	NNPP	TSSW	PROSCAN	DPFinder	Promoter 2.0	Prometheus	IDQD
S_n(%)	32	60	40	38	50	86	93
S_p(%)	34	65	56	64	54	88	92
CC	0.34	0.27	0.11	0.18	0.20	0.74	0.85

The data listed in 2 to 7 column are taken from [14]. The last column gives the prediction accuracy of our algorithm. Each value is the average over ten stochastically chosen groups in test set which are independent of training set.

可以看出，加大负集和负集中加入 Intergenic 序列，只要选取最佳阈值来判别，精度几乎无影响. 正负集数量对判别阈值影响的进一步分析将在文中 **3.3** 的讨论中给出.

2.2 对转录起始位点的识别

2.2.1 数据集的构成. 对 EPD 中给出的 1 845 个启动子序列转录起始位点 (TSS) 附近−2 bp：+2 bp (0 位点为 TSS) 的 5-mer 进行搜索，找到了 543 种 5-mer，将这 543 种 5-mer 作为潜在的 TSS (TSS 的可能候选者). 原则上序列中非 TSS 位点都可作为预测 TSS 的对照集(假集)，但为了简化讨论，我们把假集限制在潜在的 TSS 范围. 另外考虑到大部分的 TSS 实验精度为 100 bp[26, 29, 30]，只能进行间隔 100 bp 的粗粒化预测，我们由 TSS 上游 100 bp 开始向前搜索第一个潜在 TSS，以此 5-mer 为核心向前截取 250 bp，向后截取 50 bp，得到 300 bp 长的待测序列，作为假集中的一个成员(如此间包含“N”则剔除此序列)，按此方式，继续向前搜索. 在 TSS 下游也做类似的搜索. EPD 中给出的序列为 −9 999 bp：+6 000 bp (0 位点为 TSS)，这样我们搜索得到 260 023 个长度为 300 bp 的假集序列，真集序列由 1 845 个长度为 300 bp (−250 bp：49 bp)的序列组成.

2.2.2 参数定义. 由于对 TSS 识别的假集序列是在 TSS 附近选取的，比起启动子序列与 CDS 和 Intron 序列而言，真集与假集的差别更小，考虑到这些情形，进行 TSS 识别时须对参数选取做适当调整.

a. 同前，即 **2.1.1** 节的第(a)条；

b. 反映序列位点保守性的特征参数，我们分别选取−29 bp：−22 bp 固定位点的二碱基和三碱基频次，分别定义 $C_8^2\times64$ 维向量 $\boldsymbol{X}_4$ 和 $C_8^3\times16$ 维向量 $\boldsymbol{X}_5$；−1 bp：+4 bp 固定位点的单碱基、二碱基、三碱基和四碱基频次，分别定义 $C_6^1\times4$ 维向量 $\boldsymbol{X}_6$、$C_6^2\times16$ 维向量 $\boldsymbol{X}_7$、$C_6^3\times64$ 维向量 $\boldsymbol{X}_8$ 和 $C_6^4\times256$ 维向量 $\boldsymbol{X}_9$;

c. 定义 3 个非 ID 信息参数，第一个是正链和反链的顶上 30 个模体序列频数[21]，第二个是 −250 bp：−1 bp 的 *GC* 频数，第三个是−31 bp：+49 bp 的 *GC* 频数. 这 3 个参数直接作为二次判别分析向量的 3 个元素.

2.2.3 结果

对 1 845 个真集序列和 260 023 个假集序列分别进行了自洽检验和 10 折交叉检验，结果见表 4. 判别函数的阈值最佳值为 $\xi_0 = -6$，作为对照也给出 $\xi_0 = 0$ 的预测结果.

以上结果表明，在 $\xi_0 = -6$ 时 TSS 的预测精度约为 86%，而假阳性率约为每 3 000 碱基 1 个. 此结果明显优于前人的预测[4,16].

Table 4 Prediction on transcription start sites by use of IDQD

		TP	FN	TN	FP	Sn/%	Sp/%
Self-consistent	$\xi_0 = 0$	1 352	493	255 748	4 275	73.3	98.4
	$\xi_0 = -6$	1 747	98	237 207	22 816	94.7	91.2
10-fold cross-validation	$\xi_0 = 0$	111	73	25 587	415	60.3	98.4
	$\xi_0 = -6$	159	25	23 804	2 198	86.2	91.5

3 讨 论

3.1 本文依据启动子和非启动子在序列特征上的差别，应用基于多样性增量的二次判别分析算法，对人类 Pol Ⅱ启动子序列进行预测，预测结果表明，IDQD 方法能够较好地识别启动子序列和转录起始位点，优于现有的其他预测方法. 一般地，当特征可用某个频数分布来表示，并且它的意义须通过和大量标准样品(标准源)比较才能确定，这样的系统适宜于采用多样性度规，用多样性增量(ID)来进行分类. ID 是两种分布的差异性的度量，待测样本和标准样本的 ID 愈小，它属于标准样本同类的可能性愈大. 当标准样本集足够大，其分布偏离正态分布很小，这时的预测准确性就可能很高. 又由于在此途径中给出了所取样本(训练集)中每个实例之间的真实分布差别，而不是通过样本对整体的估计，这就使得用 ID 来评价两种分布的差别是确定的，不存在估计误差. 当然，这也带来本算法的一个局限，就是在小样本情形，本算法不易给出满意的预测结果.

3.2 在 IDQD 算法中，选取标准多样性源是关键的一步. 文献[25]仅使用单一标准多样性源(正集标准多样性源)，而本文则使用双标准多样性源(正集标准多样性源和负集标准多样性源)，从我们试验的结果来看，一般双源要好于单源. 表 1 的 12 个多样性源中就是一半正集，一半负集. 事实上，如果模体频次 I_{14} 也对负集适当定义，我们发现预测精度还可提高约 1 个百分点.

3.3 序列结构信息的高效抽取，加上高维信息空间映射函数 ID 和非线性判别函数 ξ 的结合，是 IDQD 算法成功的重要保证. 在此算法中，分类的最后一步是在 ξ 空间进行. 阈值 ξ_0 是经验确定的. 据我们的经验，对于一个适当的问题，如果特征(信息参数)选得合适，正负集总可在 0 附近分得很开. 在本文启动子识别的情形中，我们研究了正集序列数与 ξ_0 和负集序列数与 ξ_0 之间的关系，如图 2 所示. 在给定负集序列数时，随正集序列数的减少，最佳判别阈值逐渐在减小，而在正集序列数小于 200 条时，即使找到最佳判别阈值，也不能得到满意的判别结果. 而在给定正集序列数时，随负集序列数的减少，最佳判别阈值则向相反的方向变化，在负集序列数小于 500 条时，不能得到满意的判别结果. 在通常的预测中，正集样本数是给定的，而且不会很大，而负集样本数则是一个庞大的数字，如本文对 TSS 的预测中负集样本数约为正集样本数的 140 倍，这种情形下，最佳判别阈值 ξ_0 为负，如果正集数量也相对很大，可以预期 ξ_0 将向 0 逼近. 关于预测精度和样本数的关系，我们发现，在正样本数大于 200，负样本数大于 500，且给定正负集样本比例后的情形下，在 ξ 的最佳值 ξ_0 邻近预测，结果随样本数的变化小于 2%.

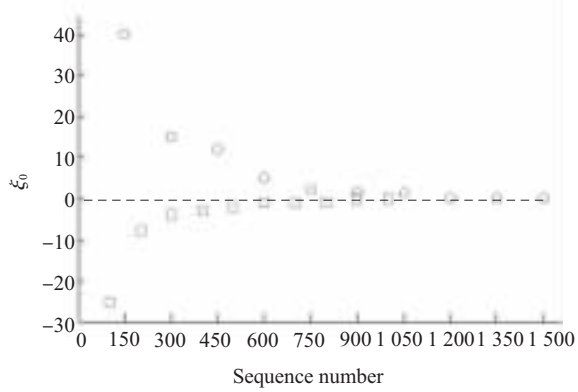

Fig. 2 Relation between the choice of ξ_0 and the sequence number in positive and negative sets

ξ_0 is the best-fit threshold for given number of sequences in positive set and that in negative set. For 1 500 sequences in negative set the ξ_0 value *vs* the number of sequences in positive set (from 100 to 1 000) is shown by □. For 1 000 seuences in positive set the ξ_0 value *vs* the number of sequences in negative set (from 150 to 1 500) is shown by ○.

3.4 根据转录起始位点的实验确定情况，可以分为三类(详细见 EPD 使用指南 http://www.epd.isb-sib.ch/current/usrman.html)：a. 单一起始位点 S 类，此类大于 90%的启动子序列的转录起始位点在 10 bp 范围内，在我们整理的 1 845 个人类启动子序列中有 348 个属于此类；b. 多起始位点 M 类，此类大于 75%的启动子序列的转录起始位点在 20 bp 范围内，有 408 个；c. 起始区 R 类,此类大于 75%的启动子序列的转录起始位点在 100 bp 范围内，有 1 089 个. 鉴于大部分为 R 类，本文设计了间隔 100 bp 搜索一个潜在 TSS 的算法. 如何进行更准确的理论预测，并为 TSS 的实验定位提供参考，有待进一步的工作.

参 考 文 献

1 Fickett J W. Finding genes by computer: The state of the art. Trends Genet, 1996, **12** (8): 316~320

2 Gelfand M S. Prediction of function in DNA sequence analysis. J Comp Biol, 1995, **2** (1): 87~115

3 Bucher P. Weight matrix description of four eukaryotic RNA polymerase Ⅱ promoter elements derived from 502 unrelated

promoter sequences. J Mol Biol, 1990, **212** (4): 563~578
4 Fickett J W, Hatzigeorgiou A C. Eukaryotic promoter recognition. Genome Res, 1997, **7** (9): 861~878
5 Knudsen S. Promoter 2.0: for recognition of Pol Ⅱ promoter sequences. Bioinformatics, 1999, **15** (5): 356~361
6 Bajic V B, Seah S H, Chong A, *et al*. Dragon promoter finder: recognition of vertebrate RNA polymerase Ⅱ promoters. Bioinformatics, 2002, **18** (1): 198~199
7 Lemon B, Tjian R. Orchestrated response: a symphony of transcription factors for gene control. Genes Dev, 2000, **14** (20): 2551~2569
8 Pedersen A G, Baldi P, Chauvin Y, *et al*. The biology of eukaryotic promoter prediction-a review. Comput Chem, 1999, **23** (3~4): 191~207
9 Smale S T. Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. Biochim Biophys Acta, 1997, **1351** (1~2): 73~88
10 Smale S T. Core promoters: active contributors to combinatorial gene regulation. Genes Dev, 2001, **15** (19): 2503~2508
11 Werner T. Models for prediction and recognition of eukaryotic promoters. Mamm Genome, 1999, **10** (2): 168~175
12 Wasserman W W, Palumbo M, Thompson W, *et al*. Human-mouse genome comparisons to locate regulatory sites. Nature Genet, 2000, **26** (2): 225~228
13 Solovyev V, Salamov A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. Proc Int Conf Intell Syst Mol Biol, 1997, **5**: 294~302
14 Gangal R , Sharma P. Human pol Ⅱ promoter prediction: time series descriptors and machine learning. Nucleic Acids Research, 2005, **33** (4): 1332~1336
15 Zhang M Q. Identification of human gene core promoters in silico. Genome Res, 1998, **8** (3): 319~326
16 Davuluri R V, Grosse I, Zhang M Q. Computer identification of promoters and first exons in the human genome. Nature Genetics, 2001, **29** (4): 412~417
17 Reese M G. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Comput Chem, 2001, **26** (1): 51~56
18 Prestridge D S. Predicting Pol Ⅱ promoter sequences using transcription factor binding sites. J Mol Biol, 1995, **249** (5): 923~932
19 Shahmuradov I A, Solovyev V V, Gammerman A J. Plant promoter prediction with confidence estimation. Nucleic Acids Research, 2005, **33** (3): 1069~1076
20 Roeder R G. The role of general initiation factors in transcription by RNA polymerase Ⅱ. Trends Biochem Sci, 1996, **21** (9): 327~335
21 Xie X H, Lu J, Kulbokas E J, *et al*. Systematic discovery of regulatory motifs in human promoters and 3'UTRs by comparison of several mammals. Nature, 2005, **434** (7031): 338~345
22 Laxton R R. The measure of diversity. J Theor Biol, 1978, **71** (1): 51~67
23 McLachlan G J. Discriminant Analysis and Statistical Pattern Recognition. New York :Wiley, 1992. 1~526
24 Zhang M Q. Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci USA, 1997, **94** (2): 565~568
25 Zhang L R, Luo L F. Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Research, 2003, **31**(21): 6214~6220
26 Schmid C D, Praz V, Delorenzi M, *et al*. The eukaryotic promoter database EPD: the impact of in silico primer extension. Nucleic Acids Research, 2004, **32**: D82~85
27 Matthias S, Andreas K, Kornelie F, *et al*. First pass annotation of promoters on human chromosome 22. Genome Res, 2001, **11** (3): 333~340
28 Luo L F, Li H, Zhang L R. ORF organization and gene recognition in the yeast genome. Comp Funct Genomics, 2003, **4** (3): 318~328
29 Suzuki Y, Yamashita R, Sugano S, *et al*. DBTSS, DataBase of transcriptional start sites: progress report 2004. Nucleic Acids Research, 2004, **32**: D78~D81
30 Suzuki Y, Taira H, Tsunoda T, *et al*. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep, 2001, **2** (5): 388~393

Human pol Ⅱ Promoter Prediction*

LÜ Jun[1)], LUO Liao-Fu**
(*Department of Physics, Inner Mongolia University, Huhhot* 010021, China)

Abstract Based on the characteristics of base distribution in promoter and non-promoter region the method of Increment of Diversity with Quadratic Discriminant analysis (IDQD) was used to predict the pol Ⅱ promoter in human genome. The prediction has attained accuracy higher than 90%. The transcription start sites have also been predicted successfully with sensitivity 86% and specificity 91% , better than other top softwares currently published.

Key words promoter, increment of diversity, quadratic discriminant analysis , transcription start site

*This work was supported by grants from The National Natural Sciences Foundation of China (90403010) and Inner Mongolia Science Foundation (200308020102).

**Corresponding author. Tel: 86-471-4992676, E-mail: lfluo@mail.imu.edu.cn
[1)]Inner Mongolia University of Technology, Huhhot 010051, China.

Received: June 13, 2005 Accepted: July 28, 2005