

人类基因 PolyA 位点预测

廖 堃 段江波 周艳红

(华中科技大学生物信息与分子成像湖北省重点实验室 武汉 430074)

摘 要 mRNA 3'端的多聚腺苷酸化是真核细胞内 mRNA 转录后处理的三个最主要步骤之一. 对 DNA 序列上发生多聚腺苷酸化的位置即 PolyA 位点的识别, 对于理解 mRNA 的形成机制以及进行基因结构预测具有重要作用. 本研究利用机器学习方法对 PolyA 位点进行预测, 其实现过程分为以下三个步骤: 特征的生成、特征的筛选、特征的综合分析聚类. 首先, 我们采取统计 k 阶核苷酸频率的方法来生成初始的特征; 然后, 通过信息学知识来对特征进行筛选; 最后, 使用 SVM(Support Vector Machines, 支持向量机)的方法进行特征的综合分析, 确定参数, 建立预测模型. 在独立的测试数据集上进行测试, 当敏感度(Sn)固定为 60% 时, 在内含子水平和外显子水平上的特异性和(Sp)分别为 71.67% 和 80.77%, 在内含子水平上的预测精度明显优于国际上的同类软件.

关键词 PolyA 信号; 机器学习; 熵; 支持向量机

中图法分类号 TP18

Prediction of Polyadenylation in Human Gene Sequences

LIAO Kun DUAN Jiang-Bo ZHOU Yan-Hong

(Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan 430074)

Abstract Polyadenylation (PolyA) occurs in mRNA 3' end is one of the three main steps of eukaryotic pre-mRNA processing. The prediction of polyadenylation sites in human DNA and mRNA sequences is very important for realizing the pre-mRNA processing and prediction of gene structure. This paper presents a machine learning method to predict polyadenylation signals (PASEs) in human DNA and mRNA sequences. This method consists of three steps of feature manipulation: Generation, selection and integration of features. In the first step, new features are generated using k -gram nucleotide acid patterns. In the second step, a number of important features are selected by an entropy-based algorithm. In the third step, support vector machines are employed to recognize true PASEs from a large number of candidates. At last, a mathematic model forms. When the sensitivity is 60%, the corresponding specificity is 71.67% on intron level, and 80.77% on exon level.

Keywords Polyadenylation signals; machine learning; entropy; support vector machines

1 引 言

前体 mRNA 3'端多聚腺苷酸化是真核细胞内 mRNA 转录后处理的三个最主要步骤(包括 5'帽子

结构的形成、内含子的剪切及 3'端的多聚腺苷酸化)之一, 与 mRNA 稳定性的调节、mRNA 的细胞内转运、翻译的起始以及一些其他的细胞机制和疾病机制有着重要关系^[1-8]. 在 3'UTR 区存在多个潜在 PolyA 位点时, 选择性多聚腺苷酸化以组织或疾

病特异性的方式影响着基因的表达^[1]. 在基因结构预测领域,对 PolyA 位点的准确识别有助于对基因 3'末端的确定,对提高基因结构预测精度有很大的帮助^[9]. 因此,对 PolyA 位点的准确识别,对于预测基因结构、理解 mRNA 的形成机制及某些疾病的分子机理具有巨大的作用.

真核生物前体 mRNA 3'端的多聚腺苷酸化包括两个步骤:特异性的核苷酸内切酶在 PolyA 位点处进行断裂和腺苷酸聚合酶在断裂位点处添加 PolyA 尾巴^[10-12]. 如图 1, PolyA 位点上游含有 PolyA 信号,一般位于 PolyA 位点上游 10~30bp 的范围内,是典型的上游作用元件,起决定下游 PolyA 位点的作用,其序列保守性比较强,一般为 AATAAA 和 ATTAAA 两种序列^[13]. PolyA 位点下游则一般含有保守性较差的 U-rich 和 G/U-rich 序列,即下游作用元件^[11,14]. PolyA 信号和下游作用元件构成了多聚腺苷酸化的顺式作用元件.

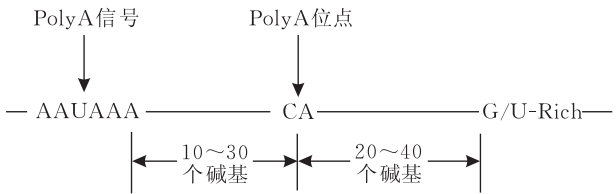


图 1 PolyA 位点、PolyA 信号和 G/U-rich 序列

在多聚腺苷酸化的反式作用元件方面,已知至少有 4 种酶参与了顺反元件的互相作用以及 3'末端的加工,分别为识别 PolyA 信号的 CPSF(Cleavage Polyadenylation Specific Factor)、识别下游作用元件的 CstF(Cleavage stimulation Factor)、负责剪接前体 mRNA 的 CFs(Cleavage Factors,包括 CFI 和 CFII)和用于加 PolyA 尾巴的 PolyA 聚合酶 PAP(PolyA Polymerase)^[11,15-16]. 多聚腺苷酸化的作用机制为:CPSF 绑定到六核苷酸的 PolyA 信号序列,CstF 识别下游的 U-rich 和 G/U-rich 序列并相互作用,CFI 在 PolyA 信号和下游作用元件之间的某个位置对前体 mRNA 进行分裂,而后在 PolyA 聚合酶的作用下添加多聚腺苷酸尾巴.

在预测 PolyA 位点方面,目前主要有两种方法:基于 EST((Expressed Sequence Tag,表达序列标签)的方法和从头预测的方法. 基于 EST 的方法主要是利用 3'EST 与基因组序列的比对信息,确定出潜在的 PolyA 位点,再结合标准的 PolyA 信号(AATAAA 和 ATTAAA)进行预测. 由于 EST 序列是由实验得到的^[17],能够在一定程度上代表基因的表达情况^[18],并且数据量非常巨大^[19],所以这类

方法具有一定的优势,但是由于 EST 序列的错误倾向性^[20-21]和基因组序列的复杂性^[22],使得这一类方法的预测精度难以有较大的提高. Kan 等人开发的工具 PASS^[23]是这一类方法的代表. 从头预测的方法一般是分析 PolyA 位点上下游序列的特征,提取有用信息,然后建立模型,确定参数,进行预测. 这一类方法的代表性预测工具有 Polyadq^[15]和 Erpin^[24]. 其中 Polyadq 分别为六核苷酸序列的 PolyA 信号和下游作用元件建立权重矩阵,利用一对二次判别式函数,以 280 条 mRNA 序列和 136 条 DNA 序列为训练集,训练模型,确定参数,用来预测 PolyA 位点. 而 Erpin 则是通过统计 PolyA 位点上下游序列一些位置特异性的二核苷酸对的出现频率,提取有用信息而训练出来的. 但是在某些情况下特别是内含子水平上,Polyadq 和 Erpin 的预测效果并不是很理想.

我们将运用机器学习的方法来解决 PolyA 位点的预测问题. 通过对 PolyA 位点形成机制的分析,我们以 PolyA 信号周围 100bp 的范围为研究对象,展开研究. 首先,我们采取统计 *k* 阶核苷酸频率的方法来生成初始的特征;然后,通过信息学上的一些知识来对特征进行筛选;最后,使用 SVM(Support Vector Machines,支持向量机)的方法进行特征的综合分析,建立模型,确定参数,用来预测 PolyA 位点.

2 数 据

数据分为训练数据和测试数据. 其中训练数据的真样本来自软件 Erpin 提供的数据,共 2327 条序列. 而假样本则来自 Genbank 下载的数据,从这些数据中扫描中央含有 AATAAA 的片断,共得到 2453 条序列,我们筛选了来自 3'UTR 区的扫描序列,以防止它对真样本产生的干扰.

测试数据也包括真样本和假样本,真样本采用 Genbank 注释的含有 PolyA 信号的序列 72 条,假样本采用人类基因结构预测程序的训练数据集,从 http://www.fruitfly.org/seq_tools/datasets/Human/ 下载,其处理方法与构建训练集假样本的方法相同,其中,外显子数据 62 条,内含子数据 64 条.

3 方 法

利用机器学习的方法来完成我们对 PolyA 位

点的分类工作,通过对 PolyA 位点上游序列和下游序列的分析来完成对 PolyA 位点的预测.本研究分为特征的形成、特征的选择、特征的综合分析,具体如下.

3.1 特征的形成

PolyA 的序列较之于其他的序列,如启动子,它的保守性弱,这样使对 PolyA 位点的位置保守性的分析产生了困难,必须建立另外的不依赖于位置保守性的特征组织方式.借鉴蛋白质序列的特征组织方式,即按内容来形成特征,具体就是以各核苷酸或核苷酸组的出现频率作为特征.数据是以 PolyA 信号为中心,前后各 100bp,一条序列共 206 个核苷酸,本研究只考虑 1 阶、2 阶、3 阶的核苷酸出现频率,如:A,CG,AGC,⋯,另外,将 PolyA 信号前的 100bp 序列和 PolyA 信号后的 100bp 序列分开考虑,同种符号特征在两个区域里的出现频率将视为两种特征.如此,在各个区域里都有 84 种特征,加起来整个序列将有 168 个特征.

3.2 特征的选择

由于数据对象的不断膨胀,使得现实的数据中产生了很多的冗余和数据噪声,所以怎样去掉这些冗余数据,和怎样消除数据噪声成为提高计算速度和计算准确率的关键.

所谓数据冗余是指对计算分析的贡献作用不大,可有可无的数据,但是数据冗余的存在会减慢运算速度,会间接地影响计算精度,所以在实际研究中,他们并不是可有可无,而是一定要去掉.

所谓噪声数据是指数据中存在着错误、或异常(偏离期望值)的数据.它将对运算的结果产生更为直接的影响.

综合以上,在运算建模的过程中采用特征选择的方法,消除冗余数据和噪声数据,将有效地提高程序的运算速度,提高结果的精度.

关于特征的选择,其方法很多^[25],这里我们用到了信息论中的一种方法来完成特征选择这项工作,就是熵理论.

下面来介绍一下熵理论的有关知识:

设 S 代表一组训练样本数据集(每个对象的类别已知),共有 m 个不同类别(class);这样 S 包含 s_i 个 C_i , $i \in [1, 2, \dots, m]$,任何一个对象属于 C_i 的概率为 s_i/s ,这里 s 为集合 S 中所有样本总数.产生相应信息需要的信息熵为

$$I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m \frac{s_i}{s} \log_2 \frac{s_i}{s}.$$

若属性 A 可以取得值为 $\{a_1, a_2, \dots, a_v\}$,它将会把对应的数据集分为 v ,即 $\{S_1, S_2, \dots, S_v\}$;其中 S_j 包含属性 A 取同一值 a_j 的数据行; S_j 包含 s_{sj} 个 C_i 数据对象.根据属性 A 的取值对当前数据集进行划分所获得的信息就称为属性 A 的熵.它的计算公式如下:

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + s_{2j} + \dots + s_{mj}}{s} I(s_{1j}, \dots, s_{mj}),$$

这样利用属性 A 对当前分支结点进行相应样本集合划分所获得的信息增益就是:

$$Gain(A) = I(s_1, s_2, \dots, s_m) - E(A).$$

换句话说, $Gain(A)$ 被认为是根据属性 A 取值进行样本集合划分所获得的(信息)熵的减少(量).选择属性时,将选择熵增益值尽可能大的属性.

根据以上所介绍的熵增益的概念,我们对最初生成的 168 个特征进行了筛选,将其中的 59 个特征删去,留下 109 个特征.即经过特征的选择这一步,我们最终确定了 109 个特征.

3.3 特征的综合分析

SVM 在生物信息学领域中的应用相当广泛,包括剪切位点和翻译起始位点等功能位点的识别;蛋白质二级结构预测;蛋白质功能预测;发现新的调控元件;蛋白质相互作用预测;基于基因芯片数据的基因功能预测;基于基因芯片数据的癌症分类.

利用 SVM 方法,在训练集独立的情况下,原始的特征空间可以转化为高维的特征空间,在新的特征空间下,样本线性可分,这样将非线性问题转化为线性问题(如图 2 所示),从而提高分类器的准确度.这种非线性变换是通过定义适当的内积实现的,其映射是由 SVM 的核函数来实现的.核函数是一个数学函数,这个函数对于所有的 x_i, x_j (x_i, x_j 都属于原始输入空间 X),满足

$$K(x_i, x_j) = f(x_i, x_j) = \phi(x_i) \cdot \phi(x_j),$$

其中, ϕ 是从输入空间到特征空间(内积空间)的映射.

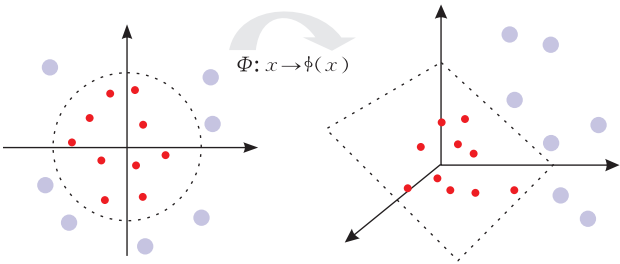


图 2 原始非线性特征空间转化为线性可分的高维特征空间

SVM 的 3 种核函数如下.

(1) 线性核函数:

$$K(x_i, x_j) = x_i \cdot x_j.$$

(2) 多项式核函数:

$$K(x_i, x_j) = [(x_i \cdot x_j) + 1]^q.$$

(3) 高斯核函数:

$$K(x_i, x_j) = \exp\left\{-\frac{|x_i - x_j|^2}{\sigma^2}\right\}.$$

其中高斯核函数比较适用于原始特征空间分布较为复杂或者无法确定的情况.

考虑到 SVM^[26-27]所具有的良好分类效果和 在生物信息学领域中的广泛应用,本模型运用 SVM 对数据进行特征聚类分析.所选取的 SVM 的判别 函数为

$$f(x) = \sum_i a_i^0 y_i K(x_i, x) + b.$$

本研究使用的是软件 SVMLight,下载地址: <http://svmlight.joachims.org/>.核函数采用的是 高斯核函数.

训练的过程中要确定两个参数 C 值和 γ 值,采 用 十字交叉的方法来实现.将训练数据分为 10 组, 以其中 9 组作为训练,一组来测评参数的效果,然后 换一组作为测评组,另 9 组训练……,在一定取值范 围内进行了 100 组 C, γ 值的测算,找出峰值的位 置,然后在峰值附近再进行一次范围缩小的测算,最 终找出一组最佳的 C, γ 值.

4 结果与分析

在本研究中,交叉验证得到一组最优的 C, γ 值, 在该值下对训练数据的测试,得到 $Sn = 77.64\%$, $Sp = 84.05\%$.测试数据如上述分为两组,分别为内 含子区域序列、外显子区域序列.通过控制 Sn (敏感 性)值得到对应的 Sp (特异性)值,它们的计算公式 分别为

$$Sn = \frac{TP}{TP + FN},$$

$$Sp = \frac{TP}{TP + FP}.$$

CC 代表相关系数,计算公式为

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

FPR 代表假阳性率:计算公式为

$$FPR = \frac{FP}{FP + TN}.$$

各符号代表的意义如下:TP 表示真阳性,FP

表示假阳性,TN 表示真阴性,FN 表示假阴性.比 较各组数据在相同 Sn 的情况下, Sp 的变化和水平, 其结果如表 1 所示(其中,Exon 表示外显子水平, Intron 表示内含子水平).

表 1 在取不同的 Sn 值时,所对应的 Sp 值,CC 值和 FPR 值

$Sn/\%$	$Sp/\%$		CC		FPR	
	Exon	Intron	Exon	Intron	Exon	Intron
40	93.33	80.00	0.4322	0.3247	0.0322	0.1093
50	85.37	74.47	0.4170	0.3333	0.0968	0.1875
60	80.77	71.67	0.4528	0.3417	0.1613	0.2656
70	79.03	68.06	0.4805	0.3364	0.2097	0.3594
80	77.33	66.67	0.5308	0.3631	0.2742	0.4531
90	73.86	63.73	0.5563	0.3719	0.3710	0.5781

从图 3 可看出,当 Sn 值上升时, Sp 值不断减 小,并且在 外显子水平上和内含子水平上 Sp 值的减 小趋势是 大致相同的.当 Sn 为 60% 时,在 内含子和 外显子水平上的 Sp 值分别为 71.67% 和 80.77%. 并且模型对外显子的区分效果要好于内含子,这可 能是因为外显子是编码蛋白质的基因序列,而 PolyA 区域与内含子都是不编码蛋白质的序列,在 序列特征上 PolyA 区域与内含子更为相似,所以在 内含子水平上的预测效果要比外显子水平上低一些.

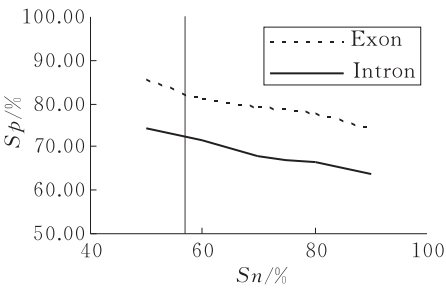


图 3 模型的 Sn - Sp 关系图

为了更好地评估模型的预测效果,将其与国际 上识别 PolyA 序列的著名软件 Erpin^[24]相比较.在 同样的测试数据集上对 Erpin 进行测试,结果如 表 2 所示.

表 2 在取不同的 Sn 值时,Erpin 软件所 对应的 Sp 值,CC 值和 FPR 值

$Sn/\%$	$Sp/\%$		CC		FPR	
	Exon	Intron	Exon	Intron	Exon	Intron
50	82.50	76.74	0.4039	0.3510	0.1166	0.1612
60	79.40	65.57	0.4030	0.2584	0.2000	0.3387
70	78.33	63.51	0.4848	0.2687	0.2166	0.4354
80	78.14	60.67	0.5732	0.2608	0.2333	0.5645
90	74.07	55.05	0.5666	0.1452	0.3500	0.7903

从图 4 可看出,Erpin 对外显子的预测效果要 好于内含子,这一点与我们的模型是一致的.当 Sn 值上升时,Erpin 软件的 Sp 值不断减小,当 Sn 为 60% 时,在 内含子和 外显子水平上的 Sp 值分别为

65.57%和 79.40%。为了进一步比较 Erpin 软件和本模型的预测效果,接下来分别在外显子水平和内含子水平上比较。

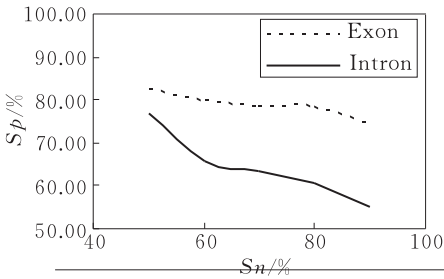


图 4 Erpin 软件的 S_n - S_p 关系图

从图 5 和图 6 可以看出,在外显子水平上,在 S_n 小于 75% 的时候,我们的模型的预测精度明显优于 Erpin;在 S_n 大于 75% 时,预测精度略低于 Erpin。在内含子水平上,在 S_n 大于 55% 的时候,我们的模型的预测精度明显优于 Erpin。

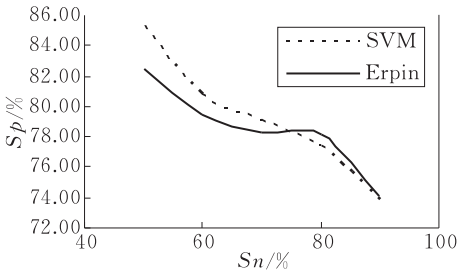


图 5 在外显子水平上比较 Erpin 和本模型

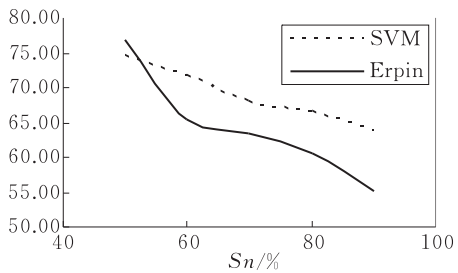


图 6 在内含子水平上比较 Erpin 和 SVM 模型

从图 7 和图 8 可以看出,在外显子水平上,Erpin 和本模型的出错率相差很小;但是在内含子水平上,我们的模型具有更小的出错率,预测效果更加理想。

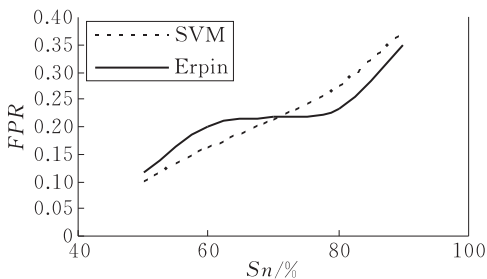


图 7 在外显子水平上比较 Erpin 和本模型的出错率

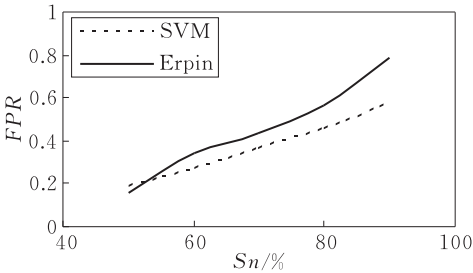


图 8 在内含子水平上比较 Erpin 和本模型的出错率

5 总 结

我们运用机器学习的方法来对人类序列的 PolyA 位点进行预测,由 PolyA 尾巴的添加机理的启发,以 PolyA 信号为中心的 206bp 长的范围内的序列为研究对象,采取基于序列内容的特征生成方法,实现了非位置性的特征生成,得到 168 个特征;然后以信息学的有关熵的一些理论为工具,展开了对生成特征的筛选工作,取得了一定的效果,删去了 59 个不符合要求的特征,保留了 109 个目标特征;然后用 SVM 对数据进行了分析聚类,完成了模型的训练.在单独的测试数据集上进行测试,当敏感度 (S_n) 固定为 60% 时,在内含子水平和外显子水平上的特异性 (S_p) 分别为 71.67% 和 80.77%。

为了进一步评估模型的预测精度,与国际上同类软件 Erpin 进行比较.选取 Erpin 进行比较的原因在于:本模型与 Erpin 是采用相同的训练数据集进行训练的,更具有可比性,比较结果也更有说服力,另外 Erpin 软件提供免费的网上服务和源代码下载。

比较结果表明,在外显子水平上,我们的模型与 Erpin 预测结果不相上下;在敏感度 S_n 小于 75% 的时候,我们的模型的预测精度优于 Erpin;在 S_n 大于 75% 时,预测精度略低于 Erpin。在内含子水平上,我们的模型的预测精度要明显优于 Erpin。经过之前的比较和分析,本模型相比 Erpin,在预测精度上有所提高可能是以下两个方面的原因:(1)在分类器的选择上,SVM 在生物信息学领域中得到广泛应用,方法已经比较成熟,相比 Erpin 可能更能体现 PolyA 位点分类的内在规律;(2)在特征的选择和筛选上,本模型进行了更加广泛的特征提取;以 PolyA 信号周围的 206bp 长的范围内的序列为研究对象,采取基于序列内容的特征生成方法,实现了非位置性的特征生成,然后以信息学的有关熵的一些理论

为工具,展开了对生成特征的筛选工作;而 Erpin 只是统计 PolyA 位点上下游序列一些位置特异性的二核苷酸对的出现频率。

此类研究的意义在于运用信息学方法对生物学问题进行分析,在大量的数据中去粗存精,为实验提供了一个良好的条件,起到了辅助的作用,减少了实验的工作量,是今后生物学发展的一个重要方向。

参 考 文 献

- [1] Edwalds-Gilbert G, Veraldi K L, Milcarek C. Alternative poly(A) site selection in complex transcription units: Mean to an end? *Nucleic Acids Research*, 1997, 25(13): 2547-2561
- [2] Wang Z, Day N, Trifillis P, Kiledjian M. An mRNA stability complex functions with poly(A)-binding protein to stabilize mRNA in vitro. *Molecular and Cellular Biology*, 1999, 19(7): 4552-4560
- [3] Decker C J, Parker R. A turnover pathway for both stable and unstable mRNAs in yeast: Evidence for a requirement for deadenylation. *Genes & Development*, 1993, 7(8): 1632-1643
- [4] Chen Z, Li Y, Krug R M. Influenza A virus NS1 protein targets poly(A)-binding protein II of the cellular 3'-end processing machinery. *EMBO Journal*, 1999, 18(8): 2273-2283
- [5] Craig A W B, Haghighat A, Yu A T K. Interaction of polyadenylate-binding protein with the eIF4G homologue PAIP enhances translation. *Nature*, 1998, 392(6675): 520-523
- [6] Zarudnaya M I, Hovorun D M. Hypothetical double-helical poly(A) formation in a cell and its possible biological significance. *IUBMB Life*, 1999, 48(6): 581-584
- [7] Gehring N H, Frede U, Neu-Yilik G, Hundsdoerfer P, Vetter B, Hentze M W, Kulozik A E. Increased efficiency of mRNA 3' end formation: A new genetic mechanism contributing to hereditary thrombophilia. *Nature Genetics*, 2001, 28(4): 389-392
- [8] Conne B, Stutz A, Vassalli J D. The 3' untranslated region of messenger RNA: A molecular 'hotspot' for pathology?. *Nature Medicine*, 2000, 6(6): 637-641
- [9] Kan Z, Rouchka E C, Gish W R et al. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Research*, 2001, 11(5): 889-900
- [10] Colgan D F, Manley J L. Mechanism and regulation of mRNA polyadenylation. *Genes & Development*, 1997, 11(21): 2755-2766
- [11] Wahle E, Kuhn U. The mechanism of 3' cleavage and polyadenylation of eukaryotic pre-mRNA. *Progress in Nucleic Acid Research and Molecular Biology*, 1997, 57: 41-71
- [12] Keller W, Minvielle-Sebastia L. A comparison of mammalian and yeast pre-mRNA 3'-end processing. *Current Opinion in Cell Biology*, 1997, 9(3): 329-336
- [13] Tabaska J E, Zhang M Q. Detection of polyadenylation signals in human DNA sequences. *Gene*, 1999, 231(1-2): 77-86
- [14] Proudfoot N. Poly(A) signals. *Cell*, 1991, 64(4): 671-674
- [15] Zhao J, Hyman L, Moore C. Formation of mRNA 3' ends in eukaryotes: Mechanism, regulation and interrelationships with other steps in mRNA synthesis. *Microbiology and Molecular Biology Reviews*, 1999, 63(2): 405-445
- [16] Zarudnaya M I, Kolomiets I M et al. Downstream elements of mammalian pre-mRNA polyadenylation signals: Primary, secondary and higher-order structures. *Nucleic Acids Research*, 2003, 31(5): 1375-1386
- [17] Adams M D, Kelley J M, Gocayne J D et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science*, 1991, 252(5013): 1651-1656
- [18] Mekhedov S, Martinez de Ilarduya O, Ohlrogge J. Towards a functional catalog of the plant genome: A survey of genes for lipid biosynthesis. *Plant Physiol*, 2000, 122(2): 389-401
- [19] Benson D A, Karsch-Mizrachi I, Lipman D J et al. GenBank: Up-date. *Nucleic Acids Research*, 2004, 32(Database issue): D23-D26
- [20] Beaudoin E, Gautheret D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Research*, 2001, 11(9): 1520-1526
- [21] Rotem S, Hershel M S. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Research*, 2003, 31(3): 1067-1074
- [22] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 2001, 409(6822): 860-921
- [23] Kan Zhengyan, Warren G et al. UTR reconstruction and analysis using genomically aligned EST sequences//Proceedings of the International Conference on Intelligent Systems for Molecular Biology, 2000, 8: 218-227
- [24] Matthieu L, Daniel G. Sequence determinants in human polyadenylation site selection. *BMC Genomics*, 2003, 4(1): 7
- [25] Liu H, Li J, Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns//Proceedings of the 13th Workshop on Genome Informatics. Tokgo, Japan, 2002: 51-60
- [26] Burges C J C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 1998, 2(2): 121-167
- [27] Platt J. Fast training of support vector machines using sequential minimal optimization//Scholkopf B, Burges C, Smola A eds. *Advances in Kernel Methods-Support Vector Learning*. MIT Press, 1998



LIAO Kun, born in 1983, master. His research interests focus on bioinformatics.

DUAN Jiang-Bo, born in 1984, master. His research interests focus on bioinformatics.

ZHOU Yan-Hong, born in 1966, Ph. D. , professor, Ph.D. supervisor. His research interests focus on bioinformatics, including pattern recognition, machine learning and its application to bioinformatics, human genome structure annotation, protein structure prediction and disease gene.

Background

Polyadenylation (PolyA) occurs in mRNA 3' end is one of the three main steps of eukaryotic pre-mRNA processing. The prediction of polyadenylation sites in human DNA and mRNA sequences is very important for realizing the pre-mRNA processing and prediction of gene structure. When 3'UTR occurs more than one latent PolyA sites, a selectivity polyadenylation will decide gene expression based on tissue and disease mechanism. For prediction of gene structure, identifying PolyA sites exactly is profitable on confirming 3' end.

In the nearly study, there are mainly two methods for PolyA site finding; EST (Expressed Sequence Tag) based method and statistics based method. The first method is mainly analyzing the EST and genomic sequences to characterize the latent PolyA sites. One of these programs is developed by Zhengyan Kan called PASS. The statistics based

method is analyzing the upstream and downstream element, profiting some useful characters to form a mathematic model for PolyA sites prediction. Polyadq and ERPIN are accomplished in this method.

This paper presents a machine learning method to predict polyadenylation signals (PASes) in human DNA and mRNA sequences. When the sensitivity is 60%, the corresponding specificity is 71.67% on intron level, and 80.77% on exon level. This study is supported by the National Natural Science Foundation of China (Main Program, Grant No. 90608020), the Specialized Research Fund for the Doctoral Program of Higher Education(Grant No. 20050487037), Program for New Century Excellent Talents in University, and National Program for Sci-Tech Basic Conditions Platform Construction of Ministry of Science and Technology of China.