基因组序列中转录始点的预测方法研究*)

张 慧 陈恩红 童 庆
(中国科学技术大学计算机系 合肥 230027)

摘 要 在生物信息学领域,基因组序列上转录始点预测是基因序列预测中重要而困难的问题。本文结合了基因序列中有重要生物意义的功能点的特征,设计了一个新算法,预测基因组序列中转录始点的位置。该算法利用权重矩阵以及判别分析方法,根据训练实例进行学习,获得判别函数用于预测。实验结果表明,本文方法具有较好的预测效果。
关键词 转录始点,判别分析,权重矩阵

A Research on Method of Predicting Transcription Start Site in Genomic Sequence

ZHANG Hui CHEN En-Hong TONG Qing
(Department of Computer Science, University of Science and Technology of China, Hefei 230027)

Abstract Prediction of transcription start site in a genomic sequence is difficult but important in the field of gene prediction in bioinformatics. A novel algorithm is presented by integrating the features of some meaningful sites in a genomic sequence to predict the position of transcription start site. Discriminant functions are got by training with the method of discriminant analysis and weight matrix and then are used for prediction. It showes that the algorithm has attained a good overall result by an experiment.
Keywords Transcription start site, Discriminant analysis, Weight matrix

1 引言

序列分析是生物信息学中的一个重要方向,它使用数据挖掘的方法来预测基因序列中重要功能点的位置,而预测转录始点(transcription start site,简称 TSS)是其中一个重要但十分困难的问题[1]。精确预测 TSS 的位置对于基因注释、理解基因转录及调整机制有重要意义。预测的困难之处在于两点:(1)生物基因的转录启动条件十分复杂,通常由多个蛋白质因子相互作用启动转录,而且不同基因需要的蛋白质因子往往不同;(2)迄今未发现能标志 TSS 的模式序列。

目前已经有一些基因整体结构预测算法[3],它们主要针对内部外显子的预测,所以预测的 TSS 往往距离真正位置十分遥远[2]。虽然专门用于预测 TSS 和其附近启动子区域的算法效率相对有所提高,但它们的共同弱点在于:当序列长度增大时,虽然它们能预测出真正的 TSS,但同时也误报了很多假因子。

本文结合序列上的多个特殊点的特征设计了一个新算法,利用权重矩阵和判别分析方法预测 TSS 的位置。其新颖之处在于:(1)结合特征翻译始点和 TATA 模式。这对于那些第一外显子包含翻译始点或者启动子存在 TATA 模式的情况,预测率将大为提高;(2)使用了近年来高效率算法中重要的特征:CpG 岛。这对于那些与 CpG 岛相关的 TSS 预测效率有显著提高。

本文首先介绍了基因预测领域的相关概念,然后简要概述了 TSS 预测领域的常用特征、方法和前人所做的工作;继而给出了本文的算法;最后是实验方法以及展望。

2 基本概念和相关工作

2.1 真核基因的组成

真核生物的基因组序列相当于一条字母表由碱基 A、C、G、T 组成的线性序列,它由多个基因和基因间区域组成。基因是 DNA 序列上基本的功能单位,它通过转录、翻译等步骤最终形成蛋白质。基因由多个外显子和内含子组成。在基因的上游区域,有一些特殊序列称为启动子,蛋白质因子与启动子结合,识别 TSS,开始启动基因的转录。启动子和 TSS 共同构成了启动子区域。翻译始点是基因编码蛋白质的开始端,它位于第一外显子中,或者内部外显子中。

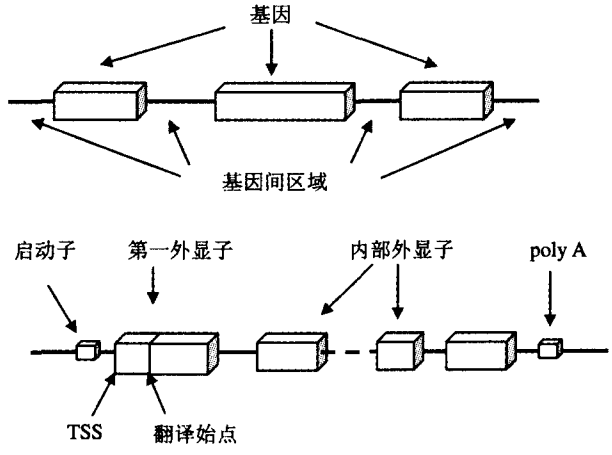

图 1 基因组结构和一个基因的结构

2.2 常用特征和方法

2.2.1 常用特征 由于 TSS 没有共同序列,因此 TSS 的预测往往与它附近功能点的预测密不可分。它的预测往往需要借助上游的启动子、附近的 CpG 岛和序列本身的特征如寡聚物频率等信息[3]。

(1)通过预测 CpG 岛的位置来判断 TSS 的位置。

*)本文研究得到国家自然科学基金资助(60005004)和安徽省自然科学基金资助(01042302)。张 慧 硕士生,主要研究方向为数据挖掘、生物信息学中的序列分析。陈恩红 博士,副教授,主要研究方向为数据挖掘、知识发现、网络信息处理。

在真核基因组中，CG 两个碱基连续出现的频率很低，因为碱基 C 容易甲基化而变成碱基 T。但统计发现在许多真核生物启动子附近的区域，CG 含量却十分丰富，这样的区域被称为 CpG 岛。若某启动子区域富含 CpG 岛，则称此启动子区域 CpG 相关。近年许多成功的算法表明，如 FirstEF[4]、DGSF[2]，结合 CpG 岛的特征可以大大提高 CpG 相关的启动子区域的预测精度。

(2)通过预测转录因子绑定点的位置来预测 TSS 的位置。

基因的转录启动是由多个转录因子相互作用，并与序列上的某些连续碱基结合开始。通过寻找在多个基因中经常出现的的绑定点的模式，如 TATA 模式、GC 模式等也有益于判断 TSS 的位置。

(3)通过寡聚物频率来判断 TSS 的位置。

统计表明，在启动子、外显子和内含子区域的寡聚物出现频率十分不同，所以可以对这些区域进行统计学习。寡聚物指序列中相连的几个碱基。如六聚物频率指连续六个相连的碱基在序列中出现的频率。

2.2.2 *常用方法* 从使用方法分类，预测 TSS 的方法与基因预测的其它领域使用的方法类似，如用权重矩阵、判别分析、神经网络等多种方法来预测 TSS。多种方法的有效结合可以提高预测效率。

(1) 权重矩阵

权重矩阵是一种表达短小、无间隙序列模式的模型[5]。当扫描基因组序列时，与共有序列模式匹配的模式将获得高分。由于某些短模式偏好使用某些碱基，但并没有完全一致的共有序列。如 TATA 模式，可能表现为 TATAAA，也可能表现为 TATAAT，根据从 TF 数据库中得到的确定的 TATA 模式出现各种不同序列的频率，为各不同序列打分。

(2) 判别分析

判别分析是基因预测中常用的统计方法。TSSW[6]用它来判断一个给定序列是 $TATA^+$ 序列还是 $TATA^-$ 序列，而 FirstEF[4]用它来发现启动子等区域。它解决的问题是：给定 N 个对象，为每个对象分类到已知的几组中，保证分类错误率最小。通常将 N 个对象分为两类。

常用的判别分析方法有线性判别分析和二次判别分析。线性判别分析相对简单，若两类分布均为正态分布，且它们的协方差矩阵相同，则可以使用它，否则使用二次判别分析。利用线性判别分析，可以得到线性判别函数；利用二次判别分析，得到二次判别函数(quadratic discriminant function)。根据判别函数得到某个阈值，可以计算未知分类样本的 QDF，根据它与阈值的大小，为其分类。

(3) 神经网络

神经网络是一种典型的分类方法。通过建立网络，使用已知分类的数据对该网络进行训练，在训练过程中，网络的参数将逐步调整。训练完成后使用该网络对未知分类的数据进行预测。Promoter2.0[8]用它和遗传算法结合来发现启动子，DGSF[2]则使用它来判断预测的 TSS 和 CpG 岛的联合是否标志着一个基因的开始。

2.3 评估标准

敏感度(Sensitivity，简称 Sn)和精确度(Specificity，简称 Sp)是目前广泛采用的两个标准[9]。Sn 是预测正确的 TSS 占所有实际 TSS 的比例，Sp 是预测正确的 TSS 占所有预测 TSS 的比例，计算如下：

$$Sn = TP/(TP+FN).$$

$$Sp = TN/(TN+FP) = TP/(TP+FP)$$

表 1 TSS 预测效率的计算参考表

TP	正确预测 TSS 的数量
TN	正确预测非 TSS 的数量
FP	将非 TSS 预测成 TSS 的数量
FN	将 TSS 预测成非 TSS 的数量

相关系数(corelation efficient，简称 cc)是另外一个较常用的标准。当 cc 为 1 时，表示预测效果完美；当 cc 为 0 时，表示预测结果差，实际上此时为随机预测。cc 计算如下：

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

2.4 现有方法的总结和分析

在 TSS 预测领域，有几种重要方法。Eponine[5]适用于哺乳动物。它使用一组权重矩阵来识别特定序列模式。每一个矩阵都与一个相对于 TSS 的位置分布相关。它识别的模式有 TATAAAA、GCGCG 和 CpG 岛。它的预测敏感度大于 50%，精确度达到 70%。CpG_promoter 在大规模序列范围内粗略定位启动子的位置。它基于 CpG 岛上的二次判别分析。此时 CpG 岛被定义成长度大于 200bp、C+G 含量超过 50%、CpG 二核苷酸的频率超过 0.6 的序列。大约有一半的哺乳动物的基因的启动子区域与 CpG 岛相关。它达到的敏感度为 0.85，精确度为 0.42。CpGProD[10]识别哺乳动物基因组中与 CpG 岛相关的启动子。它与 CpGPromoter 类似。不同之处在于，它对于不同种类的基因组使用不同的参数，而且可以预测序列启动子的方向。它的敏感度为 56%，精确度为 39%。在方向预测上，准确度可达 70%。FirstEF[4]使用二次判别分析在人类基因组中定位第一外显子的位置。它首次通过将数据分类成 CpG 相关和无关的两种，分别预测两类数据的 TSS。这样提高了 CpG 相关的启动子的预测，进而提高整个预测效率。它的敏感度达到 86%，精确度达到 83%。DGSF[2]使用神经网络预测第一外显子开始的区域。它连接了三个系统，一个是 Dragon Promoter Finder，用来在查询序列的两条带中寻找 TSS，第二个系统用来预测 CpG 岛的存在，第三个系统通过一个四层的神经网络判断 CpG 岛与 TSS 的结合是否是一个基因的开始端。它的敏感度达到 65%，对于 CpG 相关的启动子，敏感度高达 88%。

3 本文算法

3.1 采用的方法

寻找 TSS 常用的方法是在启动子和第一外显子的末尾剪接点 GT 间利用滑动窗口不断定位，同时使用 CpG 岛等特征。本算法结合了 TATA 模式和翻译始点的特征，试图减小预测的可能区间，从而能够精确定位 TSS。

判别分析是本算法使用的主要方法。为了正确的判别，需要两个条件：(1)一组特征变量。它们对于类的判别具有决定性作用。这些变量的选取往往需要对欲分类的对象的本质有深刻理解[7]。(2)判别函数 C。给定一组特征变量的值，C 将它映射到类 1 或类 2。

假定有 N 个考察对象，每个对象有 p 个特征变量，那么可以将每个对象表示为一个 p 维空间中的点。通常，假定数据来自两个不同的分布＋和－，假设先验概率为 π_+ 和 π_-，且分布概率为 $p(x\mid+)=f_+(x)$ 和 $p(x\mid-)=f_-(x)$。则序列 x 为＋的后验概率是

$$q(+\mid x)=\frac{\pi_+ f_+(x)}{\pi_+ f_+(x)+\pi_- f_-(x)}$$

则似然比 $h(x)=\ln\frac{q(+\mid x)}{q(-\mid x)}$ 可以作为判别函数。当 $h(x)>0$

时，表示 x 为＋的概率更大，判定 x 来自＋；否则当 $h(x)<0$ 时，判定 x 来自－。

判别误差为：$\int_{R^+} f^-(x)\mathrm{d}x+\int_{R^-} f^+(x)\mathrm{d}x$，其中 R^+ 指误判为＋的区域，R^- 指误判为－的区域。

为了在序列中找到第一外显子的剪接点 GT、翻译始点、TATA 模式、启动子区域，需要利用训练好的判别函数对未知序列进行不断的判别，进而推测 TSS 的位置。

同时本算法也使用了权重矩阵，在扫描序列时，用来寻找潜在的 TATA 模式，并为每个模式打分。权重矩阵中的一个值 $s(x,b)$ 表示在 x 位置碱基 b 的得分，此矩阵行数为模式长度，列数为 4。为每个位置每个碱基给出一个评分，那么对于某个长度的连续序列，将每个位置的评分相加，可得到此序列的总评分。当此评分大于阈值时，可以判断此序列为一个潜在的模式。

3.2 算法步骤

输入：一段基因组序列

输出：TSS 的位置

算法流程图：见图 2

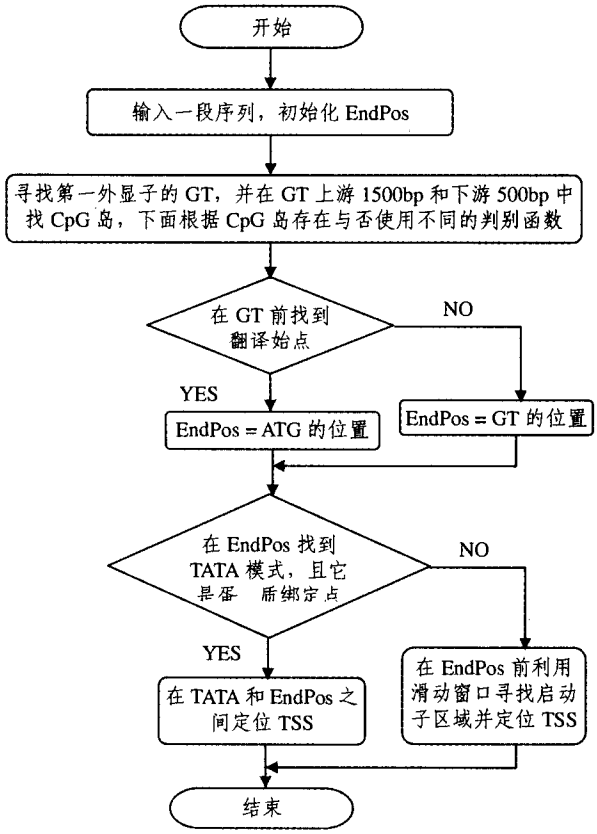

图 2　预测 TSS 的算法流程图

说明：

• EndPos 是一个变量，用来标志考察区域的一个边界，它可能为 ATG 的位置，也可能为第一外显子的 GT 的位置。

• 第一外显子的 GT 若找不到，继续找翻译始点。

• 在判断 GT 是否为第一外显子的 GT、ATG 是否为翻译始点、寻找启动子时滑动窗口是否为启动子区域、判断 TATA 模式是否为蛋白质绑定点时，分别使用训练得到的第一外显子末端 GT 的 QDF、翻译始点的 QDF、启动子的 QDF 和 TATA 模式的 QDF 进行判断。

• 在启动子中定位 TSS 时，认为 TSS 位于第 500 位，因为在训练启动子 QDF 时，选取了 TSS 上游 500bp 和 TSS 下游 70bp 作为启动子区域。

• 在 TATA 和 EndPos 间定位 TSS 时，根据训练集中 TATA 与 TSS 间距离的平均值定位。

4　实验和总结

日本通过"寡核苷酸加帽"的实验方法提供了 DBTSS 数据库，其中包含代表性全长的人类 mDNA 序列 8000 多条。将全长 mRNA 序列与基因组序列作比对，处理比对结果，可获得训练第一外显子中 GT 的正例和反例集。

本文以染色体 11 的序列为实验数据。通过如下方法构造第一外显子中 GT 的正例和反例集合：1 利用 dbtss 数据库中的 hspromoter 数据库和 reffull 的数据库文件 get_for_fasta，抽取染色体 11 的 reffull 序列 chr11；2 下载 NCBI 网站上提供的 Spidey 程序，该程序用来将 mRNA 与基因组序列比对，获得外显子在基因组上对应的坐标；3 利用 Spidey 将 chr11 与染色体 11 的基因组序列作比对。解析输出文件，由此从基因组序列中得到正例集共 496 条，获得反例集开放阅读框内的外显子共 2891 条。取 2/3 的正例和反例作为训练集，进行参数训练；另外 1/3 作为测试集，进行结果评估。

翻译始点、启动子的正例和反例在第一外显子数据集的基础上构造。

表 2　训练各功能点的正例和反例

	正例	反例
GT	第一外显子及其前后 500bp	其它外显子及其前后 500bp
TIS	第一外显子中翻译始点及其前后各 100bp	其它外显子中 ATG 及其前后各 100bp
启动子	第一外显子中，TSS 上游 500bp 和下游 70bp	其它外显子中任意选取的 570bp

注：TIS 为翻译始点

训练 QDF 时，各特征使用的变量如下：

(1)判断第一外显子中的 GT 选用的变量

坐标参照：以 GT 的 G 为＋1，GT 之前的位置为－1。

变量包括：a. 在窗口(1,200)中的六聚物评分。b. 在窗口(－200,－1)中的六聚物评分。c. 在窗口(1,64)中的三聚物评分。

(2)判断翻译始点选用的变量

坐标参照：坐标参照：以翻译始点标志序列 ATG 中 A 位置为＋1，ATG 之前碱基的位置为－1。

变量包括：a 在窗口(1,100)内的六聚物评分；b 在窗口(－100，－1)内的六聚物评分。

(4) 判断启动子区域选用的变量

坐标参照：以 TSS 前程 500bp 的位置为＋1

变量包括：a 在窗口(1,250)的六聚物评分；b 在窗口(200,450)中的六聚物评分；c 在窗口(1,50)中的五聚物评分；d 在窗口(420,500)中的五聚物评分；e 在窗口(490,570)中的五聚物评分；e 若序列是 CpG 相关的，选用窗口(1,570)中的 CpG 含量；否则，选用窗口(1,570)中的 GC 含量。

(5) 判断 TATA 模式选用的变量

a TATA 模式评分；b TATA 到 CpG 岛中心的距离(若没有 CpG 岛，定义该距离为无穷大)；c 选用 TATA 到 TSS 的距离。

n 聚物评分为窗口内所有 n 聚物的加权平均，权重为正例中 n 聚物的频率和非 n 聚物的频率的似然比。CpG 含量为 CG 连续出现的比率，GC 含量为 C 或 G 出现的比率。

实验结果如下：

表 3　预测精度

类别	Sn	Sp	CC
类 a	0.85	0.8	0.82
其余	0.66	0.58	0.61
全部	0.73	0.65	0.68

注：类 a 表示第一外显子包含翻译始点或启动子存在 TATA 模式的序列

该结果表明，对于第一外显子包含翻译始点或者启动子存在 TATA 模式的序列，本算法相对于 FirstEF[4]，TSS 预测效率有所提高。由于这部分序列预测效率的提高，序列整体的预测效果也较好。

结束语　本文提出了一个新颖的算法，通过结合翻译始点、CpG 岛、TATA 模式的特征，使用权重矩阵和判别分析在基因组序列上预测 TSS，并取得了较好的结果。这是由于结合翻译始点和 TATA 模式的特征，对于那些第一外显子包含翻译始点或者启动子存在 TATA 模式的情况，预测效率提高。同时，本算法也使用了近年来证明对于 TSS 和启动子预测效率有显著提高的 CpG 岛的特征。

由于本算法中的 QDF 训练仅基于染色体 11 的基因序列，下一步工作是对所有染色体序列进行训练，并在已经注释的染色体 21 和 22 上进行测试。

参 考 文 献

1 Michael Q. Zhang Computational Prediction of Eukaryotic Protein-coding Genes. In: Nature, 2002
2 Vladimir B. Bajic and Seng Hong Seah Dragon Gene Start Finder Identifies Approximate Locations of the 50 Ends of Genes. In: Nucleic Acids Research, 2003
3 Tao Jiang, Ying Xu, Michael Q. Zhang Computational Methods for Promoter Recognition. In: Current Topics in Computational Molecular Biology, the MITPress, 2002. 261～263
4 Davuluri R V, Grosse I, Zhang M Q. Computational Prediction of Promoters and First Exons in the Human Genome. In: Nature, 2002
5 Down TA, Hubbard TJ. Computational detection and location of transcription start sites in mammalian genomic DNA. In: Genome Res., 2002
6 Salamov A A, Solovyev V V. The Gene-Finder Computer Tools for Analysis of Human and Model Organisms Genome Sequences. In: Proc. of the Fifth Int. Conf. on Intelligent Systems for Molecular Biology, AAAI Press, 1997. 294～302
7 Michael Q. Zhang Discriminant analysis and its application in DNA sequence motif recognition. In: Brief Bioinform, 2000
8 Knudsen S. Promoter 2.0: for the Recognition of PolII Promoter Sequences. In: Bioinformatics, 1999
9 Scherf M, Klingenhoff A, Werner T. Highly Specific Localization of Promoter Regions in Large Genomic Sequences by PromoterInspector: A Novel Context Analysis Approach. In: J. Mol. Biol., 2000
10 Ponger L, Mouchiroud D. CpGProD: Identifying CpG Islands Associated with Transcription Start Sites in Large Genomic Mammalian Sequences. In: Bioinformatics, 2002

(上接第 141 页)

8 张军明，吴哲辉. 标识 S-图中同步距离的计算. 东南大学学报，1995(5)
9 袁崇义. Petri 网原理. 北京：电子工业出版社，1998
10 Peterson J. Petri net theory and the modeling of systems. 吴哲辉译. 中国矿业大学出版社，1989

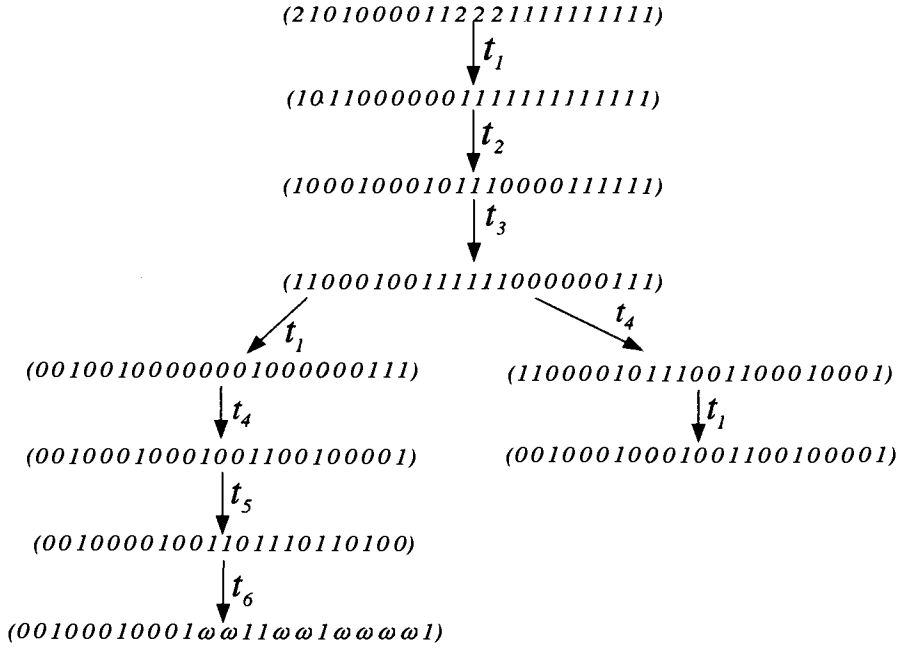

图 7　计算 Petri 网 Σ 同步距离生成树

$$
\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{array}
\begin{array}{c}
\begin{array}{ccccc} t_2 & t_3 & t_4 & t_5 & t_6 \end{array} \\
\begin{bmatrix} 1 & 1 & 2 & \omega & \omega \\ & 1 & 1 & \omega & \omega \\ & & 1 & \omega & \omega \\ & & & \omega & \omega \\ & & & & 1 \end{bmatrix}
\end{array}
$$

图 8　Petri 网 Σ 中变迁同步距离矩阵