文章编号:1001 - 2486(2006)04 - 0088 - 05

大肠杆菌转录起始位点的计算定位方法*

杜耀华,倪青山,王正志
(国防科技大学 机电工程与自动化学院,湖南 长沙 410073)

摘 要:根据已有的启动子识别算法,提出了一种基于滑动窗口的大肠杆菌转录起始位点(TSS)计算定位方法,通过在启动子信号特征中引入复合模式来改进识别分类器,并将其用于滑动窗口序列,在合理限定的 TSS 定位范围内依次计算各个序列位置的 TSS 似然得分,再利用 TSS 与翻译起始位点(TLS)的距离分布信息作为 TSS 的位置得分,两者相结合来进行位置预测。对大肠杆菌真实数据的测试表明,算法可以大幅度减少假阳性结果,实现对真实 TSS 位置的有效预测。

关键词:大肠杆菌;转录起始位点;计算定位;复合模式;滑动窗口
中图分类号:Q527 文献标识码:A

Computational Localization of Transcription Start Sites in Escherichia Coli Genomic Sequences

DU Yao-hua, NI Qing-shan, WANG Zheng-zhi
(College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Although a large number of researches have been undertaken in the area of transcription start site (TSS) localization, the problem of TSS localization has not yet been fully resolved. According to the previous promoter prediction algorithm, a new sliding window based computational localization method for E. coli TSSs is proposed. The TSS-likelihood scores of each possible position in genomic sequences are calculated by the window classifier which is improved by introducing the composite motif model in the training procedure of original promoter classifier. The distribution of distances between TSSs and translation start sites (TLSs) is also utilized to calculate the TSS-position scores. Localization results are achieved from the final score profiles which combine TSS-likelihood scores and TSS-position scores. The test results on E. coli dataset show that the method can find the putative TSSs and decrease the number of false positives efficiently.

Key words: escherichia coli; transcription start site (TSS); computational localization; composite motif; sliding window

转录起始位点(Transcription Start Site, TSS)的计算定位指的是通过计算的方法给出 TSS 在基因组序列中可能的位置。目前各种基因组数据库中的注释内容大多集中在序列的编码区域,与基因转录相关的信息还比较少。因此,TSS 的计算定位已经成为丰富基因组注释信息的重要手段和进行基因转录调控研究的基本前提。

从 TSS 上游延伸直至其下游的长度为数百碱基的序列区域通常被称为启动子(promoter)。作为与 TSS 密切关联的调控信号,启动子负责启动从 TSS 处起始的转录过程。鉴于两者在序列中的位置关系,当前的 TSS 定位基本上都依赖于对应启动子识别的结果。而针对启动子的识别,相关研究已陆续提出多种算法和工具[1-4]。然而,由于启动子信号固有的退化多变特性,即使对结构相对简单的原核启动子,识别算法的结果也不能完全令人满意[5-7]。已有的原核启动子识别方法可大致分成两类:基于组成(content)的方法和基于信号(signal)的方法。前者主要利用了启动子与背景序列在碱基组成上的差异,因此这类方法只能判断待定序列是否属于启动子区域,无法给出启动子的精确位置,更无法对 TSS 进行

* 收稿日期:2006 - 02 - 28
基金项目:国家自然科学基金资助项目(60471003)
作者简介:杜耀华(1978—),男,博士生。

定位。后者则通过发现启动子区域内特异的保守模式进行识别,可以对启动子的位置进行预测,并将识别结果的 3′端位点近似作为 TSS。困扰这类方法的问题是识别的特异性较低,会得到大量假阳性结果。例如有研究表明,每个真实大肠杆菌 σ^{70}启动子的附近平均有 38 个类似启动子信号[8]。因此,如何提高信号的特异性,降低假阳性成为提高启动子识别和 TSS 定位准确度的关键。

在先前的研究中,通过分析大肠杆菌 σ^{70}启动子的特性,提出了一种基于多种特征组合的启动子识别算法[9],在[TSS - 60…TSS…TSS + 20]的区域内分别计算启动子的组成特征、信号特征和结构特征,将特征得分组合成特征向量,再利用二次判别分析进行判别。该算法对启动子的各种特征信息进行综合,一定程度上提高了信号的特异性,在测试中获得了高于其它常用算法的识别正确率。

作为后续研究,本文基于文献[9]中介绍的启动子识别算法,提出了一种新的大肠杆菌 TSS 计算定位方法,其思想是采用形式为[S - 60…S…S + 20](长度为 81bp)的窗口,在基因翻译起始位点(Translation Start Site, TLS)上游一定范围区域内的每个碱基位置上依次滑动,利用文献[9]中的识别分类器计算窗口序列的启动子得分,将其作为窗口内位置处碱基的 TSS 似然分值,再根据整个区域似然值的分布情况确定 TSS 的位置。定位方法还对识别分类器所利用的启动子信号特征作了优化,引入了复合模式特征,进一步提高了信号的特异性,并利用 TSS 与 TLS 间的距离经验分布对 TSS 似然分值进行后验处理,减少了假阳性结果。

1 数据与方法

1.1 数据集的选取

用文献[9]中的启动子识别分类器计算窗口序列对应位置的 TSS 似然得分,因此分类器的训练数据来自文献[9]中整理的 683 条大肠杆菌 σ^{70}启动子序列集。剔除位于编码区或 TSS-TLS 距离大于 350bp 的序列,剩余的 580 条作为训练的正数据集。训练的负数据集则沿用文献[9]中非编码区负集的 612 条序列。

与真核生物相比,原核生物的 TSS 与其下游对应 TLS 之间的距离较短。实验证实,在已知的大肠杆菌 TSS 中,位于从 TLS 至其上游 350bp 区域之内的超过 90%[8,10]。因此,通常可以把 TSS 定位的范围限制在这一区域。根据这一原则,窗口[S - 60…S…S + 20]中位置 S 的可变范围为[TLS - 350…TLS],窗口序列扫过的范围则对应为[TLS - 410…TLS + 20]。按此格式,依据训练正集中提供的启动子对应 TLS 的位置信息,在大肠杆菌全基因组序列[11]中将训练正集的 580 条原长 81bp 的序列扩展为长 431bp 的序列,用以组成测试数据集,其中每条序列的 TSS 位置均为已知。

1.2 复合模式

从文献[9]中对大肠杆菌启动子各种特征的分析可知,-10 区模式和 -35 区模式两种信号特征属于较强特征,对识别的贡献较大。另有实验表明,-10 区和 -35 区模式的间隔距离大多为 16 ~ 18bp,这一距离使两个模式保持在双螺旋的同一侧,有利于与聚合酶分子相结合[12]。因此模式的间距也是一个重要的保守特征。文献[9]中计算信号特征的位置权重矩阵(Position Weight Matrix, PWM)时,将 -10 区模式和 -35 区模式作为独立的信号分别进行计算,并没有将间距信息真正结合到算法当中。为弥补这一不足,现将 -10 区模式、-35 区模式以及它们的间距组合成复合模式(composite motif),将其当作一个整体来考虑。

复合模式包含 -10/-35 区模式的 PWM 以及模式间距的分布信息,其得分 s_c 由 -10 区模式得分 s_p、-35 区模式得分 s_x 和模式间距得分 s_d 相加得到。用一种基于期望最大(Expectation Maximization, EM)思想的迭代寻优算法计算 PWM 和距离分布,以取代文献[9]中原有的计算方法。算法简述如下:

输入:训练集中的序列数目 W,-10 区模式起始位置 j 的变化区间[m,n],模式间距 l 的变化区间[p,q],迭代次数上限 T,PWM 长度 L,变化的下限 σ

1)初始化:建立 -10/-35 区模式初始 PWM E_0 和 G_0,以及模式间距初始经验分布函数 $F_0(l)$

2)循环 1,对 $t = 1,2,\cdots,T$,执行:

循环 2,对 $k = 1,2,\cdots,W$,执行:

计算复合模式得分 $s_c^t(k)$：

$$s_c^t(k) = \max_{\substack{j\in[m,n]\\ l\in[p,q]}} [s_p^t(j,k) + s_d^t(l,k) + s_x^t(j-l-L,k)] \tag{1}$$

其中：

$$s_p^t(j,k) = \sum_{i=0}^{L-1} [E_t^P(\alpha_{j+i}^k, i) - E_t^N(\alpha_{j+i}^k, i)] \tag{2-1}$$

$$s_d^t(l,k) = F_t(l) \tag{2-2}$$

$$s_x^t(j-l-L,k) = \sum_{i=0}^{L-1} [G_t^P(\alpha_{j-l-L+i}^k, i) - G_t^N(\alpha_{j-l-L+i}^k, i)] \tag{2-3}$$

记录 $s_c^t(k)$中的 -10 区模式起始位置 y_k^t 和模式间距 $d^t(k)$

循环 2 结束

对每条序列，提取从位置 y_k^t 起的 L 个碱基，重新构造 E_t；

提取从位置 $y_k^t - d^t(k) - L$ 起的 L 个碱基，重新构造 G_t；

根据 $d^t(k)$重新估计 $F_t(l)$

如果 $\| E_t - E_{t-1} \| < \sigma$ 且 $\| G_t - G_{t-1} \| < \sigma$，循环 1 结束

循环 1 结束

3) 输出 E_t、G_t 和 $F_t(l)$。

式(2-1)和(2-3)中的 $E_t(\alpha, i)$和 $G_t(\alpha, i)$分别为 E_t 和 G_i 第 i 列上碱基 α($\alpha \in \{A, C, G, T\}$)所对应的矩阵元素，上标 P 和 N 则对应训练的正负数据集。

E_0 的统计位置为[TSS-12…TSS-7]，G_0 的统计位置为[TSS-35…TSS-30]。此时两个模式的间距为 17bp，-10 区模式与 TSS 的间距为 6bp，是它们在 σ^{70}启动子中最典型的位置[9]。以此为初始点的目的是为了加速迭代过程的收敛，而 $F_0(l)$取均匀分布即可。-10 区模式起始位置的变化区间和模式间距的变化区间共同构成了最优复合模式的搜索空间，它们的大小需要利用先验知识来确定，也可在迭代过程中不断修正。

对于特定序列，利用 E_t、G_t 和 $F_t(l)$，由(1)和(2)计算复合模式得分 s_c。

引入复合模式取代 -10/-35 区模式，能更好地体现启动子信号特征的真实情况，有利于提高窗口序列对应位置 TSS 似然得分的特异性。

1.3 TSS 与 TLS 的间隔距离分布

在选取数据集时已经提到，大肠杆菌中 TSS-TLS 距离通常不超过 350bp。根据训练正集中已知的位置信息，可以计算出每条序列的 TSS-TLS 距离值，由此即可统计得到 TSS-TLS 距离在[0,350]区间上的经验概率分布。统计的直方图以及平滑后的经验分布曲线见图 1。

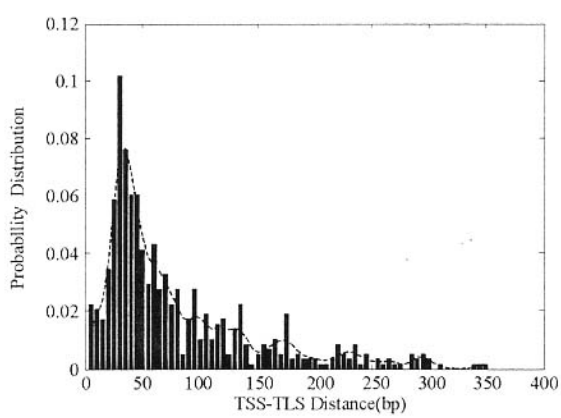

图 1 大肠杆菌 TSS-TLS 距离的直方图与经验概率分布

Fig.1 TSS-TLS distance histogram and smoothed empirical probability distribution for E. coli

由图 1 可知，TSS 在区间内并不是等概率分布的，而是更偏向位于离 TLS 较近的位置。图中的经验分布仅在整数位置取值，设此时离散经验分布函数为 $D(x)$，与相应 TLS 距离为 x 的 TSS 位置得分(position score) $s_t(x)$可由下式计算：

$$s_t(x) = \log[D(x)] \tag{3}$$

$s_t(x)$是 TSS 位置的一种先验信息，将用它来减少假阳性定位结果。

1.4 TSS 的定位

对于形式为[$S-60\cdots S\cdots S+20$]的窗口序列，用文献[9]中的二次判别函数(Quadratic Discriminant

Function, QDF)计算窗口中位置 S 处的 TSS 似然得分(likelihood score) s_l。设 S 与已知 TLS 的距离为 x,即有

$$s_l(x) = QDF(x) \tag{4}$$

将窗口沿测试序列滑动,即可得到各个位置的 TSS 似然得分。

取 TLS 上游[0,350]区间内的任一位置 x,定义其 TSS 最终得分(final score) s_f 为

$$s_f(x) = s_t(x) + s_l(x) \tag{5}$$

得到[0,350]区间内的 TSS 最终得分 $s_f(x)$之后,确立 TSS 准确位置的步骤如下:

1) 引入得分阈值 C_s,扫描整个区间,记录每个 s_f 不低于 C_s 的位置,并将满足条件的位置称为"岛";

2) 引入间隙阈值 C_a,将间隔距离未超过 C_a 的相邻岛合并;

3) 引入岛长阈值 C_l,淘汰长度小于 C_l 的岛;

4) 经过 1)、2)、3)之后,对剩余的岛,各岛内 s_f 值最高的位置即为 TSS 的预测位置。

2 测试结果

2.1 评价指标

TSS 定位常用的评价指标有敏感性 s_n 和特异性 s_p。定义 TP 为真实位点被正确定位的数目,FP 为虚假位点被定位为真实位点的数目(假阳性结果),NP 为真实位点的数目,则有

$$s_n = \frac{TP}{NP} \times 100\% \tag{6}$$

$$s_p = \frac{TP}{TP + FP} \times 100\% \tag{7}$$

2.2 阈值参数的确定

训练正集序列中的 TSS 均为经实验证实且位置已知的真实位点。为保证定位算法的敏感性,得分阈值 C_s 需要根据这些真实位点的 s_f 值来确定。

训练正集 TSS 的 s_f 值分布情况如图 2 所示。

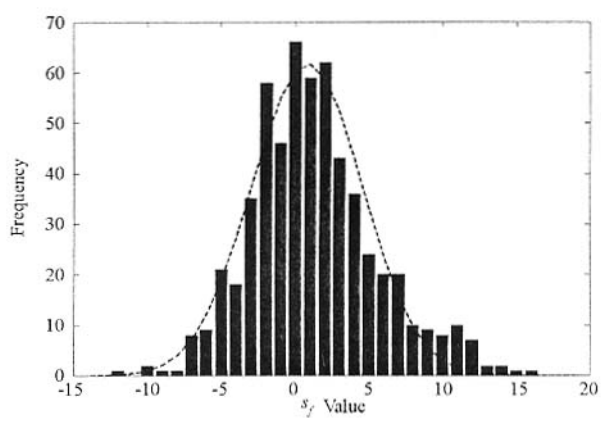

图 2 训练正集 TSS 的 s_f 值分布

Fig.2 The distribution of s_f for TSSs in positive training set

对图中的直方图进行正态概率检验证实,此经验分布近似服从正态分布。图中的虚线为对直方图做正态拟合之后得到的最优曲线,对应的分布参数均值 $\mu = 0.8$,标准差 $\sigma = 3.75$。至此,得分阈值 C_s 可取为 $\mu - \alpha\sigma$ 的形式,其中 α 为系数,可根据对算法敏感性的需求进行调整。

间隙阈值 C_a 和岛长阈值 C_l 可用刀切法(jackknife)通过在测试集上计算定位错误率的最小上界来确定。

2.3 测试结果

用 1.1 中整理的测试集(580 条序列,每条长 431bp)对定位算法进行测试,此时用刀切法求得 $C_a = 2$,$C_l = 3$。测试结果见表 1。需要指出的是,由于目前对 TSS 信号位点的认识比较有限,定位算法不可能对真实位点的位置做出完全精确的预测。即使是数据库中提供的真实位点也可能会因实验方法的局限而存在一定的位置误差。因此可以定义误差距离 d_t,只要预测位置落入区间[TSS − d_t, TSS + d_t],即被认为是正确的预测。在算法中,d_t 取 10bp。

作为对比,表 1 中分别给出了没有加入位置得分(仅似然得分)和加入位置得分(最终得分)后在不同得分阈值下的定位结果。

表1 大肠杆菌TSS定位的测试结果

Tab.1 Location results of *E. coli* TSSs

C_s	likelihood score only				final score			
	TP	*FP*	s_n(%)	s_p(%)	*TP*	*FP*	s_n(%)	s_p(%)
$\mu-0.5\sigma$	403	700	69.48	36.54	401	390	69.14	50.70
$\mu-0.75\sigma$	445	1233	76.72	26.52	449	742	77.41	37.70
$\mu-\sigma$	486	2098	83.79	18.81	488	1126	84.14	30.24
$\mu-1.25\sigma$	492	3527	84.83	12.24	493	1864	85.00	20.91

由表1可以看出,在一定范围内,随着得分阈值 C_s 的降低,定位算法的敏感性 s_n 将有所提高,但代价是假阳性结果 *FP* 也随之剧增。因此在实用中需要寻求两者之间的合理折衷,根据实际需求来确定得分阈值 C_s。从结果的对比也可以明显看出,在相近的敏感性水平下,加入位置得分使得算法减少了约45%的假阳性结果,特异性 s_p 平均提高了约11%。定位算法在 s_n 为85.00%时,其 s_p 达到了20.91%。此时平均每正确预测一个真实TSS,将得到4个假阳性位点,与每个真实位点附近有38个近似位点[8]相比,特异性得到了大幅度的提高。

3 结论

本文提出了一种基于滑动窗口的大肠杆菌TSS计算定位方法,在合理限定的TSS定位范围内,利用滑动窗口对序列进行扫描,通过窗口序列分类器计算各个位置的TSS似然得分。为了提高方法的特异性,在训练窗口分类器的启动子信号特征中引入了复合模式,并将TSS-TLS距离分布信息作为TSS的位置得分,与似然得分相结合来进行位置预测。对大肠杆菌真实数据的测试验证了定位算法预测真实TSS位置的能力和对减少假阳性结果的有效性。

应该看到,算法的实际结果与TSS的完全精确定位还有一段距离。相信随着对TSS信号研究的深入,将有更多的特征信息被发现和利用,这一差距将会不断缩小。另外,相关研究已经证实,基因组中许多基因对应的TSS并不唯一,而是存在多个备选位置。因此定位算法得到的假阳性结果中有些可能是潜在的备选TSS位点,并不是真正的假阳性,还需要进一步的研究和实验去证实。

参考文献:

[1] Werner T. Models for Prediction and Recognition of Eukaryotic Promoters [J]. Mammalian Genome, 1999, 10(2): 168 – 175.

[2] Pedersen A, Baldi P, Chauvin Y, et al. The Biology of Eukaryotic Promoter Prediction-a Review [J]. Comput. Chem., 1999, 23(3 – 4): 191 – 207.

[3] Stormo G. DNA Binding Sites: Representation and Discovery [J]. Bioinformatics, 2000, 16(1): 16 – 23.

[4] Ohler U, Niemann H. Identification and Analysis of Eukaryotic Promoters: Recent Computational Approaches [J]. TRENDS in Genetics, 2001, 17(2): 56 – 60.

[5] Vanet A, Marsanc L, Sagot M. Promoter Sequences and Algorithmical Methods for Identifying Them [J]. Res. Microbiol., 1999, 150(9 – 10): 779 – 799.

[6] Bajic V, Tan S, Suzuki Y, et al. Promoter Prediction Analysis on the Whole Human Genome [J]. Nature Biotechnology, 2004, 22(11): 1467 – 1473.

[7] 杜耀华, 王正志. 原核启动子识别研究进展 [J]. 生物技术, 2005, 15(5): 80 – 83.

[8] Huerta A, Collado-Vides J. Sigma70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals [J]. J. Mol. Biol., 2003, 333(2): 261 – 278.

[9] 杜耀华, 敖伟, 倪青山, 等. 一种基于组合特征的大肠杆菌启动子识别算法 [J]. 国防科技大学学报, 2005, 27(6): 113 – 119.

[10] Burden S, Lin Y X, Zhang R. Improving Promoter Prediction for the NNPP2.2 Algorithm: A Case Study Using Escherichia coli DNA Sequences [J]. Bioinformatics, 2005, 21(5): 601 – 607.

[11] Blattner F, Plunkett G, Bloch C, et al. The Complete Genome Sequence of Escherichia coli K – 12 [J]. Science, 1997, 277: 1453 – 1462.

[12] Lisser S, Margalit H. Compilation of E. coli mRNA Promoter Sequences [J]. Nucleic Acids Research, 1993, 21(7): 1507 – 1516.