文章编号:1001 - 2486(2005)06 - 0113 - 07

一种基于组合特征的大肠杆菌 σ^{70} 启动子识别算法*

杜耀华,敖 伟,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:启动子识别是研究基因转录调控的重要环节,但目前算法的识别正确率偏低。在深入分析启动子生物特征的基础上,提出了一种基于多种特征组合的大肠杆菌 σ^{70} 启动子识别算法,在启动子序列的组成特征、信号特征和结构特征中选取 10 种典型特征,以此为依据,对位于非编码区和编码区内部的启动子分别加以识别。首先通过特征描述模型分别计算各种特征在启动子序列和非启动子序列中的得分,将特征得分组合成 10 维特征向量,再利用二次判别分析法在特征向量集上进行训练和识别。在实际数据集中进行的刀切法测试验证了算法的有效性。对位于非编码区的启动子,平均正确率达到了 86.7%,明显优于其它算法;对位于编码区内部的启动子,平均正确率也达到了 82.4%。算法还具有良好的可扩展性,能够方便地容纳新特征,使识别性能不断提高。

关键词:大肠杆菌;σ^{70} 启动子识别;组合特征;二次判别分析法;刀切法
中图分类号:Q527 文献标识码:A

A Combined Features Algorithm for Prediction of E.coli σ^{70} Promoter Regions

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

Abstract: Promoter identification is an essential task in the research of transcription regulation, but computational prediction of promoters has been one of the most elusive problems despite considerable effort devoted to the study. A new prediction algorithm based on the combined features for E. coli σ^{70} promoters is proposed. According to their location, all promoters can be classified into two classes: promoters in non-coding regions and promoters in gene regions, and will be processed respectively. In each region, the features of primary sequence, including 1 content feature, 5 signal features and 4 structure features, are combined and defined as a 10 dimensional vector, then the vector of combined features is further used by quadratic discriminant analysis to predict the potential promoter regions. The algorithm has been trained and tested on E. coli σ^{70} promoter dataset by the jackknife method. The average prediction accuracies for "non-coding" promoters and "coding" promoters are 86.7% and 82.4%, respectively. The results indicate that our algorithm outperforms most of the existing approaches based on several performance measurements. Furthermore, algorithm framework is extendable and can accept more new features to improve the prediction results efficiently.

Key words: escherichia coli; σ^{70} promoter prediction; combined features; quadratic discriminant analysis (QDA); jackknife method

启动子的识别是构建基因表达调控网络的重要前提,已经成为计算生物学新的研究热点。目前各种基因组中启动子的可利用注释信息还很匮乏,迫切需要高精度的识别算法与方案。

将基因组序列视为由字母 $\{A, C, G, T\}$ 组成的字符串 S,已知 S 的某些特定位置上存在转录起始位点(TSS)。假设对任意一个位置 p,字符串片断 $[S_{p-U}\cdots S_p\cdots S_{p+D}]$ 包含足够的信息来判别 S_p 是不是一个 TSS,这样的片断称为待判别的启动子序列。给定一组训练数据集,包括正集(实验证实的启动子序列)和负集(不含启动子的序列),启动子识别过程就是通过训练,对任意给定的待判别启动子序列,判断其是否属于真正的启动子序列。

具有代表性的大肠杆菌 σ^{70} 启动子是原核启动子识别的主要对象之一。相关实验表明,大肠杆菌

* 收稿日期:2005 - 06 - 13
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。

σ^{70}启动子的核心区域一般从TSS上游约100bp处延伸至下游50bp左右,序列中含有若干短的保守模式(motif),其中最典型的有:-10区模式(-10 motif)、-35区模式(-35 motif)以及TSS[1]。另外,-10区和-35区模式的间隔距离也是一个重要特征[2]。

当前的原核启动子识别方法可分成两类:一类是基于组成(content)的方法,常见的有惩罚词频法[3]等。这类方法主要利用了启动子序列的碱基组成偏好特性,但由于只利用了组成信息,难以给出精确的预测位置,所以识别正确率比较低。另一类是基于信号(signal)的方法,通过发现启动子区域内的保守模式和结合位点等特征信号来进行识别。常用的解决方法有位置权重矩阵(PWM)[4]、人工神经网络(ANN)[3,5]、隐马尔可夫模型(HMM)[3,6]等。由于单一的保守模式片断太短,为了提高信号特异性,又出现了基于组合模式发现的MITRA[7]等方法。进一步的考虑是将模式发现得到的特征信号作为下一层识别模型的输入,通过整合所有特征来做出最终的识别。基于这种分层思想的方法有PWM+偏序覆盖函数(partial order cover function)[8]、比对核(sequence alignment kernel)+SVM[9]等。分层的方法综合利用了各个特征信号,并寻求特定准则下的最优决策,使得识别率有了一定提高。训练负集为编码区序列时,文献[9]中识别算法的正确率达到了81.4%,是已知方法中最高的。

然而,现有方法的识别正确率依然偏低[10]。除了特征信号本身固有的微弱多变因素,对启动子的认识不够深入,特征信息利用得不够充分是造成这种情况的主要原因之一。最近的实验发现,σ^{70}启动子区域存在一些新的特征模式,如-10区延伸模式(extended -10 motif)[11]、UP元件(UP element)[12]等等。已有的识别模型还没有利用这些特征。另外的研究证实,与其它区域序列相比,启动子区域具有较高的局部弯曲度(curvature)[13]和较低的双链稳定性(stability)[14]。虽然目前已有利用结构特征进行启动子识别的尝试,但实际的效果不能令人满意[15]。由此可知,目前启动子序列的组成特征、信号特征、结构特征在单独使用时均不能囊括全部启动子的信息,只有将这三类特征进行合理融合,才能最大限度地表征启动子的本质特性,为识别提供帮助。

启动子一般都位于转录单元(transcription unit)上游的非编码区。但原核基因组中的非编码区相对较短,下游转录单元的启动子有时会落入其紧邻上游转录单元的最后一个基因的编码区内。对位于编码区的启动子,现有的识别方法基本都是在数据准备时直接将其剔除掉,有些甚至根本不加考虑。这显然是不合理的。背景序列性质的不同会对模型参数产生较大影响,因此在识别时应该把两类启动子分开处理。

综上分析,本文提出了一种新的大肠杆菌σ^{70}启动子识别算法,对位于非编码区和编码区的启动子分别建模,并沿用了分层的思想,首先计算启动子序列的组成特征、信号特征和结构特征,然后将各个特征得分组合成高维特征空间的特征向量,再对向量进行训练和最优判别。

1 数据与方法

1.1 数据准备

1.1.1 正数据集的选取与分类

经实验证实的大肠杆菌启动子数据可以从数据库RegulonDB[16]中获取,其中σ^{70}启动子序列共有695条,每条长81bp,覆盖范围为TSS上游60bp至下游20bp,序列61bp处为TSS,格式为[TSS-60…TSS…TSS+20]。由于某些转录单元具有多个TSS,每个TSS都对应一条启动子序列,为减少数据集的冗余,当相邻TSS距离小于81bp时,下游TSS对应的启动子序列将被剔除。处理之后余下的683条序列将作为正数据集。

大肠杆菌全基因组序列数据可从GenBank中获取(AC号:U00096)。根据数据中提供的编码区位置信息对正数据集进行分类,整条序列全部落入某个编码区的启动子将被当作编码区启动子。根据这一标准,正数据集被划分为非编码区正集(612条)和编码区正集(71条)两部分。

1.1.2 负数据集的选取

负数据集应该从大肠杆菌全基因组序列中不含启动子的区域提取。但实际上并没有哪段区域被明确证明不含启动子。所以要根据转录单元的结构和启动子的分布特征尽量避开极有可能出现启动子的

区域。

非编码区按其两侧基因转录的方向可分成同向(tandem)、背离(divergent)与会聚(convergent)三类。原核生物的转录单元可以包含一个或多个基因,只有转录单元第一个基因的上游非编码区才可能含有启动子。显然,背离区肯定位于转录单元的上游,极有可能含有启动子;会聚区则肯定位于转录单元的下游,含有启动子的可能性很小;而同向区则有可能含有启动子,也可能不含。因此,在会聚的非编码区选取非编码负集,能最大限度地保证其不含启动子。大肠杆菌全基因组序列中长度不短于 81bp 的非编码序列片断共有 1966 条,其中属于会聚区的有 247 条。从会聚区序列中随机提取互不交迭的 612 条片断组成非编码区负集,每条序列长度为 81bp。

编码区启动子启动的是其下游的基因,因此一般都位于所在编码区的尾部。编码区越长,它的中前部含有启动子的可能性就越小。大肠杆菌全基因组序列中与已知 σ^{70} 启动子区域没有交迭,且长度大于 300bp(100 个密码子)的编码区共有 3845 条。从这些序列的中部随机提取 71 条长度为 81bp 的片断组成编码区负集。

负集序列的长度与格式与正集序列相同,可认为第 61bp 处为对应的虚假 TSS(nonTSS)。

1.2 特征集的计算与组合

我们选取的启动子特征集共包含分属组成、信号、结构三大类特征的 10 种信号。

1.2.1 组成特征的计算

组成特征的计算采用文献[3]中的词频分析方法。它将序列 k 的元组看作长度为 k 的单词,通过计算单词的出现频率来得到其词频得分 S_w。

计算得到的频率是单词出现概率的一种极大似然估计,数据集越大,估计值越可靠。在保证较高可靠性的前提下,词长 k 的选取将受到数据集实际大小的限制。因此,对非编码区数据集,取 k 值为 6;对编码区数据集,取 k 值为 4。

1.2.2 信号特征的计算

选取 -10 区模式、-35 区模式、TSS、-10 区延伸模式和 UP 元件共五种保守模式作为信号特征。图 1 给出了这些保守模式的一致序列在启动子区域的相对位置。

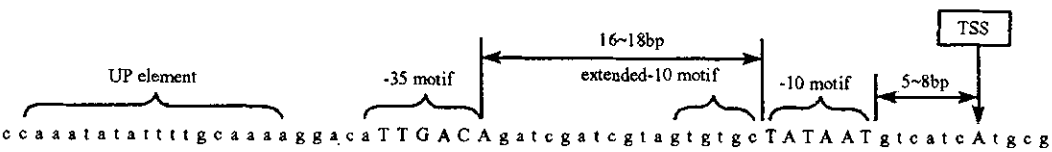

图 1 大肠杆菌 σ^{70} 启动子序列核心区域的保守模式

Fig.1 Conserved motifs in E. coli σ^{70} promoter sequences

描述保守模式的一种简单有效的模型是位置权重矩阵(PWM)。PWM 给出了保守模式每个位置上四种碱基出现频率的信息,其计算公式由文献[4]给出。如果已知 PWM,可根据公式(1)计算模式的特征得分,其中 j 为模式在序列中的起始位置,A 为 PWM 的长度,a_{j+1} 表示序列第 $j+i$ 位置处的碱基,$M^P(a,i)$ 为正集 PWM 第 i 列上碱基 a 对应的元素值,$M^N(a,i)$ 为负集 PWM 第 i 列上碱基 a 对应的元素值。

$$s_j = \sum_{i=0}^{A-1} [M^P(a_{j+i},i) - M^N(a_{j+i},i)] \tag{1}$$

TSS 的 PWM 在序列中的位置为[TSS - 2…TSS…TSS + 3],长度为 6。UP 元件 PWM 的位置为[TSS - 60…TSS - 41],长度为 20。这些位置固定的 PWM 可直接在训练集中计算得到。

由于从 RegulonDB 获得的数据正集中没有 -10 区模式和 -35 区模式的位置信息,它们的 PWM 要通过一个迭代学习过程得到,具体构造算法如下:

输入:m 条长度为 l 的序列,可能位置 j 的区间$[m,n]$,迭代次数上限 $T=50$,PWM 长度 $A=6$,变化的下限 σ。

(1)初始化:建立初始的位置权重矩阵 M_0,并设初值 $t=0$;

(2)循环1,对 $t=1,2,\cdots,7$,执行:

循环2,对 $k=1,2,\cdots,m$,执行:

$$s_k^t = \max_{j\in[m,n]}\left\{\sum_{i=0}^{A-1}\left[M_t^P(a_{j+i}^k,i) - M_t^N(a_{j+i}^k,i)\right]\right\} \tag{2}$$

$$y_k^t = \operatorname{argmax}_{j\in[m,n]}\left\{\sum_{i=0}^{A-1}\left[M_t^P(a_{j+i}^k,i) - M_t^N(a_{j+i}^k,i)\right]\right\} \tag{3}$$

循环2结束;

对每条序列,提取从位置 y_k^t 起的 A 个碱基,重新构造 M_{t+1};

如果 $\| M_{t+1} - M_t \| < \sigma$,循环1结束

循环1结束;

(3)输出 M_{t+1}。

构造 M_0 时,-10区模式的位置为[TSS-12…TSS-7],-35区模式的位置为[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 σ^{70} 启动子中最典型的位置[2]。考虑到两个模式的间距范围为16~18bp这一保守特征,对于-10区模式,j 的区间限定为[TSS-16…TSS-10];对于-35区模式,j 的区间限定为[TSS-38…TSS-32]。

-10区延伸模式紧邻在-10区模式上游,所以当-10区模式PWM构造完成以后,根据[$y_k-5\cdots y_k-1$]即可计算长度为5的-10区延伸模式PWM。

训练得到全部PWM之后,即可由公式(1)分别计算-10区模式得分 s_p、-35区模式得分 s_x、TSS得分 s_i、-10区延伸模式得分 s_l 和UP元件得分 s_u。

1.2.3 结构特征的计算

序列局部弯曲度和双链稳定性是我们要利用的结构特征。

序列的局部弯曲度可通过双螺旋结构局部偏角的变化来描述。偏角变化越大,弯曲度越高。利用文献[17]中的预测模型,可以计算转角(roll)翘角(tilt)以及扭角(twist)这三种最主要偏角的变化值。将角度转换为弧度,即可作为转角得分 s_r、翘角得分 s_t 和扭角得分 s_t。

双链稳定性可通过碱基结合的自由能(free energy)来描述。自由能越高,稳定性越低。由文献[18]中提供的预测模型可直接计算自由能得分 s_f。

1.2.4 特征的组合

算法通过一个组合的过程来实现启动子三类特征的融合,将每个特征作为一维,整个特征集就成为一个10维的特征空间,全部的特征得分可组合为一个10维特征向量 $\boldsymbol{s}$:

$$\boldsymbol{s} = [s_w, s_p, s_x, s_i, s_e, s_u, s_r, s_l, s_t, s_f]' \tag{4}$$

至此,通过计算和组合特征得分,数据集中的每条序列均可用10维特征空间的一个特征向量来表示,启动子识别问题就转换为特征空间中特征向量的判别问题。

1.3 组合特征向量的判别

根据判别分析的原理训练的分类器形式比较简单,并且在很多情况下非常有效。由于正数据集和负数据集对应的组合特征向量均近似服从多元正态分布,且协方差互不相等,所以可采用二次判别分析法(QDA),计算总体平均损失最小准则下的二次判别曲面。

设启动子向量为第1类,非启动子向量为第2类,则QDA的判别函数(QDF)为:

$$QDF = \log\frac{p_1^0}{p_2^0} - \frac{1}{2}(D_1^2 - D_2^2) - \frac{1}{2}\log\frac{|S_1|}{|S_2|} \tag{5}$$

其中,p_i^0 为第 i 类的先验概率,D_i^2 为待判别向量与第 i 类的Mahalonobis平方距离,S_i 为第 i 类的协方差矩阵,$|S_i|$ 为 S_i 的行列式。

Mahalonobis平方距离的计算公式如下:

$$D_i^2 = (\boldsymbol{x} - \boldsymbol{m}_i)' S_i^{-1} (\boldsymbol{x} - \boldsymbol{m}_i) \tag{6}$$

其中,x 为待判别向量,m_i 为第 i 类的特征均值向量,n_i 为第 i 类的向量数目。采用 Mahalonobis 平方距离的优点是可以在一定程度上克服组合特征向量各分量之间的相关性干扰,并且消除因计算模型不同而引入的量纲影响。

根据式(5),对待判别向量 x,如果 $QDF > 0$,则判别 x 为启动子向量;如果 $QDF \leqslant 0$,则判别 x 为非启动子向量。

1.4 特征的评价与筛选

如何合理地选取特征集是判别分析能否成功的关键。好的特征集中应该没有过度冗余的特征,并且尽量多的包含对判别贡献较大的特征。

特征集中如果存在过度冗余的特征,利用(6)式进行计算时,协方差阵 S 将成为病态矩阵,从而造成求逆困难,因此必须事先予以剔除。算法选取的 10 种启动子特征都有特定的生物和物理意义,虽然部分特征之间存在一定的冗余,但经检验并没有过度冗余的特征。

特征对判别贡献的大小可通过计算正负集之间的 D^2 来衡量,D^2 越大,贡献越大。表 1 给出了 10 种启动子特征在非编码区数据集和编码区数据集中的 D^2 距离平均值。

表 1 启动子特征的评价

Tab.1 The significance of features for promoter regions

Features	D^2 for non-coding promoters	D^2 for coding promoters
Word frequency	3.95	5.67
– 10 motif	2.84	3.30
– 35 motif	0.45	0.56
TSS	0.53	0.57
Extended – 10 motif	0.13	0.10
UP element	0.30	0.74
Roll	0.20	0.60
Tilt	0.11	0.10
Twist	0.44	0.61
Free energy	0.42	0.85

由表 1 可以看出,6 元(4 元)词频组成、– 10 区模式和 – 35 区模式的 D^2 距离比较大,将对判别作出主要贡献,属于强特征;而 – 10 区延伸模式和翘角的类间距较小,属于弱特征。这与前面启动子典型特征的描述基本相符。所以,选取特征集时应首先考虑 D^2 较大的特征。在精度允许的情况下,也可以去掉类间距很小的微弱特征,以提高算法的效率。

2 结果与讨论

2.1 评价指标

启动子识别常用的评价指标有敏感性 Sn、特异性 Sp、相关系数 CC 和平均正确率 AC。

定义 TP 为真实启动子被识别为真实启动子的数目,TN 为虚假启动子被识别为虚假启动子的数目,FP 为虚假启动子被识别为真实启动子的数目,FN 为真实启动子被识别为虚假启动子的数目,则有:

$$Sn = \frac{TP}{TP + FN} \tag{7}$$

$$Sp = \frac{TP}{TP + FP} \tag{8}$$

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{9}$$

$$AC\% = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \tag{10}$$

2.2 测试结果

利用1.1节中准备好的数据集,采用刀切法(jackknife)对算法进行测试,即按照每次提取一条数据作测试集,其余的数据全部作训练集的方式,依次对全部数据测试一遍,再对所有测试结果进行统计作为此数据集的测试结果。由于负数据集是在备选数据中随机提取的,为了体现算法的真实识别水平,减小偶然性,每次连续生成10组数据进行测试,将各次测试的平均结果作为算法的最终测试结果。

非编码区数据集(正集612条数据,1组;负集612条数据,10组)和编码区数据集(正集71条数据,1组;负集71条数据,10组)的测试结果见表2。

表2 大肠杆菌 σ^{70} 启动子的刀切法测试结果

Tab.2 Prediction results of E. coli σ^{70} promoters in the jackknife test

Negative datasets	Non-coding region promoters				Coding region promoters			
	Sn	*Sp*	*CC*	*AC*(%)	*Sn*	*Sp*	*CC*	*AC*(%)
1	0.91	0.84	0.74	86.8	0.82	0.87	0.69	84.5
2	0.91	0.84	0.74	86.7	0.80	0.84	0.65	82.4
3	0.92	0.84	0.74	87.1	0.80	0.83	0.63	81.7
4	0.91	0.83	0.73	86.1	0.82	0.81	0.62	81.0
5	0.91	0.84	0.73	86.6	0.80	0.86	0.68	83.8
6	0.91	0.83	0.73	86.5	0.82	0.82	0.63	81.7
7	0.91	0.83	0.73	86.4	0.80	0.80	0.61	80.3
8	0.91	0.83	0.73	86.4	0.80	0.83	0.63	81.7
9	0.91	0.86	0.76	88.0	0.83	0.86	0.69	84.5
10	0.92	0.83	0.73	86.3	0.80	0.84	0.65	82.4
Average	0.91	0.84	0.74	86.7	0.81	0.84	0.65	82.4

当利用10种特征时,对位于非编码区的启动子,算法的敏感性 *Sn* 和特异性 *Sp* 分别达到了0.91和0.84,识别的平均正确率 *AC* 为86.7%;而对位于编码区内的启动子,*Sn* 和 *Sp* 也分别达到了0.81和0.84,平均正确率 *AC* 为82.4%。

2.3 与其它方法的比较

由文献[3]中的思路发展而来的惩罚词频+SVM法和文献[9]中的比对核+SVM法分别是基于组成和基于信号两大类识别方法中的代表。它们主要识别位于非编码区的启动子,数据集也从RegulonDB中获得,与本文采用的非编码区启动子数据集基本相同。因此对于非编码区启动子,本文的组合特征+QDA法与这两种方法的识别结果具有一定的可比性,具体结果见表3。

表3 不同算法对非编码区启动子的识别结果

Tab.3 Prediction results for non-coding promoters based on different methods

Method	*Sn*	*Sp*	*CC*	*AC*(%)
Combined features + QDA(our work)	0.91	0.84	0.74	86.7
Sequence alignment kernel + SVM	0.81	0.81	0.63	81.4
Zone likelihood + SVM	0.67	0.84	0.56	77.5

从表 3 中可以看出,对于非编码区启动子,我们的算法在各个评价指标上均达到或超过了当前的两种具有代表性的算法,识别的平均正确率有了较为明显的提高。

3 结 论

本文提出了一种基于多种特征组合的大肠杆菌 σ^{70}启动子识别算法,将启动子按其与基因的相对位置划分为非编码区启动子和编码区启动子两类,分别进行训练和识别;通过深入分析启动子在基因组中可能出现的区域来合理确定背景数据的选取范围;训练中利用了启动子的多种特征(组成特征、信号特征和结构特征),其中采用了一些新特征,并运用向量组合的方式实现了多类特征的融合,将序列转换成特征向量,通过在特征空间中的统计决策来达到识别启动子的目的。启动子分类、背景数据规整以及特征组合等措施增强了启动子信号的特异性,提高了识别模型的合理性和信息利用率。在实际序列组成的数据集中对算法进行了刀切法测试,并与其它几种已有典型算法的结果进行了比较。结果显示,对位于非编码区的启动子,本文的算法在各项指标上均达到甚至超过了其它算法,识别的平均正确率有明显提高。对其它算法没有考虑的编码区内部启动子,识别的平均正确率也达到了 82.4%。

多特征组合识别算法的优点是它具有良好的可扩展性,可通过增加新特征来不断提高识别性能,而融合新特征只需在组合特征向量中相应增加新的分量,无需改变模型的结构。算法的不足之处在于识别保守模式的部分子模型过于简单,没有充分利用模式间距的信息。因此,考虑以组合模式的形式进行模式识别,寻求更精确的保守模式规律的描述方法将是后续研究工作的重点。

参 考 文 献:

[1] Harley C B, Reynolds R P. Analysis of E. coli Promoter Sequences[J]. Nucleic Acids Research, 1987, 15(5): 2343 – 2361.

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

[3] Oppon E. Synergistic Use of Promoter Prediction Algorithms: A Choice for Small Training Dataset?[D]. South Africa: Western Cape University, 2000.

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

[5] Mahadevan I, Ghosh I. Analysis of E. coli Promoter Structures Using Neural Networks [J]. Nucleic Acids Research, 1994, 22(11): 2158 – 2165.

[6] Pedersen A, Baldi P, Brunak S, et al. Characterization of Prokaryotic and Eukaryotic Promoters Using Hidden Markov Models [A]. Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology [C], 1996: 182 – 191.

[7] Eskin E, Pevzner P A. Finding Composite Regulatory Patterns in DNA Sequences [J]. Bioinformatics, 2002, 18(S1): S354 – 363.

[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] Gordon L, Chervonenkis A, Gammerman A, et al. Sequence Alignment Kernel for Recognition of Promoter Regions [J]. Bioinformatics, 2003, 19(15): 1964 – 1971.

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

[11] Mitchell J, Zheng D L, Busby S, et al. Identification and Analysis of Extended – 10 Promoters in Escherichia coli [J]. Nucleic Acids Research, 2003, 31(16): 4689 – 4695.

[12] Estrem S, Gaal T, Ross W, et al. Identification of an UP Element Consensus Sequence for Bacterial Promoters [J]. Proc. Natl. Acad. Sci., USA, 1998, 95(17): 9761 – 9766.

[13] Bolshoy A, Nevo E. Ecologic Genomics of DNA: Upstream Bending in Prokaryotic Promoters [J]. Genome Research, 2000, 10(8): 1185 – 1193.

[14] Wang H Q, Noordewier M, Benham C J. Stress-induced DNA Duplex Destabilization (SIDD) in the E. coli Genome: SIDD Sites are Closely Associated with Promoters [J]. Genome Research, 2004, 14(8): 1575 – 1584.

[15] Kanhere A, Bansal M. A Novel Method for Prokaryotic Promoter Prediction Based on DNA Stability [J]. BMC Bioinformatics, 2005, 6: 1 – 32.

[16] Salgado H, Gama-Castro S, Martínez-Antonio A. RegulonDB (Version 4.0): Transcriptional Regulation, Operon Organization and Growth Conditions in Escherichia coli K-12 [J]. Nucleic Acids Research, 2004, 32(Database Issue): D303 – D306.

[17] Goodsell D S, Dickerson R E. Bending and Curvature Calculations in B-DNA [J]. Nucleic Acids Research, 1994, 22(24): 5497 – 5503.

[18] SantaLucia J. A Unified View of Polymer, Dumbbell and Oligonucleotide DNA Nearest-neighbor Thermodynamics [J]. Proc. Natl. Acad. Sci., USA, 1998, 95(4): 1460 – 1465.