大肠杆菌内源性转录终止子的支持向量机预测方法

杜耀华[1], 王正志[2], 吴太虎[1*]

(1. 军事医学科学院 卫生装备研究所, 中国 天津 300161; 2. 国防科技大学 机电工程与自动化学院, 中国湖南 长沙 410073)

摘 要: 内源性转录终止子的计算预测是基因转录调控研究的重要内容, 但当前方法的预测特异性偏低. 在深入分析大肠杆菌内源性终止子中 RNA 发夹结构和多聚胸腺嘧啶区域等特征信号的基础上, 为内源性终止子建立了一个由 5 个特征变量组成的包含序列组分、局部构象和能量分布信息的特征集, 并根据此特征集实现了一种基于支持向量机的内源性终止子计算预测方法. 针对大肠杆菌内源性终止子数据集和编码区阴性对照集的六重交叉验证测试证实了预测方法的有效性, 对已知数据的预测平均正确率达到了 99.4%. 在对大肠杆菌全基因组限定范围内的搜索中, 该预测方法可以成功地识别出绝大多数已知内源性终止子, 与其他几种常用方法相比, 预测结果总数大幅度减少, 预测的特异性有了明显提高.

关键词: 转录终止; 内源性终止子; RNA 发夹结构; 多聚胸腺嘧啶区域; 支持向量机

中图分类号: Q615 **文献标识码**: A **文章编号**: 1007-7847(2012)06-0471-08

Prediction of Intrinsic Transcription Terminators in *Escherichia coli* by Using Support Vector Machine

DU Yao-hua[1], WANG Zheng-zhi[2], WU Tai-hu[1*]

(1. *Institute of Medical Equipment, Academy of Military Medical Sciences, Tianjin* 300161, *China*; 2. *College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha* 410073, *Hunan, China*)

Abstract: Computational prediction of intrinsic transcription terminators is an essential task in the research of transcription regulation of gene, but the performances of current methods are still unsatisfying. According to the thorough analysis about signals such as RNA hairpin structure and T-region in *E. coli* intrinsic terminators, a 5 variable-included feature set which combines the sequence content, local conformation and energy distribution information is selected. Based on such feature set, a new prediction method using support vector machine for intrinsic terminators is proposed. The favorable performance is achieved in 6-fold cross validation test on *E. coli* positive and negative control datasets, and the average prediction accuracy is 99.4%. This method is then used to scan the putative intrinsic terminators in the whole genome of *E. coli*. Comparing with several other methods, the total number of scanning hits decreases greatly when most of known intrinsic terminators are retrieved. The specificity of prediction results has been improved effectively.

Key words: transcription termination; intrinsic terminator; RNA hairpin structure; T-region; support vector machine (SVM)

(*Life Science Research*, 2012, 16(6): 471~478)

转录终止是基因表达的一个基本步骤, 与转录的起始和延伸共同组成完整的转录过程[1]. 在基因组序列中, 其转录产物能与 RNA 聚合酶或辅助蛋白因子发生交互作用, 从而使转录过程终止的特定片段称为终止子[2]. 作为转录的终止信号, 终止子在基因表达调控中起着重要作用[3]. 因此, 对

收稿日期: 2012-09-25; **修回日期**: 2012-11-18

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, 河北唐山人, 军事医学科学院卫生装备研究所助理研究员, 博士, 主要从事生物信息学与生物医学工程学研究; * 通讯作者: 吴太虎 (1962-), 男, 山西沁县人, 军事医学科学院卫生装备研究所研究员, 博士生导师, 主要从事生物医学工程学与野战卫生装备学研究, Tel: 022-84656856, E-mail: wutaihu62@gmail.com.

终止子的特征分析与计算预测将推动转录终止机制的研究和基因调控网络的构建, 丰富基因组数据库的注释信息. 在细菌等原核基因组中, 转录通常以操纵子(operon)为单位进行, 终止子的预测还有利于揭示相关操纵子的结构和功能.

鉴于真核生物转录终止过程的复杂性, 当前对其终止信号和辅助因子的了解还非常有限. 然而, 针对大肠杆菌等基因组结构相对简单的原核生物, 其转录终止规律已获得了初步的认识. 研究表明, 大肠杆菌主要有两种转录终止机制: 因子依赖型(factor-dependent)终止和非因子依赖型(factor-independent)终止[4]. 因子依赖型终止子通常需要辅助蛋白因子(主要是 rho 因子)的参与才能有效地实现转录终止[5, 6], 而非因子依赖型终止子仅凭自身的信息即可实现转录终止, 因而又称为内源性(intrinsic)终止子[7]. 两类终止子的实际区分并不严格, 对于内源性终止子, 辅助因子虽然不是必需的, 但它们的存在有助于提高其终止效率[8]. 与因子依赖型终止相比, 内源性终止的机制更为简单经济, 其终止子的结构特征因此获得了较多的研究. 相关实验证实, 大肠杆菌内源性终止子主要由可以形成 RNA 发夹(hairpin)结构的富含 G/C 的回文序列以及紧邻其后的多聚胸腺嘧啶区域(T-region)组成[2, 7], 详细的结构如图 1 所示. 终止子序列的互补链即为模板链, 其转录得到的初生 mRNA 将会形成对应的发夹结构和多聚尿嘧啶尾巴. 转录中的 RNA 聚合酶遇到发夹结构将会暂停前进, 在紧随其后的多聚尿嘧啶区域, 由于 rU·dA 的杂合能较弱, 初生 mRNA 将和 DNA 模板链完全分离, 转录终止得以实现[9~11].

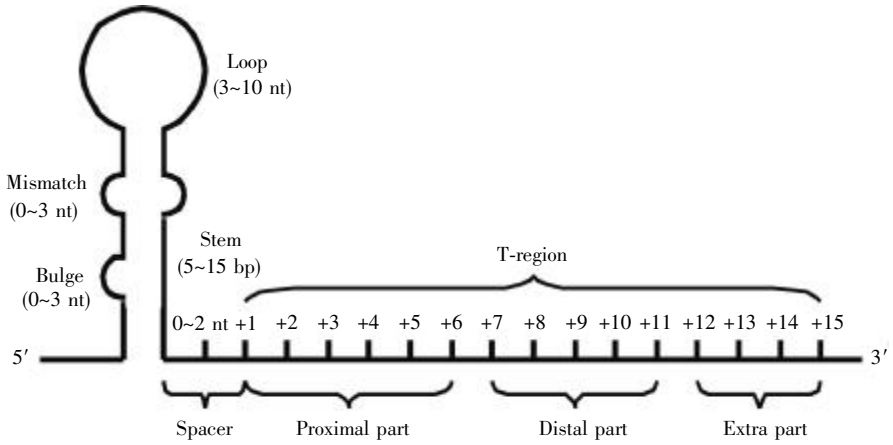

图 1 非因子依赖型内源性终止子的结构

从 5′端到 3′端依次为: (a) 发夹结构, 其中环长 3~10 个碱基, 茎长 5~15 个碱基对(茎上至多有 0~3 个碱基的错配或 0~3 个碱基的凸起); (b) 0~2 个碱基的间隔(除 T 以外的任何碱基); (c) 多聚胸腺嘧啶区域, 由 6 个碱基的近邻区、5 个碱基的远端区和 4 个碱基的额外区组成.

Fig.1 The structure of a factor-independent intrinsic terminator

Regions include from 5′ to 3′: (a) a hairpin with a loop of 3~10 nt and a stem of 5~15 bp (there may has 0~3 nt mismatches or 0~3 nt bulges in either 5′ or 3′-side of stems); (b) a spacer of 0~2 nt (any bases except T); (c) T-region divided into three parts: proximal part of 6 nt, distal part of 5 nt, and extra part of 4 nt.

随着对内源性终止子特征研究的不断深入, 相继出现了多种计算预测算法. 早期的尝试包括双核苷酸分布矩阵 (dinucleotide distribution matrix)[12, 13]和人工神经网络(artificial neural network, ANN)[14]等方法. 它们通常只利用了内源终止子序列的总体组成偏好信息, 没有考虑稳定性等方面的特征. 而随后的预测方法更多的是以内源终止子的发夹结构和多聚胸腺嘧啶区域两种局部特征作为核心信号进行预测, 区别仅在于对其描述和计算方案的不同: Carafa 等[15]用统计的方法为多聚胸腺嘧啶区域建立打分函数, 计算其组成权重, 再结合发夹结构的稳定性指标进行二元线性判别; De hoon 等[16]对多聚胸腺嘧啶区域打分函数的形式作了改进; TransTerm[17]基于同样的两类特征, 只是用逐步阈值过滤代替了二元线性判别; 另有 Yada 等[18]将这两类特征信息细化为 7 种特征参数进行多元线性判别. 这些方法均没有考虑多聚胸腺嘧啶区域的稳定性. 与之相对, RNAMotif[19]和 GeSTer[20]将结构稳定性作为形成有效内源终止子的决定因素, 根据发夹结构和多聚胸腺嘧啶区域的自由能来做出判别. 它们没有考虑序列组成方面的特征. 之后出现的 Rnall 方法[21]则较全面的利

用了局部特征区域的结构稳定性和序列组成信息, 因而获得了已知方法中最优的预测性能. 然而, 上述的各种方法要么只考虑内源性终止子的全局特征[12~14], 要么只考虑其局部特征[15~21], 信息利用得都不够充分. 这使得它们虽然能以较高的精度识别出已知的内源性终止子, 但同时也会得到大量假阳性结果, 预测的特异性较低. 显然, 最直接的改进思路是将全局特征和局部特征综合起来, 为内源性终止子选取更加全面的特征集. 另外, 近期的研究发现, 在大肠杆菌等原核生物基因组中, 与基因终止密码子相邻的下游序列区域(即终止子所在区域)通常具有比编码区和其他非编码区序列更高的弯曲度(curvature)[22, 23]. 因此, 可以将终止子序列的弯曲度作为一个新的全局特征信号. 总之, 无论是对已有特征的综合还是新特征的加入, 都可以丰富特征集的信息, 使其尽可能全面地表征内源性终止子的本质特征, 以期改善预测的特异性.

综合以上分析, 本文为内源性终止子建立了由其多种全局特征和局部特征组成的新特征集, 并根据此特征集提出了一种基于支持向量机的内源性终止子计算预测方法. 针对大肠杆菌内源性终止子数据集和编码区阴性对照集的交叉验证(cross validation) 测试证实了该预测方法的有效性. 在对大肠杆菌全基因组相关区域进行的扫描预测中, 该方法可以成功地识别出绝大多数已知内源性终止子, 与其他几种常用方法相比, 总的预测结果数目大为减少, 预测的特异性有了较大的提高.

1　数据与方法

1.1　数据集的选取与分析

1.1.1　内源性终止子数据的选取与分析

本文使用的内源性终止子数据来自文献[15]中整理的 147 条经过实验证实的大肠杆菌内源性终止子序列. 利用比对程序 BLAST[24], 在大肠杆菌全基因组 (包含全基因组序列和基因注释信息的数据文件可从 GenBank 得到, 序列 AC 号: U00096)中对这些原始序列数据进行校正, 重新获得的 138 条序列组成内源性终止子数据集(阳性集). 数据集中每条序列的长度为 55 nt, 格式为 [T−40LTL T+14]: 其中 T 为多聚胸腺嘧啶区域的起始位置. 序列的前 40 nt 对应发夹结构区域, 后 15 nt 对应多聚胸腺嘧啶区域.

图 2 给出了对阳性集中各终止子的发夹结构参数(茎长, 环长)、多聚胸腺嘧啶区域各个位置上 T 的出现频率以及多聚胸腺嘧啶区域起始位置与紧邻上游基因 3′端的距离分别进行统计的结果. 其中的发夹结构参数由 RNA 二级结构预测程序包 RNAstructure[25]预测得到, 而多聚胸腺嘧啶区域起始位置与紧邻上游基因 3′端的距离则参考了大肠杆菌全基因组序列中的基因注释信息. 由图 2(A)、(B)、(C), 结合文献[14]中对原始数据集的统计结果, 可以得到内源终止子的一些约束条件:

1) 设发夹结构的茎长为 n_s, 环长为 n_l, 发夹与多聚胸腺嘧啶区域的间距为 n_p, 则有 $5\leqslant n_s\leqslant 15$, $3\leqslant n_l\leqslant 10$, $0\leqslant n_p\leqslant 2$, 其中 n_s 的单位为 bp, n_l 和 n_p 的单位为 nt;

2) 发夹结构中至多允许 1 段错配(≤3 nt)和 1 个凸起(≤3 nt);

3) 多聚胸腺嘧啶区域的前 2 个位置均为 T, 前 4 个位置中至少有 3 个 T, 前 6 个位置中至少有 4 个 T.

为了描述多聚胸腺嘧啶区域中 T 的分布情况, 文献[15]提出了一种基于位置权重矩阵的打分函数, 用于计算整个区域的 T 组成得分 T_s:

$$T_s=\sum_{n=1}^{15}x_n \tag{1}$$

其中 x_1=0.9, 当 $n\geqslant 2$ 时, 有:

$$x_n=\begin{cases}x_{n-1}\times 0.9 & (\text{if the } n^{\text{th}} \text{ base is a T})\\ x_{n-1}\times 0.6 & (\text{if the } n^{\text{th}} \text{ base is other than T})\end{cases}$$

为了避免使用固定的权重系数, 文献[16]将(1)式改进为:

$$T_s=\sum_{n=1}^{15}\exp(-\lambda(n-1))\delta^n \tag{2}$$

其中δ^n 当第 n 个位置碱基为 T 时取 1, 不为 T 时取 0. λ 为经验系数, 可根据阳性集拟合得到. 然而, 由图 2(C)可知, 对于阳性集, T 的出现频率曲线并不符合任何已知函数形式. 为了更好地反映分布的实际情况, 直接利用 T 在各位置上的出现频率来建立打分函数:

$$T_s=\sum_{n=1}^{15}f(\mathrm{T},\ n)\delta^n \tag{3}$$

其中 f(T, n)为阳性集中多聚胸腺嘧啶区域第 n 个位置上 T 的出现频率, δ^n 的定义同(2)式. 利用(3)式计算阳性集中各终止子序列对应的 T_s, 可得到 T_s 的下界 T_s'=3.855. 因此, 对于所有的内源终止子, 应满足 $T_s\geqslant 3.855$. 将此式作为约束条件 4),

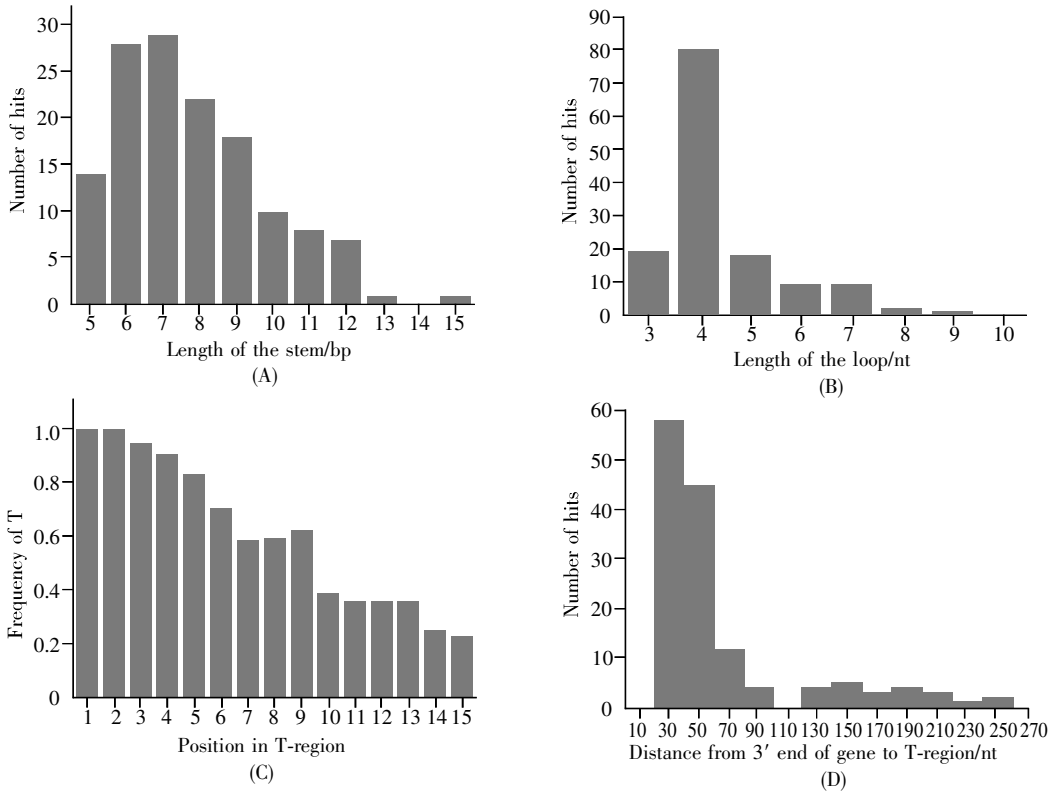

图 2 已知大肠杆菌内源性终止子的统计直方图

(A) 茎长; (B) 环长; (C) 多聚胸腺嘧啶区域各个位置上 T 的出现频率; (D) 多聚胸腺嘧啶区域起始位置与紧邻上游基因 3′ 端的距离.

Fig.2 The statistical histograms of known *E. coli* intrinsic terminators

(A) Length of the stem; (B) Length of the loop; (C) Frequency of T in each position of T-region; (D) Distance from 3′ end of gene to T-region.

可与前面提及的 3 个约束条件共同组成内源性终止子的基本约束集. 符合基本约束集中所有条件的序列称为内源性终止子的备选序列, 而计算预测的目的就是从众多备选序列中筛选出真正的内源性终止子序列.

1.1.2 阴性对照数据的选取

终止转录的功能使得终止子一般不会位于基因的编码区之内. 因此, 在基因编码区内出现的备选序列将被认为是虚假的内源性终止子序列. 这些序列可作为与阳性集相对的阴性对照集. 基于阳性集和阴性对照集, 即可训练预测分类器.

获取阴性对照数据集的基本步骤如下:

1) 根据大肠杆菌全基因组序列中的基因编码区注释信息, 提取全部的编码区序列;

2) 由于原核基因组中的非编码区相对较短, 基因编码区的两端可能与某些终止子序列存在部分交迭.可按照文献[17]中的处理方法, 去除编码区首尾各 100 nt 的序列 (长度不足 200 nt 的序列直接淘汰);

3) 在剩余的编码区序列中, 利用内源性终止子的基本约束集进行扫描, 并按照与阳性集序列相同的格式保留符合条件的备选序列. 经过计算, 最终得到的 10 932 条备选序列组成阴性对照集.

1.2 特征集的选取与计算

1.2.1 特征集的选取

本文为内源性终止子建立了一个包含 5 种特征信号的特征集. 这 5 种特征分别是终止子序列的碱基组成、终止子序列的弯曲度、发夹结构的平均碱基自由能、多聚胸腺嘧啶区域的杂合能以及发夹茎干中的 GC 碱基对含量. 前两种属于全局特征: 碱基组成反映了序列的组分信息; 弯曲度则是序列的局部构象参数, 反映的是结构信息. 而后三种属于局部特征: 发夹结构的平均碱基自由能和多聚胸腺嘧啶区域的杂合能分别描绘了内源性终止子两个局部核心信号的结构稳定性, 反映的是能量信息; 发夹茎干中的 GC 碱基对含量反映的则是发夹区域的组分信息. 对于多聚胸腺嘧啶区域的组分信息, 在内源性终止子的基本约束集中已有考虑, 此处不再单独列为特征信号.

此特征集综合了组分、结构和能量三方面的

信息, 与已有预测方法所利用的特征集相比, 更加全面地表征了内源性终止子的本质特征.

1.2.2　特征的计算

前面已经提到, 早期的双核苷酸分布矩阵预测方法[12, 13]利用的就是终止子序列的碱基组成信息. 与之类似, 本文采用 1 阶位置权重矩阵(position weight matrix, PWM) 对终止子序列的双核苷酸组成特征进行描述和计算. 对应阳性集和阴性对照集序列的格式, 可训练规模为 16×55 的 1 阶 PWM. 得到 PWM 之后, 即可计算特定序列的双核苷酸相似性得分. 训练 PWM 以及计算相似性得分的具体方法参见文献[26].

终止子序列的弯曲度可利用程序包 CURVATURE[27]进行计算. 该程序根据局部偏角的二苷参数模型估算序列轴向的弯曲度.

发夹结构的自由能可通过程序包 RNAstructure 直接计算, 能量的具体数值主要根据通用的热力学能量参数模型[28]估算得到. 然而, 由文献[15]中的分析可知, 真实终止子和虚假终止子的发夹结构自由能 ΔG 分布存在大范围的交迭, 差异并不十分明显. 但如果设 L_H 为发夹结构的序列总长, 定义平均碱基自由能 $\Delta G'$:

$$\Delta G'=\frac{\Delta G}{L_H} \tag{4}$$

将 $\Delta G'$作为特征信号, 其区分度将大幅度增强. 因此, 可根据 RNAstructure 的结构参数和自由能数值计算发夹结构的平均碱基自由能 $\Delta G'$.

由图 1 所示, 内源性终止子的多聚胸腺嘧啶区域可划分为近邻(proximal)、远端(distal)和额外(extra)3 个部分, 再加上发夹结构与近邻部分之间的间隔区域(spacer), 这 4 个部分对稳定性的贡献各不相同[19, 21]. 首先根据 RNA/DNA 杂合能二苷参数模型[29]分别计算各部分的杂合能, 再利用文献[19]中提出的公式计算整个区域的杂合能 T_h:

$$T_h=\Delta G_{\text{spacer}}+\Delta G_{\text{proximal}}+0.5\times\Delta G_{\text{distal}}+0.01\times\Delta G_{\text{extra}} \tag{5}$$

内源性终止子发夹结构的茎干中富含 GC 碱基对. 根据发夹结构的预测结果, 用茎干中实际出现的 GC 碱基对数目除以茎干的碱基对总数, 即为茎干 GC 碱基对的含量.

经过特征计算, 每条内源性终止子备选序列均可得到一个 5 维特征向量. 在后续的分类器训练和预测中, 特征向量将代替原始序列参与计算.

1.3　支持向量机预测分类器

支持向量机 (support vector machine, SVM)是一种新型统计学习方法, 它寻求结构风险最小化的准则下的最优分类面, 以获得相比传统方法更好的分类性能, 并且能在一定程度上避免过学习现象, 具有较强的泛化能力. SVM 的核心思想是通过核函数变换将输入空间的样本映射到高维特征空间, 在高维特征空间中寻找最优分类面, 从而将样本分开. 由此可见, 核函数类型的选择和确定核函数后相关参数的选取是决定 SVM 性能的关键. 由于目前还没有针对具体问题构造出合适核函数的有效方法, 实际中利用的大多还是多项式核函数、RBF 核函数、感知器核函数等标准核函数. 相关研究表明, RBF 核函数是一个普适核函数, 通过调整参数, 可以适用于任意分布的样本, 其表达形式为:

$$K(\vec{x}_i, \vec{x}_j)=\exp(-\|\vec{x}_i-\vec{x}_j\|^2/\sigma) \tag{6}$$

其中 $\vec{x}_i$, $\vec{x}_j$ 为输入向量, σ 为半径参数. σ 过大, 样本的"势力范围"会过大, 以至于一些无关的训练样本会干扰对测试样本的判断; σ 过小, 会导致分类器只有记忆功能而无法对新的样本进行判断. 在实际的计算中, 通常根据对分类器性能的具体需求来确定的 σ 大小.

因此, 本文选用基于 RBF 核函数的 SVM 来对内源性终止子特征向量进行训练和预测, 其具体计算过程可利用通用软件包 LIBSVM[30]来实现. 由于特征向量中的各个分量值来自不同的计算模型, 为了消除量纲差异带来的影响, 在输入 SVM 之前先要对其进行标准化处理. 对于训练集, 设 s_{ij} 和 s^*_{ij} 分别为其中第 i 个特征向量的第 j 个分量的原始数值和标准化数值, 则有:

$$s^*_{ij}=\frac{s_{ij}-\bar{s}_j}{\sigma_j}\quad \begin{matrix} i=1, 2, \cdots, N; \\ j=1, 2, \cdots, 5 \end{matrix} \tag{7}$$

其中 $\bar{s}_j$ 和 σ_j 分别为特征向量第 j 个分量的平均值和标准差, N 为训练集特征向量的总数. 对于测试集, 处理方法与(7)式相同, 但对应的 $\bar{s}_j$ 和 σ_j 仍需使用根据训练集特征向量算得的数值.

2　预测结果

2.1　特征的评价

按照 1.2.2 中的方法分别对内源性终止子的阳性集和阴性对照集序列进行特征计算, 得到特征向量 5 个分量值在两类数据集中的分布直方图(图 3). 显然, 分量值在两类数据集中的分布总体

差异越大, 其对应特征的区分能力也就越强, 对预测的贡献也就越大. 由图 3 可知, 序列的双核苷酸相似得分和发夹结构的平均碱基自由能所对应的分布差异比较明显, 是内源性终止子特征集中的显著特征 (prominent feature), 将在预测中起主要作用; 而其余三个特征分量则属于微弱特征 (weak feature), 在预测中起次要作用.

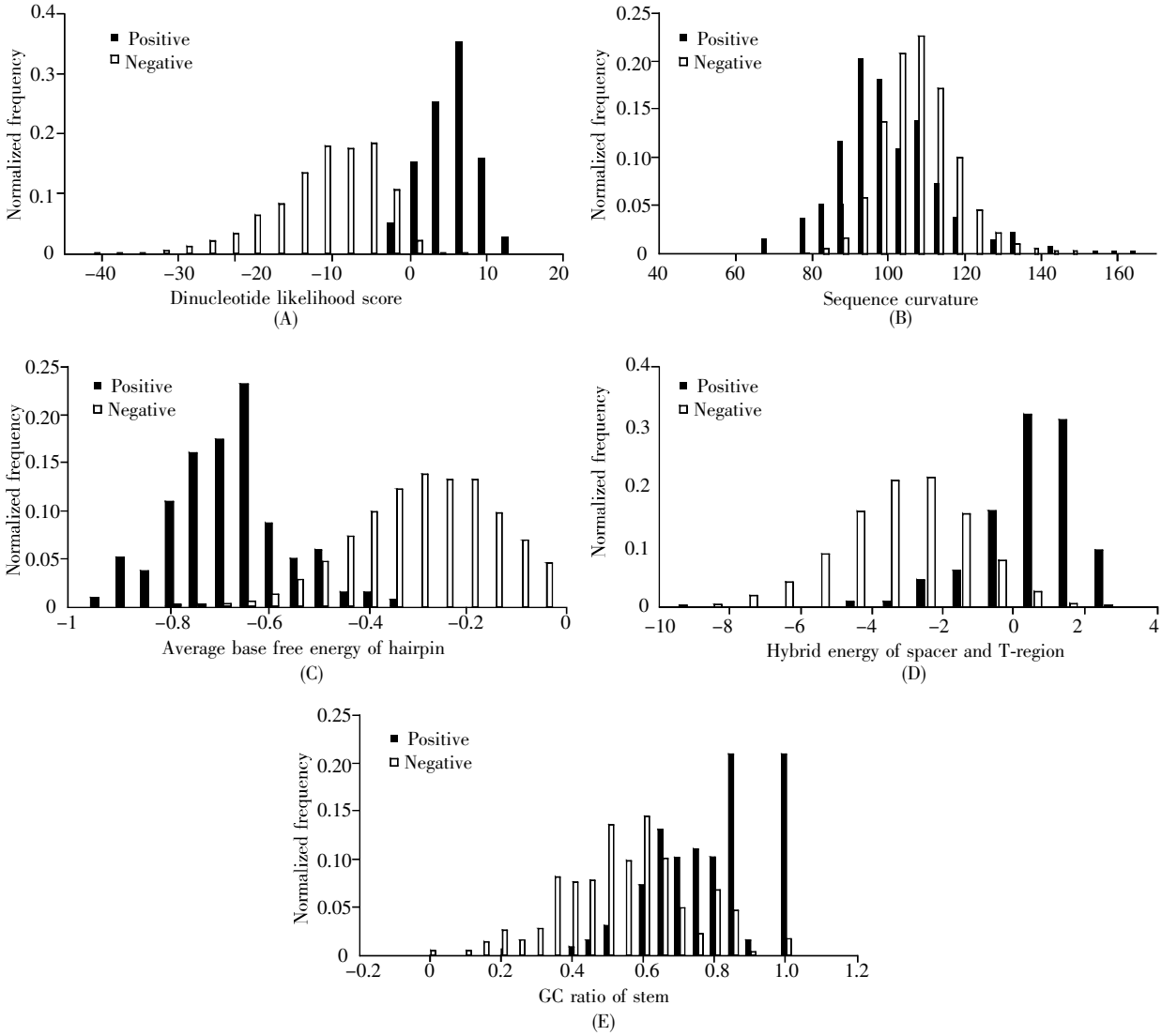

图 3 特征量值在阳性集和阴性对照集中的分布直方图

(A) 双核苷酸相似性得分; (B) 序列弯曲度; (C) 发夹结构的平均碱基自由能; (C) 多聚胸腺嘧啶区域的杂合能; (E) 茎干的 GC 碱基对含量.

Fig.3 The distributional histograms of feature values in positive and negative control datasets

(A) Dinucleotide likelihood score; (B) Sequence curvature; (C) Average base free energy of hairpin; (D) Hybrid energy of spacer and T-region; (E) GC ratio of stem.

2.2 交叉验证测试

内源性终止子预测的常用性能评价指标有敏感性 *Sn*(sensitivity)、特异性 *Sp*(specificity)、假阳性率 *FP*%(false positive rate)和平均正确率 *AC*%(average accuracy rate). 定义 *TP* 为真阳性数目, *TN* 为真阴性数目, *FP* 为假阳性数目, *FN* 为假阴性数目, 则有:

$$Sn=\frac{TP}{TP+FN} \tag{8}$$

$$Sp=\frac{TP}{TP+FP} \tag{9}$$

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

$$AC\%=\frac{TP+TN}{TP+FN+FP+TN}\times 100\% \tag{11}$$

根据上述评价指标对内源性终止子的阳性集

(138 条序列)和阴性对照集(10 932 条序列)进行六重交叉验证测试. 通过调整 RBF 核函数的半径参数 σ, 可以得到多组测试结果, 对应不同的指标水平. 图 4 给出了根据测试结果绘制的 ROC(receiver operating characteristics)曲线图. 由图可知, 敏感性指标 *Sn* 和特异性指标 *FP*%是互斥的, 不能同时达到最佳值, 需要根据实际需求在两者之间寻找折衷. 而平均正确率 *AC*%则为两者提供了一个综合性的指标. 在本文的测试中, 当 σ=25 时, *AC*%取得最大值, 对应的最优结果见表 1. 此时的预测方法对已知内源性终止子的敏感性为 0.93, 同时能拒绝 1–*FP*%=99.5%的编码区虚假终止子, 预测的平均正确率达到了 99.4%.

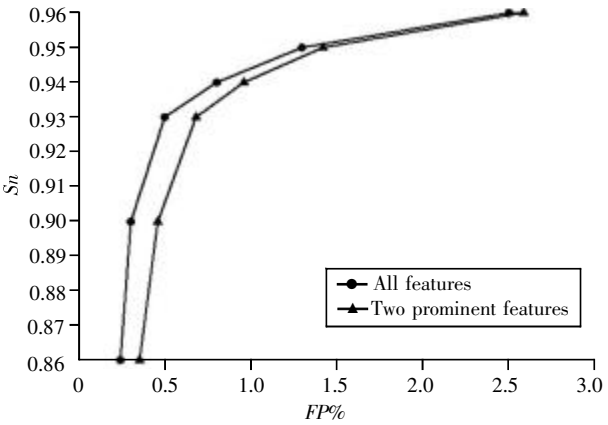

图 4　基于五重特征向量的六重交叉验证测试 ROC 曲线图
Fig.4　ROC curves of our prediction method in 6-fold cross validation test based on 5 different features

表 1　六重交叉验证测试(σ=25)的预测结果
Table 1　The results of our prediction method in 6-fold cross validation test(σ=25)

Dataset	*Sn*	*Sp*	*FP*%	*AC*%
1	0.87	0.74	0.38	99.5
2	0.96	0.67	0.60	99.3
3	0.96	0.69	0.55	99.4
4	0.91	0.72	0.44	99.5
5	0.96	0.69	0.55	99.4
6	0.91	0.70	0.49	99.4
Average	0.93	0.70	0.50	99.4

2.3　大肠杆菌全基因组搜索预测

基于内源性终止子的阳性集和阴性对照集训练 SVM 分类器, 对大肠杆菌全基因组进行搜索预测, 有可能发现更多的未知内源性终止子. 相关分析表明, 绝大多数的已知内源性终止子位于紧邻基因 3′端的非编码区内, 极少数与下游基因有交迭[16, 17, 19]. 由图 2(d)可知, 138 条已知内源性终止子的多聚胸腺嘧啶区域起始位置与紧邻上游基因距离均没有超过 260 nt. 因此, 可以把内源性终止子的搜索预测范围限制在基因 3′端上游 100 nt 至下游 300 nt 的区间之内 (当基因间隔区域较短时, 搜索预测范围可能会包含下游基因的序列. 为与选取编码区阴性对照集时的范围相对应, 限定搜索预测范围至多包含下游基因 5′端的 100 nt 序列, 此时的预测范围以实际的序列长度为准). 预测时, 首先在搜索范围内对满足内源性终止子基本约束集的所有备选序列进行定位, 然后逐一计算其对应的标准化特征向量, 将其输入 SVM 分类器, 以确定是否将其归类为内源性终止子.

经过计算, 当 σ=25 时, 本文的方法可在大肠杆菌全基因组限定的序列范围内预测出 610 个可能的内源性终止子, 其中包括 136 个已知内源性终止子, 针对阳性集的敏感性 *Sn* 约为 0.99. 这一结果与常用方法 RNAMotif、GeSTer 和 Rnall 的搜索预测结果的比较见表 2. 表 2 中 4 种方法都利用了文献[15]中的已知内源性终止子数据, 具有基本相同的阳性集. 由表 2 可以看出, 与其他方法相比, 本文的方法在更高的阳性集敏感性水平和更宽的序列搜索范围内却得到了更少的预测结果, 预测的特异性有明显提高. 除去已知的 136 条序列, 预测结果中剩余的 474 条序列很有可能是尚未发现的真实内源性终止子, 它们是否具有实际的生物功能还需要后续实验的进一步证实.

表 2　不同预测方法在大肠杆菌全基因组中的搜索结果
Table 2　The scanning results of different prediction methods in *E.coli* genome

Method	Scanning interval	*Sn* for positive dataset	Number of prediction
RNAMotif	[−10, +60]	0.81	1 075
GeSTer	[−20, +270]	0.90	1 883
Rnall	[−50, +280]	0.92	1 193
Our work	[−100, +300]	0.99	610

3　讨论

假阳性结果过多导致特异性偏低是当前大肠杆菌内源性终止子计算预测方法存在的主要问题. 针对这一情况, 本文首先对已有的各种特征信息进行合理综合, 并引入序列弯曲度特征, 为大肠杆菌内源性终止子建立了一个包含 5 个特征变量的更加全面的特征集, 然后在新特征集上实现了基于支持向量机的内源性终止子预测方法. 该预测方法在对大肠杆菌已知数据集进行的交叉验证测试和对大肠杆菌全基因组限定范围内的搜索预测中均获得了优于其他几种常用方法的性能评价, 预测的特异性有了大幅度提高.

内源性终止子特征集中的每个特征对预测的贡献是不相同的. 本文根据特征向量各分量值在阳性集和阴性对照集中的分布差异对相应特征的强弱进行了粗略的评价, 将其划分为显著特征和微弱特征两类. 由图 4 中的比较测试结果可知, 仅利用两个显著特征, 预测方法就可以获得与利用全部特征时相近的性能. 可见, 特征对预测的贡献与其强弱程度密切相关. 除了定性分析, 还可以通过引入诸如类间距离等参数来对特征的强弱程度进行定量计算. 当备选特征较多时, 利用这些参数从中筛选出有效特征组成特征集, 可以提高预测的效率. 另外, 基于特征集的预测方法框架具有良好的可扩展性. 随着研究的不断深入, 会有更多有效的内源性终止子特征信号被发现[31~33]. 它们丰富了特征集包含的信息, 预测方法的性能也会因此得到持续的改善.

相关研究指出, 部分内源性终止子缺少多聚胸腺嘧啶区域, 这一局部特征信号可能并不是内源性终止所必需的[20, 34]. 本文的预测方法虽然是针对具有“发夹-多聚胸腺嘧啶区域”结构的内源性终止子提出的, 但只要重新选取特征集, 并对备选序列扫描等环节进行相应的调整, 无须对方法框架进行大的修改, 就能够用于只有发夹结构的更为一般性的内源终止子预测.

目前对具有发夹结构的内源性终止子相关特征的认识主要来自大肠杆菌的研究. 虽然发夹结构并不是原核生物转录终止的普适机制[35], 但具有发夹结构的内源性终止子仍然在终止子中占有重要地位. 考察它们在不同物种中的特征信息变化, 并将预测方法扩展应用到其他原核物种基因组序列将是后续研究的重点.

参考文献(References):

[1] von HIPPEL P. An integrated model of the transcription complex in elongation, termination and editing[J]. Science, 1998, 281(5377): 660-665.

[2] PLATT T. Transcription termination and the regulation of gene expression[J]. Annual Review of Biochemistry, 1986, 55: 339-372.

[3] RICHARDSON J. Transcription termination[J]. Critical Reviews in Biochemistry and Molecular Biology, 1993, 28(1): 1-30.

[4] HENKIN T. Control of transcription termination in prokaryotes[J]. Annual Review of Genetics, 1996, 30: 35-57.

[5] BOGDEN C, FASS D, BERGMAN N, *et al*. The structural basis for terminator recognition by the rho transcription termination factor[J]. Molecular Cell, 1999, 3(4): 487-493.

[6] SKORDALAKES E, BERGER J. Structure of the rho transcription terminator: mechanism of mRNA recognition and helicase loading[J]. Cell, 2003, 114(1): 135-146.

[7] YARNELL W, ROBERTS J. Mechanism of intrinsic transcription termination and anti-termination[J]. Science, 1999, 284(5414): 611-615.

[8] DAS A. Control of transcription termination by RNA-binding proteins[J]. Annual Review of Biochemistry, 1993, 62: 893-930.

[9] GUSAROV I, NUDLER E. The mechanism of intrinsic transcription termination[J]. Molecular Cell, 1999, 3(4): 495-504.

[10] LEE S, KANG C. Opposite consequences of two transcription pauses caused by an intrinsic terminator oligo (U): antitermination versus termination by bacteriophage T7 RNA polymerase[J]. Journal of Biological Chemistry, 2011, 286(18): 15738-15746.

[11] HUANG Y, WENG X, RUSSU I M. Structural energetics of the adenine tract from an intrinsic transcription terminator[J]. Journal of Molecular Biology, 2010, 397(3): 677-688.

[12] BRENDEL V, TRIFONOV E. A computer algorithm for testing potential prokaryotic terminators[J]. Nucleic Acids Research, 1984, 12(10): 4411-4427.

[13] BRENDEL V, HAMM G, TRIFONOV E. Terminators of transcription with RNA polymerase from *Escherichia coli*: what they look like and how to find them[J]. Journal of Biomolecular Structure & Dynamics, 1986, 3(4): 705-723.

[14] NAIR T, TAMBE S, KULKARNI B. Application of artificial neural networks for prokaryotic transcription terminator prediction[J]. FEBS Letters, 1994, 346(2-3): 273-277.

[15] CARAFA Y, BRODY E, THERMES C. Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures[J]. Journal of Molecular Biology, 1990, 216(4): 835-858.

[16] de HOON M, MAKITA Y, NAKAI K, *et al*. Prediction of transcriptional terminators in *Bacillus subtilis* and related species[J]. PLoS Computational Biology, 2005, 1(3): e25.

[17] ERMOLAEVA M, KHALAK H, WHITE O, *et al*. Prediction of transcription terminators in bacterial genomes[J]. Journal of Molecular Biology, 2000, 301(1): 27-33.

[18] YADA T, NAKAO M, TOTOKI Y, *et al*. Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models[J]. Bioinformatics, 1999, 15(12): 987-993.

[19] LESNIK E, SAMPATH R, LEVENE H, *et al*. Prediction of rho-independent transcriptional terminators in *Escherichia coli*[J]. Nucleic Acids Research, 2001, 29(17): 3583-3594.

[20] UNNIRAMAN S, PRAKASH R, NAGARAJA V. Conserved economics of transcription termination in eubacteria[J]. Nucleic Acids Research, 2002, 30(3): 675-684.

[21] WAN X F, XU D. Intrinsic terminator prediction and its application in *Synechococcus sp*. WH8102[J]. Journal of Computer Science and Technology, 2005, 20(4): 465-482.

[22] HOSID S, BOLSHOY A. New elements of the termination of transcription in prokaryotes[J]. Journal of Biomolecular Structure & Dynamics, 2004, 22(3): 347-354.

[23] KOZOBAY-AVRAHAM L, HOSID S, BOLSHOY A. Involvement of DNA curvature in intergenic regions of prokaryotes[J]. Nucleic Acids Research, 2006, 34(8): 2316-2327.

[24] ALTSCHUL S, MADDEN T, SCHÄFFER A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25(17): 3389-3402.

[25] MATHEWS D, DISNEY M, CHILDS J, *et al*. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure[J]. Proceedings of the National Academy of Sciences USA, 2004, 101(19): 7287-7292.

(下转第 495 页)