基于 SVM 方法构建细菌 sRNA靶标预测模型

技术方法

赵雅琳,李 华,侯妍妍,查 磊,曹 源,王立贵,应晓敏*,李伍举*
(军事医学科学院基础医学研究所计算生物学中心,北京 100850)

[摘要] **目的**:为实验方法鉴定细菌 sRNA靶标和研究 sRNA功能提供生物信息学支持。**方法**:首先以实验证实的 132个 sRNA与靶标相互作用数据为训练集,其中包含 46个阳性数据和 86个阴性数据;其次,以实验证实的 22个阳性数据和随机生成的 1 700个阴性数据为测试集;最后以 RNA二级结构谱等特征为变量,运用支持向量机(SVM)方法构建 sRNA靶标预测数学模型。**结果和结论**:构建的模型对训练集的敏感性和特异性均为 100%,对测试集的敏感性和特异性分别为 72.73%和 80.65%。所构建的数学模型为实验发现 sRNA靶标提供了生物信息学支持。

[关键词] sRNA;靶标;预测;机器学习; SVM
[中图分类号] Q811.4 [文献标识码] A
[文章编号] 1000-5501(2008)04-0375-04

Construction of a model for prediction of bacterial sRNA targets using support vector machines

ZHAO Ya-Lin, LI Hua, HOU Yan-Yan, ZHA Lei, CAO Yuan, WANG Li-Gui, YING Xiao-Min, LI Wu-Ju**
(Center of Computational Biology, Institute of Basic Medical Sciences, Academy of Military Medical Sciences, Beijing 100850, China)

[Abstract] **Objective:** To provide bioinformatics support for experimental identification of bacterial sRNA targets and for the study of sRNA functions. **Methods:** To construct a model for prediction of bacterial sRNA targets, 132 sRNA-mRNA interactions verified by experiments were collected first as the training dataset, which contained 46 positive samples and 86 negative samples. Then, 22 sRNA-mRNA interactions verified by experiments as the positive test dataset and 1700 randomly-generated sRNA-mRNA interactions as the negative test dataset were selected. Finally, support vector machines (SVM) were used to construct the model with the profile of sRNA-mRNA secondary structure as the features. **Results and Conclusion:** The model's sensitivity and specificity were 100.00% and 100.00% for the training data, and 72.73% and 80.65% for the test dataset, respectively. Therefore, the model provides bioinformatics support for experimental identification of sRNA targets.

[Key words] sRNA; target; prediction; machine learning; support vector machines

目前研究发现,在细菌中存在着许多小的非编码 RNA(small RNA, sRNA)。以研究最为深入的大肠杆菌(*E. coli*)为例,经实验证实的 sRNA数目已达到 70多条[1]。随着高通量实验技术和生物信息学方法在 sRNA识别上的应用,将有越来越多的 sRNA被证实。这些 sRNA长度为 40～500 nt不等,不编码蛋白质,功能也与 rRNA、tRNA不同,大部分 sRNA通过与靶标结合在基因转录后调控中起着重要作用。到目前为止,尽管有一些 sRNA的功能得到证实,但还是有相当一部分的 sRNA功能是未知的,因此,识别 sRNA的靶标对研究 sRNA的功能具有重要意义。

通过分析 *E. coli*中经实验证实的 sRNA靶标,结果表明大部分 sRNA结合于 mRNA靶标的 5′UTR区附近,并以碱基互补配对方式调控其靶基因的转录后表达,此过程常需要伴侣蛋白 Hfq的参与[2～4]。此外,根据 sRNA结合于 mRNA靶标序列位置的不同,调控结果可分为正调控和负调控两类。在正调控中,sRNA在靶标上的结合位点常位于起始翻译位点上游的 90～120 nt之间,促进靶基因的表达;在负调控中,sRNA在靶标上的结合位点位于 mRNA的 SD序列附近,sRNA的结合则会阻碍核糖体与 mRNA序列的结合[5],对 mRNA的表达起阻遏作用,或者使 sRNA与 mRNA同时降解[6,7]。到目前为止,鉴于正调控的例子很少,经实验证实

[收稿日期] 2008-04-30
[基金项目] 国家"863"高技术项目(2006AA02Z323);国家自然科学基金项目(30500105, 30470411)
[作者简介] 赵雅琳(1974-),女,吉林省长春市人,硕士研究生,研究方向为计算生物学。
*通讯联系人,李伍举(Tel: 010-66931324, E-mail: liwj@bmi.ac.cn);应晓敏(Tel: 010-66932301, E-mail: yingxm@bmi.ac.cn)

的只有两对 (DsrA-rpoS; RprA-rpoS)[8,9],所以我们只研究负调控的情况。

Vogel等[10]系统论述了细菌 sRNA靶标发现的生物信息学和实验方法,尽管 sRNA靶标最终需要实验来验证,但 sRNA靶标的计算识别方法为实验验证提供了一种快捷的方式,到目前,主要发展了 3个数学模型来预测 sRNA靶标。

Zhang等[11]构建的细菌 sRNA靶标预测模型主要是通过在 Smith-Waterman局部序列比对算法基础上添加了一些特征信息得到的。这些信息包括 sRNA的二级结构、Hfq结合蛋白的结合位点、候选靶标 mRNA的翻译起始区 -35～25 n间的序列片段和 sRNA与候选 mRNA靶标在大肠杆菌 K-12及相邻 8个菌株中的保守谱等。该模型对每一对 sRNA与 mRNA片段进行比较并打分,分数高的就被认为是 sRNA可能的靶标。在已知的 10对经实验证实的 sRNA靶标中,有 7对的得分位于前 50名。由此可见,对于训练集来说,预测精度为 70%。因为此算法加入了保守谱这一特征,所以不适用于某些大肠杆菌中不保守的 sRNA或其他细菌的 sRNA的靶标预测。另外,由于该算法只考虑了 sRNA的二级结构,而忽略了 sRNA与 mRNA相结合后的二级结构特征,所以其预测结果可能产生偏差。

在预测模型 TargetRNA中,Tjaden等[12]建立了两个 sRNA靶标预测模型,分别命名为单碱基模型和碱基堆积模型。预测时,首先选择一种模型对相互作用的 sRNA和 mRNA进行打分;而后,假定分数服从极值分布,对每一个 mRNA,均得到一个 *P*值与之对应;最后利用训练集对涉及的翻译起始区的大小及核心匹配 (seed match)片段的长度进行优化,获得的翻译起始区大小为 -30～20 nt区域,完全匹配片段的长度为 9 nt。利用此模型对训练集进行判别时,12对数据中有 8对正确,分类精度为 66.67%。

在 Cossart等[13]提出的预测模型中,利用 4对经实验证实的 sRNA靶标,对相关的热动力学参数进行优化,包括碱基堆积作用、凸环和内部环的罚分,然后利用这些参数给 sRNA与 mRNA相互作用形成的双链区打分。在预测 sRNA的靶标时,同时考虑了两个区间片段,一个是序列 5′端的 -140～90 nt区间的序列片段,一个是序列 3′端的终止密码子上游 60 nt到下游 90 nt区间的序列片段。最后,应用该模型对新发现的 9条 sRNA进行靶标预测,并对某些结果进行了实验证实。

总的来说,以上 3种方法所包含的阳性数据都比较少。例如,TargetRNA模型包含样本数据最多,但只有 12对。另外,这 3个模型考虑的序列 5′端区间分别为 -35～15 nt、-30～20 nt和 -140～90 nt,但是,未能指出哪一个区间是最优区间。为了解决上述问题,我们首先系统收集了目前文献报道的 sRNA及其靶标,然后利用 sRNA靶标二级结构谱等特征为变量,运用支持向量机 (support vector machines, SVM)方法构建了细菌 sRNA靶标预测模型。

1 材料与方法

以研究最为深入的 *E. coli* K-12基因组为研究对象,收集经实验验证的 sRNA与 mRNA靶标序列,以 sRNA靶标二级结构谱等参数为特征变量构建分类器。为此,首先需要构建合适的数据集并提取相关的特征变量,具体过程如下。

1.1 训练数据集和测试数据集

为了构建细菌 sRNA靶标的预测模型,我们从文献中收集到 46对阳性样本数据 (经实验证实通过碱基互补配对方式发生相互作用的 sRNA与 mRNA序列)及 86对阴性样本数据 (经实验证实不发生相互作用的 sRNA与 mRNA序列),构成训练集。另外,由文献 [11,12]可知,当 sRNA对其靶标 mRNA进行负调控时,结合位点通常位于 mRNA序列 5′端 -35～20 nt区域,为了考虑 sRNA的所有负调控情况,我们将考虑的区间扩大到 -80～50 nt区域。最后从 NCBI上下载 *E. coli* K-12的相关注释文件 (NC_000913.ptt, NC_000913.fna),以此为依据,提取训练集中涉及的 mRNA序列片段,构成训练集样本数据。

关于测试集,我们将从文献中另外收集到的 22对经实验证实发生相互作用的 sRNA与 mRNA序列,作为测试集中的阳性数据集,其中 6对来自大肠杆菌[14],16对来自沙门菌属[15～18]。另外,对于训练集中涉及到的 17条 sRNA,针对每一条 sRNA,我们从 *E. coli* K-12基因组注释的 4 131条 mRNA序列中随机抽取 10条,组成 170对数据样本,作为测试集的阴性数据集,共重复 10次,最终获得 1 700个阴性样本。在随机抽取过程中,若发生抽取到训练集中已知靶标的情况,则放弃不用,重新再取,直到全部不为已知的 sRNA靶标为止。

1.2 特征提取

由文献 [12]可知,在描述 sRNA与靶标 mRNA序列相互作用时,sRNA与 mRNA相结合的热动力学稳定性是一个重要的指标。但在该文献构建的模型中,每次仅考虑了翻译起始区的一个区域。这里我们拟同时考虑多个区域的情况,对于每一个候选的靶标 mRNA序列,首先提取其翻译起始区 -80～50 nt区域、长度为 130 nt的序列片段;然后围绕核心区域 -30～30 nt提取所有可能的 1 000个子序列片段 (图 1);最后,对于每一个子序列,用“LLLLLL”将其与 sRNA序列相连,形成“sRNA-LLLLLL-mRNA”和“mRNA-LLLLLL-sRNA”两种情况,这里“LLLLLL”代表两个核酸序列之间的一个连接序列 (linker sequence),其中的 L表示除去 A、U、C和 G 4种碱基之外的符号,不参与 RNA二级结构的形成,因此,在计算过程中不对其进行分析,最后利用 RNAfold程序分别计算它们的二级结构自由能,并假定能量最低的为真实的结合情况,在此基础上提取以下 10个特征变量。

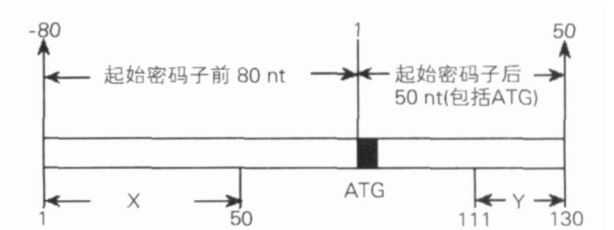

图 1 基于区间 [X, Y]提取所有可能的 1 000个子序列模式图
$1 \le X \le 50$, $111 \le Y \le 130$

以 sRNA与 mRNA子序列片段相连的最低自由能为标准,找出此种状态的二级结构特征。其中内部环、凸环、发夹

环、螺旋区及多分支环中包含的碱基数占相连序列总碱基数的百分比,为相应的第 1到第 5个特征变量;第 6个特征变量是单碱基自由能,因为训练集中每条 sRNA序列的长度都不相等,单个碱基的能量值在阳性样本与阴性样本数据之间更具有可比性,由公式 $\triangle G_m / L_m$ 计算,其中 $\triangle G_m$ 为 sRNA与相应的 mRNA子序列片段相连后具有的二级结构自由能,L_m 则为二者相连后的序列长度;第 7个特征变量则是两条序列结合前后的能量差值,由公式 $\triangle G_m - \triangle G_s - \triangle G_T$ 得出,其中 $\triangle G_m$ 为 2条序列相连后的二级结构自由能,$\triangle G_s$ 和 $\triangle G_T$ 分别代表 sRNA与 mRNA子序列片段各自的二级结构自由能;此外,根据 TargetRNA预测模型[12]得知,两条序列相互作用时的核心匹配片段长度是一个很重要的特征,因此将这个长度值作为要提取的第 8个特征变量;最后,根据 Hfq蛋白通常结合于 RNA序列单链区的 AU富集区,我们分别计算 sRNA与 mRNA子序列片段在形成各自的二级结构时,单链区中碱基 A+U所占的百分比,作为第 9和第 10个特征变量。

因为每一条候选的 mRNA靶标序列均提供 1 000个子序列,而每一对 sRNA与 mRNA子序列片段又会产生相应的 10个特征变量,因此,每对相互作用的 sRNA-mRNA可用 10 000个特征变量来描述,对于包含 132个样本的训练集来说,得到了一个大小为 10 000 ×132的数值矩阵,称这一数值矩阵为 sRNA-mRNA相互作用的二级结构谱。显然,从数据结构上来说,这个数值矩阵与基因表达谱非常相似。因此,许多应用于基因表达谱的方法和软件工具可以用来分析这个数值矩阵。鉴于我们的目标是预测 sRNA与 mRNA之间的相互作用,因此采用机器学习方法来构建模型。

1.3　t检验分析

在构建判别模型之前,先对特征变量集进行 t检验分析,以了解每个特征变量间的差异,结果见表 1。结果显示共有 4 293个特征变量的 P值小于 0.01。按 P值大小以升序排列,前 40个参数($P \leq 10^{-10}$)皆为不同区间的第 7个参数,也就是两条序列结合前后的能量差值。由此可知,sRNA序列、mRNA序列片段及它们的相连序列的二级结构自由能在 sRNA靶标预测中具有重要作用。

表 1　不同 P值区间所包含的特征变量数

概率区间	特征变量数	概率区间	特征变量数
$[0, 10^{-10}]$	40	$(10^{-5}, 10^{-4}]$	223
$(10^{-10}, 10^{-9}]$	124	$(10^{-4}, 10^{-3}]$	1 082
$(10^{-9}, 10^{-8}]$	274	$(10^{-3}, 10^{-2}]$	1 203
$(10^{-8}, 10^{-7}]$	154	$(10^{-2}, 10^{-1}]$	986
$(10^{-7}, 10^{-6}]$	484	$(10^{-1}, 1]$	4 721
$(10^{-6}, 10^{-5}]$	709	$[0, 1]$	10 000

1.4　分类器的构建

机器学习方法很多,在此选取 SVM方法来构建分类器。SVM方法是由 Vapnik领导的贝尔实验室研究小组在 1963年提出的一种分类技术[19],是在统计学理论的基础上发展起来的一种有监督的机器学习方法,具有很好的逼近和泛化能力。我们选用 libSVM包来实现这个工作。libSVM是林智仁博士(Lin Chih-Jen)等开发设计的一个简单、易于使用和快速有效的 SVM模式识别与回归软件包,可以在 http://www.csie.ntu.edu.tw/～cjlin/免费获得。

基于 SVM构建模型的关键在于核函数的选择。在 SVM理论中,采用不同的核函数将导致不同的分类性能。目前 SVM的核函数用得较多的主要有 3种:多项式核函数、径向基函数(radial basis function, RBF)和 Sigmoid核函数。在此以 RBF为核函数。

由于 SVM分类器的性能受惩罚参数 C和 RBF核参数 γ的影响很大,因此我们采用 libSVM包中提供的网格搜索策略搜索近优的 C和 γ,并以此为基础构建分类模型,步骤如下。

(1)将训练集中的样本描述为特征向量,而后采用 libSVM中的 svm-scale将特征向量归一化至 [-1, +1]区间;

(2) SVM的核函数采用 RBF函数,惩罚参数 C和 RBF核参数 γ采用网格搜索近优参数。

(3)采用 libSVM中的 svm-train,根据搜索得到的 C和 γ训练分类模型,使用"-b 1"参数,以计算每个样本的概率估计。

因为 SVM本身是不进行变量选择的一类机器学习方法,我们根据上面提到的 t检验结果,根据 P值分别选取 3个特征集合 SET1, SET2,和 SET3, SET1包含了所有 10 000个参数, SET2则以 P值小于 0.001为标准,共含有 3 090个参数,而 SET3中的参数共有 1 785个,其 P值小于 0.00001。

针对这 3个数据集,我们分别采用网格搜索策略搜索近优的惩罚参数 C和 RBF核参数 γ,并以此为基础构建分类模型 sRNA TargetSVM1、sRNA TargetSVM2和 sRNA TargetSVM3。

2　结果与讨论

2.1　分类器的性能

将得到的 3个分类模型分别用于训练集进行判别分析,结果见表 2。从表 2中可看出,sRNA TargetSVM1的分类效果是最好的,精度为 100%。但一个分类模型的好坏,不能仅从它在训练集上的表现来定论,为了客观地评价我们构建的分类模型的性能,分别将上述 3个模型应用于独立的测试集,结果如表 3所示。从表 3中可看出,分类模型 sRNA TargetSVM1的效果最佳,其分类精度、敏感性和特异性分别为 80.55%、72.73%和 80.65%。

2.2　与其他细菌 sRNA靶标预测模型比较

为了与细菌 sRNA靶标预测模型 TargetRNA进行比较,将测试集中阳性样本涉及的 4个 sRNA提交到 TargetRNA网络服务器,并采用该分类器默认参数进行判别分析,结果表明,在这 22个阳性样本中,仅有 4个被正确预测,分别为 micA-ompA、RybB-ompN、GcvB-dppA和 GcvB-oppA,而利用所构建的模型 sRNA TargetSVM,有 16个样本被正确预测,在阳性样本检出率上,所构建的分类模型要高于模型 TargetRNA。

表 2 运用 SVM 构建的 3个分类模型对训练集的判别结果

特征变量集	特征变量数	C	γ	精度(%)	敏感性(%)	特异性(%)
SET1	10 000	32.0	1.2207×10^{-4}	100.00	100.00	100.00
SET2	3 090	2.0	9.7656×10^{-4}	91.67	95.35	84.78
SET3	1 785	8.0	9.7656×10^{-4}	84.09	98.84	56.52

表 3 各分类模型对测试集的判别结果

分类模型	TP	TN	FP	FN	精度(%)	特异性(%)	敏感性(%)
sRNATargetSVM1	16	1 371	329	6	80.55	80.65	72.73
sRNATargetSVM2	13	1 103	597	9	64.81	64.88	59.09
sRNATargetSVM3	3	1 392	308	19	81.01	81.88	13.64

3 结论

目前,sRNA的功能研究正成为细菌基因组学的热点之一,而识别 sRNA靶标是研究 sRNA功能的关键步骤。但由于目前所证实的阳性数据比较少,这使得现有分类器的预测结果的假阳性率较高。由此,我们在系统收集经实验验证的 sRNA及其靶标的基础上,以 sRNA靶标二级结构谱等特征为变量,采用 SVM方法构建了 sRNA靶标预测模型。从训练集和测试集的判别效果上可以看出,所构建的分类器在分类精度方面要优于目前已有的分类器,但特异性只达到 80.65%,仍然有待进一步提高。我们将从两个角度对分类器的性能做进一步优化,首先是进一步收集经实验证实的 sRNA靶标,其次是探讨不同的机器学习方法和分类策略在细菌 sRNA靶标预测模型中的应用,最终为实验发现 sRNA靶标提供更好的生物信息学支持。

需要指出的是,在构建分类模型 sRNATargetSVM的过程中,所用的特征变量都是与 RNA二级结构相关的数据,使得该分类器不仅适用于大肠杆菌,还适用于其他细菌的 sRNA靶标的预测,例如,在独立测试集的 22个阳性样本中,有 16个阳性样本来自沙门菌属,其中 12个样本被正确判别。

[参考文献]

[1] Vogel J, Sharma CM. How to find small non-coding RNAs in bacteria [J]. Biol Chem, 2005, 386(12): 1219 - 1238.

[2] Brennan RG, Link TM. Hfq structure, function and ligand binding[J]. Curr Opin Microbiol, 2007, 10(2): 125 - 133.

[3] Geissmann TA, Touati D. Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator[J]. EMBO J, 2004, 23(2): 396 - 405.

[4] Brescia CC, Mikulecky PJ, Feig AL, *et al* Identification of the Hfq-binding site on DsrA RNA: Hfq binds without altering DsrA secondary structure[J]. RNA, 2003, 9(1): 33 - 43.

[5] Storz G, Opdyke JA, Zhang A. Controlling mRNA stability and translation with small, noncoding RNAs[J]. Curr Opin Microbiol, 2004, 7(2): 140 - 144.

[6] Masse E, Gottesman S. A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*[J]. Proc Natl Acad Sci USA, 2002, 99(7): 4620 - 4625.

[7] Vanderpool CK, Gottesman S. Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system[J]. Mol Microbiol, 2004, 54(4): 1076 - 1089.

[8] Majdalani N, Cunning C, Sledjeski D, *et al* DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription[J]. Proc Natl Acad Sci USA, 1998, 95(21): 12462 - 12467.

[9] Majdalani N, Hernandez D, Gottesman S. Regulation and mode of action of the second small RNA activator of RpoS translation, RprA[J]. Mol Microbiol, 2002, 46(3): 813 - 826.

[10] Vogel J, Wangner EG. Target identification of small noncoding RNAs in bacteria[J]. Curr Opin Microbiol, 2007, 10(3): 262 - 270.

[11] Zhang Y, Sun S, Wu T, *et al* Identifying Hfq-binding small RNA targets in *Escherichia coli*[J]. Biochem Biophys Res Commun, 2006, 343(3): 950 - 955.

[12] Tjaden B, Goodwin SS, Opdyke JA, *et al* Target prediction for small, noncoding RNAs in bacteria[J]. Nucleic Acids Res, 2006, 34(9): 2791 - 2802.

[13] Mandin P, Repoila F, Vergassola M, *et al* Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA targets[J]. Nucleic Acids Res, 2007, 35(3): 962 - 974.

[14] Masse E, Vanderpool CK, Gottesman S. Effect of RyhB small RNA on global iron use in *Escherichia coli*[J]. J Bacteriol, 2005, 187(20): 6962 - 6971.

[15] Udekwu KI, Darfeuille F, Vogel J, *et al* Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA[J]. Genes Dev, 2005, 19(19): 2355 - 2366.

[16] Bossi L, Figueroa-Bossi N. A small RNA downregulates LamB maltoporin in *Salmonella*[J]. Mol Microbiol, 2007, 65(3): 799 - 810.

[17] Sharma CM, Darfeuille F, Plantinga TH, *et al* A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites[J]. Genes Dev, 2007, 21(21): 2804 - 2817.

[18] Papenfort K, Pfeiffer V, Mika F, *et al* SigmaE-dependent small RNAs of *Salmonella* respond to membrane stress by accelerating global omp mRNA decay[J]. Mol Microbiol, 2006, 62(6): 1674 - 1688.

[19] Cortes C, Vapnik V. Support vector network[J]. Machine Learning, 1995, 20(3): 273 - 297.

(本文编辑 杨兆弘)