微生物学报 *Acta Microbiologica Sinica*
49(1):1-5;4 January 2009
ISSN 0001-6209;CN 11-1995/Q
http://journals.im.ac.cn/actamicrocn

细菌 sRNA 基因及其靶标预测研究进展

王立贵,赵雅琳,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:细菌 sRNA 是一类长度在 40~500 nt 之间的非编码 RNA,主要以不完全碱基配对方式与靶标 mRNA5′端相互作用进而发挥其生物学功能。鉴于预测方法可以为细菌 sRNA 及其靶标的实验发现提供指导,因此,细菌 sRNA 与靶标预测研究受到了广泛重视。文章首先将 sRNA 预测方法分为 3 类,分别是基于比较基因组学的预测方法、基于转录单元的预测方法和基于机器学习的预测方法;其次,将 sRNA 靶标预测方法分为 2 类,分别是序列比较方法与基于 RNA 二级结构的预测方法;最后对各类方法的原理、核心思想、优点和局限性进行了分析,并探讨了进一步的发展方向。

关键词:sRNA;预测;靶标;细菌
中图分类号:Q933　　**文献标识码**:A　　**文章编号**:0001-6209(2009)01-0001-05

细菌 sRNA(small RNA)是一类长度在 40~500 nt之间,以 RNA 形式发挥作用的一类分子,主要位于基因间[1],但也有位于编码基因 5′和 3′UTR 区的报道[2],由 sRNA 基因转录而来[3]。sRNA 与动植物中具有广泛调控功能的 microRNA(miRNA)不同,microRNA 的显著特点是其前体折叠形成茎环或类似茎环的二级结构,其成熟体的长度约为 21 nt,而细菌 sRNA 却不具有保守的二级结构特征,其长度变化也比较大,有关 microRNA 特征及预测研究,请参见我们最近发表的一篇综述[4]。另外,反义 RNA(antisense RNA)是一种人为设计的用于沉默某个特定靶基因的 RNA 片段,其长度在 10~30 nt 之间[5],这 3 类 RNA 的共同特点是均可以抑制靶基因的表达。现有研究表明,细菌 sRNA 功能多样,在细菌的转录调节、RNA 加工与修饰、mRNA 稳定性与翻译、以及蛋白质降解、质粒复制和细菌感染等方面发挥重要功能[6-8]。其中一种主要作用方式是通过不完全碱基互补配对形式与靶标 mRNA 的 5′端结合来调控靶基因的表达[9-10],进而在细菌与环境的相互作用中发挥重要功能。例如,MicF 是大肠杆菌中长度为 93 nt 的 sRNA,可以抑制外膜蛋白 ompF 的表达[11];长度为 90 nt 的 RyhB 可以调控细胞中铁离子的使用,对若干功能为铁离子存储或铁离子使用的蛋白起着调控作用[12-13]。

但是,由于 sRNA 基因较小,且不受移码或无义突变的影响,很难用遗传筛选方法来识别。因此,为实验发现细菌 sRNA 开发基因组水平的生物信息学预测方法非常重要,而大量物种的全基因组测序工作的完成也为 sRNA 基因的理论预测提供了数据基础。迄今为止,已提出了许多 sRNA 基因预测算法,虽然它们一般都是为某个特定的基因组而设计,且可靠性和效率远低于蛋白编码基因的识别算法,但是它们在 sRNA 基因的发现方面发挥越来越重要的作用。在 Jonathan 等的关于细菌 sRNA 识别的综述中[7],列出了从 1966~2006 年用实验和生物信息学方法所发现的 sRNA 各自所占的比重,可以发现生

基金项目:国家"863 计划"(2006AA02Z323);国家自然科学基金(90608004,30470411)
* 通信作者。Tel:+86-10-66931324;Fax:+86-10-68213039;E-mail:liwj@nic.bmi.ac.cn
作者简介:王立贵(1983-),男,江苏沭阳人,硕士研究生,研究方向为计算生物学。
收稿日期:2008-07-04;修回日期:2008-08-29

物信息学方法在 sRNA 发现过程中起着重要作用，在 2000 年以前仅有 13 条大肠杆菌 sRNA 在实验中被偶然发现[14]，而从 2001 年开始，随着生物信息学预测方法的应用，仅大肠杆菌中就至少有 80 条 sRNA 被发现。

随着越来越多的 sRNA 被证实，接下来急需解决的问题是研究这些 sRNA 的功能。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 sRNA 的功能是未知的，因此识别 sRNA 的靶标对研究 sRNA 的功能具有重要意义。目前的 sRNA 靶标预测研究还处于起步阶段，预测方法主要有基于序列比较方法和基于 RNA 二级结构的预测方法。鉴于开展细菌 sRNA 基因及其靶标预测研究具有重要意义，为此，本文就其相关研究进展进行综述。

1 细菌 sRNA 的预测方法

随着大规模细菌基因组测序的完成，目前已发展了多种细菌 sRNA 预测方法，主要分为 3 类，分别是比较基因组学方法、转录单元预测方法和机器学习方法。

1.1 比较基因组学方法

基于比较基因组学方法来识别 sRNA 的主要原理是，作为一个 sRNA 基因，其在相近物种的基因组中具有较高的序列保守性和结构保守性。为此，Rivas 等开发了一个用于 sRNA 预测的概率模型 QRNA[15-16]，首先基于序列比较，找出相近物种基因组序列中保守的基因间区，然后，运用概率模型 QRNA 判断该基因间区的类型：蛋白编码区、sRNA 或者是其它类型。通过分析大肠杆菌 23000 个保守的基因间区，找出了 275 个可能的 sRNA，而用作实验验证的 49 个片段中至少有 11 个被证实是 sRNA。该方法的主要缺点是必须有用于比较的相近物种的基因组序列；其次，相关的 sRNA 序列必须有保守的二级结构；再次，识别出的候选 sRNA 序列还可能包含其它类型的 RNA 序列；最后，该方法不能识别出一个物种特有的 sRNA[17]。

1.2 转录单元预测方法

转录单元预测方法是在比较基因组学方法的基础上，通过基因间区预测和转录单元的识别来进行 sRNA 预测，即通过在基因间区寻找启动子或是终止子或是完整的转录本来寻找 sRNA。

Argaman 等[18]在大肠杆菌中运用转录单元预测方法取得了很好的结果，首先，他们在大肠杆菌的基因间区通过寻找 -10 区，-35 区的保守序列来预测启动子，然后，通过终止子的发夹结构特征和自由能特征来预测 ρ-非依赖型终止子，并在去冗余基础上预测出具有完整转录单元的基因间区，最后，通过与相近的基因组比对，找出保守的基因间区作为候选的 sRNA 序列，在预测的 24 个候选片段中，有 14 条被证实为 sRNA。虽然假阳性检出率比较低，但是由于所用的限制条件非常严格而漏检了许多 sRNA。

Chen 等提出的细菌 sRNA 预测方法[20]，基本思想和 Argaman 等提出的算法类似，首先，在大肠杆菌的基因间区用谱检索算法来预测启动子（http://www.isrec.isb-sib.ch/ftp-server/pftools/）[22]，然后，用 RNA motif 算法在基因间区预测 ρ-非依赖型终止子[23]，再由启动子和终止子所确定并满足一定长度的序列作为候选的 sRNA，共获得 227 个 sRNA 候选基因，然后去掉已知的基因和包含较长的 ORFs 基因，最终获得 144 个 sRNA 候选基因。由于长度的限制，这种方法对预测较长的 sRNA 显得无能为力。

Livny 等[24]构建了称为 sRNAPredict 的算法，这个算法主要是通过相近物种的保守性和预测 ρ-非依赖型终止子来快速预测 sRNA，结果 9 个候选基因中 6 个被实验证实。这个算法可以根据使用者的不同情况灵活地改变一些输入参数，另外一个特点是它的预测速度比较快。主要缺点是被预测的片段要在相近物种间有保守性，具有物种特异性的 sRNA 则很难被预测出来。

虽然转录单元预测方法在预测细菌 sRNA 方面取得了很好的结果，但是它也有自身缺点。首先，它只能预测 ρ-非依赖型终止子的 sRNA，而对 ρ-依赖型终止子却无能为力；其次，这些算法都是通过预测启动子和终止子来预测 sRNA 的，只是在启动子和终止子预测算法上有不同，但它们所得到的候选集的彼此覆盖性却非常差；最后，在基因间区预测 sRNA 必将遗漏位于 UTR 等区域的 sRNA。

1.3 机器学习方法

利用机器学习方法进行 sRNA 预测主要包含 3 个基本步骤，首先是构建包含阳性和阴性数据的训练集，然后是基于样本数据提取特征变量，最后是利用机器学习方法构建分类模型，进而预测新的 sRNA。目前主要有两篇文献论述了机器学习方法在 sRNA 预测中的应用研究。

Carter 等[19]用神经网络方法来预测细菌和古菌的 sRNA，他们首先构建由 86 条 tRNA、22 条 rRNA 和 11 条 sRNA 组成的阳性数据集，然后由注释好的基

因组提取相应的基因间区作为阴性数据集，再次对阳性数据和阴性数据进行步长为 80 nt 重叠 40 nt 进行滑窗并提取 26 个特征变量，其中 4 个是一联碱基组分，16 个是二联碱基组分，6 个是“结构模体”，最后基于这些特征，利用神经网络构建的分类器在基因间区扫描，结果在细菌中的交叉检验精度达到 80%～90%，在嗜热古细菌中的交叉检验精度达到了 90%～99%。虽然没有用实验验证其候选集，但是利用文献报道的新发现 19 条大肠杆菌 sRNA 作为独立测试集，有 17 条被正确预测，同时有些候选集与相近物种有高度保守性，都说明了这个算法的合理性。

Saetrom 等[21]则用基于遗传算法的机器学习方法来预测大肠杆菌的 sRNA，基本思想和 Carter 等[14]类似，他们首先构建由 86 条 tRNA、22 条 rRNA 和 46 条 sRNA 组成的阳性数据集，然后由注释好的基因组提取相应的基因间区作为阴性数据集，再次对阳性数据和阴性数据进行步长为 50 nt 重叠 25 nt 进行滑窗，以模体为特征构建细菌 sRNA 预测模型，最后利用此模型预测基因间区，最终获得了 306 条可能的 sRNA 序列。此外，利用 Northern 方法，作者从中随机选择 16 条序列进行验证，证实其中 12 条序列具有杂交信号。

机器学习方法虽然在一定程度上克服了前两类预测方法的一些缺点，如可以预测出细菌特异性的 sRNA，对 ρ-非依赖型终止子或对 ρ-依赖型终止子的 sRNA 均可以预测等，但是机器学习方法也有自身的缺点。目前的主要缺点是阳性数据较少，以目前研究较多的大肠杆菌来说，也只有 80 多条 sRNA 被证实，而对于其它细菌来说，实验证实的 sRNA 则更少；其次是特征变量的选择问题，如何选择较好的特征变量来描述 sRNA 也是一个难点。随着数据的积累和研究的深入，这些困难有望被逐渐克服，可以预期基于机器学习的 sRNA 预测方法将越来越重要。

2 细菌 sRNA 的靶标预测方法

在 Vogel 等[27]最近发表的一篇综述中，系统地论述了目前关于识别 sRNA 靶标的实验及生物信息学方法。尽管 sRNA 靶标的最终识别需要经过实验来证实，但生物信息学方法仍然为实验验证提供了一种快捷的方式。到目前为止，主要发展了 2 类 sRNA 靶标预测方法，分别是序列比较方法与基于 RNA 二级结构的靶标预测方法。

2.1 基于序列比较的 sRNA 靶标预测方法

在 smith-waterman 局部序列比对算法基础上，Zhang 等[25]构建了细菌 sRNA 靶标预测模型，并在模型中融合了下列信息：sRNA 的二级结构特征、伴侣蛋白 Hfq 在 RNA 序列上的结合位点、候选靶标 mRNA 的起始密码子上游 – 35 nt 到下游 25 nt 间的序列片段、以环区为中心的扩展序列比对和 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。对每一个 sRNA，该模型考虑基因组中每一个候选的 mRNA 与之比对情况并打分，然后将所有的 mRNA 按分数排序，分数高的 mRNA 被认为是 sRNA 可能靶标。在已知的经实验证实的 10 对 sRNA 与 mRNA 相互作用中，有 7 对的打分分数位于前 50 名中，由此可见，对于训练集来说，该预测模型的精度为 70.00%。由于该模型加入了保守谱这一因素，所以不适用于某些大肠杆菌中不保守的 sRNA 的靶标预测或其它细菌中 sRNA 靶标的预测。另外，由于该算法只考虑了 sRNA 的二级结构特征，而忽略了 sRNA 与 mRNA 相结合后的二级结构特征，这使得预测的结果可能会有偏差。

2.2 基于 RNA 二级结构的细菌 sRNA 靶标预测方法

在预测模型 TargetRNA 中[10]，Tjaden 等建立了两个 sRNA 靶标预测模型，分别命名为单碱基模型和碱基堆积模型。单碱基模型是通过为 Smith-Waterman 算法引进新的比较积分系统来实现的，适用于 sRNA 与 mRNA 序列间相互作用区域较短的情况；碱基堆积模型是运用 RNA 二级结构自由能的计算规则来实现的，所采用的方法是动态规划算法，适用于 sRNA 与 mRNA 序列间相互作用区域较长的序列。在进行靶标预测时，首先根据情况选择一种模型对候选的 sRNA 和 mRNA 序列进行打分，并假定分数遵从极值分布，于是每一个候选靶标得到一个 P 值，P 值越小，相应的 mRNA 越可能是靶标；然后利用训练集对翻译起始区的大小及核心匹配片段的长度进行优化，获得的最优值分别为：翻译起始区的取值范围从起始密码子上游 – 30 nt 到下游 20 nt，核心匹配片段的长度为 9 nt；最后利用此模型对训练集进行判别时，12 对数据中有 8 对正确，预测精度为 66.67%。

在由 Cossart 等[26]提出的预测模型中，利用 4 对属于不同细菌基因组且经实验证实相互作用的 sRNA 及其 mRNA 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,该模型同时考虑了两个区间片段,一个是序列 5′端的起始密码子上游 – 140 nt 到下游 90 nt 区间的序列片段,一个是序列 3′端的终止子上游 – 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。

最近,基于实验证实的 46 对 sRNA 与靶标相互作用与 86 对 sRNA 与靶标不发生相互作用数据集,我们利用机器学习方法构建了 2 个 sRNA 靶标预测模型 sRNATargetNB 和 sRNATargetSVM,其在训练集上的 LOOCV 预测精度分别为 91.67%和 100.00%。为了评价模型的泛化能力,我们构建了一个独立的包含 22 个阳性样本和 1700 个随机生成的阴性样本的测试集[28],其预测精度分别为 93.03% 和 80.55%,最终为 sRNA 靶标的实验发现提供了生物信息学支持。

3 总结和展望

本文系统地总结了细菌 sRNA 及其靶标的预测方法,在 3 类细菌 sRNA 预测方法中,由于机器学习方法具有许多优点,如不需要考虑细菌 sRNA 在不同基因组中的保守性和终止子是否为 ρ-非依赖型等,因此,随着实验数据的积累,基于机器学习的 sRNA 预测将会发挥越来越重要的作用。在细菌 sRNA 靶标预测方面,目前还刚刚起步,据我们所知,目前只发展了 4 个细菌 sRNA 靶标预测模型,如何提高敏感性与特异性是以后开展 sRNA 靶标预测的一个重要方向。随着研究的深入,我们相信将会有愈来愈多的 sRNA 及其靶标被证实,进而为细菌 sRNA 及其靶标预测提供很好的数据资源,拓展人们对细菌生命活动的理解,最终造福于人类。

参考文献

[1] Hershberg R, Altuvia S, Margalit H. A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Research*, 2003, 31(7): 1813 – 1820.

[2] Kawano M, Reynolds AA, Miranda-Rios J, et al. Detection of 50- and 30-UTR-derived small RNAs and cis-encoded antisense RNAs in *Escherichia coli*. *Nucleic Acids Research*, 2005, 33(3): 1040 – 1050.

[3] Wassarman KM. Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes. *Cell*, 2002, 109(2): 141 – 144.

[4] 侯妍妍, 应晓敏, 李伍举. microRNA 计算发现方法研究进展. 遗传(*Hereditas*), 2008, 30(6): 687 – 696.

[5] Camps-Valls G, Chalk AM, Serrano-López AJ, et al. Profiled support vector machines for antisense oligonucleotide efficacy prediction. *BMC Bioinformatics*, 2004, 5: 135 – 144.

[6] Gisela S. An Expanding Universe of Noncoding RNAs. *Science*, 2002, 296(5571): 1260 – 1263.

[7] Jonathan L, Matthew KW. Identification of small RNAs in diverse bacterial species. *Current Opinion in Microbiology*, 2007, 10(2): 96 – 101.

[8] Vogel J, Sharma CM. How to find small non-coding RNAs in bacteria. *Biological Chemistry*, 2005, 386: 1219 – 1238.

[9] Gottesman S. The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annual Review of Microbiology*, 2004, 58: 303 – 328.

[10] Tjaden B, Goodwin SS, Storz G, et al. Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Research*, 2006, 34(9): 2791 – 2802.

[11] Delihas N, Forst S. MicF. an antisense RNA gene involved in response of *Escherichia coli* to global stress factors. *Journal of Molecular Biology*, 2001, 313(1): 1 – 12.

[12] Masse E, Gottesman S. A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, 99(7): 4620 – 4625.

[13] Masse E, Vanderpool CK, Gottesman S. Effect of RyhB small RNA on global iron use in *Escherichia coli*. *Journal of Bacteriology*, 2005, 187(20): 6962 – 6971.

[14] Masse E, Majdalani N, Gottesman S. Regulatory roles of small RNAs in bacteria. *Current Opinion in Microbiology*, 2003, 6(2): 120 – 124.

[15] Rivas E, Klein RJ, Eddy SR, et al. Computational identification of noncoding RNAs in E. coli by comparative genomics. *Current Biology*, 2001, 11(17): 1369 – 1373.

[16] Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2001, 2: 8.

[17] Eddy SR. Computational genomics of non – coding RNA genes. *Cell*, 2002, 109(2): 137 – 140.

[18] Argaman L, Hershberg R, Vogel J, et al. Novel small RNA – encoding genes in the intergenic regions of *Escherichia coli*. *Current Biology*, 2001, 11(12): 941 – 950.

[19] Carter RJ, Dubchak I, Holbrook SR. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Research*, 2001, 29(19): 3928 – 3938.

[20] Chen S, Lesnik EA, Hall TA, et al. A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems*, 2002, 65(2-3): 157-177.

[21] Saetrom P, Sneve R, Kristiansen KI, et al. Predicting non-coding RNA genes in *Escherichia coli* with boosted genetic programming. *Nucleic Acids Research*, 2005, 33(10): 3263-3270.

[22] Bucher P, Karplus K, Moeri N, et al. Aflexible search technique based on generalized profiles. *Computers & Chemistry*, 1996, 20(1): 3-24.

[23] Lesnik EA, Sampath R, Levene HB, et al. Prediction of Rho-independent transcription terminators in *Escherichia coli* genome. *Nucleic Acids Research*, 2001, 29(17): 3583-3594.

[24] Livny J, Fogel MA, Davis BM, et al. sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Research*, 2005, 33(13): 4096-4105.

[25] Zhang Y, Sun S, Wu JT, et al. Indentifying Hfq-binding small RNA targets in *Escherichia coli*. *Biochemical and Biophysical Research Communications*, 2006, 343(3): 950-955.

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

[27] Vogel J, Wagner EG. Target identification of small noncoding RNAs in bacteria. *Current Opinion in Microbiology*, 2007, 10(3): 262-270.

[28] Zhao YL, Li H, Hou YY, et al. Construction of two mathematical models for prediction of bacterial sRNA targets. *Biochemical and Biophysical Research Communications*, 2008, 372(2): 346-350.

Research progress of prediction of bacterial sRNA genes and their targets - A review

Ligui Wang, Yalin Zhao, Wuju Li*

(Center of Computational Biology, Beijing Institute of Basic Medical Sciences, Beijing 100850, China)

Abstract: Bacterial sRNAs are a class of non-coding RNAs with 40-500 nucleotides in length. Most of them function as posttranscriptional regulation of gene expression through binding to the translation initiation region of their target mRNAs. In view that prediction of sRNAs and their targets provides support for experimental identification, some prediction methods have been developed for both of them in recent years. In this review, we firstly gave an overview of methods for prediction of sRNA genes, which are classified into three categories, namely, comparative genomics-based, transcription units-based and machine learning-based prediction methods. Secondly, the methods for sRNA target prediction are classified into two types, which are sequence alignment-based method and prediction of RNA secondary structure-based method, respectively. Finally, the principles, advantages and limitations of each kind of method are discussed, and perspectives for prediction methods of sRNA and their targets is pointed out.

Keywords: sRNA; prediction; target; bacteria

(本文责编 :张晓丽)

Supported by the National Program for High Technology Research and Development of China (2006AA02Z323) and the National Sciences Foundation of China (90608004,30470411)

* Corresponding author. Tel: +86-10-66931324; E-mail: liwj@nic.bmi.ac.cn

Received: 4 July 2008/Revised: 29 August 2008