• 生物信息专栏 •

DBEncRNA：细菌必需非编码 RNA 数据库

叶远浓[1,2]，梁定发[1]，曾 柱[2*]

(1. 贵州医科大学大健康学院 贵阳 550025；
2. 贵州医科大学环境污染监测与疾病控制教育部重点实验室 贵阳 550025)

【摘要】细菌非编码 RNA(non-coding RNA, ncRNA) 是近年来在细菌基因组内新发现的一类基因表达调控因子，与必需基因概念类似，有一部分 ncRNA 是生物体生存所必不可少的，称之为“必需非编码 RNA”。因此，细菌的必需 ncRNA 可以作为药物开发的潜在靶标，以降低致病菌的耐药性。同时，必需 ncRNA 也成为最小基因组研究的重要对象之一。目前已经通过湿实验系统地确定了 10 余种细菌的必需 ncRNA，然而还没有一个专门的必需 ncRNA 数据库，导致对必需 ncRNA 的研究远远跟不上科学研究和药物设计的需要。因此，该研究构建了一个专门的细菌必需 ncRNA 数据库 DBEncRNA，以帮助研究人员开发高效的必需 ncRNA 计算机识别方法，用于进一步研究抗菌药物靶标发现和最小基因组。DBEncRNA 数据库可以通过 http://yeyn.group:86/免费访问使用。

关 键 词 抗菌药物靶标； 数据库； 必需非编码 RNA； 最小基因组； 致病菌
中图分类号 TP391; Q615 **文献标志码** A doi:10.12178/1001-0548.2021389

DBEncRNA：Database of Bacterial Essential ncRNA

YE Yuannong[1,2], LIANG Dingfa[1], and ZENG Zhu[2*]

(1. School of Big Health, Guizhou Medical University Guiyang 550025;
2. Key Laboratory of Environmental Pollution Monitoring and Disease Control, Ministry of Education, Guizhou Medical University Guiyang 550025)

Abstract Bacterial non-coding RNA (ncRNA) is a class of gene expression regulators in bacteria in recent years. Similar to the concept of essential genes, a part of ncRNA is indispensable for the survival of organisms, which is called essential ncRNA. Therefore, the bacterial essential ncRNAs could be used as potential targets for drug development in order to reduce drug resistance in pathogenic bacteria. The essential ncRNA has become one of the important objects of minimal genome research. At present, the essential ncRNAs of more than a dozen bacteria have been systematically identified by experimental method. However, there is no specialized essential ncRNAs database. And hence, the research on essential ncRNAs is far behind the needs of scientific research and drug design. In consequence, in this study, a special database of bacterial essential ncRNA (DBEncRNA) is proposed and constructed to help researchers develop the efficient recognition methods for essential ncRNA *in silico*, for further studies including antimicrobial target discovery and minimal genome research. The DBEncRNA database can be accessed by http://yeyn.group:86/ freely.

Key words antimicrobial drug targets; database; essential non-coding RNA; minimal genomic; pathogenic bacterium

细菌非编码 RNA(non-coding RNA, ncRNA) 是近年来在细菌基因组内新发现的一类基因表达调控因子，分子大小为 40～500 个核苷酸，在 RNA 的转录调节、染色体复制、RNA 加工与修饰、mRNA 翻译与稳定性、蛋白质降解与转运和细菌感染等生物过程中扮演着重要角色[1]。随着被发现的细菌 ncRNA 数目迅速增加，及其在生物体内的重要作用，细菌 ncRNA 已成为微生物的研究热点之一[2]。由于 ncRNA 在生物体内扮演重要角色，新 ncRNA 的识别具有重要的科学意义和极大的商业价值。

在生物体所包含的 ncRNA 中，与必需基因概念类似，有一部分 ncRNA 是生物体生存所必不可少的，称之为“必需非编码 RNA”(必需 ncRNA，essential ncRNA)[3]。虽然必需 ncRNA 不能像必需

收稿日期：2021－12－26；修回日期：2022－02－17
基金项目：国家自然科学基金 (61803112, 32160151)；贵州省科技支撑计划 (黔科合支撑 [2019]2811号)
作者简介：叶远浓 (1985－)，男，博士，副教授，主要从事生物信息学方面的研究.
*通信作者：曾柱，Email：zengzhu@gmc.edu.cn

基因一样编码蛋白，但其在生物学上的研究地位与必需基因同等重要，具有重要的理论研究和实际应用价值。如大部分抗生素以基本的细胞过程为靶标，而细菌的 ncRNA 在细菌生命活动中发挥着极为广泛的作用，包括结构调节到催化作用，影响各种加工过程，如细菌毒性、发育控制、mRNA 稳定性与蛋白质降解等[4]，因此细菌的必需 ncRNA 可以作为药物开发的潜在靶标，以降低致病菌的耐药性。同时，对必需 ncRNA 的理论研究有助于理解和确定最小基因组的构成和功能作用，如文献 [5-6] 认为一个完整的最小基因组除了编码蛋白，还需包括调控和结构原件，如 5’-UTRs 和 ncRNA。文献 [7] 报道了一个包含必需 ncRNA 的最小细胞。文献 [8] 在构建细菌最小基因集算法中也提出一个最小基因组，除了最小基因集，还应包含最小非编码 RNA 集。

文献 [9-10] 确定了一个新的 miRNA 为 ncRNA，最早提出“必需 ncRNA (essential non-coding RNA)”的概念。文献 [6] 使用 428 735 个 Tn5 转座子插入测定新月柄杆菌 (*Caulobacter crescetus*) 的基因组时，除了确定 480 个必需基因外，还确定了 29 个必需 tRNA 和 8 个必需小 ncRNA。在肺结核分支杆菌 (*Mycobacterum tuberculosis*) 中，文献 [11] 使用 36 788 个转座子插入方法在确定必需基因的同时发现了 25 个必需基因组片段，包括 10 个 tRNA 和参与 tRNA 过程的 RNaseP 的 RNA 催化单元。文献 [12] 用类似的方法在鼠伤寒沙门氏菌 (*Salmonella enterica serovars*) 中发现了 15 个必需 ncRNA。值得注意的是，RNaseP 再次被确定为必需 ncRNA，因此它可能是一个在细菌中普遍存在的必需 ncRNA。

文献 [13] 测试了一些 ncRNA 对毒性效应具有 niche-specific 的作用的假说，因为越来越多的证据表明 ncRNA 参与致病菌致病过程，该文献首次用 RNA-seq 技术确定了一种肺炎病原体——肺炎链球菌 (*Streptococcus pneumoniae*) 的全套 ncRNA，包含 89 个 ncRNA。文献 [14] 重新确认了酵母的 180 个必需 ncRNA。

正是由于细菌 ncRNA 在细菌生长、侵染宿主和致病机理过程中发挥着极为广泛的调控作用，对细菌 ncRNA，特别是必需 ncRNA 的干扰会使其失去调控作用，从而影响到细菌的生长、侵染宿主的能力。在细菌耐药性问题日益突出的今天，亟待积极研发新型抗菌靶点和药物。基于细菌必需 ncRNA 为靶点的新型药物开发，有助于降低细菌耐药性问题，所以亟需发展细菌必需 ncRNA 的高效识别、鉴定方法。

ncRNA 在合成生物学研究领域也具有不可或缺的地位。在现阶段，定义一个能够维持生物体存活的最小基因组是生物学的主要挑战之一。目前大部分关于最小基因组的研究主要基于传统的蛋白编码基因，而忽略了 ncRNA，这种基于不完整的注释，导致最小基因组的准确性受到了限制[15]。针对这一问题，文献 [7] 以注释较为完整、本身具有较小基因组的细菌——肺炎支原体 (含有 694 个 ORF、311 个 ncRNA、43 个编码 RNA) 作为研究对象，首次获得了一个既包含编码基因，又包含 ncRNA 的最小细胞。

总的来说，研究基因组中的必需基因组元件，如必需 ncRNA 等，在生物学研究中具有重要的科学意义和应用价值，包括从合成生物学到抗病原菌的药物靶标确定。因此，必需 ncRNA 应该如必需基因概念一样，成为最小基因组研究的重要对象之一。为达到这一目标，亟需确定细菌的必需 ncRNA，这就需要发展快速确定必需 ncRNA 的计算机识别算法，因此收集细菌的必需 ncRNA 作为算法开发数据集显得及其重要和必要。

目前，还没有专门的必需 ncRNA 数据库。天津大学生物信息中心构建的必需基因数据中虽然收集了目前测序的必需 ncRNA，但是该数据库仅收集了必需 ncRNA 的序列信息[16-19]，这对于开发高效的必需 ncRNA 计算机识别方法是不足的。基于此，本研究构建了专门的细菌必需 ncRNA 数据库 DBEncRNA(database of bacterial essential ncRNA)，更便于进一步研究抗菌靶标发现和最小基因组。

1　材料与方法

1.1　微生物必需 ncRNA 数据来源

目前在 12 种细菌中，必需 ncRNA 已经被系统地实验确定。虽然必需 ncRNA 的数据量相较必需基因要少很多，但没有一个真正的必需 ncRNA 数据库跟得上科学研究和药物设计的需要。本研究收集测序的细菌基因组中包含了和人类疾病密切相关的细菌必需 ncRNA。目前，DEG 数据库收录了部分细菌的必需 ncRNA 数据[16]，如表 1 所示。

此外，为了使得构建 DBEncRNA 数据库包含的物种和序列更全面，除了上表所列数据，本文还通过“essential”、“ncRNA”、“non-coding RNA”、“essentiality”、“microorganism”、“bacteria”

等关键字的组合在 Google、Pubmed 等数据库上进行检索，将检索到的符合要求的序列作为 DBEncRNA 数据库的来源。

表 1 来源于 DEG 数据库的细菌必需 ncRNA 数据

物种	必需序列数
Acinetobacter baumannii ATCC 17978[20]	59
Acinetobacter baumannii ATCC 17978[20]	1
Agrobacterium fabrum str. C58[21]	11
Bacillus subtilis[22]	2
Brevundimonas subvibrioides ATCC 15264[21]	35
Caulobacter crescentus[6]	532
Mycobacterium tuberculosis H37Rv III[11]	35
Salmonella enterica serovar Typhi Ty2[12]	24
Salmonella enterica serovar Typhi SL1344[12]	23
Sphingomonas wittichii RW1[23]	32
Synechococcus elongatus PCC 7942[24]	34
Streptococcus pneumoniae[13]	72

1.2 必需 ncRNA 二级结构数据来源

必需 ncRNA 是从功能上来定义的，而功能与结构是密切相关的[2, 25]，因此对 RNA 分子结构的研究就成为分子生物学的一个重要领域，其中 RNA 二级结构预测被作为研究 RNA 分子结构的主要手段。因此为了方便用户使用 DBEncRNA 数据库，本文用 RNAfold 工具对每一个收集的必需 ncRNA 进行了二级结构预测[26]。同时为了方便用户直观地观察 ncRNA 的二级结构，本文调用了 RNA 二级结构可视化工具 Forna[27]。

1.3 序列比对

在生物信息学中，通常认为序列相似则功能相似，为了帮助用户挖掘其余未经实验确定的必需 ncRNA，DBEncRNA 数据库引入 BLAST 序列比对功能，帮助使用者基于 DBEncRNA 数据库通过同源序列比对发现其感兴趣的 ncRNA 序列[28]。

2 结果与讨论

2.1 DBEncRNA 数据库内容

DBEncRNA 数据库的原始必需 ncRNA 数据来源于 DEG 6.5 和关键字爬取，在获得原始数据后进行以下处理：首先，因为 DBEncRNA 数据库提供了必需 ncRNA 的二级结构信息，因此剔除没有核酸序列的 ncRNA 信息；其次，根据 DBEncRNA 数据库的使用功能，筛选保留描述 ncRNA 的相关信息，如表 2 所示。最终获得了一个含有 20 株细菌，共包含 884 条必需 ncRNA 序列及相关信息的数据库，如表 3 所示。

表 2 DBEncRNA 数据库细菌必需 ncRNA 信息

字段名	具体信息
Accession Number	DBEncRNA数据库编号
RefSeq	基因组在genbank的登录号
Category	ncRNA所属类
Condition	培养条件
Cross-Ref	该序列在其他数据库中登录号
Description	功能描述
Organism	来源物种
Reference	参考文献
Date	发表日期
Nucleotide Sequence	核酸序列

表 3 DBEncRNA 数据库数据统计信息

物种名	基因组编号	培养条件	必需ncRNA数目/个
Caulobacter crescentus	NC_011916	完全培养基	532
Acinetobacter baumannii ATCC 17978	NC_009085	完全培养基	60
Escherichia coli O157:H7 str. EDL933	NZ_CP008957	LB糖培养基	37
Synechococcus elongatus PCC 7942	NC_007604	完全培养基	34
Mycoplasma pneumoniae M129	NC_000912	LB糖培养基	34
Sphingomonas wittichii RW1	NC_009511	完全培养基	32
Brevundimonas subvibrioides ATCC 15264	NC_014375	完全培养基	31
Mycobacterium tuberculosis H37Rv III	NC_000962	完全培养基	29
Providencia stuartii strain BE2467	NZ_CP017054	LB糖培养基	25
Salmonella enterica serovar Typhi Ty2	NC_016810	完全培养基	24
Salmonella enterica serovar Typhimurium SL1344	NC_016810	完全培养基	23
Streptococcus mutans UA159	AE014133	血培养基	6
Agrobacterium fabrum str. C58 chromosome linear	NC_003063	完全培养基	6
Agrobacterium fabrum str. C58 chromosome circular	NC_003062	完全培养基	5
Mycobacterium tuberculosis H42Rv III	NC_000962	完全培养基	1
Mycobacterium tuberculosis H38Rv III	NC_000962	完全培养基	1
Mycobacterium tuberculosis H39Rv III	NC_000962	完全培养基	1
Mycobacterium tuberculosis H41Rv III	NC_000962	完全培养基	1
Mycobacterium tuberculosis H40Rv III	NC_000962	完全培养基	1

其中新月柄杆菌 (*Caulobacter crescentus*) 的必需 ncRNA 数目占数据库总数的近 61%，其次是鲍氏不动杆菌 (*Acinetobacter baumannii* ATCC 17978) 的必需 ncRNA 数目，占近 7%。实验确定必需 ncRNA 的培养条件总共有 5 种，其中主要以完全培养基 (rich medium) 条件为主，占 75%，这是在充足生长条件下确定必需基因和必需 ncRNA 的常用培养条件。根据 ncRNA 所属类别可将 ncRNA 分为 10 大类，如图 1 所示，属于启动子类型的 ncRNA 将近一半，其次是属于 tRNA 类型的 ncRNA。

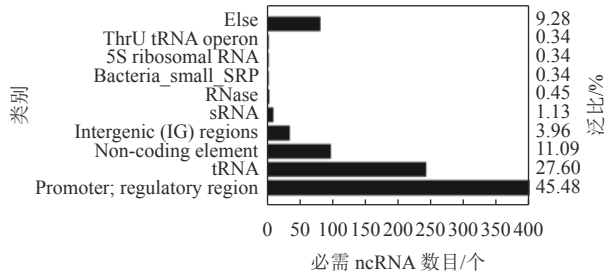

图 1　DBEncRNA 数据库必需 ncRNA 类别分布图

2.2　ncRNA 二级结构

为了方便用户使用 DBEncRNA 数据库，本文用 RNAfold 软件数据库收集的每个必需 ncRNA 进行二级结构预测，对于每一条必需 ncRNA，RNAfold 采用两种方法对其进行预测，分别是基于最小自由能的预测方法 (minimum free energy) 和基于热力学的预测方法 (thermodynamic ensemble)，对于每一种预测的二级结构，均给出该结构下的最小自由能等信息。

将预测出的每种二级结构以及对应的分子结构注释信息导入到 DBEncRNA 数据库，同时，引入可视化插件，使用人员可以按需查看其二级结构。

2.3　DBEncRNA 数据库构建

DBEncRNA 的数据主要包括 884 个 ncRNA 及其预测的分子结构和注释信息，所有数据被整理并存储在关系型数据库 MYSQL 上，可通过 http://yeyn.group:86 免费访问，DBEncRNA 经测试可在不同的操作系统 (如 Windows、Linux 和 Mac) 以及各种浏览器 (如 Internet Explorer、Mozilla Firefox、Google Chrome) 上使用。

2.4　通过序列比对预测必需 ncRNA 与数据下载

通常认为序列相似则功能相似，为了帮助用户确定其感兴趣的 ncRNA 是否属于必需 ncRNA，将 BLAST 序列比对工具引入 DBEncRNA 数据库。用户可以通过提交序列预测其必需性，该功能可以通过点击 DBEncRNA 数据库页面上的“BLAST”链接实现。

为了方便用户使用，本文还提供 DBEncRNA 数据库的数据下载，用户可以根据研究需要，下载必需 ncRNA 的核酸序列和二级结构数据。

3　结 束 语

当前，必需 ncRNA 的数据量持续增加，但还没有一个真正的必需 ncRNA 数据库。这远远跟不上科学研究和药物设计的需要，急需开发出专门的数据库并在此基础上开发必需 ncRNA 识别的计算机软件去识别更多的必需 ncRNA。因此，本研究通过收集已经测序的细菌基因组中包含的必需 ncRNA，构建了必需 ncRNA 数据库。基于该数据库的数据，生物信息人员后续可以开发基因序列组成和序列衍生信息的必需 ncRNA 识别算法，同时可以利用其二级结构数据以提高相关算法的准确性。

DBEncRNA 数据库能对抗菌药物靶标发现和对合成生物学研究提供数据支撑。除此之外，对病原菌必需 ncRNA 的深入研究也将推动开发新的致病菌快速检测系统。DBEncRNA 数据库有助于设计针对特定致病菌高度特异和高度敏感的 RNA 探针，而后者可应用于临床快速检测系统。总之，利用 DBEncRNA 数据有助于开发出预测每种致病菌特有必需 ncRNA 的方法，也有助于发展新的致病菌特异性预防和治疗方法。

参　考　文　献

[1] DAR D, SOREK R. Bacterial noncoding RNAs excised from within protein-coding transcripts[J]. Mbio, 2018, 9(5): e0173018.

[2] CUI Z Q, ZHANG Y, KAKAR K U, et al. Involvement of non-coding RNAs during infection of rice by Acidovorax oryzae[J]. Env Microbiol Rep, 2021, 13(4): 540-554.

[3] ZENG P, CHEN J, MENG Y, et al. Defining essentiality score of protein-coding genes and long noncoding RNAs[J]. Front Genet, 2018, 9: 380.

[4] ROMBY P, VANDENESCH F, WAGNER E G. The role of RNAs in the regulation of virulence-gene expression[J]. Curr Opin Microbiol, 2006, 9(2): 229-236.

[5] GIL R, SILVA F J, PERETO J, et al. Determination of the core of a minimal bacterial gene set[J]. Microbiol Mol Biol Rev, 2004, 68(3): 518-537.

[6] CHRISTEN B, ABELIUK E, COLLIER J M, et al. The essential genome of a bacterium[J]. Mol Syst Biol, 2011, 7: 528.

[7] LLUCH-SENAR M, DELGADO J, CHEN W H, et al. Defining a minimal cell: Essentiality of small ORFs and

ncRNAs in a genome-reduced bacterium[J]. Mol Syst Biol, 2015, 11(1): 780.
[8] YE Y N, MA B G, DONG C, et al. A novel proposal of a simplified bacterial gene set and the neo-construction of a general minimized metabolic network[J]. Sci Rep, 2016, 6: 35082.
[9] HANNON G J, RIVAS F V, MURCHISON E P, et al. The expanding universe of noncoding RNAs[J]. Cold Spring Harb Symp Quant Biol, 2006, 71: 551-564.
[10] DUBESSAY P, RAVEL C, BASTIEN P, et al. The switch region on Leishmania major chromosome 1 is not required for mitotic stability or gene expression, but appears to be essential[J]. Nucleic Acids Res, 2002, 30(17): 3692-3697.
[11] ZHANG Y J, IOERGER T R, HUTTENHOWER C, et al. Global assessment of genomic regions required for growth in Mycobacterium tuberculosis[J]. PLoS Pathog, 2012, 8(9): e1002946.
[12] BARQUIST L, LANGRIDGE G C, TURNER D J, et al. A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium[J]. Nucleic Acids Res, 2013, 41(8): 4549-4564.
[13] MANN B, OPIJNEN T V, WANG J, et al. Control of virulence by small RNAs in Streptococcus pneumoniae[J]. PLoS Pathog, 2012, 8(7): e1002788.
[14] PARKER S, FRACZEK M G, WU J, et al. A resource for functional profiling of noncoding RNA in the yeast Saccharomyces cerevisiae[J]. RNA, 2017, 23(8): 1166-1171.
[15] AUSLANDER S, AUSLANDER D, FUSSENEGGER M. Synthetic biology-the synthesis of biology[J]. Angew Chem Int Ed Engl, 2017, 56(23): 6396-6419.
[16] LUO H, LIN Y, LIU T, et al. DEG 15, an update of the database of essential genes that includes built-in analysis tools[J]. Nucleic Acids Res, 2020, 49(D1): D677-D686.
[17] ZHANG R, OU H Y, ZHANG C T. DEG: A database of essential genes[J]. Nucleic Acids Res, 2004, 32(Database issue): D271-272.
[18] LUO H, LIN Y, GAO F, et al. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements[J]. Nucleic Acids Res, 2014, 42(Database issue): D574-580.
[19] GAO F, LUO H, ZHANG C T, et al. Gene essentiality analysis based on DEG 10, an updated database of essential genes[J]. Methods Mol Biol, 2015, 1279: 219-233.
[20] WANG N, OZER E A, MANDEL M J, et al. Genome-wide identification of Acinetobacter baumannii genes necessary for persistence in the lung[J]. Mbio, 2014, 50(3): e01163.
[21] CURTIS P D, BRUN Y V. Identification of essential alphaproteobacterial genes reveals operational variability in conserved developmental and cell cycle syst[J]. Mol Microbiol, 2014, 93(4): 713-735.
[22] KOBAYASHI K, EHRLICH S D, DEUERLING E. Essential Bacillus subtilis genes[J]. Proceedings of the National Academy of Sciences, 2003, 100(8): 4678-4683.
[23] ROGGO C, CORONADO E, MORENO S, et al. Genome-wide transposon insertion scanning of environmental survival functions in the polycyclic aromatic hydrocarbon degrading bacterium Sphingomonas wittichiiRW1[J]. Environ Microbiol, 2013, 15(10): 2681-2695.
[24] RUBIN B E, WETMORE K M, PRICEM N, et al. The essential gene set of a photosynthetic organism[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 2015: 6634-6643.
[25] CHAO Y, VOGEL J. The role of Hfq in bacterial pathogens[J]. Curr Opin Microbiol, 2010, 13(1): 24-33.
[26] DENMAN R B. Using RNAflod to predict the activity of small catalytic RNAs[J]. Biotechniques, 1993, 15(6): 1090-1095.
[27] KERPEDJIEV P, HAMMER S, HOFACKER I L. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams[J]. Bioinformatics, 2015, 31(20): 3377-3379.
[28] TATUSOVA T A, MADDEN T L. BLAST2Sequences, a new tool for comparing protein and nucleotide sequences[J]. FEMS Microbiol Lett, 1999, 174(2): 247-250.

编 辑 刘飞阳