微生物学通报
Microbiology China
tongbao@im.ac.cn

Dec. 20, 2014, 41(12): 2583–2592

DOI: 10.13344/j.microbiol.china.140280

主编点评文章

利用不同 G+C 含量细菌基因组评估细菌 ncRNA 基因预测工具

刘林梦　温权　欧竑宇*

(上海交通大学 微生物代谢国家重点实验室 生命科学技术学院　上海　200030)

摘　要：【目的】为识别已完成全测序细菌基因组中的 ncRNA 基因，对 3 个常用 ncRNA 预测工具 sRNAPredict、PORTRAIT 和 sRNAscanner 进行评估。【方法】选择了细菌 ncRNA 数据库 BSRD 收录的含有已知 ncRNA 基因数目大于 30 的 9 个细菌基因组，并按基因组 G+C 含量进行分类，比较 sRNAPredict 和 PORTRAIT 工具的预测准确性。提取不同 G+C 含量基因组中 ncRNA 基因转录起始和终止区的序列特征，对 sRNAscanner 预测结果进行评估。【结果】sRNAPredict 对细菌 ncRNA 基因的预测特异性和阳性检出率均高于 PORTRAIT，而敏感性则较差；两种工具预测效果均随基因组 G+C 含量不同而产生明显变化。在不同 G+C 含量的细菌基因组中，ncRNA 基因启动子和终止子区域的序列特征有明显差异。利用这些序列特征能提高 sRNAscanner 预测 ncRNA 基因的平均水平。【结论】3 种 ncRNA 基因工具预测效果随基因组 G+C 含量变化而不同。不同 G+C 含量基因组中 ncRNA 基因的转录起始和终止区特征可作为 ncRNA 基因预测的重要参数之一。

关键词：细菌 ncRNA 基因，基因组 G+C 含量，转录起始区域，转录终止区域，预测

Assessment of bacterial ncRNA gene prediction tools using bacterial genomes with different G+C content

LIU Lin-Meng　WEN Quan　OU Hong-Yu*

(*State Key Laboratory for Microbial Metabolism*, *School of Life Sciences & Biotechnology*, *Shanghai Jiao Tong University*, *Shanghai* 200030, *China*)

Abstract: [Objective] Bacterial ncRNAs are a versatile class of non-coding RNA which plays an important role in the process of microbial life. In this study, we assess three ncRNA gene-prediction tools used frequently with the different bacterial genomes. **[Methods]** Prediction tools representing the position weight matrix method (sRNAscanner), comparative genomics method (sRNAPredict) and machine learning method (PORTRAIT) were tested by using 7 BSRD-archived bacterial genomes with low, middle and high G+C contents, each of which contains more than 30 experimentally verified ncRNA genes. A set of genomic G+C content-associated position weight matrixes of transcription initiation and termination regions of ncRNA genes was generated and

基金项目：国家自然科学基金项目(No. 31170082)；高等学校博士学科点专项科研基金项目(No. 20130073110062)
*通讯作者：Tel：86-21-62932943；✉：hyou@sjtu.edu.cn
收稿日期：2014-04-06；接受日期：2014-05-21；优先数字出版日期(www.cnki.net)：2014-05-26

employed to test sRNAscanner prediction. **[Results]** The sRNAPredict tool had higher specificity and positive prediction value, but lower sensitivity than PORTRAIT. The performance of both tools varied with the selected strains of different G+C contents. The obtained G+C content-associated matrix slightly improved the average accuracy of sRNAscanner. **[Conclusion]** The changing accuracy of the bacterial ncRNA gene detection tools under study was attributed to genomic G+C heterogeneity. Conserved sequence features of ncRNA gene promoters and terminators in genomes sharing similar G+C contents may be helpful to enhance bacterial ncRNA genes prediction.

Keywords: Bacterial ncRNA gene, Genomic G+C content, Transcriptional initiation region, Transcriptional termination region, Prediction

细菌非编码 RNA (ncRNA)是一类不编码蛋白质的 RNA 分子，长度一般在 40−500 nt 之间[1]。研究表明，ncRNA 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、蛋白质降解与转运以及细菌感染等许多过程中都发挥着重要作用。细菌 ncRNA 基因可借助 Northern 印记杂交、RT-PCR、5′RACE、RNA-Seq 转录组测序及生物信息学预测或发现。生物信息学方法需要预先采用不同细菌中已知 ncRNA 基因进行训练[2]。目前，关于细菌 ncRNA 基因的生物信息学预测方法主要包括比较基因组学方法如 QRNA[3]、转录单元预测方法如 sRNAScanner[4]、基于二级结构以及能量稳定性的预测方法如 RNAz[5]、机器学习从头预测法 PORTRAIT[6]等。其中比较基因组学和转录单元预测方法是目前最常用的方法。这两种方法的结合使用也取得了较好的效果。如目前评价较好的 sRNAPredict[7]则是基于亲缘菌 ncRNA 基因的序列保守性和转录终止信号来预测基因组序列中的 ncRNA 基因。它拥有较高的特异性，但灵敏度较低。它不能够预测菌株特异的 ncRNA 基因，同时也不能预测转录终止信号不明显的 ncRNA 基因。而利用机器学习法构建 ncRNA 基因预测模型的最大难点在于两个方面：(1) 已知的细菌 ncRNA 基因太少不能构建合适的训练集；(2) 不能够提取种属特异性的 ncRNA 样本特征。

本文分别选取了目前评价较好的基于比较基因组学和转录单元方法的综合性预测工具 sRNAPredict，以及基于机器学习的从头预测工具 PORTRAIT 作为代表，用 G+C 含量不同的细菌基因组对其预测效果进行评估。大肠杆菌中基因的转录起始区域有典型的保守序列特征，如−35 box 和−10 box。目前基于−35 box 和−10 box 的位置权重算法已经应用在细菌 ncRNA 基因的预测中，如 sRNAscanner。但这种方法的准确性依赖于不同细菌可靠的−10 box 和−35 box 训练集，因此对不同物种细菌的预测效果差异很大。

本文按基因组 G+C 含量将细菌分为高、中、低三类，提取这三类基因组中已知 ncRNA 基因的转录起始和终止区域的序列特征。根据 BSRD 数据库收集已实验验证的 ncRNA 数据集[8]，用三类 G+C 含量不同的细菌基因组来评价分别基于比较基因组学和机器学习法的 ncRNA 基因预测工具 sRNAPredict 和 PORTRAIT 的准确性。此外，捕捉细菌 ncRNA 基因转录起始区域和转录终止区域中保守序列特征，作为 sRNAscanner 工具的训练集对不同 G+C 含量细菌基因组中的 ncRNA 基因进行预测评价。这些分析将能够为不同细菌的 ncRNA 基因预测提供一些有益的参考。

1 材料与方法

1.1 数据集

将 BSRD 数据库中收录的 1 008 个来自 63 个细菌菌株的已验证 ncRNA 基因作为原始数据集。按照其基因组平均 G+C 含量分为三类：低 G+C (LowGC)，即 G+C 含量小于 45%，共 310 个 ncRNA 基因；中 G+C (MidGC)，即 G+C 含量为 45%−60%，共 357 个 ncRNA 基因；高 G+C (HighGC)，即 G+C 含量大于 60%，共 341 个 ncRNA 基因。对 BSRD 数据库中每个菌株中 ncRNA 基因的数目进行统计，发现数目大于 30 的菌株有 9 个(表 1)。本文分

表 1 BSRD 数据库中收录的含有大于 30 个 ncRNA 基因的细菌基因组
Table 1 BSRD-archived bacterial genomes contained more than 30 ncRNA genes

编号 No.	菌株 Strains	NCBI 登录号 Accession No.	基因组 G+C 含量 G+C content (%)	ncRNA 基因数目 ncRNA gene number
1	*Staphylococcus aureus* subsp. *aureus* N315	NC_002745	33	57
2	*Helicobacter pylori* 26695	NC_018939	39	80
3	*Synechocystis* sp. PCC 6803	NC_000911	48	75
4	*Escherichia coli* K-12 MG1655	NC_000913	51	63
5	*Salmonella enterica* serovar Typhimurium SL1344	NC_016810	52	113
6	*Salmonella enterica* serovar Typhimurium LT2	NC_003197	52	68
7	*Mycobacterium bovis* BCG str. Pasteur 1173P2	NC_008769	66	36
8	*Pseudomonas aeruginosa* UCBPP-PA14	NC_008463	66	66
9	*Pseudomonas aeruginosa* PAO1	NC_002516	67	117

别选取金黄色葡萄球菌(*Staphylococcus aureus* subsp. *aureus* N315，NCBI RefSeq 登录号 NC_002745，基因组 G+C 含量为 33%)和幽门螺旋杆菌(*Helicobacter pylori* 26695，NC_018939，基因组 G+C 含量为 39%)作为低 G+C 含量基因组的代表；选取集胞藻(*Synechocystis* sp. PCC 6803，NC_000911，基因组 G+C 含量为 48%)、大肠杆菌(*Escherichia coli* K-12 MG1655，NC_000913，基因组 G+C 含量为 51%)、沙门氏菌(*Salmonella enterica* serovar Typhimurium SL1344，NC_016810，基因组 G+C 含量为 52%；*Salmonella enterica* serovar Typhimurium LT2，NC_003197，基因组 G+C 含量为 51%)作为中 G+C 含量代表基因组；选取铜绿假单胞菌(*Pseudomonas aeruginosa* UCBPP-PA14，NC_008463，基因组 G+C 含量为 66%；*Pseudomonas aeruginosa* PAO1，NC_002516，基因组 G+C 含量分别为 67%)和牛分枝杆菌(*Mycobacterium bovis* BCG Pasteur 1173P2，NC_008769，基因组 G+C 含量为 66%)作为高 G+C 含量基因组的代表。即选取含大于 30 个已验证 ncRNA 基因的共 9 个细菌基因组作为检测基因组，用来对常用 ncRNA 预测工具进行评估。

1.2 ncRNA 预测工具评估

为了探究不同生物信息学工具预测不同 G+C 含量基因组 ncRNA 基因的准确性，本文对基于比较基因组学和转录单元特征的预测工具 sRNAPredict，以及基于机器学习从头预测方法的工具 PORTRAIT 进行评估。(1) 获取阳性数据集(Positive dataset)。分别提取 BSRD 数据库中 9 个检测菌株的已验证 ncRNA 基因及其上游 70 nt 和下游 50 nt 序列作为阳性数据集，阳性 ncRNA 数目如表 2 所示。(2) 获取阴性数据集(Negative dataset)。分别提取 9 个检测基因组中小于 700 nt 的 NCBI Refseq 注释的蛋白编码基因(Protein-coding sequences，CDS)作为阴性数据集，阴性数据 CDS 数目如表 2 所示。大多 ncRNA 基因长度在 40−500 nt，加上上游 70 nt 和下游 50 nt，因此阳性数据集的长度都比较短；因此我们选取了长度小于 700 nt 的 CDS 作为阴性数据集。(3) 选择敏感性(Sensitivity)、特异性(Specificity)和阳性检出率(Positive prediction value，*PPV*)作为主要评价指标，对 sRNAPredict 和 PORTRAIT 的预测结果进行评价。其计算公式如下。其中 TP 为真阳性(True positive)、TN 为真阴性(True negative)、FP 为假阳性(False positive)和 FN 为假阴性(False negative)[9]。

$$\text{Sensitivity}=TP/(TP+FN)\times 100\% \quad (1)$$

$$\text{Specificity}=TN/(TN+FP)\times 100\% \quad (2)$$

$$PPV=TP/(TP+FP) \quad (3)$$

表 2 细菌 ncRNA 预测工具 sRNAPredict 和 PORTRAIT 的评估
Table 2 Prediction accuracy of sRNAPredict and PORTRAIT for bacterial ncRNA genes

菌株(G+C 含量) [阳性数据集][阴性数据集] Strain (G+C content) [Positive data][Negative data]	结果 Results	sRNAPredict		PORTRAIT	
		ncRNA	CDS	ncRNA	CDS
Staphylococcus aureus N315 (33%) [57][2 486]	ncRNA	TP=5	FP=0	TP=41	FP=94
	CDS	FN=52	TN=2 486	FN=16	TN=2 392
	Sensitivity	8.77%		71.03%	
	Specificity	100%		96.22%	
	PPV	100%		30.37%	
Helicobacter pylori 26695 (39%) [80][1 504]	ncRNA	TP=2	FP=0	TP=71	FP=107
	CDS	FN=78	TN=1 504	FN=9	TN=1 397
	Sensitivity	2.50%		88.75%	
	Specificity	100%		92.89%	
	PPV	100%		39.89%	
Synechocystis sp. PCC 6803 (48%) [75][2 981]	ncRNA	TP=0	FP=0	TP=60	FP=78
	CDS	FN=75	TN=2 981	FN=15	TN=2 903
	Sensitivity	0		80%	
	Specificity	100%		97.38%	
	PPV	0		43.48%	
Escherichia coli K-12 MG1655 (51%) [63][3 925]	ncRNA	TP=32	FP=0	TP=48	FP=94
	CDS	FN=31	TN=3 925	FN=15	TN=3 831
	Sensitivity	50.79%		76.19%	
	Specificity	100%		97.61%	
	PPV	100%		33.80%	
Salmonella enterica serovar Typhimurium SL1344 (52%) [113][4 217]	ncRNA	TP=−	FP=−	TP=89	FP=100
	CDS	FN=−	TN=−	FN=24	TN=4 117
	Sensitivity	−		78.76%	
	Specificity	−		97.63%	
	PPV	−		47.09%	
Salmonella enterica serovar Typhimurium LT2 (52%) [68][4 227]	ncRNA	TP=51	FP=0	TP=52	FP=107
	CDS	FN=17	TN=4 227	FN=16	TN=4 120
	Sensitivity	75.00%		76.47%	
	Specificity	100%		97.47%	
	PPV	100%		32.70%	
Mycobacterium bovis BCG str. Pasteur 1173P2 (66%) [36][3 719]	ncRNA	TP=13	FP=0	TP=33	FP=89
	CDS	FN=23	TN=3 719	FN=3	TN=3 630
	Sensitivity	36.11%		91.67%	
	Specificity	100%		97.61%	
	PPV	100%		27.05%	
Pseudomonas aeruginosa UCBPP-PA14 (66%) [66][5 551]	ncRNA	TP=8	FP=0	TP=38	FP=79
	CDS	FN=58	TN=5 551	FN=28	TN=5 472
	Sensitivity	12.12%		57.58%	
	Specificity	100%		98.58%	
	PPV	100%		32.48%	
Pseudomonas aeruginosa PAO1 (67%) [117][5 253]	ncRNA	TP=20	FP=0	TP=77	FP=41
	CDS	FN=97	TN=5 253	FN=40	TN=5 212
	Sensitivity	17.09%		65.81%	
	Specificity	100%		99.22%	
	PPV	100%		65.25%	

Note: TP: True positive; TN: True negative; FP: False positive; FN: False negative. *PPV*: Positive prediction value. Positive data: ncRNA genes collected by BSRD; Negative data: All the annotated CDS in the size of less than 700 nt. −: No sRNAPredict-prediction results found for *Salmonella enterica* serovar Typhimurium SL1344.

1.3 ncRNA 基因转录起始和终止区域特征分析

1.3.1 不同 G+C 含量基因组转录单元特征：BSRD 数据库中 1 008 条已验证 ncRNA 按照其所在基因组的 G+C 含量分为低 G+C (LowGC)、中 G+C (MidGC)和高 G+C (HighGC)三组。另将 BSRD 数据库中大肠杆菌(*Escherichia coli*) (属于 MidGC)的 67 个 ncRNA 基因作为大肠杆菌 ncRNA 数据集。分别提取三组数据及大肠杆菌数据集中 ncRNA 基因的转录起始位点上游 70 nt 作为转录起始区域序列训练集；提取转录终止位点上下游各 50 nt 共 101 nt 作为转录终止区域序列训练集。使用 MEME[10]工具对这四类数据集进行序列保守性分析，捕捉转录起始和终止区域的特征序列。

1.3.2 sRNAscanner 预测工具验证转录单元特征：为探究不同 G+C 含量基因组中 ncRNA 基因转录起始和终止区域特征对 ncRNA 基因预测的影响，本文选取依赖于转录起始和终止区域的特征矩阵进行 ncRNA 预测的工具 sRNAscanner 作为评估对象，采用两种方法对 9 个检测基因组进行 ncRNA 基因预测。(1) sRNAscannerI：用 sRNAscanner 的缺省矩阵(用已实验验证的 *E. coli* 10 个启动子和 21 个非 *rho* 依赖型的转录终止子来训练)和缺省参数对检测基因组进行预测，得到 A 类数据集；(2) sRNAscannerII：首先提取 MEME 得到的转录起始区域特征和终止区域特征的相应序列，然后转换成位置权重矩阵(Position weight matrix，PWM)[11]，并替换 sRNAscanner 缺省的矩阵，同时相应更改 sRNAscanner 的矩阵距离、滑动窗口等参数设置，分别对 9 个检测基因组进行预测，得到 B 类数据集。BSRD 数据库中收录的这 9 个检测基因组中各自的 ncRNA 基因个数如表 1 所示，这类数据为 C 类数据集。BSRD 中所有已验证 ncRNA 作为 D 类数据集，共 1 008 条 ncRNA 基因。为了对两种方法的预测结果进行评价，做了如下分析：(1) 分别将 A、B 类数据集与 C 类数据集做交集，评价 sRNAscanner 的阳性检出情况，当两个 ncRNA 基因有≥30%的重叠时认为它们是同一个 ncRNA 基因。没有匹配到 C 类数据集的 ncRNA 基因分别作为 E 和 F 类数据集。(2) 将 E、F 分别与 D 类数据集进行 BLASTn[12]比对，取 *E*-value≤0.000 1，*H*-value≥0.32 (*H*-value=一致性分数×比对长度比例，$H=i\times(lm/lq)$，*i* 表示 Identities 百分比，*lm* 表示得分最高的匹配区域的长度(含空缺位置)，*lq* 表示输入序列的长度)[13]，考察这些数据与 BSRD 中已收录 ncRNA 数据集的同源性。(3) 将 A 和 B 类数据集做交集，评价两种方法得到结果的重合性，判断标准同(1)。

2 结果与讨论

2.1 ncRNA 基因长度分布

将 BSRD 中所有已验证 ncRNA 基因进行长度统计(图 1)，发现大于 97%的 ncRNA 基因长度在 40−601 nt 之间，长度大于 601 nt 和小于 40 nt 的 ncRNA 基因分别占 2.0%和 0.7%。这是由于细菌中除了存在大量长度在 40−500 nt 之间的 ncRNA 外，还存在一部分较长的 ncRNA (Long non-coding RNA，lncRNA)[14]，与 40−500 nt 的 ncRNA 特征差异较大。此外，该数据集中还有极小一部分长度小于 40 nt 的 ncRNA。本文选取的 9 个不同 G+C 含量基因组在 BSRD 数据库中的 ncRNA 长度均在 40−600 nt 之间。

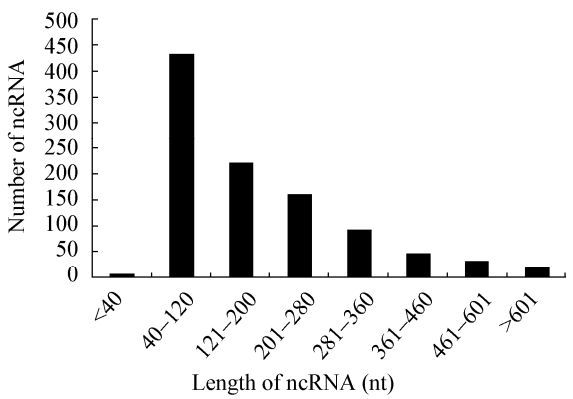

图 1 BSRD 数据库中收录的 1 008 个 ncRNA 基因的长度分布

Figure 1 Length distribution of BSRD-archived 1 008 bacterial ncRNA genes

2.2 不同 G+C 含量基因组评估 ncRNA 预测工具

基于机器学习法开发的工具 PORTRAIT 的敏感性较高，并且随基因组 G+C 含量的变化呈一定幅度波动，高 G+C 时明显降低(表 2)。低 G+C 含量和中等 G+C 含量基因组预测的平均敏感性分别为 79%和 78%，而高 G+C 含量基因组预测的平均敏感性则降为 72%。而基于转录单元特征和比较基因组学相结合开发的软件 sRNAPredict，对中等 G+C 含量基因组的平均敏感性为 42%；而对高 G+C 含量基因组的平均敏感性为 22%；对低 G+C 基因组敏感性则仅为 6%。随基因组 G+C 含量变化，sRNAPredict 的敏感性变化显著，且与 PORTRAIT 相比均较差。这表明两种工具的 ncRNA 预测效果受所预测细菌基因组 G+C 含量的影响。

此外，PORTRAIT 虽然灵敏度高，可以预测到更多 ncRNA 基因，但阳性检出率低于 65.25%，预测结果准确性较难保证。这与 sRNAPredict 高达 100%的阳性检出率形成鲜明对比。但 sRNAPredict 的敏感性却远远不如 PORTRAIT。由此可见，基于比较基因组学和转录单元特征的方法预测 ncRNA 基因，虽然阳性检出率高，但在发现潜在 ncRNA 基因方面表现并不突出。而机器学习法虽然敏感性较高，但其阳性检出率较低，即预测结果中含大量假阳性结果，不利于后期进行实验验证。

经以上分析，基于转录单元特征预测 ncRNA 的方法受细菌基因组 G+C 含量的影响非常明显。如对中、高、低 G+C 基因组预测的敏感性分别为 42%、22%和 6%。基于机器学习法的 PORTRAIT 由于没有用到转录单元特征，随着基因组 G+C 含量的变化，其阳性检出率在 30.37%−65.25%之间波动。且敏感性变化幅度也较小。我们推测 ncRNA 基因转录区域的序列特征随着细菌基因组 G+C 含量的变化而呈现出不同特征，这种序列特征在预测 ncRNA 基因方面有一定贡献性。

2.3 ncRNA 基因转录起始和终止位点序列特征与细菌基因组 G+C 含量关系

2.3.1 ncRNA 基因转录起始区域序列特征：通过分别提取高、中、低 G+C 含量基因组及大肠杆菌(属于中 G+C)中已验证 ncRNA 基因的转录起始区域序列后，使用 MEME 分析其特征，得到相应的序列信号，且每个信号均覆盖了 98%以上的输入序列。比较大肠杆菌(图 2A)和中 G+C (图 2B)基因组 ncRNA 基因转录起始区域的保守序列，发现它们均对 A、G 和 T 具有一定核苷酸偏好性。大肠杆菌 ncRNA 基因转录起始区域特征与中等 G+C 含量基因组中 ncRNA 在核苷酸偏好性方面是一致的，但并没有呈现出与 sRNAscanner 工具使用的大肠杆菌−10 box 和−35 box 完全一致的信号[4]。

高、中、低 G+C 含量基因组的 ncRNA 基因转录起始区域均能捕捉到明显的序列信号特征(图 2)。与同属中 G+C 的大肠杆菌基因组相比(图 2A)，中等 G+C 含量基因组的 ncRNA 基因转录起始区域 A+G 核苷酸偏好性的信号更加明显，且噪音信号更弱；噪音信号的减弱可能与输入的 ncRNA 数目(357 条序列)大幅增加有关。低 G+C 含量菌株

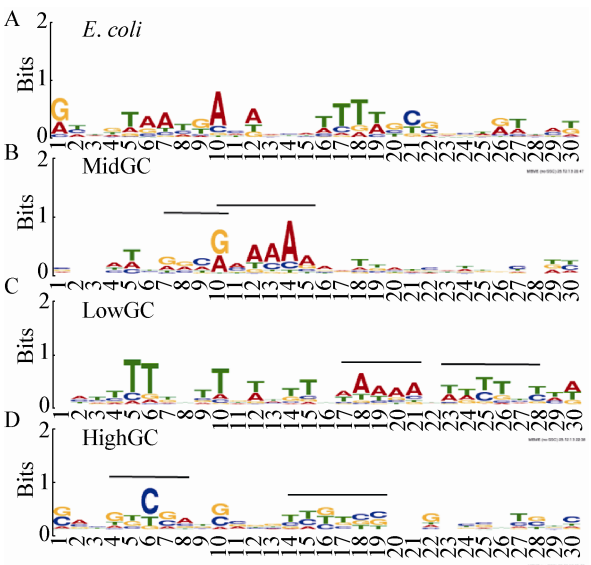

图 2 细菌 ncRNA 基因的转录起始区域信号特征

Figure 2 Conserved bases in the promoter regions of the bacterial ncRNA genes classified by genome G+C content

Note: A: *E. coli*; B: Genome with middle G+C content (45%−60%); C: Genome with low G+C content (<45%); D: Genome with high G+C content (>60%). Conserved sequence patterns used by sRNAscannerII in this study are under the black lines.

ncRNA 基因转录起始区域对 A 和 T 具有很强的偏好性(图 2C)，同时高 G+C 含量菌株对 G、T 和 C 具有一定核苷酸偏好性(图 2D)。除核苷酸偏好性呈明显不同外，G+C 含量不同的基因组，其 ncRNA 基因转录起始区域的高度保守序列 motif 各不相同。以大肠杆菌启动子区中的−10 box 和−35 box 为参照，从高、中、低 G+C 三个组别的转录起始区域特征中，分别选取两个 4−6 nt 的强信号保守区域，作为转录起始区域的特征模型进行下一步分析。如图 2 中横线所标位置，中等 G+C 含量的两个信号特征区域在图中相对坐标分别为 7−10 nt 和 10−15 nt；低 G+C 含量的相对坐标分别为 17−21 nt 和 23−28 nt；高 G+C 含量的相对坐标分别为 4−8 nt 和 14−19 nt。上述分析说明 ncRNA 基因转录起始区域特征与基因组 G+C 含量紧密相关。基因组 G+C 含量不同，ncRNA 基因转录起始区域序列呈现不同且具有代表性的特征。

2.3.2 ncRNA 基因转录终止区域特征：目前基于转录终止区域特征预测 ncRNA 的软件只能识别不依赖于 *rho* (ρ)的终止子[9]，这也是影响 ncRNA 预测准确性的一个要素。为了探究转录终止区域在不同 G+C 含量基因组中是否存在较大差异，我们同样分析了其特征信号。使用 MEME 对高、中、低 G+C 含量基因组以及大肠杆菌(属于中 G+C)中 ncRNA 基因的转录终止区域特征进行分析，得到相应的信号特征，且每个信号均覆盖了 98%以上的输入序列。对结果进行比较，发现大肠杆菌(图 3A)与中 G+C (图 3B)基因组 ncRNA 基因转录起始区域的保守序列具有高度一致性。

高、中、低 G+C 含量基因组的 ncRNA 基因转录终止区域均能捕捉到较强的信号特征(图 3)。中 G+C 基因组 ncRNA 基因转录终止区域的保守序列中含有长度为 3 nt 的 G 核苷酸偏好性的序列，其下游含多个 T 核苷酸(图 3B)。低 G+C 组别的转录终止区域保守序列在末端有多个 T 核苷酸，而其上游则对 T 和 G 具有一定核苷酸偏好性(图 3C)。在高 GC 的转录终止区域中，多个 T 核苷酸区的上游特征不明显，其下游的序列则呈现一定规律性(图 3D)。由此可见，细菌 ncRNA 基因转录终止区域的特征与细菌基因组 G+C 含量高度相关。

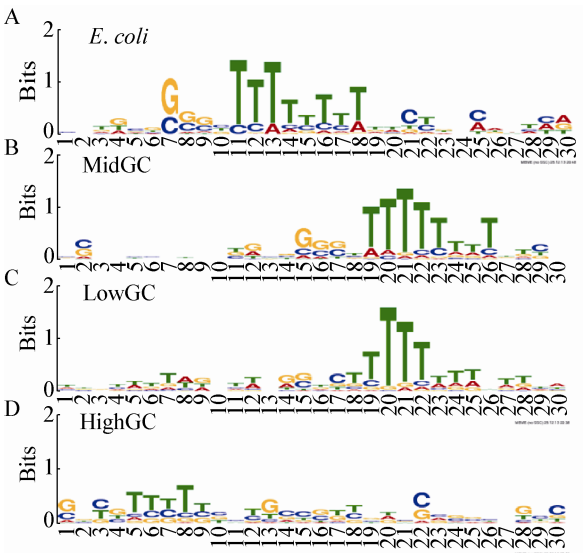

图 3 细菌 ncRNA 基因的转录终止区域信号特征

Figure 3 Conserved bases in the terminator regions of the bacterial ncRNA genes classified by genome G+C content

Note: A: *E. coli*; B: Genome with middle G+C content (45%–60%); C: Genome with low G+C content (<45%); D: Genome with high G+C content (>60%).

但是目前一些基于转录单元特征预测 ncRNA 基因的方法或软件[4]，在基因间序列利用权重矩阵等方法寻找转录起始和终止区域时，并没有考虑基因组 G+C 含量不同导致转录起始和终止区域序列特征不同这一因素，因此在预测不同 G+C 含量的物种时，敏感度会大幅下降。这也是限制目前软件普适性的重要因素之一。

2.3.3 ncRNA 基因转录起始和终止区序列特征对 sRNAscanner 预测效果有很大影响：sRNAscanner 软件根据大肠杆菌转录起始和终止区域的特征序列矩阵来预测 ncRNA 基因[4,9]，在对大肠杆菌及其相近菌属的预测中表现优异。用 sRNAscannerI 和 sRNAscannerII 两种方法分别对 9 个检测基因组进行 ncRNA 预测，结果如表 3 所示。低和高 G+C 组别的预测结果中，sRNAscannerII 比 sRNAscannerI 预测的 ncRNA 数目大幅增加。对于低 G+C 的金黄色葡萄球菌和幽门螺旋杆菌，sRNAscannerII 预测

表 3 sRNAscannerI 和 sRNAscannerII 两种方法对 ncRNA 基因的预测结果
Table 3 ncRNA gene prediction results of sRNAscannerI and sRNAscannerII

菌株(G+C 含量) Strains (G+C content)	基因组大小 Genome size (Mb)	ncRNA 基因数目 ncRNA gene No.			
		方法 I sRNAscannerI	方法 II sRNAscannerII	BSRD 中对应菌株 The same strain in BSRD	方法 I 与方法 II 重叠 Overlapping by sRNAscannerI and II
Staphylococcus aureus N315 (33%)	2.9	123[21][a]	349[29][a]	57	83
Helicobacter pylori 26695 (39%)	1.7	71[0][a]	201[22][a]	80	0
Synechocystis sp. PCC 6803 (48%)	3.6	26[0][a]	11[0][a]	75	2
Escherichia coli K-12 MG1655 (51%)	4.7	83[15][a]	48[9][a]	63	12
Salmonella enterica serovar Typhimurium SL1344 (52%)	5.0	74[11][a]	60[9][a]	113	10
Salmonella enterica serovar Typhimurium LT2 (52%)	4.9	83[1][a]	45[6][a]	68	0
Mycobacterium bovis BCG str. Pasteur 1173P2 (66%)	4.4	1[0][a]	33[0][a]	36	0
Pseudomonas aeruginosa UCBPP-PA14 (66%)	6.6	10[3][a]	44[5][a]	66	1
Pseudomonas aeruginosa PAO1 (67%)	6.4	12[2][a]	45[9][a]	117	0

Note: [a]: Number of predicted ncRNA genes matching to the ncRNA genes collected by BSRD.

的 ncRNA 比 sRNAscannerI 分别多了 184%和 183%；对于高 G+C 的牛分枝杆菌和两株铜绿假单胞菌则分别多预测了 3 200%、275%和 340%。然而对于中 G+C 组别的集胞藻、大肠杆菌和两株沙门氏菌，sRNAscannerII 比 sRNAscannerI 分别少预测了 58%、42%、18%和 46%的 ncRNA。这极有可能是由于 sRNAscanner 预测模型来源于大肠杆菌基因数据，因而对大肠杆菌及其相近菌属的预测效果优异。这些结果部分证明根据基因组 G+C 含量不同细分 ncRNA 基因转录起始和终止区域特征，从而进行 ncRNA 基因预测具有一定合理性。

为考察这些预测结果的可信度，进一步将预测到的 ncRNA 基因与 BSRD 中相同菌株的 ncRNA 基因进行比较。预测的 ncRNA 中没有被 BSRD 数据库收录的部分，则与整个 BSRD 数据库进行序列相似性比对。以此得到这些预测 ncRNA 基因中被实验验证的情况(图 4)。sRNAscannerI 预测结果中，中等 G+C 含量的大肠杆菌基因组中 83 个潜在的 ncRNA 基因，被 BSRD 收录的个数为 12，没有被收录的 71 个 ncRNA 基因与 BSRD 数据库中 ncRNA 基因同源的有 3 个，共占约 18%。而 sRNAscannerII 预测的 48 个潜在 ncRNA 基因中分别为 7 和 2，共占约 19%。同时，在已验证的 63 个 ncRNA 中，两种方法分别预测出了 12 和 7 个。同样对于中 G+C 的两株沙门氏菌，sRNAscannerI 预测的 74 和 83 个 ncRNA 中，分别共有 11 和 1 个与 BSRD 数据库 ncRNA 重叠或同源，各占约 15%和 1%。但 sRNAscannerII 预测的 60 和 45 个 ncRNA 中各占 15%和 20%，比 sRNAscannerI 的敏感度高。而对于中等 G+C 含量的集胞藻基因组，sRNAscannerI 和 sRNAscannerII 分别预测了 26 个和 11 个 ncRNA 基因，且与 BSRD 数据库 ncRNA 重叠或同源的均为 0。而对于低 G+C 含量的幽门螺旋杆菌基因组，sRNAscannerI 方法预测的 71 个 ncRNA 没有被 BSRD 收录，且与 BSRD 中 ncRNA 基因同源的 ncRNA 为 0，但是 sRNAscannerII 预测的 201 个 ncRNA 被 BSRD 收录以及与其同源的均为 11 个，共约占预测总数的 11%，预测效果大大提高。低 G+C 的金黄色葡萄球菌 sRNAScannerI 和 sRNAscannerII 分别预测到的 123 和 349 个

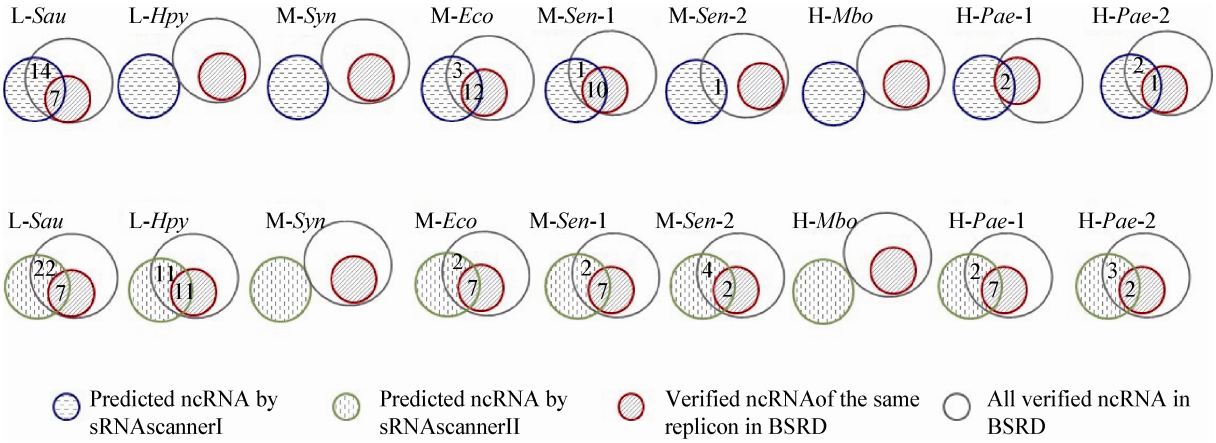

图 4　sRNAscanner 两种方法预测 G+C 含量不同基因组的 ncRNA 基因结果示意图

Figure 4　Schematic diagram of ncRNA prediction results by using sRNAscanner for the genomes with different G+C content

Note: L: Genome with low G+C content (<45%); M: Genome with middle G+C content (45%−60%); H: Genome with high G+C content (>60%). *Sau*: *Staphylococcus aureus* subsp. *aureus* N315; *Hpy*: *Helicobacter pylori* 26695; *Syn*: *Synechocystis* sp. PCC 6803; *Eco*: *Escherichia coli* K12 MG1655; *Sen*-1: *Salmonella enterica* serovar. Typhimurium SL1344; *Sen*-2: *Salmonella enterica* serovar Typhimurium LT2; *Mbo*: *Mycobacterium bovis* BCG str. Pasteur 1173P2; *Pae*-1: *Pseudomonas aeruginosa* PAO1; *Pae*-2: *Pseudomonas aeruginosa* UCBPP-PA14.

ncRNA 中，与 BSRD 重叠的 ncRNA 均为 7 个，与 BSRD 同源的 ncRNA 分别为 14 和 22 个，所占预测总数比例分别约为 17%和 8%。对于高 G+C 含量的两株铜绿假单胞菌，sRNAscannerI 预测的 ncRNA 与 BSRD 重叠或同源的共为 2 和 3 个，分别约占预测总数的 17%和 30%；sRNAScannerII 的则为 3 和 5 个，分别约占预测总数的 20%和 11%。而同样属于高 G+C 含量的牛分枝杆菌，sRNAscannerI 和 sRNAscannerII 分别预测了 1 和 33 个 ncRNA 基因，且与 BSRD 数据库 ncRNA 重叠或同源的均为 0。以上说明，sRNAscannerII 预测的 ncRNA 基因总数基本上有所增加，同时在结果可靠性上并不比 sRNAscannerI 逊色。

从上述分析中发现，对大肠杆菌基因组的 ncRNA 基因预测中，用中等 G+C 含量基因组的 ncRNA 基因的转录单元特征预测得到的结果，没有用 sRNAscanner 缺省的位置权重矩阵，即针对大肠杆菌而建立的矩阵预测的效果好。一方面可能是由于 MEME 捕捉保守序列存在一些不稳定性，需要进一步使用更加精细的手段进行转录单元特征信号的捕捉，或者进行特征信号放大或噪音消除处理；另一方面说明，不同种属的 ncRNA 基因转录单元特征可能存在差异性，这方面还需要进行更加深入的研究。

3　小结

大多数基于基因组学和转录单元特征方法的细菌 ncRNA 基因预测工具普适性很低，而基于机器学习方法的预测工具受阳性训练集以及特征向量参数的制约，假阳性很高同时阳性检出率很低。本文通过对不同 G+C 含量基因组中 ncRNA 基因转录起始和终止区域的序列特征进行捕捉，证明随基因组 G+C 含量不同，ncRNA 基因转录单元特征存在明显差异。为探究这些特征信号对 ncRNA 基因预测的影响，用这些特征评估基于转录起始和终止区的 ncRNA 基因预测工具 sRNAscanner。结果发现在预测的 ncRNA 基因总数增加以及可靠性方面，绝大多数基因组至少有一个方面 sRNAscannerII 是优于 sRNAscannerI 的。这说明不同 G+C 含量基因组中 ncRNA 基因转录单元特征是 ncRNA 基因预测的重要参数之一，而 RNA-Seq 转录组测序获得菌株特异的转录单元特征将为提高

ncRNA 基因的准确预测提供新的思路。

参 考 文 献

[1] Li W, Ying X, Lu Q, et al. Predicting sRNAs and their targets in bacteria[J]. Genomics, Proteomics & Bioinformatics, 2012, 10(5): 276-284.

[2] Pichon C, Felden B. Small RNA gene identification and mRNA target predictions in bacteria[J]. Bioinformatics, 2008, 24(24): 2807-2813.

[3] Rivas E, Klein RJ, Jones TA, et al. Computational identification of noncoding RNAs in *E. coli* by comparative genomics[J]. Current Biology, 2001, 11(17): 1369-1373.

[4] Sridhar J, Sambaturu N, Sabarinathan R, et al. sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes[J]. PLoS One, 2010, 5(8): e11970.

[5] Gruber AR, Neuböck R, Hofacker IL, et al. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures[J]. Nucleic Acids Research, 2007, 35: W335-W338.

[6] Arrial RT, Togawa RC, Brigido MM. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus *Paracoccidioides brasiliensis*[J]. BMC Bioinformatics, 2009, 10(1): 239.

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

[8] Li L, Huang D, Cheung MK, et al. BSRD: a repository for bacterial small regulatory RNA[J]. Nucleic Acids Research, 2013, 41: D233-D238.

[9] 刘倩, 应晓敏, 吴佳瑶, 等. 基于转录终点序列特征预测大肠杆菌 sRNA[J]. 生物物理学报, 2011, 27(3): 257-264.

[10] Bailey TL, Williams N, Misleh C, et al. MEME: discovering and analyzing DNA and protein sequence motifs[J]. Nucleic Acids Research, 2006, 34: W369-W373.

[11] Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences[J]. Bioinformatics, 1999, 15(7/8): 563-577.

[12] Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25(17): 3389-3402.

[13] Shao Y, He X, Harrison EM, et al. mGenomeSubtractor: a web-based tool for parallel *in silico* subtractive hybridization analysis of multiple bacterial genomes[J]. Nucleic Acids Research, 2010, 38: W194-W200.

[14] Mellin JR, Cossart P. The non-coding RNA world of the bacterial pathogen *Listeria monocytogenes*[J]. RNA Biology, 2012, 9(4): 372-378.