·论 著·

鼠疫耶尔森菌基因组核糖体结合部位识别

王宇萌，俞东征

中国疾病预防控制中心传染病预防控制所鼠疫室，传染病预防控制国家重点实验室，北京 102206

摘要：目的 通过统计分析鼠疫耶尔森菌（鼠疫菌）CO92株中核糖体结合位点（RBS）与翻译起始位点的距离，探索最适合翻译起始的RBS与起始位点的距离。**方法** 通过对鼠疫菌所有拷贝的16S rRNA 3′端，及已经通过实验确定的177处RBS进行分析，确定鼠疫菌RBS的序列特征。统计鼠疫菌CO92株中存在的所有RBS序列的位置和数量，及RBS序列与翻译起始序列ATG（GTG，TTG，CTG）的距离。观察已经公布的CO92株的编码序列（CDS）片段前20个碱基序列的特征。**结果** 在CO92的全基因组序列中搜索，共发现可能的RBS结构5081个，2909个后面一段距离内存在开放读码框架，其中1541与标注的CDS相符，535与标注的CDS终止位点相同，但起始位点不同。CO92序列中的3887 CDS前面20个碱基中，57.7%含有RBS序列。**结论** RBS序列与起始位点间隔7个碱基和8个碱基出现的次数最多；RBS可能作为基因识别的重要指征。

关键词： 核糖体结合位点；翻译起始；鼠疫耶尔森菌

中图分类号：R254.8 **文献标志码：**A **文章编号：**1003-4692（2012）06-0503-04

Identification of ribosomal binding sites in *Yersinia pestis* genome

WANG Yu-meng, YU Dong-zheng

State Key Laboratory for Infectious Diseases Prevention and Control, *National Institute for Communicable Disease Control and Prevention*, *Chinese Center for Disease Control and Prevention*, *Beijing* 102206, China

Supported by the Major National Science and Technology Projects of China（No. 2011ZX10004-001）and Grant from the Ministry of Health of the People's Republic of China（No. 201202021）

Abstract: Objective To determine the optimal distance between ribosomal binding site (RBS) and translation initiation site (TIS) by statistical analysis of RBS-TIS distances in *Yersinia pestis* CO92. **Methods** All copies of 3′ -terminal 16S rRNA and 177 annotated RBS sequences in CO92 were analyzed to identify the characteristics of *Y. pestis* RBS sequences. The positions and number of all RBS sequences in CO92 and their distances from the TIS ATG (GTG, TTG, CTG) were determined by statistical analysis. The characteristics of the 20 upstream bases in the annotated coding DNA sequences (CDSs) in CO92 were observed. **Results** A total of 5081 potential RBS sequences were found in CO92 genome. Of the 5081 RBS sequences, 2909 had downstream open reading frames (ORFs); of the 2909 RBS sequences, 1541 had the ORFs identical to the annotated CDSs, and 535 had the same termination sites as the annotated CDSs but had different initiation sites. In the 20 upstream bases of 3887 annotated CO92 CDSs, 57.7% contained RBS sequences. **Conclusion** The most frequent distances between RBS and TIS were 7 and 8 bases in CO92 genome. RBS can be an important gene index for identification.

Key words: Ribosomal binding site; Translation initiation; *Yersinia pestis*

分子流行病学是近十几年才迅速发展起来的一门流行病学新分支，它是由传统流行病学学科发展的强烈需求和分子生物学理论及技术取得的巨大成就相结合的产物。实验室检测是分子流行病学的研究方法之一，其中DNA测序近年来发展很快，通过对造成某一疾病流行的病原体，在基因水平上分析其特征，从而更准确地解决传染源和传播途径等有关流行病学问题。

随着英国sanger中心2001年第1株鼠疫菌全基因组序列测定的完成，目前GeneBank上已经公布7株鼠疫菌的全基因组序列。如何充分利用这些巨大和宝贵的生物学数据库，是目前生物学、分子流行病学面临的挑战和难题。基因是生物基因组中的主要组成部分，微生物中的基因改变决定着疾病能否发生以及是否流行，确定微生物基因组中突变是否在基因之内是其中的关键，这就要求对基因准确定位，这是研究基因功能、表达和它们之间关系以及如何控制基因转录等工作的基础。

目前，基因的识别使用按照马尔科夫模型设计的软件进行，只有那些与试验确定的基因接近到一定程度，起始密码与终止密码间距离为3的整数倍的片段，才被命名为编码序列（coding sequence，CDS）[1]。软件需要馈入已确定的基因资料进行训练。因此，不同时

基金项目：国家科技重大专项课题（2008ZX10004-001）；卫生行业科研专项鼠疫流行病学新技术研究与应用（201202021）

作者简介：王宇萌（1987-），女，实习研究员，主要从事病原微生物学研究工作。Email: mach9446@sina.com

期测定的细菌全基因组序列，由于软件的训练程度不同，读出的CDS长度可能不同。对于发展与微生物自然方式相符的基因确定方法需求迫切。马尔科夫模型显然不是生物识别其基因的方式。实际上蛋白质的合成方式在分子生物学发展早期即已确定：在蛋白质编码序列翻译开始之前，mRNA和tRNA都结合到核糖体30S的正确位置才能使核糖体有效识别起始密码子[2-5]。序列在翻译起始时，核糖体选择翻译起始位点和开放读码框架（open reading frame，ORF）的正确与否依赖于起始密码子上游的核糖体结合位点（ribosome-binding site，RBS序列）和mRNA的互补结合。也曾有人试图使用这种方式来确定基因[6]，但由于RBS和基因间的距离变化，该方法未能得到广泛采用。借此对鼠疫菌全基因组序列进行比较的机会，本研究比较了鼠疫菌基因组中的16S rRNA基因序列，对已经注出的RBS序列进行分析，试图建立一种依靠RBS识别基因的方法。

1 材料与方法

1.1 基因组序列 由NCBI上下载鼠疫菌CO92株的全序列，为英国SANGER中心2001年公布的第1株鼠疫菌全基因组序列[7-8]。

1.2 已注释RBS搜索 使用Editplus软件，搜索上述CO92序列中的RBS注释，记录注释的RBS序列，以及该序列末端距起始密码的距离。

1.3 16S rRNA基因序列比较 搜索CO92序列中的16S rRNA基因，用EditSeq截取其序列，在Megalign上对齐。

1.4 未标注RBS搜索 在Editseq软件中打开CO92序列，使用搜索功能，逐一搜索GGAG、GAGG、AGGA[9] 3种四联体，由GGAGG的A开始间隔6～10 bp的起始密码ATG；将光标移至ATG前，再用该软件的ORF搜索功能搜索ATG后的ORF，记录发现ORF的ATG位置。

在Editplus软件中打开CO92序列，检验前面搜索的ORF序列的终止密码是否与已标注的CDS重合。

1.5 分析已标注的CDS前20个碱基的特征 截取CO92序列中已标注的3887段CDS前的20个碱基，在这些片段中搜索RBS序列，记录RBS序列和起始位点间隔的距离以及RBS序列的长度，即组成它们的碱基数量。

2 结 果

2.1 鼠疫菌已标注RBS分析 使用Editplus软件搜索，在CO92基因组中共发现177处已标注的RBS序列。除2处未发现与CO92标注的CDS联系外，对175 RBS序列分析的结果，发现RBS序列至少需要包括GGAGG中间的A在内的连续4个与SD序列互补的碱基。这些RBS与起始密码间的间隔列于表1。这段距离在4～13核苷酸范围内，呈现以7为中值的正态分布趋势。由表2可以看出，16S rRNA与mRNA间的互补结合区域以4 bp占绝大多数，越长越少。

表1 已标注的175段RBS与CDS的关系

Table 1 Distances between 175 annotated RBS sequences and CDSs

间隔碱基数	频数	间隔碱基数	频数
4	3	10	23
5	5	11	13
6	24	12	7
7	46	13	2
8	23	合计	175
9	29		

表2 已标注的175段RBS长度分布

Table 2 Length distribution of 175 annotated RBS sequences

RBS长度（bp）	出现频数
4	109
5	33
6	21
7	12

2.2 鼠疫菌16S rRNA 3′末端序列分析 截取CO92序列中所有的6拷贝16S rRNA序列，比较发现所有拷贝均长1543核苷酸。3′末端为CCTCCTTA，由于CO92只有6拷贝的rRNA基因簇。另与KIM的7个拷贝比较[10]，也与此相同。10株鼠疫菌比较仅发现5处SNP，均不在此范围内。与作为RBS研究经典的大肠埃希菌16S rRNA 3′末端比较，只有最后一个碱基差异。说明在鼠疫菌中只存在一套RBS序列。

2.3 CO92全序列中RBS统计 在CO92全基因组序列中搜索RBS结构，结果见表3。

表3 CO92基因组中发现的RBS

Table 3 RBS sequences found in CO92 genome

A～A间隔数	RBS-ATG总数	与已有CDS符合数	终止位点相同但起始位点不同	新ORF	RBS后无ORF
6	928	351	63	218	296
7	1132	371	161	141	459
8	1148	340	50	181	577
9	824	295	68	110	351
10	1049	184	193	183	489
合计	5081	1541	535	833	2172

四联体后间隔6～10 bp的ATG共5081个，其中可以引导一段ORF起始2909个，占四联体的57.3%。与CO92标注的3887段CDS完全相符的1541个，占全部3887段CDS的39.6%；与标注的CDS终止位点相同，但起始位点不同的共有535个，占全部3887段CDS的

13.8%，两者合计占53.4%。

2.4 CO92中RBS序列与翻译起始位点间的距离 CO92的3887段已标注CDS前面的20个碱基中，含有可能为RBS序列的共2245段，占57.7%，RBS与起始密码的距离见表4。

表4 RBS序列与起始位点不同间隔的频率分布

Table 4 Frequency distribution of different distances between RBS and TIS

间隔碱基数	频数	频率(%)	累积频数	累积频率(%)
0	8	0.36	8	0.36
1	16	0.71	24	1.07
2	15	0.67	39	1.74
3	21	0.94	60	2.68
4	27	1.20	87	3.88
5	59	2.63	146	6.51
6	388	17.28	534	23.79
7	411	18.31	945	42.10
8	406	18.08	1351	60.18
9	325	14.48	1676	74.66
10	201	8.95	1877	83.61
11	108	4.81	1985	88.42
12	104	4.63	2089	93.05
13	50	2.23	2139	95.28
14	30	1.34	2169	96.62
15	30	1.34	2199	97.96
16	22	0.98	2221	98.94
17	16	0.71	2237	99.65
18	7	0.31	2244	99.96
19	1	0.04	2245	100.00
合计	2245	100.00		

在2245段中，含有RBS序列的CDS起始于RBS间隔碱基数成正态分布趋势，且间隔5～13个碱基数的累积频率为91.32%。起始位点前含有RBS序列但间隔＜4或＞13的CDS共166个。

2.5 CO92中RBS的长度统计 2245段RBS序列的长度4～7个碱基不等，长度为4的RBS出现频率最高，长度为5～7的RBS出现的频率依次降低(表5)。

表5 RBS序列长度的频率分布

Table 5 Frequency distribution of RBS sequence lengths

RBS长度	频数	频率(%)	累积频数	累积频率(%)
4	1342	59.78	1342	59.78
5	658	29.31	2000	89.09
6	193	8.60	2193	97.68
7	52	2.32	2245	100.00
合计	2245	100.00		

3 讨 论

人们最初希望按照实际的蛋白质合成方式来确定基因组中的未知基因，引入马尔科夫模型是迫不得已的办法。主要有两个因素阻碍了利用RBS定位基因。其一是人们认为从起始密码到RBS应该有一个固定的距离，结果发现这一区域内的碱基变化很大。本研究结果表明16S rRNA与mRNA间的互补结合，需要一段准确的对应关系。但无论是RNA还是核糖体都是相对柔性的结构，因此，RBS与起始密码间的距离在一定程度间变化，并不影响tRNA与起始密码在核糖体空隙中的结合。影响RBS作为基因定位因素的另一个障碍，是人们以为RBS也像基因密码一样，在所有细菌间通用。但现已知道，rRNA序列在不同生物种间存在差异，因此，不同细菌的RBS也不尽相同。幸运的是，鼠疫菌属于肠杆菌科，其基因组都具有很高程度的同源性。因此，早期在大肠埃希菌中获得的RBS研究结果，能很容易应用于鼠疫菌。

形成互补关系所需的4个相邻核苷酸分布于整个基因组，然而，只有与起始密码发生一定的关系，它们才可能起RBS的作用。在本研究中，尽管只搜索了其中频度最高的部分，发现可能位点的数量也远超出已标注的CDS数量。然而，这样的序列结构后面跟随着ORF的数量，却只占已标注CDS的一部分。将本研究的第3部分与第4部分比较，标注的CDS前具有可能的RBS结构比例，从53.4%上升至57.6%，说明还需要对四联体包括的范围及与起始密码间的距离做一定的调整，RBS也可能成为实用的基因定位指标。

本研究还发现，可能的RBS结构开始翻译过程的作用并不是等效的。mRNA与16S rRNA 3′末端的互补段越长，其结合就越牢固，翻译起始作用就越可靠。在本研究中，不论是对已标注的177 RBS位点的分析，还是对全基因组的搜索，都发现一些基因前的RBS可长达7个核苷酸，这可能是最有效的RBS结构。而RBS与起始密码间的距离，在177已标注的RBS中7个核苷酸出现频率最高，而且全基因组搜索中也以7个与8个核苷酸的出现频率最高。因此，在已经标注的CDS中，可能符合这些条件的基因为最有效表达的基因。而不仅基因本身的突变、改变RBS或者RBS至起始密码距离的突变，也会影响到基因的表达。

以RBS作为基因定位的指标，还存在着一些问题。在本研究中，还出现了相当数量的符合蛋白质合成开始条件却未在基因组注释范围之内的ORF，其中绝大多数是不满100个氨基酸的小肽。这种短肽，是原来被排除在CDS标注范围之外的。它们是否实际存在，需要通过实验研究来确定。

总而言之，本研究所做的工作还非常有限，仅对1株鼠疫菌进行分析，进一步的工作需对更多菌株进行

(下转第511页)

2005, 43(5):2286-2290.

[6] Clarridge JE 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious disease[J]. Clin Microbiol Rev,2004,17(4):840-862.

[7] Woo PC, Ng KH, Lau SK, et al. Usefulness of MicroSeq 500 16S ribosomal DNA-based bacterial identification system for identification of clinically significant bacterial isolates with ambiguous biochemical profiles[J]. J Clin Microbiol,2003,41(5):1996-2001.

[8] Fontana C, Favaro M, Pelliccioni M, et al. Use of the MicroSeq 500 16S rRNA gene-based sequencing for identification of bacterial isolates that commercial automated systems failed to identify correctly[J]. J Clin Microbiol,2005,43(2):615-619.

[9] Carretto E, Barbarini D, Couto I, et al. Identification of coagulase-negative staphylococci other than *Staphylococcus epidermidis* by automated ribotyping [J]. Clin Microbiol Infect, 2005, 11 (3): 177-184.

[10] Layer F, Ghebremedhin B, Moder KA, et al. Comparative study using various methods for identification of *Staphylococcus* species in clinical specimens[J]. Clin Microbiol,2006,44(8):2824-2830.

[11] Jones SW, Franceseoni SC. DNA assays for detection, idcntification, and individualization of select agent microorganisms[J]. Croat Mcd J,2005,46(4):522-529.

[12] Noller HF, Hoang L. The 30S ribosomal P site: a function of 16S rRNA[J]. FEBS Lett,2005,579(4):855-858.

[13] Kibe R, Sakamoto M, Hayashi H, et al. Maturation of the murine cecal microbiota as revealed by terminal restriction fragment length polymorphism and 16S rRNA gene clone libraries [J]. FEMS Microbiol Lett,2004,235(1):139-146.

[14] Petrosino JF, Highlander S, Luna RA, et al. Metagenomic pyrosequencing and microbial identification[J]. Clin Chem, 2009, 55(5):856-866.

[15] Peterson DA, Frank DN, Pace NR, et al. Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases [J]. Cell Host Microbe,2008,3(6):417-427.

[16] Sönksen UW, Christensen JJ, Nielsen L, et al. Fastidious Gram-Negatives: identification by the Vitek 2 Neisseria - Haemophilus card and by partial 16S rRNA gene sequencing analysis[J]. Open Microbiol J,2010,4:123-131.

[17] Zbinden A, Böttger EC, Bosshard PP, et al. Evaluation of the colorimetric Vitek 2 card for identification of gram - negative nonfermentative rods: comparison to 16S rRNA gene sequencing [J]. J Clin Microbiol,2007,45(7):2270-2273.

[18] 陈佳木,李亚伦,高思维,等. 马尾口岸蝇类生物多样性和携带病原微生物情况的研究[J]. 检验检疫学刊,2009,19(6):21-27.

[19] Tóth EM, Schumann P, Borsodi AK, et al. *Wohlfahrtiimonas chitiniclastica* gen. nov., sp. nov., a new gammaproteobacterium isolated from *Wohlfahrtia magnifica* (Diptera: Sarcophagidae)[J]. Int J Syst Evol Microbiol,2008,58(4):976-981.

[20] Rebaudet S, Genot S, Renvoise A, et al. *Wohlfahrtiimonas chitiniclastica* bacteremia in homeless woman[J]. Emerg Infect Dis, 2009,15(6):985-986.

[21] Almuzara MN, Palombarani S, Tuduri A, et al. First case of fulminant sepsis due to *Wohlfahrtiimonas chitiniclastica*[J]. J Clin Microbiol,2011,49(6):2333-2335.

收稿日期:2012-07-16

(上接第505页)

抽样和统计分析,以求得到RBS在鼠疫菌中与基因翻译起始更加明确的关系。并且核糖体结合部位的识别是一种复杂的受多因素影响的过程,对其进行准确可靠的预测,仍需不断纳入新方法,不断提高预测模型的精度,使预测核糖体结合部位、基因识别更加准确。

参考文献

[1] Salzberg SL, Delcher AL, Kasif S, et al. Microbial gene identification using interpolated Markov models[J]. Nucleic Acids Res,1998,26(2):544-548.

[2] Korostelev A, Trakhanov S, Asahara H, et al. Interactions and dynamics of the Shine - Dalgarno helix in the 70S ribosome [J]. PNAS,2007,104(43):16840-16843.

[3] Haentjens-Sitri J, Allemand F, Springer M, et al. A competition mechanism regulates the translation of the *Escherichia coli* operon encoding ribosomal proteins L35 and L20[J]. J MoL Biol,2008,375 (3):612-625.

[4] Winkler WC, Breaker RR. Regulation of bacterial gene expression by riboswitches[J]. Annu Rev Microbiol,2005,59:487-517.

[5] Yusupova G, Jenner L, Rees B, et al. Structural basis for messenger RNA movement on the ribosome [J]. Nature, 2006, 444 (7117): 391-394.

[6] Hayes WS, Borodovsky M. Deriving ribosomal site(RBS) statistical models from unannotated DNA sequences and the use of RBS model for N-terminal prediction[J]. Pac Symp Biocomput,1998:279-290.

[7] Parkhill JBW, Wren NR, Thomson RW, et al. Genome sequence of *Yersinia pestis*, the causative agent of plague[J]. Nature,2001,413 (6855):523-527.

[9] Ma J, Campbell A, Karlin S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures[J]. J Bacteriol,2002,184(20):5733-5745.

[8] 申小娜,海荣,俞东征. 鼠疫菌基因组学研究进展[J]. 疾病监测, 2009,24(6):440-445.

[10] Deng W, Burland V, Plunkett G, et al. Genome sequence of *Yersinia pestis* KIM[J]. J Bacteriol,2002,184(16):4601-4611.

收稿日期:2012-04-29