利用蛋白质同源性搜索检验细菌预测基因的起始位点

夏 伟[1,2]， 周大为[1,2]， 李炜疆*[1,2]

（1. 江南大学 工业生物技术教育部重点实验室，江苏 无锡，214122；2. 江南大学 生物工程学院，江苏 无锡 214122）

摘要：细菌基因组中基因密度高、结构相对简单，基因预测可以达到较高的准确度。细菌基因预测的一个重要问题是难以准确判断基因起始位点，目前应用最广的基因组自动注释方法给出的结果中，占总数约 1/5 的基因的起始位点有分歧，尚无利用基因结构的统计特征改进起始位点预测的有效方法。通过蛋白质同源性搜索，为推断起始位点提供新的线索。对氧化葡萄糖酸杆菌 *Gluconobacter oxydans* 621H 全基因组的分析表明，这种方法可以实质性改善起始位点预测结果。

关键词：基因组注释；基因预测；同源性搜索

中图分类号：Q 753　**文献标志码**：A　**文章编号**：1673－1689(2012)08－0865－06

Examining Start Sites of Predicted Bacterial Genes Through Protein Homology Search

XIA Wei[1,2]， *ZHOU Da-wei*[1,2]， *LI Wei-jiang**[1,2]

(1. Key Laboratory of Industrial Biotechnology, Ministry of Education, Jiangnan University, Wuxi 214122; 2. School of Biotechnology, Jiangnan University, Wuxi 214122, China)

Abstract: Due to high density and relatively simple structure, genes in bacterial genomes can be predicted with high accuracy. An important remaining issue in bacterial gene prediction is to precisely identify the gene start sites. At present, the most widely used automatic genome annotation pipelines result in the disagreed prediction of start sites for about 1/5 of the total genes. There is no effective method available to improve the start site prediction by using statistical characteristics of gene structure. This manuscript attempts to find new clues for the deduced start sites through protein homology search. Our analysis of the genome of *Gluconobacter oxydans* 621H shows that this method can substantially improve the start site prediction.

Key words: genome annotation, gene prediction, homology

近年来测序技术的发展非常迅速，使用最新的测序方法，微生物的完整基因组能够以较低的成本、在较短的时间内完成，目前在 GenBank 中存储着＞1 500 个已完成细菌基因组，另外还有近 5 000 个细菌基因组正在进行测序分析(http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi)。因而利用完整基因组数据探索理解微生物的生化生理性质已经成为生物技术有效而必须的途径。

收稿日期：2011－07－02

*通信作者：李炜疆(1964－)，男，陕西榆林人，理学博士，教授，博士研究生导师，主要从事生物信息学与计算分子生物学研究。
E-mail:wjlee01@gmail.com

获得基因组序列后的第一项基因组学分析任务就是寻找基因并预测所编码蛋白质的生物学功能，主要是通过适当的计算方法自动完成的[1]。基因发现对于原核生物和真核生物存在着巨大的本质性差异，这不仅仅由于基因组尺度（典型的原核基因组碱基对数不及典型真核基因组的 1/100），更重要的是因为基因组织结构的不同。真核生物基因组主要由大量非编码序列构成，基因密度很低，例如人基因组的 3%编码蛋白质基因，并且基因由不连续片段构成，一个完整基因常常被大段的内含子分割为许多外显子，这为真核基因识别带来很大困难[2-3]。

相比之下，原核生物的基因发现要简单很多。原核基因组一般大小为 2～5 Mbp，有很高的基因密度，通常整个基因组的 80%以上是蛋白质编码基因。对基因发现最有利的是，基因是连续的编码序列，中间没有插入片段，因而基因组中所有起始密码与终止密码之间可能的阅读框架就构成了全部蛋白质编码基因的候选集合，原核基因预测方法就是根据适当的序列特征判断哪些候选者是真正的基因[4-5]。

由于原核基因组相对简单，以基因预测方法为主要支撑的完全自动注释已具有很高的实用性，有很多机构提供免费、快捷的 Web 注释服务，其中最著名的、代表当前注释水平的是 NCBI 的 genbank 数据库、美国能源部联合基因组研究所（JGI）和劳伦斯伯克力美国国家实验室（BDMTC）联合主办的 IMG、J. Craig Venter 研究所的 JCVI 和国家病原微生物资源（NMPDR）的 RAST。这些服务给出的自动注释结果有相当高的一致性，然而也存在显著的差别，主要体现在基因起始位点预测分歧较大，解决这些分歧将有助于提高自动注释质量。利用基因结构的概率统计特征提高起始位点预测准确率，需要对基因以及基因上游（如核糖体结合位点）结构设计更好的识别模型，依靠基因预测方法本身无法对分歧作出评判。作者通过蛋白质序列同源性查询，从基因组以外的知识库中寻找准确起始位点的证据。

1 材料与方法

1.1 材料

1.1.1 基因组数据 作者以氧化葡糖杆菌 *Gluconobacter oxydans* 621H 基因组（GenBank 存取码 NC_006677）为基本素材。氧化葡糖杆菌 *G. oxydans* 是革兰氏阴性菌，属于醋酸杆菌（Acetobacteraceae）科，广泛出现于蜂蜜、果汁、啤酒、葡萄酒中，可以不完全氧化糖、酒精和酸，具有重要的工业用途[6]。它的全基因组序列在 2005 年发表[7]，该基因组共 2 702 173 bp，GC 含量为 61.1%，其在 NCBI 中 Genbank 的原始注释共 2 501 个蛋白质编码基因。

1.1.2 基因组注释服务器 原核生物经常使用的注释软件有：Glimmer [8]、GeneMarkS [9]、EasyGene 等。本文使用的基因自动注释服务器是 IMG（http://img.jgi.doe.gov）、RAST（http://rast.nmpdr.org/）、JCVI（http://www.jcvi.org/cms/research/projects/annotation－service/），Glimmer 和 GeneMarkS 都是利用统计学模型的从头预测，不依赖任何先验知识和经验参数对基因序列的预测，其中 IMG 基于 GeneMarkS，RAST 和 JCVI 基于 Glimmer 对全基因组进行注释，这个阶段主要是对实际编码蛋白基因的识别；接下来搜寻已被确切注释的亲缘关系相近的基因组，对之前预测的基因片段进行修饰。IMG 可允许用户直接在其服务器上查看分析结果，而 RAST 和 JCVI 需提交基因组数据等待服务器的反馈结果。

1.2 方法

1.2.1 获取基因组注释数据 在线提交全基因组序列的 FASTA 格式文件至 3 个注释服务器，服务器指定任务代码，等待一到两个月后下载注释结果，其中包含预测的基因起始位点和终止位点、蛋白质功能、RNA 基因，提供格式为 excel 和 txt 格式。

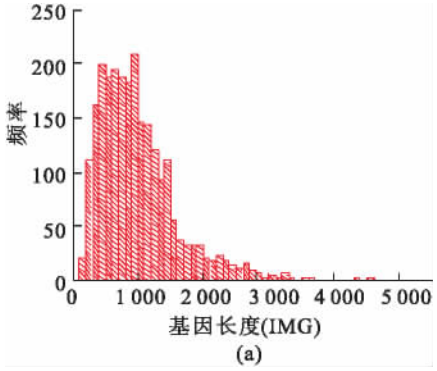
1.2.2 同源性查询 本文同源性查询是提交 NCBI 提供的 Blast 服务（http://blast.ncbi.nlm.nih.gov/Blast.cgi）完成，为了尽可能不遗漏同源性信息，作者将 E-value 阈值设置为很宽松的 0.01。3 种注释服务器对同一基因组注释的结果是不完全相同，本研究集中比较三者中具有不同起始位点、相同终止位点的基因，然后取最长的氨基酸序列 BLAST[10]。由于作者的目的是从已知的蛋白质中寻找有关待查询基因的准确长度的线索，所以查询目标库选择为 Swiss－Prot. Swiss－Prot 数据库是全部已知蛋白质库 Uniprot（www.uniprot.org）的一个子集，其中的内容都经过研究人员通过计算机工具和查

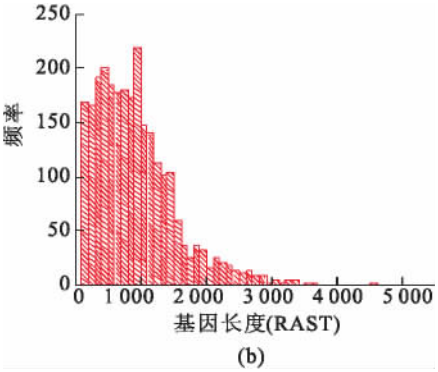
阅相关资料文献仔细核实的,可以看作是已经确认的蛋白质的数据库[11]。

2 结果与分析

2.1 注释概况

在本研究中,综合运用了序列相似性搜索、基因预测软件的一致性策略的方法对 DNA 序列进行预测。经比较发现,GenBank 文件中的注释与 IMG 注释结果完全一致,说明 NCBI 是直接引用 IMG 服务器的注释结果的,而通过当下一种注释服务器来分析一个基因组是无法得到非常准确的结果的,最大化降低注释错误的有效方法是对多种注释服务器的结果进行综合比较。由于不同的注释方法和标准,不同的注释服务器预测的基因也是不同的。JCVI 预测的基因数目最多,为 2 664 个;其次是 RAST,共 2 632 个;IMG 预测的基因最少,2 432 个。3 个注释服务器预测的最长基因都是 5 253 bp,最短基因则有所不同:IMG 预测的最短基因长度为 81 bp,RAST 和 JCVI 的是 114 bp。图 1 是预测基因的具体长度分布,可以看出,IMG 预测的短基因明显少于另外两种注释。IMG 注释的基因平均长度为 1 001 bp,远高于 RAST 注释的 932 bp 和 JCVI 注释的 923 bp,表明 IMG 更倾向于避免短基因。





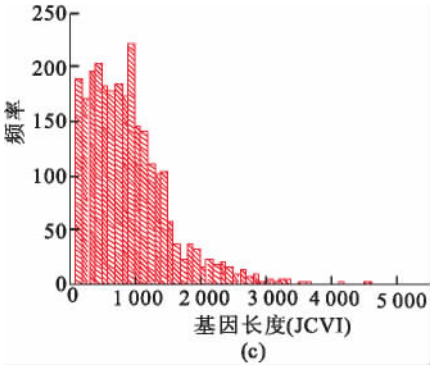


图 1 基因的长度分布

Fig. 1 Gene length distribution

同样,即便是同一个基因片段,服务器各自不同的注释标准也可能导致编码基因具有不同的起始位点或终止位点,由此引发的争议对接下来进一步的基因预测带来很大的难度[12]。在 3 种注释服务器预测的所有蛋白质编码基因,其中具有相同的终止位点的基因有 2 356 个;再比较精确匹配,也就是依据相同的起始位点和终止位点发现,RAST 和 JCVI 之间拥有与 IMG 更多的精确匹配的基因,IMG 则拥有最多的特异性起始位点,因为 RAST 和 JCVI 都是基于 Glimmer 软件预测基因的,而 IMG 则是基于 GeneMarkS。在 3 种注释服务器的预测中,有 1 686 个蛋白质编码基因是完全匹配的(Figure 2),这部分由 3 种服务器共同预测的完全匹配的编码基因的可靠性是相对较高的,但其余很大比例的编码基因的起始位点是无法确定,故基于同源性搜索来判断编码基因不同起始位点的准确性十分必要。

2.2 用同源性判断差异基因起始位点

原核基因预测是先确定基因的终止位点,然后再根据基因序列的概率模型以及基因上游区域(如核糖体结合位点)的统计特征寻找与之匹配的起始位点。一个基因在两种注释中完全吻合意味着两种注释具有完全相同的终止和起始位点;如果两种注释只具有相同的终止位点但起始位点不同,则表明这两种注释预测了同一基因,但基因长度有分歧。图 2 是对 G. oxydans 621H 基因组注释的结果统计分析,表明 IMG、RAST 和 JCVI 3 种注释中完全一致的基因有 1 686 个;而注释不一致的基因有 806 个,其中完全不同的基因很少,只有 136 个,绝大部分分歧表现为基因起始位点不同,这些基因共有 670 个。由此可见,识别正确的基因起始位点

是当前原核基因组注释面临的重要挑战。

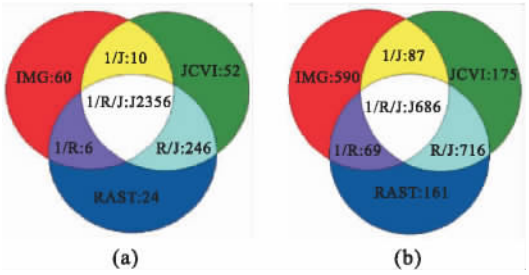


图 2 (a)左图代表 3 种注释服务器预测具有相同终止位点的蛋白质编码基因的数目,重叠区域代表 3 种服务器预测的彼此具有相同的终止位点;(b)右图显示 3 种注释服务器预测具有相同起始位点和终止位点的蛋白质编码基因数目,重叠区域代表三者之间预测的基因是精确匹配

Fig. 2 (a) The left diagram shows the number of predicted protein coding genes that share stop sites with the other annotations. Overlapping regions indicate genes having same stop sites between annotation. (b) The right diagram shows the number of predicted protein coding genes that share start site and stop site with the other annotations. Overlapping regions indicate genes having exact matches between annotations

为了更进一步分析起始位点注释的差异,我们罗列了每一个预测基因的起始密码子。在所有的基因注释中,ATG 是最普遍被利用的起始密码子,最多可占所有起始密码子的 84%。而 RAST 则包含最大比例的替代起始密码,服务器的基因预测 22%的起始密码子并不是 ATG,IMG 和 JCVI 使用相对较低比例的代替起始密码子,分别为 16%和 21%(Table1)。

表 1 基因组注释结果概况

Tab. 1 Overview of the genome annotation results

注释服务器	基因数量	平均基因长度/bp	ATG起始密码子/%	TTG起始密码子/%	GTG起始密码子/%
IMG	2432	1001	84	4.0	12
RAST	2632	932	78	6.0	16
JCVI	2664	923	79	6.0	15

以尿苷酸激酶为例,3 种注释服务器预测的基因终止位点是相同的,但起始位点都不相同,起始密码子也不同,见表 2。

以 3 种服务器预测的 670 个相同阅读框的基因序列作为数据来源对其进行本地批量 BLAST 处理,依据数据库同源性搜索已有证据对其起始位点的准确性进行推测。一般情况下,一致性在 25~30%为序列同源,相似性高于一定程度后,可推测序列为同源序列,在本次结果中,为提高数据准确性,将标准提高到 50%,由于 3 种服务器预测的基因起始位点不一样,其对应翻译的氨基酸序列长短也不相同,如果氨基酸数差异较小,也很难对其作出明确判断,故选取氨基酸序列长度差异大于 20 的基因进行分析,这一部分是可以做出判断的(Figure 3)。事实上,提供同源性方面的证据,是最直接也最准确的支持,通过比对编码基因氨基酸序列,我们选取了其中具有代表性的 8 个基因做同源匹配图,对 BLAST 结果作出细致分析。为了凸现差异,图中显示了序列中的前 120 个氨基酸匹配位置,也对 3 种服务器预测的基因的差异序列已做出明显标记,如果差异序列能够搜索到匹配,推测 3 者中长的氨基酸的起始位点是正确的,反之,在同源性标准较高的情况下,则推测是较短基因的起始位点是正确的(Figure 4)。

表 2 比较尿苷酸激酶的起始位点

Tab. 2 Comparison of uridylate kinase start sites

注释服务器	DNA编码区	起始密码子	长度/bp
IMG	1985747…1986577 (+)	TTG	831
RAST	1985735…1986577 (+)	GTG	843
JCVI	1985837…1986577 (+)	ATG	741

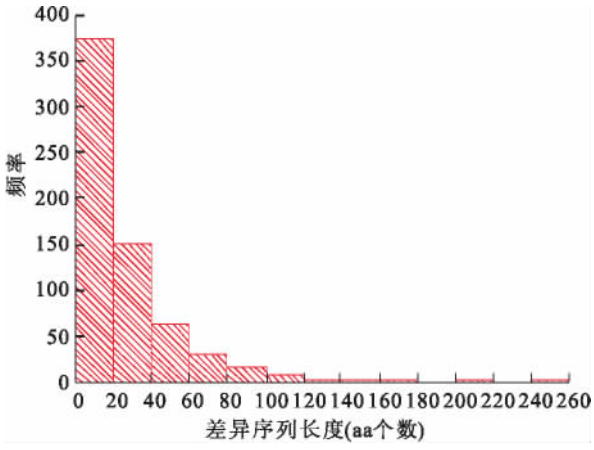


图 3 差异序列片段的长度分布

Fig. 3 Length distribution of the disagreedly predicted gene fragments



(a)



(b)

图 4 (a)代表以长片段基因序列的起始位点为依据的联配比较；(b)代表以短片段基因序列的起始位点为依据的联配比较(query 表示三种服务器预测的各自不同起始位点、相同终止位点，编码蛋白的功能，sequence1 表示预测编码基因的前 120 个氨基酸序列，sequence2－6 表示数据库中序列匹配的相同部分，ID 表示 Swiss－Prot 库中同源匹配序列名称，＋表示与之匹配的是化学性质类似的氨基酸，红色加粗字体代表各服务器预测不同起始位点的差异片段)

Fig. 4 (a)shows alignment comparison based on the start site of long fragment sequence; (b)shows alignment comparison based on the start site of short fragment sequence(Query shows that the different start sites and the same stop site by three annotation servers prediction, protein－coding function, sequence1 shows that the first 120 amino acid sequence of the gene encoding, sequence2 to 6 shows the same matching part in database, ID represent the name of sequence homologous in Swiss－Prot Database, and ＋ means the match part is chemically similar amino acids, red bold font server on behalf of the predicted start sites between the different fragments).

3 结 语

对代表当前水平的 3 个注释服务器的结果分析表明，目前细菌基因预测已达到较高的一致性水平，但不同注释服务器预测编码基因的差异是显著的，大片段差异可能会影响到后续对蛋白质功能的分析，为了更有效的从基因组注释中得到益处[13]，故解决起始位点问题不能忽视。利用蛋白质同源性搜索可以在推测起始位点中发挥重要作用，而目前存在的问题是评估同源性搜索结果的系统机制较复杂，如何实现全自动分析是将来研究的重点[14－17]。

参考文献(References):

[1] 张顺达,鞠守勇.微生物基因组的注释[J].科技创新导报.2008,(011):195－197.
ZHANG Shun-da, JU Shou-yong. The annotation of Microbial genome[J]. **Science and Technology Innovation Herald**, 2008,(011):195－197. (in Chinese)

[2] 陈铭.后基因组时代的生物信息学[J].生物信息学,2004,2(2):29－34.
CHEN Ming. Bioinformatics in the post－genomic era[J]. **China Journal of Bioinformatics**,2004,2(2):29－34. (in Chinese)

[3] 方刚,陈蕴佳,高歌.基因组数据库简介[J].遗传,2003,25(4):440－444.
FANG Gang, CHEN Yun－jia, GAO Ge. Introduction to Genome Databases[J]. **Hereditas**, 2003,25(4):440－444. (in Chinese)

[4] 刘贯山.烟草基因组知识篇:3.基因组注释[J].中国烟草科学,2010,31(01):82－82.
LIU Guan-shan. The Genome annotation[J]. **Chinese Tobacco Science**,2010,31(01):82－82. (in Chinese)

[5] 张恩民,海荣,俞东征.基因预测方法的研究进展[J].中国媒介生物学及控制杂志,2009,20(3):271－273.
ZHANG En-min, HAI Rong, YU Dong-zheng. Rseearch progress of gene prediction methods[J]. **Chin J Vector Bio & Control**,2009, 20(3):271－273. (in Chinese)

[6] Arun Gupta ,Vinay K. Singh ,G. N. Qazi ,et al. Gluconobacter oxydans: its biotechnological applications[J]. **J. Mol. Microbiol. Biotechnol**,2001 ,(3.3): 445－456.

[7] Prust C , Hoffmeister M , Liesegang H , et al. Complete genome sequence of the acetic acid bacterium Gluconobacter oxydans[J]. **Nat Biotechnol**,2005 , (23):195－200.

[8] Delcher AL, Harmon D, Kasif S, White O , et al. Improved microbial gene identification with GLIMMER[J]. **Nucleic Acids Res**,1999 ,27(23):4636－41.

[9] Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes Implications for finding sequence motifs in regulatory regions[J]. **Nucleic Acids Research**, 2001 ,29(12):2607.

[10] Christiam C,George C,Vahram A. BLAST+:architecture and applications[J]. **BMC Bioinformatics**, 2009,(10): 1186/1471－2105－10－42.

[11]李盛钰,赵玉娟,张雪 .鼠李糖乳杆菌 JAAS8 胞外多糖生物合成基因的克隆及生物信息学分析[J].食品与生物技术学报,2011,30(2):250－254.
LI Sheng-jue,ZHAO Yu-juan,ZHANG Xue. Cloning and bioinformatics analysis of exopolysaccharide biosynthesis genes from lactobacillus rhamnosus JAAS8[J]. **Journal of Food Science and Biotechnology**, 2011,30(2):250－254.

[12] Bakke P, Carney N, DeLoache W, et al. Evaluation of three automated genome annotations for Halorhabdus utahensis [J]. **PloS one**,2009, 4(7):6291.

[13]周大为,于雪飞,李炜疆.16S rDNA 序列关键位点的信息论分析[J].食品与生物技术学报,2010,29(5):782 －787.
ZHOU Da-wei ,YU Xue-fei , LI Wei-jiang. Information-Theory Analysis for Key Sites of 16S rDNA Sequence[J]. **Journal of Food Science and Biotechnology**, 2010,29(5):782－787.

[14] Yukawa H, Omumasaba C, Nonaka H et al. Comparative analysis of the Corynebacterium glutamicum group and complete genome sequence of strain R[J]. **Microbiology**,2007, 153(4):1042.

[15] Poptsova M, Gogarten J. Using comparative genome analysis to identify problems in annotated microbial genomes[J]. **Microbiology**,2010 ,156(7):1909－1912.

[16] Warren A, Setubal J. The Genome Reverse Compiler: an explorative annotation tool[J]. **BMC Bioinformatics**, 2009 ,10 (1):35－39.

[17] Zhu H Q, Hu G Q, Ouyang Z Q , et al. Accuracy improvement for identifying translation initiation sites in microbial genomes[J]. **Bioinformatics**, 2004 ,(18):3308－17.