doi:10.3969/j.issn.1005-3026.2016.10.021

Paenibacillus Shenyangensis 的 DNA 序列拼接与组装

付丽丽[1,2], 姜彬慧[1], 胡筱敏[1]
(1. 东北大学 资源与土木工程学院, 辽宁 沈阳 110819; 2. 辽宁石油化工大学 石油天然气工程学院, 辽宁 抚顺 113001)

摘 要: 在已有测序数据基础上,利用三种常见的序列组装软件对 *Paenibacillus Shenyangensis* 全基因组测序结果进行拼接组装,分析比较了不同软件在各自最优参数条件下 DNA 序列的组装数据,并与 NCBI 数据库中类芽孢杆菌属其他近缘种进行基因比对与预测. 结果表明,SOAPdenovo 的组装结果最优,在 k-mer 为 23 时,组装基因组总长和 N50 分别为 5 501 467 和 293 864 bp,预测的 4 800 个基因中有 4 393 个与 NCBI-Nr 数据库比对并注释成功.

关 键 词: 序列拼接;基因组;絮凝菌;高通量测序;类芽孢杆菌
中图分类号: Q 78 **文献标志码**: A **文章编号**: 1005-3026(2016)10-1465-04

DNA Sequence Assembly of *Paenibacillus Shenyangensis*

FU Li-li[1,2], *JIANG Bin-hui*[1], *HU Xiao-min*[1]
(1. School of Resources & Civil Engineering, Northeastern University, Shenyang 110819, China; 2. College of Petroleum Engineering, Liaoning Shihua University, Fushun 113001, China. Corresponding author: HU Xiao-min, E-mail: hxmin_jj@163.com)

Abstract: On the basis of existing sequencing data, three sequence assembling programs were utilized to assemble the genome sequence of *Paenibacillus Shenyangensis*. The assembly data of DNA sequence were analyzed and compared with different software in their optimal parameters, and were compared with genes of the other species of the *Paenibacillus* in the NCBI database. The results showed that SOAPdenovo is the most appropriate assemble software. When k-mer is 23, the total genome length and N50 are 5 501 467 and 293 864 bp, respectively. 4 393 of the total 4 800 genes are successfully matched and annotated.

Key words: sequence assembly; genome; flocculants bacteria; high-throughput sequencing; *Paenibacillus*

第二代测序技术使得 DNA 测序进入高通量、低成本时代,直接通过聚合酶或者连接酶进行体外合成测序,一次能对几十万到几百万条 DNA 分子进行序列测序,使得对一个物种的转录组测序或基因组深度测序变得方便易行[1-3]. 第二代测序平台主要包括罗氏 454 公司的 GS-FLX 测序平台、Illumina 公司的 Solexa Genome Analyzer 测序平台和 ABI 公司的 SOLID 测序平台. 这些新的测序技术产生的数十亿短片段也给传统的从头测序序列的拼接与组装带来了巨大挑战. 序列拼接与组装任务是将测序生成的 reads 短片段拼接起来,恢复出原始序列,拼接质量直接影响到序列标注、基因预测、基因组比较等后续工作. 目前,Velvet, ABySS, SOAPdenovo, VCAKE, SPAdes 等[4-6]多种与二代测序技术相匹配的 de novo 组装工具应运而生,而如何在众多组装工具中,根据序列属性和具体要求来选择与分析组装工具的实用性,对组装最佳结果及后续信息分析尤为重要.

类芽孢杆菌属(*Paenibacillus*)是 1993 年由 Ash 等[7]将 11 个菌种从芽孢杆菌中分出来的,是

收稿日期: 2015-10-27
基金项目: 国家科技重大专项(2013ZX07202-010);国家自然科学基金资助项目(51278090).
作者简介: 付丽丽(1982-),女,辽宁沈阳人,东北大学博士研究生;胡筱敏(1958-),男,江西婺源人,东北大学教授,博士生导师.

芽孢杆菌属分类上的新发展. 类芽孢杆菌在工农业、医药、化工等领域具有重要意义,具有生物防治、产抗菌蛋白、固氮、絮凝等重要功能[8]. 目前对类芽孢杆菌的研究中以模式种多黏类芽孢杆菌(*Paenibacillus polymyxa*)居多,大多涉及多黏类芽孢杆菌的分离筛选、生长条件、代谢产物的分离鉴定及农业医学等方面的应用[9-10]. 目前,类芽孢杆菌属已有多黏类芽孢杆菌 *Paenibacillus polymyxa* E681 和 *Paenibacillus polymyxa* SC2 等 29 个菌种完成了全基因组测序工作,而对于测序后的基因工程学还有待深入研究.

Paenibacillus Shenyangensis 是本课题组从桃树栽植土壤中距表层深 10 cm 处土样中经过分离纯化得到的高效微生物絮凝剂产生菌,并鉴定为类芽孢杆菌属新种. 本课题组已完成高效产絮菌种的分离筛选、培养条件优化、结构检测和机理分析等研究. 本文在 *Paenibacillus Shenyangensis* 基因组测序数据基础上,应用目前国际上常用的 AbySS,SPAdes 和 SOAPdenovo 三种拼接组装软件对原始测序数据进行拼接比较,同时改变与优化拼接参数以获得最佳拼接效果,并于最佳拼接条件下预测基因,为后续生物信息学分析提供基础数据.

1 材料与实验方法

1.1 菌株与培养基

本实验所用菌种 *Paenibacillus Shenyangensis* 由东北大学课题组从果树种植土壤分离纯化得到[11],现保存于中国科学院微生物研究所菌种保藏中心(CGMCC2040),通过 16S rRNA 序列检测分析及理化性质检测确定为类芽孢杆菌属(*Paebubacillus sp.*)微生物的新种. 将纯化菌株接种于发酵培养基中,在 30 ℃,150 r/min 的摇床中发酵培养 36 h,其絮凝率可达到 90% 以上.

1.2 基因组测序与序列拼接

采用 Bacteria DNA Kit (OMEGA) 提取菌种基因组 DNA, TBS - 380 fluorometer (Turner BioSystems Inc., Sunnyvale, CA) 定量后取高质量 DNA (OD260/280 = 1.8 ~ 2.0, > 6 μg) 用于建库分析,并于 Illumina Hiseq 2000 测序平台进行双端测序.

用 paired - end 测序数据及软件 SolexaQA 进行低质量 reads 过滤,得到的 clean reads 用于序列拼接与组装. 应用目前国际上常用的 AbySS, SPAdes 和 SOAPdenovo 三种拼接组装程序对测序数据进行拼接比较,同时改变与优化拼接参数以获得最佳拼接结果.

1.3 基因序列比对

以 NCBI 中下载的两株同属近缘种的基因组(表 1)作为参考序列,用 Mauve 与之进行比对分析,并用 BRIG(BLAST ring image generator)可视化展示拼接基因组与类芽孢杆菌属其他两种基因序列相似性的程度.

表 1 类芽孢杆菌属参考基因组信息

Table 1 Reference genome information of *Paenibacillus sp.*

种名	GC 的质量分数/%	序列总长度 Mbp	基因数
Paenibacillus sp. JDR - 2	50.30	7.18	6 322
Paenibacillus sp. Y412MC10	51.20	7.12	6 302

1.4 基因预测与功能注释

用 Glimmer 3.0 预测基因信息并统计预测结果,将预测到的基因序列或蛋白序列分别与 NCBI - NR, Swissprot 和 KEGG 数据库进行 blast 比对,将数据库中匹配最好的基因(e - value < 1e - 5)进行功能注释.

2 结果分析

2.1 基因序列拼接优化结果

用三种拼接软件对絮凝菌测序序列进行拼接组装,通过改变频数统计关键参数 k - mer 对 scaffold 组装数量、最大 scaffold 长度、基因组总长和 N50 的影响,综合筛选出三种软件的最优拼接结果,其中 N50 和 Max scaffold 是评估组装软件优越性的重要指标. 拼接结果如表 2 所示.

表 2 三种软件最优条件下的拼接结果

Table 2 Genome assembly results of three software programs in optimal conditions

组装软件	最优 k - mer 值	组装总长度/bp	scaffold		N50/bp	错配率/%
			最大长度/bp	数量		
SOAPdenovo	23	5 501 467	337 819	64	293 864	0.77
ABySS	25	5 523 772	322 581	49	196 783	24.73
SPAdes	29	5 607 712	426 344	195	203 036	28.41

在 *k* – mer 分别为 23，25 和 29 时，三种软件的拼接结果分别达到最优. 就 scaffold 的数量而言，SOAPdenovo 和 ABySS 较低，分别是 64 个和 49 个，而 SPAdes 结果较差，有 195 个；从拼接的最大长度而言，SOAPdenovo 与 ABySS 相近，分别为 337 819 和 322 581 bp，小于 SPAdes 的 426 344 bp，但是三者组装的 scaffold 总长度相近；N50 作为拼接质量的重要评价标准，SOAPdenovo 的最长，为 293 864 bp，质量最好；就拼接序列的错配率而言，SOAPdenovo 也是最低的，每 100 kb 的长度错配率只有 0. 77%，远低于 ABySS 和 SPAdes.

2. 2　与类芽孢杆菌其他种参考基因组的比较

絮凝菌基因组序列与 NCBI 数据库中类芽孢杆菌中参考基因组 *Paenibacillus sp.* JDR – 2、参考基因组 *aenibacillus sp.* Y412MC10 比对的结果如图 1 所示. 在图 1a 中，每个刻度表示基因组上 500kb，环状结构由内向外分别为：*Paenibacillus Shenyangensis* 拼接得到的基因组 scaffolds、参考基因组 *Paenibacillus sp.* JDR – 2、参考基因组 *Paenibacillus sp.* Y412MC10、最外圈用交替的蓝色和红色作为 scaffolds 的分隔，序列比对时相似性（sequence identity）越高紫色和绿色环的颜色会越深. 絮凝菌与 *Paenibacillus sp.* Y412MC10 和 *Paenibacillus sp.* JDR – 2 分别有 833 和 643 个基因相同（图 1b），其中三菌种共有基因 456 个，在后续基因预测与注释工作中，这些共有基因为基因功能注释提供一定依据.

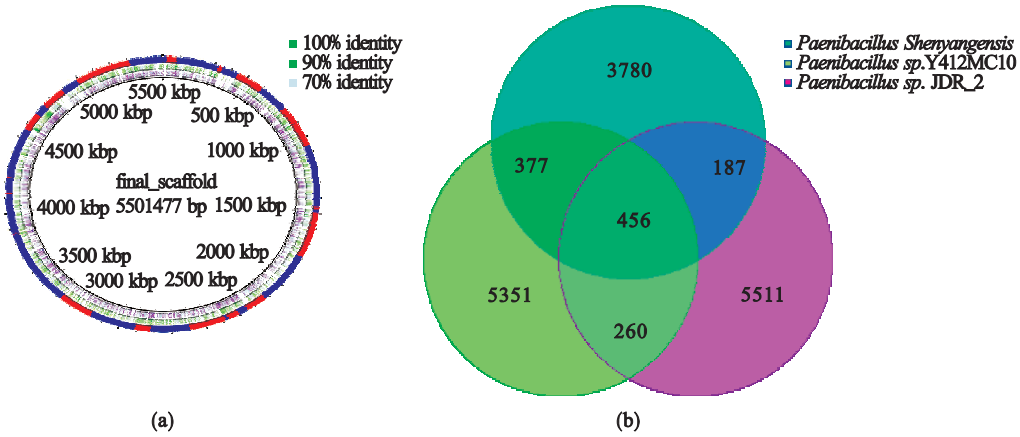

图 1　类芽孢杆菌属菌种基因组比对图

Fig. 1　Genome comparison of *Paenibacillus sp*

（a）—环形比对图；（b）—共有基因韦恩图.

2. 3　基因预测与功能注释

用 Glimmer 3. 0 预测 AbySS，SPAdes 和 SOAPdenovo 三种软件拼接序列中的基因数量并统计预测结果，如图 2 所示，不管从基因总数还是从大于 300，1 500，3 000 bp 的不同长度基因来比较，后两者之间数据相差不大，而 SOAPdenovo 的预测结果均优于另外两种，共预测成功 4 800 个基因.

将 SOAPdenovo 预测成功的 4 800 个基因分别与 NCBI – Nr（National Center for Biotechnology Information，non – redundant），KEGG（Kyoto encyclopedia of genes and genomes）数据库，SwissProt（SwissProt protein databases）蛋白数据库进行比对，分别比对成功 4 393（91. 52%），3 920（81. 67%）和 3 293（68. 60%）. 其中 NCBI – Nr 数据库比对成功率最高，比对结果统计如图 2 所示.

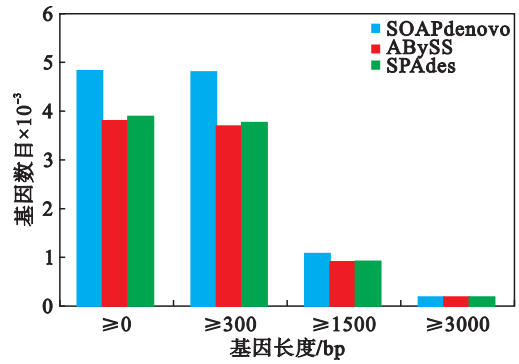

图 2　三种软件拼接的基因数量

Fig. 2　Assembled gene number of three software programs

与 NCBI – Nr 比对成功的 4 393 个基因中，有 49. 2% 的比对数据是来自类芽孢杆菌属（图 3b），包括 *paenibacillus polymyxa* SC2，*paenibacillus sp.* Y412MC10，*paenibacillus lactis* 154，*paenibacillus terrae* HPL – 003，*paenibacillus polymyxa* E681 和

paenibacillus sp. oral taxon 786 str. D14 等常见类芽孢杆菌. 比对结果的 *E* – value 值(图 3a)表示匹配假阳性的概率,该值越小表示匹配的可信度越高,预测基因与 NCBI – Nr 比对成功的 4 393 个基因中,*E* – value 值小于 e – 15 占比 85. 9%,本比对数据可信度较高.

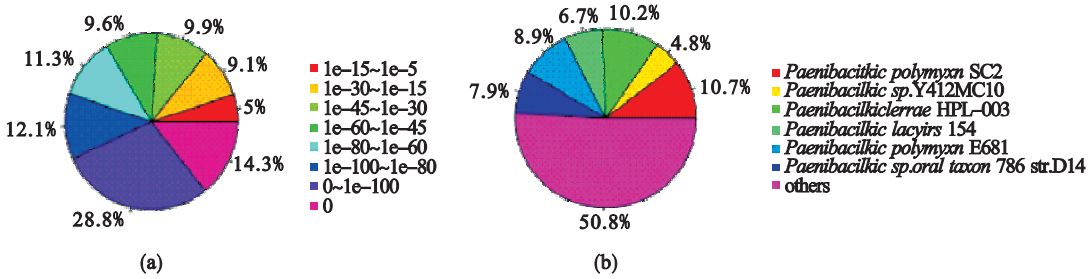

图 3 *Paenibacillus Shenyangensis* **与** NCBI – Nr **数据库比对结果统计**

Fig. 3 Results of genes mapped with *Paenibacillus Shenyangensis* and NCBI – Nr

(a) —*E* – value **值分布;**(b) **—比对种属分布**.

3 结 论

本研究通过 AbySS,SPAdes 和 SOAPdenovo 对絮凝菌 *Paenibacillus Shenyangensis* 基因组进行拼接与比较分析可知,不同的拼接软件对絮凝菌的组装与拼接结果存在一定差异,其中 SOAPdenovo 软件拼接结果最优,其 N50,Max scaffold 和错配率分别为 293 864 bp,337 819 bp 和 0. 77,更适合于本实验基因组的拼接. 预测的 4 800个基因中有 4 393 个与 NCBI – Nr 比对成功,其匹配假阳性概率 *E* – value 值小于 e – 15 占比 85. 9%,本实验数据可信度较高,也证明了 SOAPdenovo 组装的优势. 因此选择合适的组装软件和参数,充分利用测序的数据对于基因组的拼接组装是十分重要的,也为进一步的酶学及代谢工程的研究提供理论基础与基本数据.

参考文献:

[1] Scheibye A K, Hoffmann S, Frankel A, et al. Sequence assembly [J]. *Computer Biology and Chemistry*, 2009, 33 (2): 121 – 136.

[2] Stephan C S. Next-generation sequencing transforms today's biology [J]. *Nature Methods*, 2008, 5 (1): 16 – 23.

[3] Olena M, Marco A M. Applications of next-generation sequencing technologies in functional genomics [J]. *Genomics*, 2008, 92 (5): 255 – 264.

[4] Anton B, Sergey N, Dmitry A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing [J]. *Journal of Computational Biology*, 2012, 19 (5): 455 – 477.

[5] Li R Q, Zhu H M, Ruan J, et al. Denovo assembly of human genomes with massively parallel short read sequencing [J]. *Genome Research*, 2009, 20 (2): 265 – 272.

[6] Giuseppe N, Bud M. Comparing de novo genome assembly: the long and short of it [J]. *Plos One*, 2011, 6 (4): 1 – 14.

[7] Ash C, Priest F G, Collins M D. Molecular identification of rRNA group 3 bacilli using a PCR probe test [J]. *Antoni Van Leeuwenhoek*, 1993, 64 (3): 253 – 260.

[8] Naghmouchi K, Hammami R, Fliss I, et al. Colistin A and colistin B among inhibitory substances of *Paenibacillus polymyxa* JB05-01-1 [J]. *Archives of Microbiology*, 2012, 194 (5): 363 – 370.

[9] Sadhana L, Silvia T. Ecology and biotechnological potential of *Paenibacillus polymyxa*: a minireview [J]. *Indian Journal of Microbiology*, 2009, 49 (1): 2 – 10.

[10] Mokaddem H, Sadaoui Z, Boukhelata N, et, al. Removal of cadmium from aqueous solution by polysaccharide produced from *Paenibacillus polymyxa* [J]. *Journal of Hazardous Materials*, 2009, 172 (2/3): 1150 – 1155.

[11] Jiang B H, Zhao X, Liu J L, et al. *Paenibacillus shenyangensis sp. nov.*, a bioflocculant-producing species isolated from soil under a peach tree [J]. *International Journal of Systematic an Evolutionary Microbiology*, 2015, 65 (1): 220 – 224.