高通量基因组测序中结构性错误纠错研究

唐静波[1] 何献忠[1] 任力锋[1] 郭 妍[2] 黄远飞[1] 彭 健[1]

[1]（中南大学生物医学工程研究院 长沙 410008）

[2]（中国科学院深圳先进技术研究院 深圳 518055）

摘 要 伴随着高通量测序技术的发展，展开许多新的应用研究。DNA测序技术成为最重要的生物医学研究手段之一。但基因组测序中的结构性错误仍是尚待解决的难题。针对基因组组装序列的结构性错误，提出利用mate-pair的插入长度信息识别和纠错的方法。

关键词 高通量；测序；纠错

Research of Correct Structure Error of High-Throughput Genome Sequencing

TANG Jing-bo[1] HE Xian-zhong[1] REN Li-feng[1] GUO Yan[2] HUANG Yuan-fei[1] PENG Jian[1]

[1](*Institute of Biomedical Engineering of Central South University, Changsha* 410008, *China*)

[2](*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen* 518055, *China*)

Abstract The continuing improvements to high-throughput sequencing have begun to unfold a lots of new applications. DNA sequencing technology has played important roles in biological research. Correct structure error of genome sequencing remains an important problem. We develop the method to detect and correct structure error use of insert size information of mate-pair.

Keywords high-throughput; sequencing; error correction

1 引 言

在过去的三十年中，DNA测序技术成为最重要的生物医学研究手段之一，伴随着高通量测序技术的迅速发展，其数据产出能力呈指数增长[1,2]。对于每个生物体来说，基因组包含了整个生物体的遗传信息。测序技术能够读取基因组DNA上的遗传信息，进而比较全面地揭示基因组的复杂性和多样性，为生物学研究提供原始数据支持[3-5]。但由于基因组中的重复序列（duplications），会导致基因组序列组装中产生结构性错误，从而严重影响后期的的生物学研究[6-8]。本文就针对基因组组装序列的结构性错误，提出识别和纠错的方法。

2 基因组denovo组装

目前，高通量基因组测序序列组装算法主要有两类：基于Hamilton路径的组装算法和基于Euler路径的组装算法[9,10]。

基于Hamilton路径的组装算法将所有待拼接的DNA片段构成一个有向图，每个片段看成一个结点[11-13]。首先“overlap”通过比对，寻找片段问的重叠信息；然后“layout”，根据重叠信息，排列所有片段；最后“consensus”，确定最终序列。统称为OLC（Overlap Layout Consensus）[14]。

基于Euler路径的组装算法，根据DeBruijn图，把DNA序列拼接问题转化为Euler超路径问题[15-19]。首先将所有待拼接的DNA片段切割，然后根据这些切割

作者简介：唐静波，博士生，研究方向为生物信息学，Email:tangjingbo2000@126.com；何献忠，研究生，研究方向为生物医学工程；任力锋，教授，研究方向为生物医学工程；郭妍，助理研究员，研究方向为生物医学工程；黄远飞，研究生，研究方向为生物信息医学应用；彭健，副教授，研究方向为生物医学工程。

片段的重叠信息，叠加为一个重叠群（contig）[20]，最后根据这些contig之间的连接关系(如mate-pair)，把这些contig构建成scaffold。

3 Duplications

Duplications指生物基因组中的重复片段，它的类型有两种：简单重复，具有单一结构的重复片段；复杂重复，它除了本身是一个重复片段之外，还包括重复次数不同的子重复片段[21,22]。Duplications会导致序列组装中产生结构性错误[23]。

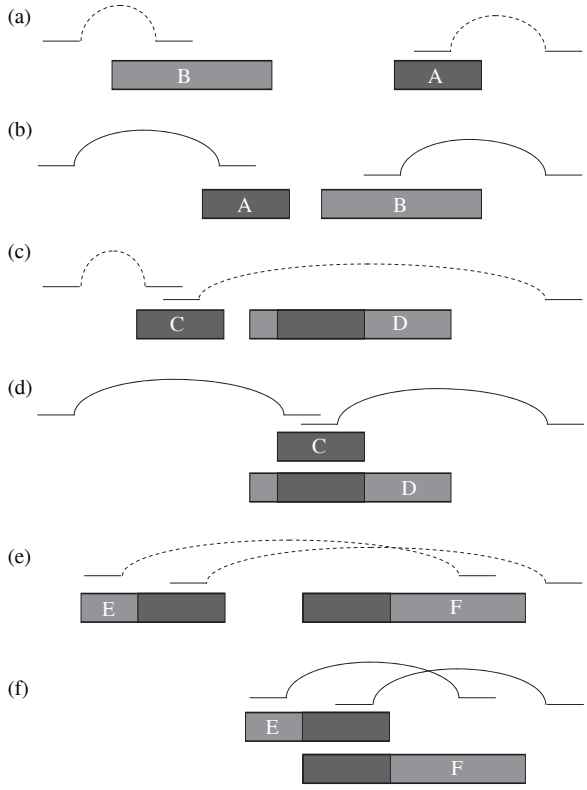

图1 结构性错误示意图

如图1所示：(a),(c),(e)为错误排列位置的DNA片段，(b),(d),(f)为它们正常的排列位置。图中连接的虚线为错误的mate-pair连接，实线为正常的mate-pair连接。

4 结构性错误的识别和纠错

高通量基因组测序技术能产生mate-pair reads，一对mate-pair reads能产生一个forward read 和一个reverse read，它们之间的距离称为插入长度（insert size），同一次建库产生的DNA片段的insert size 应该是大致相等的。如果scaffold中出现大量的冲突序列（insert size差异性很大，如图1中 (a),(c),(e)），就可以识别出此处构建scaffold的contig连接存在结构性错误，打断并标记这些错误的contig连接，重新构建scaffold，就可以实现纠错的效果。

具体步骤为：

（1）把所有测序reads按照kmer大小逐个碱基滑动切割，得到固定长度的短串序列。并添加滑动中相邻短串左右的连接关系及连接数量。例如基因组序列为AGATCTCTTTATTAGATCTCTTATTAGGA，使用kmer大小为5bp，第一次滑动得到kmer1‘AGATC’，它存在正向的边指向下一个kmer2 ‘GATCT’，继续滑动到‘GATCT’，该kmer的正向边指向了‘ATCTC’，反复滑动到最后一个kmer‘TAGGA’，它不存在正向延伸的边。

（2）我们按这种方法滑动基因组正链和互补序列。如果在一个滑动中得到的kmer在之前的滑动中存在，那么它所延伸的边就合并在已存在的顶点上。最终得到了DeBruijn图。

（3）把DeBruijn图中的contig，按照连接关系支持和pair end 信息构建成scaffold。此时的scaffold是初次组装的基因组序列。

（4）把reads通过比对软件（如BWA）maping到scaffold上，根据mate-pair 的insert size信息，分析insert size的冲突情况。

（5）根据insert size的冲突判定，打断并标记这些错误的contig连接。

（6）再次把DeBruijn图中的contig，按照新的连接关系，构建成scaffold。

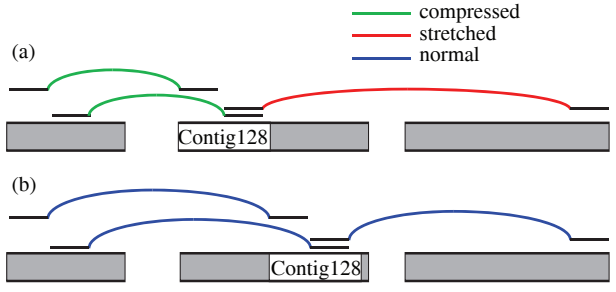

图2 结构性错误识别纠错示意图

如图2所示：Contig128 处于基因组组装序列(a)位置时，reads 的insert size有明显冲突，绿色表示压缩（compressed），红色表示拉伸（stretched），蓝色表示正常（normal）。打断contig连接，重新构建scaffold后，insert size全部正常。表明此处纠错有明显效果。

表1　E. coli 基因组组装纠错前后对比

	Genome coverage	Correct number	Correct length
纠错前	4031168(95.6%)	69(77.5%)	4021738(96.6%)
纠错后	4125027(97.9%)	134(87.0%)	4116896(98.9%)

5　评价结果

对大肠杆菌（E.coli，基因组4639675bp）基因组测序数据，使用基因组组装软件进行组装，然后利用mate-pair的插入长度信息对基因组组装序列进行识别和纠错。

如表1所示，纠错后基因组序列的基因组覆盖度（Genome coverage）、数量正确率（Correct number）、长度准确率（Correct length）都有提高，说明纠错效果明显。

6　结论和展望

Duplications的处理方法决定着基因组测序组装结果的精度，而组装结果的优劣又直接影响着序列的生物学分析，所以Duplications是DNA序列研究不可忽视的重要方面[24]。利用mate-pair 的插入长度信息对基因组组装序列的结构性错误进行识别和纠错，可以极大提高基因组测序组装的精度。但仍然无法完全纠正所有Duplications引起的组装错误。随着测序技术的发展，reads越来越长，精度越来越高，希望能寻找到新的组装纠错算法更好的解决这个问题。

参 考 文 献

[1] Yu N, Jensen-Seaman M I, Chemnick L, et al. Low nucleotide diversity in chimpanzees and bonobos [J]. Genetics, 2003, 164:1511-1518.

[2] Zerbino D R, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs [J]. Genome Research, 2008, 18:821-829.

[3] Cheung J, Estivill X, Khaja R, et al. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence [J]. Genome Biology, 2003, 4:336-342.

[4] Nicholas T J, Cheng Z, Ventura M, et al. The genomic architecture of segmental duplications and associated copy number variants in dogs [J]. Genome Research, 2009, 19:491-499.

[5] Butler J, MacCallum I, Kleber M, et al. De novo assembly of whole-genome shotgun microreads [J]. Genome Research, 2008, 18:810-820.

[6] Salzberg S L, Yorke J A. Beware of mis-assembled genomes [J]. Bioinformatics, 2005, 21:4320-4321.

[7] Fleischmann R D, Adams M D, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd [J]. Science, 1995, 269:496-512.

[8] Kim J H, Waterman M S, Li L M. Diploid genome reconstruction of Ciona intestinalis and comparative analysis with Ciona savignyi [J]. Genome Research, 2007, 17:1101-1110.

[9] Barriere A, Yang S P, Pekarek E, et al. Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes [J]. Genome Research, 2009, 19:470-480.

[10] Holt R A, Subramanian G M, Halpern A, et al. The genome sequence of the malaria mosquito Anopheles gambiae [J]. Science, 2002, 298:129-149.

[11] Jones T, Federspiel N A, Chibana H, et al. The diploid genome sequence of Candida albicans [J]. Proceedings of the National Academy of Sciences, 2004, 101:7329-7334.

[12] Vinson J P, Jaffe D B, O'Neill K, et al. Assembly of polymorphic genomes: algorithms and application to Ciona savignyi [J]. Genome Research, 2005, 15:1127-1135.

[13] Bailey J A, Church D M, Ventura M, et al.Analysis of segmental duplications and genome assembly in the mouse [J]. Genome Research, 2004, 14:789-801.

[14] Sharp A J, Locke D P, McGrath S D, et al. Segmental duplications and copy-number variation in the human genome [J]. The American Journal of Human Genetics, 2005, 77:78-88.

[15] Teichmann S A, Babu M M. Gene regulatory network growth by duplication [J]. Nature Genetics, 2004, 36:492-496.

[16] Varki A, Altheide T K. Comparing the human and chimpanzee genomes: searching for needles in a haystack [J]. Genome Research, 2005, 15:1746-1758.

[17] Conrad B, Antonarakis S E. Gene duplication: a drive for phenotypic diversity and cause of human disease [J]. Annual Review of Genomics and Human Genetics, 2007, 8:17-35.

[18] De Gobbi M, Viprakasit V, Hughes J R, et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter [J]. Science, 2006, 312:1215-1217.

[19] Phillippy A M, Schatz M C, Pop M. Genome assembly forensics: finding the elusive mis-assembly [J]. Genome Biology, 2008, 9:R55.

[20] Choi J H, Kim S, Tang H, et al. A machine-learning approach to combined evidence validation of genome assemblies [J]. Bioinformatics, 2008, 24:744-750.

[21] Zimin A V, Smith D R, Sutton G, et al. Assembly reconciliation [J]. Bioinformatics, 2008, 24:42-45.

[22] Zimin A V, Delcher A L, Florea L, et al. A whole-genome assembly of the domestic cow, Bos Taurus [J]. Genome Biology, 2009, 10:R42.

[23] The Chimpanzee Sequencing and Analysis Consortium: Initial sequence of the chimpanzee genome and comparison with the human genome [J]. Nature, 2005, 437:69-87.

[24] Elsik C G, Tellam R L, Worley K C, et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution [J]. Science, 2009, 324:522-528.