第 33 卷第 5 期　　河 北 科 技 大 学 学 报　　Vol. 33, No. 5
2012 年 10 月　　Journal of Hebei University of Science and Technology　　Oct. 2012

文章编号：1008-1542(2012)05-0384-02

基因组装配中存在重复序列叠加时重叠群计数的推广的 Lander-Waterman 定理

黄海云[1]，张　屹[2]

（1. 河北科技大学图书馆，河北石家庄　050018；2. 河北科技大学理学院，河北石家庄　050018）

摘　要：针对基因组组装算法理论进行了改进，该研究对于经典的 Lander-Waterman 定理在repeat collapse 存在的情况下进行了推广，对于判断基因组组装的 contig 的个数是否合理，组装质量是否可靠有重要的参考价值。

关键词：Lander-Waterman 公式；重复序列；基因组组装；重叠群

中图分类号：O29　　**MSC(2010)主题分类号**：60A10　　**文献标志码**：A

Generalized Lander-Waterman theorem under repeat collapse in genome assembly

HUANG Hai-yun[1], ZHANG Yi[2]

(1. Library, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China; 2. College of Sciences, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China)

Abstract: We improved the theory of algorithm of genome assembly which is a generalized Lander-waterman theorem under repeat collapse. It is important for appraising the rationality of the number of contig and the quality of genome assembly.

Key words: Lander-Waterman formula; repeat sequence; genome assembly; contig

第二代测序技术为人们提供了基因组测序的新思路。从数学上说，就是通过把基因组打成多重的小片段，然后用算法把它们组装在一起来得到全基因组序列。但是由于重复序列和杂合现象的存在，使得生成的 de-Brujin 图巨大复杂，难以压成一条序列[1-2]。更严重的是，重复序列会使 contig 的数量虚高，以至于无法估计真正的 contig 数量是多少[3-4]。

1　存在 repeat 叠加时重叠群个数的数学期望

基于 2000 年 Science 的刊文[3]中关于 A-statistics 的论述，笔者假设建立的某个生物的全基因组测序片段库中共有 F 个片段需要组装，基因组的大小已被事先用 k-mer 方法估计为 G 个 AGCT 字符。通过组装，得到若干个无法再通过重叠操作来加长的大片段，这些大片段叫做重叠群(contig)。设一个由 k 个片段组成的 contig 长为 r 个 AGCT 字符，即从这个 contig 的第一个组装片段的第一个字符到最后一个组装片段的最后一个字符之间的距离是 r 个字符。假设这个 contig 没有被重复取样(可按 blast 去冗余来保证这一点)，

收稿日期：2012-09-02；责任编辑：张士莹
基金项目：国家自然科学基金资助项目(11171088)；河北省自然科学基金资助项目(A2011208002)
作者简介：黄海云(1969-)，女，内蒙古通辽人，馆员，主要从事生物信息学方面的研究。
通讯作者：张　屹副教授。E-mail：zhaqi1972@163.com

则按照概率论的知识，在长为 r 的序列中发现 $k-1$ 个片段起点的概率为 $[(rF/G)^k/k!]e^{-rF/G}$。

但是，如图 1 所示，当 2 个片段 R1，R2 为重复片段时，这 2 个重复片段将被所有的算法（soapdenovo 算法也一样）当成同一段序列的不同拷贝而被以高分组装成一个片段，而他们之间原来的片段会因为 blast 分值较低被挤走，成为单独的一个 contig。这样，由于 repeat 的存在会使得组装之后的 contig 的个数与原来公式估计的不一样了。依据文献[3]中的结果，如果某个 contig 是 2 个 repeat 叠加的结果，则在长为 r 的序列中发现 $k-1$ 个片段起点的概率为 $[(2rF/G)^k/k!]e^{-2rF/G}$。按这样计算，如果这个 contig 是 x 个 repeat 片段叠加的结果，则这个概率应该是 $[(xrF/G)^k/k!]e^{-xrF/G}$。同时，每次 repeat 的叠加都可以挤出一个 contig[4]，则 x 个不可区分的重复片段将产生 $x-1$ 个多余的 contig。笔者所在研究组的飞蝗基因组的重复片段占整个基因组的 1/2 以上，repeat 对于 contig 计数的影响是巨大的和不可忽视的。

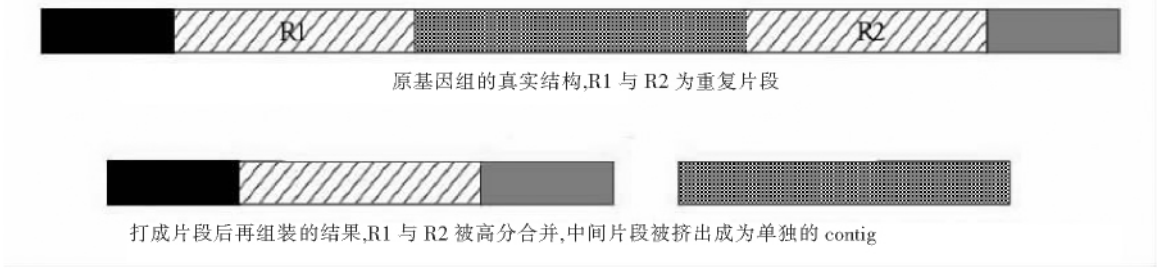

图 1　在组装时，重复片段 R1 与 R2 的叠加会引起 contig 个数的增加

Fig. 1　Repeat collapse of R1 and R2 increases a contig in assembly

显然，有 x 个 repeat 片段进行叠加的情况下，组装结果中 contig 增加的个数的数学期望为 $[(xrF/G)^k/k!]e^{-xrF/G}\times(x-1)$。基于 A-Brujin 图的组装算法中对 repeat 的分类原则[5]，总可以把一个基因组中全部的 repeat 分成共 z 个类，每个类中含有的 repeat 片段不可辨别，因为这个相似性的阈值是可控的。这 z 个 repeat 类分别含有 $w_1, w_2, \cdots, w_z$ 个 repeat。因此，打断之后再组装，由于 repeat 的叠加而生成的多出的 contig 的个数的数学期望是 $\sum_{i=1}^{z}[(w_irF/G)^k/k!]e^{-w_irF/G}(w_i-1)$。利用这个结果来推广经典的 Lander-Waterman 公式。

2　Lander-Waterman 的原定理

1988 年，LANDER 和 WATERMAN 给出了基因组组装时 contig 的分布定理[6]。他们假设 2 个片段之间至少要有全长的 θ 比例的片段重叠才能连接在一起，而且这个标准要足够严格以保证较小的假阳性出现。另外，假设基因组被打断后形成的片段集是完备的，覆盖整个基因组的。定理中所用的变量如下：

G 为基因组的总长字符数；

L 为片段平均长度；

N 为全部克隆片段的个数，对应着 A-statistics[3] 中的 F；

$\alpha=N/G$ 为一个碱基是一个片段的起始点的概率，对应着 A-statistics[3] 中的 F/G；

T 为 2 个片段被确定相连接并且可以组装在一起时，需要的最小重叠长度；

$\theta=T/L$；

$c=LN/G$ 是覆盖的冗余度，对应于 A-statistics 中的 LF/G；

r 为 contig 的长度的期望。

Lander-Waterman 的原定理 1：contig 个数的数学期望值为 $Ne^{-c(1-\theta)}$。

3　笔者推广的 Lander-Waterman 定理 1

设有 z 个 repeat 类，contig 的个数的数学期望为

$$Fe^{-\frac{LF}{G}(1-\theta)}+\sum_{i=1}^{z}[(w_irF/G)^k/k!]e^{-w_irF/G}(w_i-1)。\tag{1}$$

（下转第 463 页）

确定最优培养条件为 $A_1B_2C_2D_2$，即培养时间为 28 h，初始 pH 值为 7.3，装液量为 120 mL，接种量为 1.5%(体积分数)。根据表中极差值大小，排列出影响试验的因素的主次顺序：C(装液量)＞ D(接种量)＞B(初始 pH 值)＞A(培养时间)。

正交试验的方差分析如表 3 所示。

表 3　正交试验方差分析

Tab. 3　Variance analysis of orthogonal test

因素	平方和	自由度	方差	F 临界值
A	0.107	2	0.100	4.460
B	0.247	2	0.231	
C	3.687	2	3.445	
D	0.240	2	0.224	
误差	4.28	8		

通过表 3 的方差分析再次证明，装液量对结果影响最大，其他影响因素与之相比影响较小。

综上所述，确定了门多萨假单胞菌 DS04-T 菌株对 Poly(3HB-co-4HB)的最佳产酶条件：培养温度为 30 ℃，培养时间为 28 h，培养基初始 pH 值为 7.3，摇床转速为 150 r/min，培养基装液量为 120 mL(250 mL 三角瓶)，接种量为 1.5%(体积分数)。在此优化条件下菌株产生的 Poly(3HB-co-4HB)降解酶活力可达(26.2±0.7)U·mL^{-1}。

参考文献：

[1] PAPANEOPHYTOU C P, VELALI E E, PANTAZAKI A A. Purification and characterization of an extracellular medium-chain length polyhydroxyalkanoate depolymerase from Thermus thermophilus HB8 [J]. Polym Degrad and Stabil, 2011, 96(4): 670-671.

[2] LUCAS N, BIENAIME C, BELLOY C, et al. Polymer biodegradation: Mechanisms and estimation techniques [J]. Chemosphere, 2008, 73(4): 429-442.

[3] WEN Xing, LU Xiu-ping. Microbial degradation of poly(3-hydroxybutyrate-co-4-hydroxybutyrate) in soil[J]. J Polym Environ, 2012, 20(2):381-387.

[4] CHANPRATEEP S. Current trends in biodegradable poly-hydroxyalkanoates[J]. J Biosci Bioeng, 2010, 110(6): 621-632.

[5] 刘宝友，罗湘南，邢质贤，等. 微生物降解含煤油废水的研究[J]. 河北科技大学学报(Journal of Hebei University of Science and Technology), 2007, 28(4): 314-316.

[6] 王瑜瑜，张春晓，郭建博，等. 耐盐基因工程菌降解偶氮染料特性研究[J]. 河北科技大学学报(Journal of Hebei University of Science and Technology), 2010, 31(5): 483-486.

[7] WANG Zhan-yong, GAO Jia, LI Lin-lin, et al. Purification and characterization of an extracellular poly(3-hydroxybutyrate-co-3-hydroxyvalerate) depalymerase from *Acidouorax sp*. HB01[J]. World J Mierob Biot, 2012, 28(6): 2 395-2 402.

[8] 刘天识. 可控降解 PHB 为单体及其发酵条件的探索研究[D]. 长春：东北师范大学, 2007.

(上接第 385 页)

实际上，A-statistic[3] 中的 r 是与 Lander-Waterman 定理[6] 中的 T，L 和 θ 相关的，$r=(k-1)(1-\theta)L$，因此 contig 的个数也可以写成

$$Fe^{-\frac{LF}{G}(1-\theta)}+\sum_{i=1}^{z}[(w_i(k-1)(1-\theta)LF/G)^k/k!]e^{-w_i(k-1)(1-\theta)LF/G}(w_i-1)。\quad(2)$$

在 Lander-Waterman 定理[6] 中有几个公式，其他的几个公式都可以据此推广到有重复片段叠加的一般情况，由于篇幅限制，本文只给出第一个公式的推广。

4　在基因组中的应用说明

基于给出的式(1)和式(2)，可以计算出正确的 contig 的个数，可与实际组装生成的 contig 的个数相比较，来评价组装的质量以及受 repeat 叠加影响的严重程度。

参考文献：

[1] LI R, FAN W, TIAN G, et al. The sequence and de novo assembly of the giant panda genome[J]. Nature, 2010, 463: 311-317.

[2] PEVZNER P A, TANG H, WATERMAN M S. An Eulerian path approach to DNA fragment assembly[J]. Proc Natl Acad Sci, 2001, 98: 9 748-9 753.

[3] EUGENE W, MYER S. A whole-genome assembly of drosophila[J]. Science, 2000, 287: 2 196.

[4] STEVEN L, SALZBERG J A. Beware of mis-assembled genomes[J]. Bioinformatics, 2005, 21: 4 320-4 321.

[5] PAUL A P, HAIXU T, GLENN T. De novo repeat classification and fragment assembly[J]. Genome Research, 2004, 14: 1 786-1 796.

[6] ERIC L, MICHAEL S W. Genomic mapping by fingerprinting random clones: A mathematical analysis[J]. Genomics, 1988, 2: 231-239.