第 35 卷 第 4 期 计 算 机 工 程 2009 年 2 月
Vol.35 No.4 Computer Engineering February 2009
·软件技术与数据库· 文章编号：1000—3428(2009)04—0088—03 文献标识码：A 中图分类号：N945

DNA 片段拼接中的预归并重复序列屏蔽方法

蔡 葵，杨进才
(华中师范大学计算机科学系，武汉 430079)

摘 要：针对 DNA 片段拼接中的重复序列识别及屏蔽问题，提出一种预归并重复序列屏蔽方法。在片段拼接前通过扫描子串标识出可能存在重叠关系的 shotgun 片段，利用子串归并该相关片段，标识出重复序列的位置信息，达到屏蔽的目的。计算机模拟分析表明，该方法识别重复序列的错误率低，通过预归并有效缩减了 shotgun 集合的规模，降低了拼接时的计算复杂度。

关键词：片段拼接；预归并；重复序列；屏蔽

Pre-merged Repeats Masking-off Method in DNA Fragment Assembly

CAI Kui, YANG Jin-cai
(Department of Computer Science, Huazhong Normal University, Wuhan 430079)

【Abstract】This paper proposes a pre-merged repeats masking-off method by studying repeats analysis in DNA fragment assembly. The method can recognize and merge the different shotgun fragments owning the same overlap substring by scanning the shotgun set, and mark the position of the repeats and masking-off them before DNA fragment assembly. Simulations show that the rate of false repeats recognition with the method is descended, and CPU time of DNA fragment assembly is reduced because of pre-merged method.

【Key words】 fragment assembly; pre-merged; repeats; masking-off

1 概述

将 shotgun 集合中的 DNA 片断拼接成目标序列的主要难点在于大量重复序列(repeats)的存在会产生许多错误的重叠，从而导致结果的严重偏差[1]。所以，在拼接前屏蔽重复序列对于提高序列拼接的精确度和减少错误率非常重要[2]。通过研究 DNA 目标序列和 shotgun 集合，RePS 重复序列屏蔽方法[3]首先提出定长子串的概念，用定长为 20 的序列子串在 shotgun 片段集合中出现的次数来确定此子串是否为 repeats 的一部分。文献[4-6]都以此为基础，把定长 20 换为 k (k 由 shotgun 集合中片段数目 n 确定)，提出了各自的重复序列屏蔽方法。

以上方法都需扫描 shotgun 集合中的所有片段，但在识别重复序列之余，扫描得到的信息并没有得到充分利用。本文方法进一步利用了这些信息，根据具有相同特征子串的 shotgun 片段来自目标序列同一个位置或者相同 repeats 的原理，将这些 shotgun 片段进行预归并操作，以减少 shotgun 集合中的片段数目并且精确标识出原目标序列中的重复序列，从而降低之后拼接时的计算复杂度。

2 片段预归并重复序列屏蔽方法

DNA 目标序列上每个长为 k 的字符串称为 k-mer 子串[4]，若它也是某个 shotgun 片段的一部分，则称为此片段的 k-mer 子串；当它作为某个片段的标识性信息时，就称为该片段的特征子串[5]。选择恰当的 k 值，使得除了 repeats 中包含的 k-mer 子串之外，目标序列中的 k-mer 子串互异；设 n 为 shotgun 集合中的片段总数，经过数学推导得到 k 与 n 的关系为[5]：$k \geq \log_4 n$。

k-mer 子串和特征子串如图 1 所示。

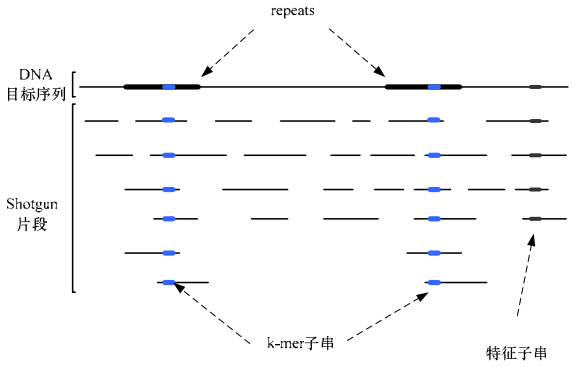

图 1 k-mer 子串和特征子串示例

2.1 基本思想

根据 k-mer 子串间的互异性，如果 shotgun 片段集合是由目标序列的 C 条克隆随机打散而成，那么目标序列中不属于 repeats 的 k-mer 子串在 shotgun 集合片段中最多出现 C 次；属于重复 R 次的 repeats 中的 k-mer 子串在 shotgun 集合片段中最多出现 $C \times R$ 次。设 T 为判定某一 k-mer 子串是否属于 repeats 的阈值，可以取 $T=C$。在 shotgun 集合片段中出现次数小于或等于 T 的 k-mer 子串可以当成特征子串；出现次数大于 T 的，可以看作原目标序列中 repeats 的一部分。

方法主要步骤如下：

(1)对 shotgun 集合中所有片段编号(1~n)，根据 n 的数目确定一个 $k(k \geq \log_4 n)$；初始 k-mer 子串统计表 S 为空，

作者简介：蔡 葵(1977－)，男，硕士研究生，主研方向：生物信息计算；杨进才，副教授、博士

收稿日期：2008-04-20 **E-mail：**kuicai@tom.com

S={k-mer 子串, 出现次数, 对应片段位置}。

(2)按编号顺序扫描所有片段，对每个片段按由首到尾顺序统计其所有 k-mer 子串，同时可以计量该片段长度并存于 L_i 中(i 为片段编号)：

1)如果某个 k-mer 子串在 S 中不存在，则将该“ k-mer 子串 ”按字符顺序插入 S 中，同时将其“ 出现次数 ”置为 1，在“ 对应片段位置 ”中记录当前 shotgun 片段编号及 k-mer 子串在该片段中的起始位置，如(1, 7)表示 k-mer 子串起始于 1 号 shotgun 片段的第 7 个字符。

2)如果此 k-mer 子串在 S 中已存在，则直接将其“ 出现次数 ”值加 1，同时将当前 shotgun 片段编号及 k-mer 子串在该片段中的起始位置记录添加到“ 对应片段位置 ”中。

扫描完成后，删除表 S 中“ 出现次数 ”仍为 1 的 k-mer 子串。一个 DNA 目标序列及其 shotgun 集合如图 2 所示，其对应的 k-mer 子串统计表 S 如表 1 所示。

图 2　DNA 目标序列及其 shotgun 集合示例

表 1　k-mer 子串统计表 S 示例

k-mer 子串	出现次数	对应片段位置
GTAAT	6	(1,7);(4,3);(7,6);(3,2);(6,2);(9,7)
CGTAT	3	(2,4);(5,2);(8,3)

3)根据表 S 对 shotgun 集合中的片段进行预归并和重复序列识别操作。

2.2　特征子串的预归并

(1)顺序扫描表 S，其中，“ 出现次数小于等于 T ”的特征子串的“ 对应片段位置 ”中记录了含有该特征子串的所有 shotgun 片段编号及特征子串起始位置，记录这些信息到 R 中。找到其中起始位置值最大的 shotgun 片段(m, t)，它是特征子串左邻最长的片段，左邻长为 t-1；同时找到特征子串右邻最长的片段(f, e)，右邻长为 L_f–e–k+1。

(2)根据特征子串的定义，直接拼接“ m 的左邻+特征子串+f 的右邻 ”，用得到的大片段替换原来的 m 片段(编号仍为 m)。更新 L_m 为此大片段的长度 t+L_f–e，同时在 shotgun 集合中删掉 R 中记录的除 m 之外的 shotgun 片段及它们在 L 中的长度数据，在表 S 中删除该特征子串记录。

(3)顺序扫描表 S 中所有的“ 对应片段位置 ”项目，如果片段编号在 R 中(设在 R 中为(g, j)，当前为(g, u))，就更新其编号为 m，其对应的 k-mer 子串在 m 中起始位置为 t–j+u，即(m, t–j+u)。每个 k-mer 子串“ 对应片段位置 ”中只需更新第 1 个编号在 R 中的片段，其他编号在 R 中的片段可直接删除，每删除一个，就将该 k-mer 子串的“ 出现次数 ”减 1，如果“ 出现次数 ”值最后等于 1，则在表 S 中删除该 k-mer 子串记录。

(4)遍历表 S，对其中所有特征子串都做如上处理后，表 S 中剩余的记录实际上都是属于 repeats 的 k-mer 子串，这些 k-mer 子串“ 对应片段位置 ”经过上述的(编号, 起始位置)更新之后，相当于进行了初步预归并处理，为后面的重复序列识别打下了基础。

2.3　重复序列的预归并与识别

顺序扫描表 S，每个 k-mer 子串的“ 对应片段位置 ”中记录了含有该 k-mer 子串的所有 shotgun 片段编号及 k-mer 子串起始位置，分别记录这些信息到 A 和 B 中。这些片段源于 DNA 目标序列中不同位置的相同 repeats，它们的预归并分向左和向右两部分进行。

(1)向左

置 j 的初值为 1，找到 A 中 k-mer 子串左邻最长的 shotgun 片段(m, t)，将其记录到 A_j 中，同时在 A 中删除，然后把它与 A 中第 1 个片段，设为(f, e)，由起始位置开始向左同步对比：1)如果一直到 f 片段的最左端都匹配，那么把 f 也记录到 A_j 中，同时在 A 中删除。2)如果还没有到达 f 片段的最左端就发现不匹配，则将 j 值加 1；如果 j>2，还得将 f 片段与 A_2~A_{j-1} 中的第 1 个片段做同样的比对操作，当与其中 A_p(1<p<j)的第 1 个片段直到最左端都匹配时，把 f 记录到 A_p 中，否则，把 f 记录到 A_j 中，并在 A 中删除。

继续取 A 中下一个片段，仍设为(f, e)，重复上述操作，直到 A 为空。

(2)向右

置 j 初值为 1，找到 B 中 k-mer 子串的右邻最长的 shotgun 片段(m, t)，将其记录到 B_j 中并同时在 B 中删除，然后把它与 B 中第 1 个片段，设为(f, e)，由结束位置开始向右同步比对：1)如果一直到 f 片段的最右端都匹配，那么把 f 也记录到 B_j 中，同时在 B 中删除。2)如果还没有到达 f 片段的最右端就发现不匹配，则将 j 值加 1；如果 j>2，还得将 f 片段与 B_2~B_{j-1} 中的第 1 个片段做同样的比对操作，当与其中 B_p(1<p<j)的第 1 个片段直到最右端都匹配时，把 f 记录到 B_p 中，否则，把 f 记录到 B_j 中，并在 B 中删除。

继续取 B 中下一个片段，仍设为(f, e)，重复上述行为，直到 B 为空。

完成后得到 q 个片段集 A_1~A_q 和 q 个片段集 B_1~B_q，原 DNA 序列中就有 q 个包含此 k-mer 子串相同的 repeats。对每个 A_i 和 B_i(1≤i≤q)以当前 k-mer 子串为基础分别向左归并(右侧无法归并的部分舍弃)、向右归并(左侧无法归并的部分舍弃)其中所有的片段。对临近的具有相同重合 shotgun 片段的 k-mer 子串的实时更新可以参照 2.2 节特征子串的处理方法施行。最后得到 q 对片段 a_1,b_1~a_q,b_q，表明该 repeats 在原 DNA 目标序列中出现 q 次，具体 repeats 的构成可以由任一对 a_i 和 b_i 归并得到，在后面的拼接时屏蔽此 repeats 即可。

2.4　本方法的计算复杂性

本方法的主要过程是对 shotgun 集合中的所有片段进行扫描和查找，以及对表 S 中各条记录的归并操作和实时更新。可以估算扫描 shotgun 集得到表 S 的复杂性为 $O(l{\cdot}n{\cdot}k)$，其中，n 为片段总数目；l 为片段平均长度；k 为子串长度。对表 S 中各条记录的归并及更新操作的复杂性都在 $O(l{\cdot}n{\cdot}4^k)$之内，表 S 的记录数最多为 4^k，所以，归并复杂性为 $O(l{\cdot}n{\cdot}4^{2k})$。两者综合得本方法的复杂性为 $O(l{\cdot}n{\cdot}4^{2k})$，比文献[4-6]方法的复杂性 $O(l{\cdot}n{\cdot}4^k)$稍大，但 l 和 k 均有上界，两者都可简化为 $O(n)$。另外，本方法的预归并处理还可以大大减少 shotgun 集合中

的片段数目 n，减小其后的拼接处理复杂性，计算机模拟分析也验证了这一点。

3 计算机模拟分析

从 GenBank 数据库[7]取 5 条 DNA 序列，编号和长度见表 2。随机分解每条序列的 C=3 个克隆构成 shotgun 片段集合，片段数目为 n，子串定长 k=[$\log_4 n$]+2，T=C=3。本文方法命名为 PreMerged，所有实验均在 Intel Core2 E4300 (1.8 GHz)、1 GB 内存(667 MHz)的联想 PC 上完成。

表 2 RECON 和 PreMerged 的识别结果比较

序列编号	序列长度/bp	克隆条	分解片段数目	定长	识别 repeats 个数		识别率/(%)	
					RECON	PreMerged	RECON	PreMerged
GI:3492857	157 653	3	1 633	7	33	159	5.6	26.9
GI:160914588	365 126	3	3 987	7	68	352	6.3	32.6
GI:154147533	593 963	3	5 869	8	89	605	6.9	46.9
GI:160904213	767 118	3	8 698	8	101	787	7.1	55.3
GI:12539724	988 176	3	11 396	8	150	1 253	7.3	60.2

用 PreMerged 与目前流行的重复段处理方法 RECON 分别处理这 5 个 shotgun 片段集合，得到的结果比较如表 2 所示，两者运行时间比较如图 3 所示。

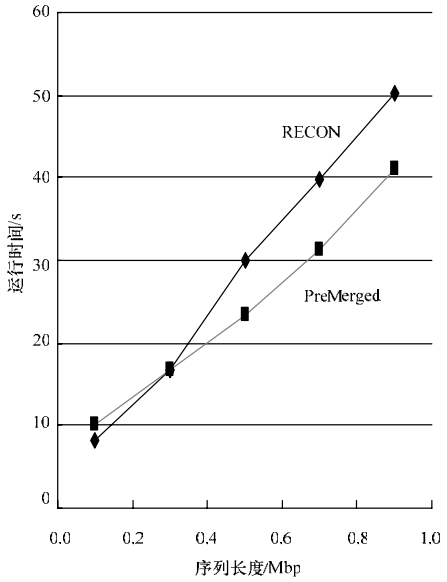

图 3 RECON 和 PreMerged 的运行时间比较

从表 2 可以看出， PreMerged 识别重复段的能力明显强于 RECON。从图 3 可知，PreMerged 对较长序列(长于 0.4 Mbp)生成的较大规模 shotgun 集合的计算速度也快于 RECON。对较小规模 shotgun 集合，本文方法不优是在 n 不大的情况下由 4^{2k} 因子的比重偏大造成的。另外，序列长度与运行时间之间呈现大致的线性关系，这些都验证了方法计算复杂性分析的结果。

这 5 个 shotgun 片段集合分别用 RECON 和 PreMerged 做预处理后，再用当前流行的 Phrap 工具拼接，运行时间比较如图 4 所示。

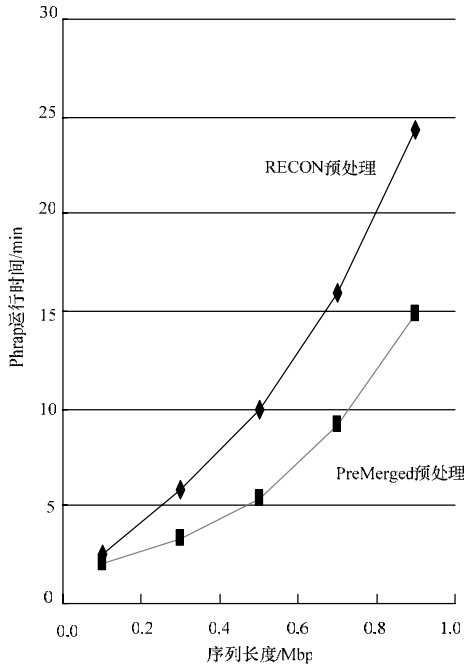

图 4 RECON 和 PreMerged 预处理后的 Phrap 拼接时间比较

可以看出，经过 PreMerged 预归并后再用 Phrap 工具拼接的速度明显较快，也验证了方法计算复杂性分析的结果。

4 结束语

本文的预归并重复序列屏蔽方法充分利用了扫描 shotgun 集合片段得到的信息，根据子串的唯一性进行片段预归并以及重复序列识别工作。不足之处在于：因为要求精确的子串匹配，所以对测序中产生的一些误差序列处理能力有限，影响了该方法的应用。本文进一步的工作是研究非精确的子串匹配，提高识别特征子串和重复序列的精度，以及研究该方法针对海量片段的大规模并行处理技术。

参考文献

[1] International Human Genome Sequencing Consortium. Initial Sequencing and Analysis of the Human Genome[J]. Nature, 2001, 409(6822): 860-921.

[2] Kececioglu J D, Meyers E W. Combinatorial Algorithms for DNA Sequencing Assembly[J]. Algorithmica, 1995, 13(1/2): 7-15.

[3] Wang Jun, Wong Gane Ka-Shul. RePS: A Sequence Assembler That Masks Exact Repeats Identified from the Shotgun Data[J]. Genome Research, 2002, 12(5): 824-831.

[4] 王 磊, 张祖平, 陈建二. DNA 片段拼接中重复序列算法研究[J]. 计算机科学, 2006, 33(7): 164-170.

[5] 张博峰, 王正华. DNA 片段拼接中基于定长特征子串的重复序列信息屏蔽方法[J]. 国防科技大学学报, 2002, 24(6): 67-70.

[6] 涂俐兰, 王能超. DNA 序列拼接中重复序列屏蔽的一种新方法[J]. 华中科技大学学报: 自然科学版, 2004, 32(8): 107-109.

[7] GenBank Database[Z]. [2008-06-10]. http://www.ncbi.nlm.nih.gov/Genbank/index.html.