# 算法设计课程项目报告——宏基因组组装

15307130072 陈杨栋

**【问题介绍】**

本次课程项目的问题可以描述为给定大量从一个母串 G 上复制下来的较短的子串，顺序可能是原始方向，反方向或者相应方向按基因互补的串，并且存在一定的错误率，子串的形式可能是较多重复率较高的短串或者较少重复率较低的长串，需要根据这些子串，尽可能长、准地拼接出更长的母串，主要评价标准为组装好的串占母串的比例和与母串的匹配程度，同时要尽量保证错误率和重复匹配母串同一处更低。

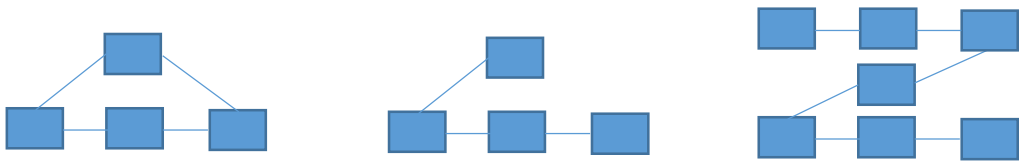
**【解决方案】**

经过查阅资料，发现目前针对这一问题已有较为成熟的方案，将切割的子串称为 read，拼接的串称为 contig，针对给出长 read 的情形，采取重叠图的算法；针对较多短 read 的情形，提出了更新颖的 de brujin 图方案。对于这两种方案，各种文献普遍看好后者，而我从信息含量的角度考虑，后一种切割方案，在信息保留方面有更大的优势，而且错误率相较前者会更低一些，故本次 PJ 我采取了 de brujin 图算法进行序列拼接。

首先介绍一下 de brujin 图（后面简称 DBJ）算法，DBJ 图由节点和有向边构成，每个节点是一个长度为 k 的串，称为 kmer，若一个 kmer 的后 k-1 个字符与另一个 kmer 的前 k-1 个字符相同，且这两个 kmer 在至少一条 read 中相邻，则前一个 kmer 向后一个连一条边，将每一个长度为 L 的 read 逐字符切割为 L-k+1 个 kmer，并根据 read 的重叠来累加节点和边的权重，就可以构建出一个节点和边都带权的 DBJ 图，为了得到拼接后的 contig，需要在图中找到合适的入口，并且

从入口开始寻找一条较为合适的路径，根据走过的路径即可生成一条 contig，为使 contig 较长并且尽可能少的重复，我认为这条路径应该类似于欧拉回路。

下面考虑生成初始的图之后的一些优化、纠错操作，查阅文献可知，由于 read 错误引起的 DBJ 图错误主要有 bubble、tip 和 erroneous connection，bubble 错误是指 a 的情况，在正常序列中浮出几个多余的 kmer 连在 contig 上，tip 错误是指类似 b 中在正常序列上伸出一个较短的边，erroneous connection 是类似 tip，但是伸出的点又连到了另一个序列上，三种情况如下图所示



(a)bubble　　(b)tip　　(c)erroneous connection

为了消除这三种情况，可以先进行一些预处理，设定一些阈值，然后对图进行宽度优先搜索，这样在每次分支的时候可以记录分支的长度，对于低于阈值的分支链直接删除，可以去掉 tip 并且不影响图的连通情况，对于气泡，可以记录分支时的分支节点，在搜索到后一个并入的节点时，可以留下上一次分支的标记，这样可以找到 bubble 的位置，然后可以删去权重较低的一边，对于误连接的情况，同样是需要通过遍历算法找到并入节点，然后回溯到分支处，通过节点权重判断跨越两条链的边和节点，将之删去。在实际操作中，为了代码编写方便，我采取了比较第一个分支处的权重，然后根据较大的权重采取一个阈值，删去分支中低于阈值的点，直到重新找到一个高于阈值的点，为了避免误删，这个阈值可以设置一个较小的值。

在进行错误筛除之前，为了遍历的时空效率，还可以事先对 DBJ 图进行一次化简，主要方法为将连续的入度和出度均为 1 的节点进行合并，这一操作的正确性

显而易见，如果一个 kmer 只属于一个 read，那么可能这个 read 的这一段只有这一次出现，那其属于的 read 将是关键信息，将其合并并不会影响图的连通性和形状，同时每次遍历到这一位置都可以减少相应的搜索代价。

在进行完上述操作之后，就可以开始搜索路径了，从问题出发，我选取了所有入度为 0 的节点作为入口，并且通过实验，对于长度为 100 的 read，选取长度 k 为 30-40 的效果最好，而且这一规模的 kmer 多样性也有保障，保证真实的入口被其他 read 中的前驱污染入度的概率较低。为了处理串顺序的问题，我把原 read 正反互补的四种情况都放入图中进行搜索，使用邻接表存储图，STL_map 存储每个 kmer 对应节点的序号，这样 map 建好后可以据此搜索出每个节点的后继，从而构建整张 DBJ 图，然后在搜索路径的过程中，我采取了贪心向的决策，采用出度最高的后继作为下一步走的方向，直到没有下一步或者当前 contig 已经超过母串长度，则停止寻路。后来发现贪心算法的覆盖率不够，尝试使用暴力深搜寻找欧拉路径，可以一定程度提高覆盖率，并且复杂度也可以接受。在 data4 中，由于数据规模原因，仍采取了贪心方案。

最终，为了输出的有效性和降低 duplication ratio，我规定舍弃 0.8*母串长度以下的 contig。在重复率较低，虚假入口较多的情况下，随机丢弃一些入口。

**【实验结果】**

实验测试了不同 kmer 长度 k 的表现，在 50 以上，能拼出的 contig 的长度急剧减少，甚至没有足够长能输出的 contig。而在 30-40 这一范围效果较好，如果 kmer 更小，则 kmer 本身的多样性会受到一定的影响，错误的拼接数量会急剧增加，contig 的质量会受到比较大的影响。

**【参考资料】**

https://wenku.baidu.com/view/1290edc5172ded630b1cb683.html

http://blog.sciencenet.cn/blog-759995-908468.html