HomeWork 12.11

邹翔宇 | 2410833001

假设一次 DNA 测序的实验中得到如下三条短读长:，第一条读长：ATTGCCTA，第二条读长：GCCTACAG，第三条读长：ACAGTTGA

问题 1：利用读长之间两两重叠的方法把这三条读长连接起来成为一条基因组序列。

1. **分析第一条读长（ATTGCCTA）和第二条读长（GCCTACAG）的重叠情况**
   - 从第一条读长的末尾开始尝试寻找与第二条读长开头的重叠部分。可以发现第一条读长的后 5 个碱基“GCCTA”与第二条读长的前 5 个碱基“GCCTA”是重叠的。
   - 基于此重叠，将两条读长连接起来得到：ATTGCCTACAG。
2. **分析已连接得到的序列（ATTGCCTACAG）和第三条读长（ACAGTTGA）的重叠情况**
   - 同样从已连接序列的末尾开始找与第三条读长开头的重叠部分，已连接序列的后 4 个碱基“ACAGT”与第三条读长的前 4 个碱基“ACAGT”是重叠的。
   - 把第三条读长连接上去，最终得到的基因组序列为：ATTGCCTACAGTTGA。

问题 2：利用 de Bruijn 图构建方法把这三条读长连接起来成为一条基因组序列（设 k-mer 的长度为 5）。

1. **提取 k-mers（长度为 5 的子序列）**
   - 对于第一条读长（ATTGCCTA），可以提取出以下 3 个 k-mers：
     - ATTGCC
     - TTGCCT
     - TGCCTA
   - 对于第二条读长（GCCTACAG），可以提取出以下 4 个 k-mers：
     - GCCTA
     - CCTAC
     - CTACA
     - TACAG
   - 对于第三条读长（ACAGTTGA），可以提取出以下 3 个 k-mers：
     - ACAGT
     - CAGTT
     - AGTTG
2. **构建 de Bruijn 图**
   - 以这些 k-mers 为节点，当两个 k-mers 有长度为 (k - 1)（这里 (k = 5)，即长度为 4）的重叠部分时，就在它们之间构建有向边。例如：
     - “ATTGC”和“TTGCCT”有长度为 4 的重叠部分“TTGCC”，所以从“ATTGC”到“TTGCCT”构建有向边。
     - “TTGCCT”和“TGCCTA”有长度为 4 的重叠部分“TGCCT”，所以从“TTGCCT”到“TGCCTA”构建有向边，依次类推构建整个图。
   - 构建好的 de Bruijn 图中，节点和边的关系大致如下（简化示意）：
     - “ATTGC” -> “TTGCCT” -> “TGCCTA” -> “GCCTAC” -> “CCTACA” -> “CTACAG” -> “ACAGT” -> “CAGTT” -> “AGTTG”
3. **通过遍历 de Bruijn 图得到基因组序列**
   - 从图中的起始节点（一般选择入度为 0 的节点，这里可以从“ATTGC”开始），沿着有向边依次遍历节点，将节点对应的 k-mers 进行拼接（去掉相邻节点之间重复的 (k - 1) 长度部分），最终可以得到基因组序列：ATTGCCTACAGTTGA。

综上，无论是利用读长之间两两重叠的方法还是利用 de Bruijn 图构建方法（k-mer 长度为 5），最终连接得到的基因组序列均为 ATTGCCTACAGTTGA。