算法设计与分析 PROJECT

一、预备知识

1.1 DNA 基础

DNA，全称脱氧核糖核酸，是生物体内重要的大分子结构，通常是由两条脱氧核苷酸链围绕一个共同的中心盘绕，立体上构成双螺旋结构。

抽象来说，DNA 的一条链可以理解为是由字符集$\Sigma = \{A,T,C,G\}$组成的字符串。由于 DNA 双链是互补的（AT 互补，CG 互补），因此通常情况下只需要其中一条链既可以代表该 DNA 所蕴含的遗传信息。

下图是对于单链 DNA 方向示例描述：

DNA 双链	------ATTAGCCCAA------ ------TAATCGGGTT------
原始方向	------ATTAGCCCAA------
相反方向	------AACCCGATTA------
互补方向	------TAATCGGGTT------
反向互补方向	------TTGGGCTAAT------

1.2 SV 基础

SV，即染色体结构变异，通常就指基因组上大长度的序列变化和位置关系变化。类型很多，包括长片段序列插入（Insertion）或者删除（Deletion）、串联重复（Tandem Duplication）、染色体倒位（Inversion）、染色体内部或染色体之间的序列易位（Translocation）以及形式更为复杂的嵌合性变异。

在本次 PROJECT 中，只需要了解下述类型的 SV 即可，对于每种 SV 给出一个例子，红色部分体现为 SV 部分：

（1）INS，长片段的插入

ref	------ATTAGCCCAA------
sv-ins	------ ATTAGTTTTCCCAA ------

（2）DEL，长片段的缺失

ref	------ATTAGTTTTCCCAA------
sv-del	------ ATTAGCCCAA ------

（3）DUP，简单的串联重复

ref	------ATTAGCCCAA------
sv-dup	------ ATTAGTAGCCCAA ------

（4）INV，染色体倒位，在单链上体现为反向互补序列

ref	------ATTAGCCCAA------
sv-inv	------ ATGCTACCAA ------

（5）TRA，染色体间异位，体现为 ins 和 del 的复合变异

ref 两条正确原始链	------ATTAGCCCAA------ ------AGAGAATTC------
sv-tra 易位后两条错误链	------ ATTAGAA ------ ------AGAGCCCAATTC------

1.3 DNA 测序

DNA 测序,即从 DNA 分子中得到 DNA 的序列信息。而由于 DNA 分子小，长度极长，想要准确无误的得到 DNA 序列是十分困难的。目前，测序技术主要有经历了三个阶段，以 sanger 测序为代表的的第一代测序技术，以 illumina 测序为代表的第二代测序技术，以 pacbio 测序为代表的第三代测序技术。

通常测序会有一定错误主要是单个位置碱基的增添、缺失或者替换。通过大量测序信息可以进行纠错，因而通常使用高覆盖度的数据前可以先进行纠错。

下面给出三种测序方法的比较：

方法	读长	准确度	每小时输出碱基数	每百万碱基价格
sanger	600~1000	99.9%	1.3e5	$2400
illumina	100~200	99.9%	7.5e10	$0.007
pacbio	~20000	85%	2.5e8	$0.08

二、任务设置

2.1 任务说明

考虑到实验内存限制，我们使用较小的 DNA 序列来模拟真实情况。一共有两组任务，每组任务模拟了共 50 个 SV，长度 50-1000 碱基随机，包括 INS-DEL20 个、INV10 个、DUP10 个、TRA10 个。注意两个任务的参考基因不同。

Task1

提供 ref.fasta 文件，表示原始的参考基因。

提供 sv.fasta 文件，表示生成若干 sv 后的基因数据。

任务要求根据这两个文件得到 sv 的信息。

Task2

提供 ref.fasta 文件，表示原始的参考基因。

提供 long.fasta 文件，表示模拟 pacbio 测序方式对生成若干 sv 后的基因测序得到的测序数据文件。

提供 report.txt 文件，包含测序数据的重要信息，包括测序数据的平均长度，测序的覆盖度，测序的错误率等信息。具体说明会在 report.txt 中描述。

2.2 测试数据

为了大家测试代码的性能，对每个任务提供一个包含答案的版本，可以自行和答案比对来评估正确率，包含答案的版本和任务本身没有关系。

答案包含在 sv.bed 文件中，格式即为正式 task 应输出的格式。

需要说明的是，在测试版本的正确率不能够实际反映出在真实任务的正确率。但方法足够好的情况下，是会呈正相关关系的。

三、文件格式

3.1 输入文件

输入文件均为.fasta 格式文件，每两行表示一串 DNA 序列，其中第一行为字符'>'加上序列 ID，第二行为 DNA 序列信息。

3.2 输出文件

有统一的输出格式。

输出文件的每一行代表一个 SV 的信息，每行的开头为 SV 的类型，对于 5 种类型的 SV 使用 INS、DEL、INV、DUP、TRA 指代。

对于 INS，DEL，INV，DUP 类型的 SV，随后应跟随该 SV 发生的位置，依次有 3 个信息输出，为 DNA 序列的 ID，发生 SV 的起始位置，SV 的起始位置加上 SV 的长度，使用用空格间隔开。

TYPE ID POS POS+LEN

对于 TRA 类型的变异，需要说明清楚 SV 从哪里删除，并增加到哪里。依次输出 6 个信息，即 DEL 位置 3 个信息与 INS 位置的 3 个信息。

TRA ID1 POS1 POS1+LEN ID2 POS2 POS2+LEN

四、得分方式

4.1 总体概况

对于每个 task 需要提交实验报告和代码，并提交 task1 和 task2 的实验结果，对于测试数据不需要提交其结果。

PROJECT 一共 15 分，task1 报告及代码 5 分，结果 2 分，task2 报告及代码 5 分，结果 2 分，独立工作体现 1 分。

4.2 报告要求

要求简明扼要地说明算法的思路，体现出对问题的思考与分析。报告字数不宜过多，每个 tast 不超过 3 页内容。

4.3 结果评估标准

每个 task 有 50 个 SV，在提交结果中输出的 SV 数目不应超过 120 个，超过 120 个只取前 120 个结果评估。

对于每个 SV 正确的位置区间，对于在结果集合中存在一个 SV 位置区间，记该 SV 长度 len1，结果集合中 SV 长度 len2，两个 SV 位置交集长度 len3，如果 len3/max{len1, len2}>0.6，那么认为正确的找到了该 SV。对于 TRA 类型要满足两个位置区间同时满足要求。

对 project 有任何疑问可以邮件联系 20210240014@fudan.edu.cn