算法设计课程项目

染色体结构变异

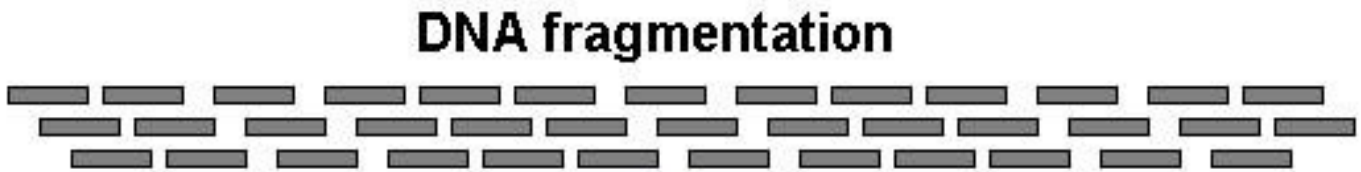
Background——DNA

- 生物的世界
  - 脱氧核糖核酸，是一种分子，双链螺旋结构，由脱氧核糖核苷酸（组成。可组成遗传指令，引导生物发育与生命机能运作。

- 算法的世界
  - DNA序列为字符集为$\Sigma = \{A, C, G, T, N\}$ 的字符串(N表示未知，可能为ACGT中任一个）。人类某一染色体

Background——DNA测序

- 生物的世界
  - DNA测序，即从DNA分子上测定DNA序列的技术。由于技术原因，目前主要使用的方法是霰弹测序法。其先通过PCR将DNA复制若干份，再将它们随即切割为长度接近的小段，分别进行测序。若有需要再将这些互相交叠的小段拼接起来，得到完整的DNA序列。



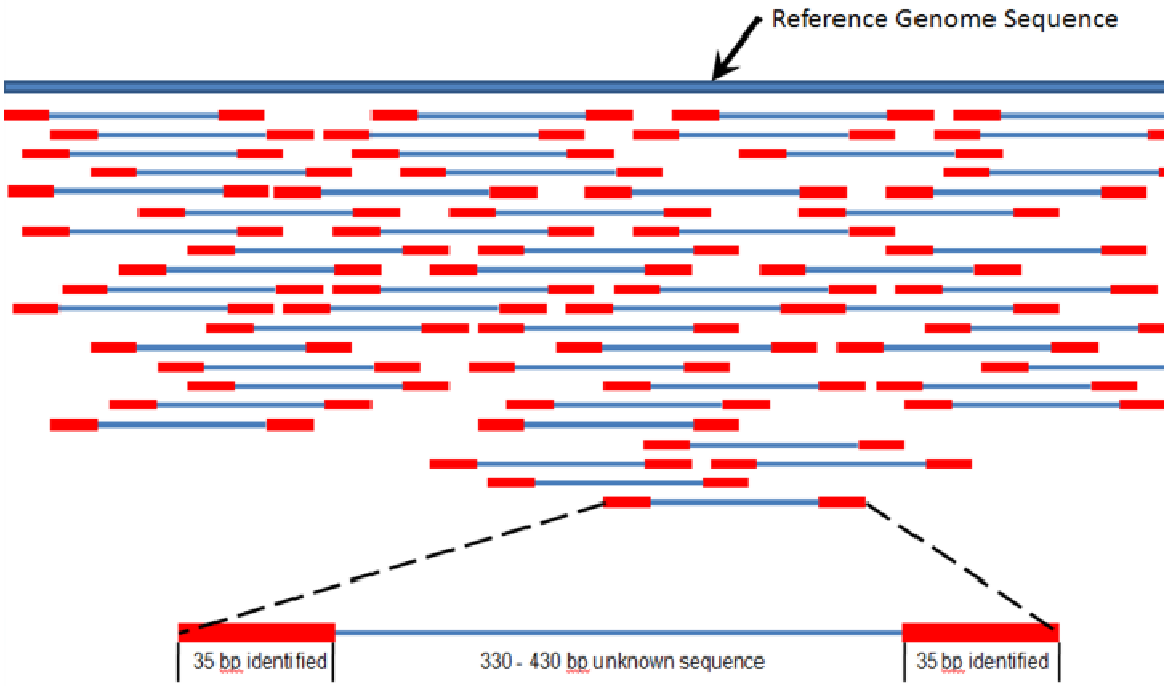
Background——DNA测序

- 算法的世界
  - 我们得到的测序数据是DNA原字符串的一些子串
    - 1. 子串长度接近
    - 2. 子串均匀覆盖分布于整个字符串中
    - 3. 子串数量= $\frac{\text{复制次数}\times\text{原字符串长度}}{\text{子串平均长度}}$

Background——Paired end

- 生物的世界
  - 在测序过程中，我们可以先将序列均匀分解成较长的片段，并对片段两端进行标记，再将其分解成小片段进行测序。这样就可以知道结果中被标记的两段的距离，方便拼接。
- 算法的世界
  - 测序结束后，除了前述的子串内容之外，我们还可以得到某些子串间在原串上的距离。

Background——Paired end



图中DNA被切割为约450长度的片段，对片段两端各35长度的片段进行标记。这样测序完成后，就可以得到被标记片段间长度约为350的额外信息。

Background——DNA测序错误

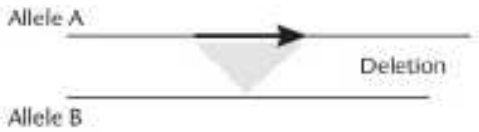
- 由于技术原因，DNA测序得到的子串会有一定几率错误（几率约为每个字符1%）
- 测序错误体现为
  - 1. 替换：某一位置字符替换为其他字符
  - 2. 插入：两字符间插入一字符
  - 3. 删除：某字符未被测出而缺失。

Background——结构变异

- 生物的世界：
  - 染色体结构变异（SV）是染色体变异的一种。
  - 在自然条件或人为因素的影响下，染色体发生的结构变异主要有4种：
    - 1. 缺失
    - 2. 重复
    - 3. 倒位
    - 4. 易位

Background——缺失

- 生物的世界：
  - 染色体中某一片段的缺失 例如，猫叫综合征是人的第5号染色体部分缺失引起的遗传病，因为患病儿童哭声轻，音调高，很像猫叫而得名。果蝇的缺刻翅的形成也是由于一段染色体缺失造成的。
- 算法的世界：
  - 假设原串$s = s_1 + s_2 + s_3$
  - 则发生缺失变异后的字符串$s' = s_1 + s_3$

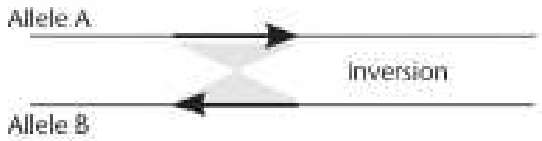


Background——重复

- 生物的世界
  - 染色体增加了某一片段 果蝇的棒眼现象就是X染色体上的部分重复引起的。
- 算法的世界

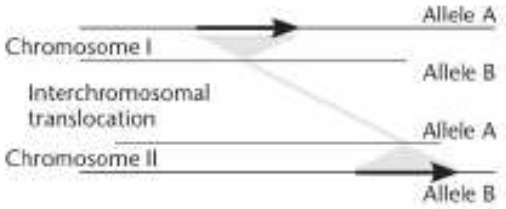
Background——倒位

- 生物的世界
  - 染色体某一片段的位置颠倒了180度，造成染色体内的重新排列 如女性习惯性流产（第9号染色体长臂倒置）。
- 算法的世界
  - 假设原串$s = s_1 + s_2 + s_3$
  - 则发生重复变异后的字符串$s' = s_1 + s_2' + s_3$
  - 其中s_2'为s_2的逆向字符串



Background——移位

- 生物的世界
  - 染色体的某一片段移接到另一条非同源染色体上或同一条染色体上的不同区域 如惯性粒白血病（第14号与第22号染色体部分易位，夜来香也经常发生这样的变异。
- 算法的世界
  - 假设原染色体1：$s_1 = s_{11} + s_{12} + s_{13}$
  - 染色体2：$s_2 = s_{21} + s_{22}$
  - 变异后染色体1：$s_1' = s_{11} + s_{13}$
  - 染色体2：$s_2' = s_{21} + s_{12} + s_{22}$



Problem

- 给出一参考DNA序列——长字符串
- 及比较DNA序列的测序结果——若干短字符串，及某些字符串间距离
- 求测序结果相对参考DNA序列发生的缺失、重复、倒位结构变异，不考虑跨染色体的移位变异。

Problem

- 输入格式：
  - 1. “genome.txt”
    - 参考DNA序列的长字符串。
  - 2. “reads.txt”
    - 测序结果的短字符串。
    - 文件中每行包括成对的两个子串及距离

Problem

- 输出格式：
  - “out.txt”
  - 每行代表一个回报的结构变异，包括以空白字符分隔的
    - 1. 字符串：deletion/inversion/duplication 代表结构变异种类
    - 2. 数字：参考串发生结构变异的起始位置
    - 3. 数字：参考串发生结构变异的结束位置
- 注意：
  - 允许一定限度内的多报，请输出所有你认为可能的结构变异。

Problem

- 测试数据
  - 对于所有数据保证：
    - 复制次数<=30，即子串长度之和小于30倍参考串长度。
    - 比较序列子串长度<=105
    - 成对子串距离∈ (100,1000)
  - 测试数据包含模拟数据与真实数据

Problem

- 模拟数据：
  - 对于所有模拟数据，保证比较DNA序列直接经由参考DNA序列发生若干次结构变异产生。
  - 保证结构变异不相互重叠。
  - 包含以下类型：
    - 1. 参考序列长度<=3000, 无测序错误
    - 2. 参考序列长度<=3000, 存在1%的测序错误
    - 3. 参考序列长度<=20M, 无测序错误
    - 4. 参考序列长度<=20M, 存在1%的测序错误
    - 5. 参考序列长度<=200M, 存在1%的测序错误

ONE
TWO
THREE

ONE
TWO
THREE

测试

- 近期会公布一些1，2类模拟数据及答案
- 期中进行一次3，4类模拟数据的测试（不计分）
- 期末进行所有模拟数据及真实数据的测试，决定成绩
- 此外，期末提交程序代码及项目报告。
- 测试分数主要取决于覆盖率，即正确回报占所有正确结构变异的比例。其次考虑准确度，即正确回报占回报总数量的比例。
- 恶意回报大量错误结果将记为0分

附录

- 如有问题请联系：
  - 黄剑铮 14307130264@fudan.edu.cn
- 参考论文：
  - Rausch, Tobias, et al. "DELLY: structural variant discovery by integrated paired-end and split-read analysis." *Bioinformatics* 28.18 (2012): i333-i339.
  - http://bioinformatics.oxfordjournals.org/content/28/18/i333.full