王 冰

Date：1992.10 E-mail：wangbing@ict.ac.cn
Phone：(+86)156-3885-9198 Blog: http://blog.csdn.net/zzucaicai

教育背景

中国科学院计算技术研究所（保送 前 2%） 2014.09 - 2017.07
工学硕士，前瞻研究实验室生物信息课题组 研究方向：基因组序列拼接算法研究

郑州大学 2010.09 - 2014.07
工学学士，计算机科学与技术专业 专业排名：**2 / 89**

项目经历

基因组序列拼接算法的研究及拼接软件 ARCS 的实现 2015.08 - 今
核心开发人员 **中科院计算所**

- 简介：从百万量级的 DNA 序列片段中恢复出原始 DNA 序列。该问题可以形式化为一个求解序列 overlap 图的汉密尔顿回路问题，图中点即为序列，两条序列的重叠区域大于某一值则连边。由于汉密尔顿路径问题是 NP 完全问题，将其转化为求解 De bruijn 图的超欧拉回路问题。再使用 pair-end 序列信息（已知长 DNA 两端碱基序列及距离）对解 De bruijn 图生成的序列做二次拼接得到更长的序列。
- 职责：
  - 负责设计根据 pair-end 连接边的距离信息判定序列是否是重复序列的算法。使用改进的混合高斯模型对距离信息聚类，并使用 BIC 准则判定聚类中心的个数，根据聚类中心个数及连接边的情况判定重复序列。
  - 负责根据 pair-end 连接边的距离信息，确定序列在原始 DNA 序列的位置。将问题形式化为线性规划问题并使用 glpk 求解，从而得到问题的**全局最优解**。
  - 完成 ARCS 使用 pair-end 信息部分代码编写。使用多线程优化大批量数据处理。
- 成果： ARCS 实际测得序列 N50（所有长度大于 N50 的序列长度之和为所有序列长度和的 50%）优与主流软件约 10%，并与北京基因组所合作将 ARCS 投入使用。
- **关键字**：欧拉回路，混合高斯模型，线性规划，c++，多线程，boost

基因组酶切位点拼接算法的研究及拼接软件 nanoARCS 的实现 2015.12 - 今
核心开发人员 **中科院计算所**

- 简介：某些酶可以识别特定的 DNA 短序列并记录其位置信息（位点），酶切位点拼接即为从大量的位置信息中拼接出原始 DNA 序列的位点信息。与 DNA 序列拼接区别在于：1，错误酶切位点比例很高（13%），且位置信息不精确，不能精确匹配；2，酶切位点的数目较少，贪心方法求得结果通常比较好。
- 职责：
  - 设计数据结构及算法，确定位点序列之间的相似性及相对距离。将位点序列拆分成短序列，根据已有算法对短序列相似性打分，使用 **p-value** 判定分值的显著性。由于短序列数目多，设计先由窗口粗分类后由分数细分类的聚类策略对短序列聚类，以获取准确性和性能的平衡。后根据聚类结果确定原始位点序列的连接关系。
  - 设计贪心策略以得到最终拼接结果。根据连接信息，拼接 A，B 序列，当且仅当 B 为 A 的最佳后继且 A 为 B 的最佳前驱(最佳策略验证中)。
- 代码开源：https://github.com/zzucainiao/nanoARCS

实习经历

2013.10 - 2013.12 **金山云** **分布式文件系统测试及性能优化**

- 职责：安装配置 **MooseFS**，测试各个参数对性能的影响，查找性能瓶颈。用汇编语言改写 crc 校验部分代码。
- 收获：crc 校验速度提升 40%，对分布式文件系统有整体了解，熟练使用常用 linux 命令。

个人技能

- 编程能力：熟悉 C++，了解面向对象基本思想及常用设计模式，知道常用 boost 库；了解 Java 编程。
- 算法能力：良好的数据结构和算法基础，了解基本机器学习算法。

获奖情况及其它

- 2012/2013 ACM-ICPC 国际大学生程序设计竞赛亚洲区金华站 **银奖；** 长沙站、成都站 **铜奖**
- 2015 中国科学院“三好学生” 2015 中国大学生程序设计竞赛 **银奖**
- 2012 国家奖学金（**前 1%**） 2010/2011/2012 郑州大学一等奖学金（**三次**）（**前 5%**）