哈尔滨工业大学计算机科学与技术学院

实验报告

课程名称：生物信息学

课程类型：专业必修课

实验项目名称： 大规模基因组序列表示与索引系统设计与实现

实验题目：

班级：

学号：1180300903

姓名：陈翔

|  |  |  |
| --- | --- | --- |
| 设计成绩 | 报告成绩 | 指导老师 |
|  |  |  |

1. 实验目的

给定基因组序列（大肠杆菌E. Coli, 长度~5Mbp）作为输入

建立算法和系统，可以为基因组序列建立自索引表示

自索引结构可以选择CSA或者BWT

1. 实验原理

BWT是一种以数据块为操作对象的可你的数据变换方法，其核心思想是对字符串循环移动后得到字符矩阵进行排序和变换。也就是对参考基因组进行了一次有规律的重新排序，变换的目的就是为了方便后续进行查找。

 一阶BWT构造和查找的具体过程如下：

一、BW变换（编码）：

1.在源文本T后面添加一个额外的字符‘$’，定义该字符比源文件中所有字符都要小。

2.文本T$做循环移动（每次一个字符）形成新的文本，原始文本T$和新产生的文本组成一个文本集。

3.对文本集中的文本按照字典顺序排序（A-Z）。

4.抽取步骤3中排序后形成矩阵的最后一列L，且由于矩阵中的每一列都是源文本的一个置换，所以L也是源文本的一个置换

*源文本：ATTCGCTT*

步骤1：

    添加一个额外的字符‘$’，ATTCGCTT$。

步骤2：

    每一次变换中，首字符移动到末字符，形成新文本（字符串阵列），

步骤3：

    按照字典顺序进行排序（按照首位从小到大的顺序排序，若首位字符相同，则看下一位，直至不同位），

步骤4：

    最后一列，T$TGCTTCA。

SA后缀数组、Occ数组和BWT串是BWT索引算法的3个辅助数组结构。

其中，SA数组保存了BWT矩阵中$符号前面的后缀在参考基因组中出现的真实位置。如，A[7]=6表示BWT矩阵的第7行的后缀‘TT’在X中出现的位置是6。Occ[i]表示第i行的BWT串BWT[i]在整个BWT串种出现的位次，例如，Occ[2]=1表示BWT[2]='T'在索引T$TGCTTCA中是第二次出现的“T”。

对序列建立索引以后，假定字符串W是X的一个子串，则W在X中每次出现的起始位置集合是后缀数组SA中的一个连续区间，这是因为所有包含W并以其为前缀的子序列都已经被按照字典序排列过

于这个发现，可以定义2个变量sp和ep，分别表示SA中W出现的起始和结束位置（即为BWT矩阵中的行号）：

            sp(W)=min{k|所有X中以W为前缀的位置}

            ep(W)=max{k|所有X中以W为后缀的位置}

特殊的，当W为空时，sp(W)=1,ep(W)=n-1。W在X中所有出现的位置集合为：{SA(k):sp(W)<=k<=ep(W)}。例如，‘TT’的SA区间为[7,8]，对应到X上的真实位置是6和1.

实际应用中，SA后缀数组和Occ数组隔32行存一次，以便减少辅助数据结构所占用的空间，得到数据压缩的效果。当带查找的序列结束的位置不在SA后缀数组中时，接着往下做sp和ep的查找直至找到存储在SA后缀数组中的某一元素，让其减去该查找步骤进行的步数，即可得到原查找串的真实位置

1. 测试结果及分析
2. 经验体会

【】

1. 附录：源代码（带注释）