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：限选

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

班级：1803401

学号：1180300906

姓名：张飞扬

|  |  |  |
| --- | --- | --- |
| 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。

实验要求

自索引结构可以选择CSA或者BWT

自索引构建算法可以参照以下论文之一

Liu and Zhu, 2016（Advised）

Hon et al., 2007（Advised）

Karkkainen, 2007

其他相关论文

高分要求：

完整地完成算法所有模块

程序运行正确

程序代码优化较好

所有文档完整且写作规范

1. 实验原理

本次实验采用增量构建的方法来构建 CSA。

首先，定义字母表 ∑，在DNA序列中，∑ = {A, C, G, T}。定义 $ 为字符串末尾字符。

规定字典序如下：$<A<C<G<T。 设 T 是一个 n 长的定义在 ∑ 上的字符串。T[0..n] = T[0]T[1] . . . T[n], 并规定 T[n] = $。规定 Ti = T[i..n] = T[i]T[i + 1] . . . T[n] ，即 T 的一个从第 i 位开始至最后一位的后缀。

首先得到后缀数组 SA ，SA是将 T 的 n 个后缀排序后的序列，数组SA[i] 存储 T 的第 i 小的后缀在后缀数组中的位置。

定义 SA^-1[i] 为 T 的第 i 个后缀在后缀数组中存储的位置，即 Ti 的顺序。

定义函数 Ψ：Ψ[i] = SA^-1[SA[i] + 1] for i = 1, 2, . . . , n，并规定 Ψ[0] = SA^-1[0]。不难看出 Ψ[i] 存储的是第 i 小的后缀在 T 上的下一个后缀在 SA 中存储的位置，或者说是第 i 小的后缀去掉首字符之后在 SA 的位置，我们可以把 Ψ 理解为一个指向 SA 的指针。CSA 中存储就的是数组 Ψ[0..n]，共 n+1 个整数。

1. 算法设计

将 T 划分为 n/l 个连续的区域，记为 T1 , T2 , . . . , T n/l，l = O( log nn )。通过 Base Step 和 Merge Step 两个部分来逐步构建整个 T 上的 CSA。

3.1 Base Step

主要构建 T^n/l 的 Ψ。首先通过排序得到 T^n/l 的后缀数组，这个操作耗费 O(l log l) 的时间。计算得到 SA^-1 耗费 O(l) 时间，然后得到 Ψ 需要 O(l) 时间。因此 Base Step 的时间复杂度是 O(l log l) 的。

3.2 Merge Step

这一步的主要思想是基于之前的排序结果，将新的串插入进来。记 Ti = c1c2 · · · cl , i ∈ 1, 2, . . . , n/l-1，设 suf1, suf2, . . . , sufl 是 TiT′ 的 l 个最长的后缀，显然 sufk = ckck+1 · · · clT′。

首先，将这 l 个最长后缀排序，这一步和 Base Step 中的排序并无区别。然后对每一个 sufi，要得到其在 T′ 中的顺序 order(sufi , T′)。然后就可以构造 TiT′ 的 Ψ。

1. 测试结果分析
2. 经验体会
3. 附录：源代码（带注释）