软件工程课程设计项目 “DNA 序列的 k-mer index 问题”

一、项目设计要求：

1. 针对课题问题，分析并撰写“项目需求规约”。
2. 设计算法：要求每个小组设计至少两种不同的算法，以解决这个问题并对这两种算法的效率和效用做出评价。
3. 编写描述系统架构和详细设计的相关文档。
4. 编写及调试代码，以实现解决这个问题的可运行之项目原型。

二、问题描述

2.1 项目背景 (来源：2015 年“深圳杯”数学建模夏令营)

这是一个来自 DNA 序列的 k-mer index 问题。

给定一个 DNA 序列，这个系列只含有 4 个字母 ATCG，如 S = “CTGTACTGTAT”。给定一个整数值 k，从 S 的第一个位置开始，取一连续 k 个字母的短串，称之为 k-mer（如 k= 5，则此短串为 CTGTA），然后从 S 的第二个位置，取另一 k-mer（如 k= 5，则此短串为 TGTAC），这样直至 S 的末端，就得一个集合，包含全部 k-mer。如对序列 S 来说，所有 5-mer 为：

{CTGTA，TGTAC，GTACT，TACTG，ACTGT，TGTAT}

通常这些 k-mer 需一种数据索引方法，可被后面的操作快速访问。例如，对 5-mer 来说，当查询 CTGTA，通过这种数据索引方法，可返回其在 DNA 序列 S 中的位置为 {1，6} 。

2.2 所需解决的问题

现在以文件形式给定 100 万个 DNA 序列，序列编号为 1-1000000，每个基因序列长度为 100 。

（1）要求对给定 k， 给出并实现一种数据索引方法，可返回任意一个 k-mer 所在的 DNA 序列编号和相应序列中出现的位置。每次建立索引，只需支持一个 k 值即可，不需要支持全部 k 值。

（2）要求索引一旦建立，查询速度尽量快，所用内存尽量小。

（3）给出建立索引所用的计算复杂度，和空间复杂度分析。

（4）给出使用索引查询的计算复杂度，和空间复杂度分析。

（5）假设内存限制为 8G，分析所设计索引方法所能支持的最大 k 值和相应数据查询效率。

（6）按重要性由高到低排列，将依据以下几点，来评价索引方法性能

- 索引查询速度。
- 索引内存使用。
- 8G 内存下，所能支持的 k 值范围。
- 建立索引时间。

三、项目数据文件

下载链接：http://math.tongji.edu.cn/model/camp2015B.html

(©)