**2015年“深圳杯”数学建模夏令营**

B题：DNA序列的k-mer index 问题

这个问题来自 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｝。

**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