2019 年“认证杯”数学中国数学建模网络挑战赛

第二阶段

B 题 外星语词典

我们发现了一种未知的语言，现只知道其文字是以 20 个字母构成的。我们已经获取了许多段由该语言写成的文本，但每段文本只是由字母组成的序列，没有标点符号和空格，无法理解其规律及含义。我们希望对这种语言开展研究，有一种思路是设法在不同段文本中搜索共同出现的字母序列的片段。语言学家猜测：如果有的序列片段在每段文本中都会出现，这些片段就很可能具备某种固定的含义 (类似词汇或词根)，可以以此入手进行进一步的研究。在文本的获取过程中，由于我们记录技术的限制，可能有一些位置出现了记录错误。可能的错误分为如下三种：

1. 删失错误：丢失了某个字母；
2. 插入错误：新增了原本不存在的字母；
3. 替换错误：某个字母被篡改成了其他的字母。

第一阶段问题： 假设我们已经获取了 30 段文本，每段文本的长度都在 5000–8000 个字母之间。我们希望找到的片段的长度在 15–21 个字母之间。为简单起见，我们假设文本中出现的错误只有替换错误，而且对我们要找的片段而言，在文本中每次出现时，最多只会出现 4 个字母的替换错误。请设计有效的数学模型，快速而尽可能多地找到符合要求的字母片段，并自行编撰算例来验证算法的效果。

第二阶段问题： 现假设我们已经获取了 30 段文本，每段文本的长度都在 5000–8000 个字母之间。我们希望找到的片段的长度为 15 个字母。由于技术

的限制，当我们在记录每个字母时，都可能有五分之一的概率发生错误。错误类型可能为删失错误、插入错误或替换错误，每个错误只涉及一个字母，且每个错误的发生是独立的。请你设计合理的数学模型，快速而尽可能多地找到符合要求的片段，并自行编撰算例来验证算法的效果。如果我们事先不知道所寻找的片段的长度，算法又应当进行什么改进呢？