思考题 6（习题 15-5）

15-5 （**编辑距离**） 为了将一个文本串 $x[1..m]$转换为目标串 $y[1..n]$，我们可以使用多种变换操作。我们的目标是，给定 x 和 y，求将 x 转换为 y 的一个变换操作序列。我们使用一个数组 z 保存中间结果，假定它足够大，可存下中间结果的所有字符。初始时，z 是空的，结束时，应有 $z[j]=y[j]$，$j=1, 2, \cdots, n$。我们维护两个下标 i 和 j，分别指向 x 中位置和 z 中位置，变换操作允许改变 z 的内容和这两个下标。初始时，$i=j=1$。在转换过程中应处理 x 的所有字符，这意味着在变换操作结束时，应有 $i=m+1$。

我们可以使用如下 6 种变换操作：

复制(copy)——从 x 复制一个字符到 z，即进行赋值 $z[j]=x[i]$，并将两个下标 i 和 j 都增 1。此操作处理了 $x[i]$。

替换(replace)——将 x 中一个字符替换为另一个字符 c，$z[j]=c$，并将两个下标 i 和 j 都增 1。此操作处理了 $x[i]$。

删除(delete)——删除 x 中一个字符，即将 i 增 1，j 不变。此操作处理了 $x[i]$。

插入(insert)——将字符 c 插入 z 中，$z[j]=c$，将 j 增 1，i 不变。此操作未处理 x 中字符。

旋转(twiddle，即交换)——将 x 中下两个字符复制到 z 中，但交换顺序，$z[j]=x[i+1]$且$z[j+1]=x[i]$，将 i 和 j 都增 2。此操作处理了 $x[i]$和 $x[i+1]$。

终止(kill)——删除 x 中剩余字符，令 $i=m+1$。此操作处理了 x 中所有尚未处理的字符。如果执行此操作，则转换过程结束。

下面给出了将源字符串 algorithm 转换为目标字符串 altruistic 的一种变换操作序列，下划线指出执行一个变换操作后两个下标的位置：

操 作	x	z
初始字符串	algorithm	_
复制	algorithm	a_
复制	algorithm	al_
替换为 t	algorithm	alt_
删除	algorithm	alt_
复制	algorithm	altr_
插入 u	algorithm	altru_
插入 i	algorithm	altrui_
插入 s	algorithm	altruis_
旋转	algorithm	altruisti_
插入 c	algorithm	altruistic_
终止	algorithm_	altruistic_

注意，还有其他方法将 algorithm 转换为 altruistic。

每个变换操作都有相应的代价。具体的代价依赖于特定的应用，但我们假定每个操作的代价是一个已知的常量。我们还假定复制和替换的代价小于删除和插入的组合代价，否则复制和替换操作就没有意义了。一个给定的变换操作序列的代价为其中所有变换操作的代价之和。在上例中，将 algorithm 转换为 altruistic 的代价为

(3 · cost(复制)) + cost(替换) + cost(删除) + (4 · cost(插入)) + cost(旋转) + cost(终止)

a. 给定两个字符串 $x[1..m]$ 和 $y[1..n]$ 以及变换操作的代价，x 到 y **编辑距离**(edit distance)是将 x 转换为 y 的最小代价的变换操作序列的代价值。设计动态规划算法，求 $x[1..m]$ 到 $y[1..n]$ 的编辑距离并打印最优变换操作序列。分析算法的时间和空间复杂度。

编辑距离问题是 DNA 序列对齐问题的推广(参考其他文献，如 Setubal 和 Meidanis [310，3.2 节])。已有多种方法可以通过对齐两个 DNA 序列来衡量它们的相似度。有一种对齐方法是将空格符插入到两个序列 x 和 y 中，可以插入到任何位置(包括两端)，使得结果序列 x' 和 y' 具有相同的长度，但不会在相同的位置出现空格符(即不存在位置 j 使得 $x'[j]$ 和 $y'[j]$ 都是空格符)。然后为每个位置"打分"，位置 j 的分数为：

- +1，如果 $x'[j]=y'[j]$ 且不是空格符。
- −1，如果 $x'[j]\neq y'[j]$ 且都不是空格符。
- −2，$x'[j]$ 或 $y'[j]$ 是空格符。

对齐方案的分数为每个位置的分数之和。例如，给定序列 $x=$ GATCGGCAT 和 $y=$ CAATGTGAATC，一种对齐方案为

```
G ATCG GCAT
CAAT GTGAATC
-*++*+*+-++*
```

+ 表示该位置分数为 +1，− 表示分数为 −1，* 表示分数为 −2，因此此方案的总分数为 $6\cdot 1-2\cdot 1-4\cdot 2=-4$。

b. 解释如何将最优对齐问题转换为编辑距离问题，使用的操作为变换操作复制、替换、删除、插入、旋转和终止的子集。