

COMP3600/COMP6466 in 2006 – Lab Two

QUESTION ONE

RNA is a polymer of nucleotides. Each nucleotide contains one of the four bases: Adenine (A), Guanine (G), Thymine (T), and Cytosine (C).

Two RNA sequences can be aligned in many ways by inserting gaps. For example, $X = GACGGATTAGAA$ and $Y = GATCGGAATAG$ can be aligned like this:

```
G A - C G G A T T A G A A
G A T C G G A A T A G - -
```

The score of the above alignment is 2, calculated thus:
1 for a match, -1 for a mismatch, -2 for a gap. In the above example:

```
G A - C G G A T T A G A A
G A T C G G A A T A G - -
-----
1 1 -2 1 1 1 1 -1 1 1 1 -2 -2   Total = 2.
```

The problem is to find an alignment whose score is the greatest.

Let $X = x_1x_2\dots x_m$ and $Y = y_1y_2\dots y_n$ be two RNA sequences with lengths m and n respectively. Let $C(i, j)$ be the score of the best alignment of the i -th prefix $X_i = x_1x_2\dots x_i$ of X with the j -th prefix $Y_j = y_1y_2\dots y_j$ of Y , $0 \leq i \leq m$ and $0 \leq j \leq n$. Then, the best score $C(i, j)$ for aligning X_i with Y_j satisfies a recurrence as follows.

$$C(i, j) = \max \begin{cases} 0 & i = j = 0 \\ C(i-1, j) - 2 & \text{align } x[i] \text{ with a gap } (i > 0) \\ C(i, j-1) - 2 & \text{align } y[j] \text{ with a gap } (j > 0) \\ C(i-1, j-1) + 1 & \text{align } x[i] \text{ with } y[j], \text{ match } (i, j > 0, x[i] = y[j]) \\ C(i-1, j-1) - 1 & \text{align } x[i] \text{ with } y[j], \text{ mismatch } (i, j > 0, x[i] \neq y[j]) \end{cases}$$

The file <http://cs.anu.edu.au/student/comp3600/alignment.c> contains source code for computing the best score of an alignment of two RNA strings X and Y . It also computes an array W that can be used to determine an example of an optimal alignment (read the comments in the program).

Your task is to finish the part of the program that constructs the padded strings for the optimal alignment.