**Needleman-Wunsch algorithm handout**

MIP-280A4

*To find the optimal alignment between two sequences S1 and S2 of length M and N:*

**Matrix creation and calculation:**

1. Create an empty table with M+1 columns and N+1 rows.
   1. The rows of this table are indexed by *i:* for the first row: *i*=1
   2. The columns of this table are indexed by *j:* for the first column: *j*=1
   3. Cells are designated C(*i,j*)
2. Starting with the 2nd column, label the columns using the letters in S1
3. Starting with the 2nd row, label the rows using the letters in S2
4. Write a 0 in the upper-left cell: C(1, 1)
5. Calculate a score for the rest of the cells in the table:

Score for a cell = the maximum of:

Up: the score of cell above + the gap cost.

Left: the score from the cell to the left + the gap cost.

Diag: the score from the cell above-left + a match reward or mismatch cost.

Note that:

Up will result in a gap in the upper sequence in the alignment

Left will result in a gap in the left sequence in the alignment

Diag will result in a non-gap in the alignment (a match or a mismatch)

If there is a tie between Up, Left, or Diag. Keep track of this fact. This corresponds to two alignments of equal score.

1. Note that the first row and the first column are special: there are no cells above the first row and no cells to the left of the first column. So you fill these out first. The only option for cells in the first column is “Up”. The only option for cells in the first row is “Left”. The first row and first column correspond to alignments consisting of all end gaps.
2. Moving from top to bottom, and left to right, beginning with cell C(2,2), calculate the score for the remaining cells as the maximum of Up, Left, or Diag.
3. As you fill in the value for each cell, keep track of whether the maximum value came from Up, Left, or Diag. This is the “traceback direction” for each cell.
4. Stop once you’ve reached the bottom-right cell. *The value in this cell is the score of the highest-scoring alignment.*

**Traceback**: This part of the algorithm allows you to write out the highest scoring alignment from the matrix.

**Some points about the traceback:**

* Traceback starts at the bottom right cell and ends when you get back to the upper left cell (except for the upper-left 0), so to: C(2,2).
* Traceback doesn’t use the values in the cells: it only uses the traceback directions.
* During traceback, you will write out the higest scoring alignment from right to left.

1. For each cell in the traceback path, examine the traceback direction. If the traceback direction is:
   1. **Diagonal**: correspond to a non-gap in the alignment. Write the corresponding letter from S1 in the top row of the alignment and the corresponding letter from S2 in the bottom of the alignment.
   2. **Up**: corresponds to a gap in the upper sequence (S1). Write a gap “-“ in the top row of the alignment and the corresponding letter from S2 in the bottom row of the alignment.
   3. **Left**: corresponds to a gap in the left sequence (S2). Write the corresponding letter from S1 in the top row of the alignment and a gap “-“ in the bottom row.
2. Follow the traceback direction to the next position in the alignment until you reach the upper-left cell.

**Exercise: align the following sequences using the N-W algorithm and the appropriately-sized table (matrix) provided:**