

## The Needleman-Wunsch Algorithm Exercise – Instructor’s Guide

**Class level:** This exercise is used in a sophomore-junior level bioinformatics course. It typically has both Bioinformatics majors and Biology majors in the course, all of whom have had a year of General Biology. Most of the Bioinformatics majors have had at least one programming course, but typically have not had a course that involves algorithms. The Biology majors have not had any programming courses.

**Class presentation:** I typically introduce the algorithm on the whiteboard, explaining how it works by drawing a matrix similar to the one used in the exercise. I then have them spend about 15 minutes or so filling out the diagram. Depending on the time I will go over the answer on the board then or at the beginning of the next class period.

**Student preparation:** When we do this students have typically been doing some sequence alignment exercises, such as BLAST searches, and have used different scoring systems to score different short alignments, so they are familiar with sequence alignments and the concept of scoring the alignments, but they have not been exposed to how to generate an alignment. This is also typically the first time they have seen any type of algorithm. They have been asked to read the section in the textbook that introduces this, which also includes some computer pseudocode.

## The Needleman-Wunsch Algorithm Exercise

One of the fundamental techniques in bioinformatics is the pairwise alignment of protein sequences. This basic procedure underlies a large number of bioinformatics applications, and can be used to accomplish a wide range of tasks. While it is often straightforward to align sequences that are both short and fairly similar, most sequences are far too long to align manually, and if they are not highly similar it is not straightforward to determine the best way to align the two sequences. This is where alignment algorithms come in.

Algorithms are “recipes” for solving particular problems in mathematics or computer science. The first published sequence alignment algorithm was the Needleman-Wunsch (NW) algorithm, and it has provided the foundation for a large number of subsequent alignment algorithms, such as that used by the BLAST program. Understanding how the Needleman-Wunsch algorithm works provides a foundation for you to understand a variety of other bioinformatics algorithms. In this exercise you will implement a NW algorithm using pen and paper as preparation for being able to implement it in a computer program.

The NW algorithm uses a matrix approach, placing the letters for the two sequences to be aligned along the rows and columns of a matrix. It then involves pre-computing alignment scores at each position in the matrix, then tracing back through the matrix to determine the highest scoring, or optimal, alignment between the two sequences. (The process of pre-computing certain values and then determining the optimal result is known as *dynamic programming*, and is a common technique in computer science.) The algorithm produces one or more **optimal alignments** (there can be more than one alignment with the same score). It produces a **global** alignment, requiring the sequences to be aligned over their complete lengths.

Using the Needleman and Wunsch dynamic programming method outlined below, construct the alignment score matrix on the following page for the following two sequences, using the following scoring parameters: match score = +1, mismatch score = 0, gap penalty = -1.

ACGT and ACGGT

Show the traceback path on the matrix and write the resulting alignment below the matrix.

Carry out the following steps on the attached matrix:

- 1.) Start in the upper-left corner of the matrix and put a zero in the circle.
- 2.) Every horizontal or vertical move corresponds to introducing a gap in the alignment, so for each such move in the first row or first column add the gap penalty to the score and write the number in the circle.
- 3.) Every diagonal move corresponds to aligning two nucleotides (or amino acids). For the first diagonal down and to the right of the upper left circle

determine whether the two nucleotides are the same or different and calculate the score after adding the match or mismatch score.

4.) Each circle after the first row and column has three arrows pointing to it. For each circle, calculate the score for each of the three arrows and write it next to the arrow head. For example, the three arrows for the second circle in the second row are a -2 (from the left – zero plus a down gap and an across gap), 1 (from the diagonal – zero plus a match), and -2 (from above – zero plus an across gap and a down gap). Write the largest of the three numbers inside the circle. Continue this until all the circles are filled in.

5.) When you get to the bottom right circle, the number in the circle is the optimal alignment score.

6.) To determine the traceback path, start in the bottom right circle. Determine which of the three arrows had the highest score and draw an arrow in the opposite direction parallel to that arrow. Keep repeating that process for each circle until you end up back at the upper left circle. You should now have a path leading through the matrix from the bottom right circle to the top left circle.

7.) Convert the path to an alignment by aligning the corresponding letters in any diagonal move. Write out the alignment right-to-left. Put the top sequence on top, and the vertical sequence on the bottom. For horizontal moves, insert a gap (a dash) in the bottom sequence. For vertical moves, insert a gap in the top sequence.

