

Ps5: Needleman-Wunch DNA analysis programs

Assignment Description:

Ps5 tasked us to create a program that checks for the similarity of 2 strings using the Needleman-Wunch algorithm to obtain the edit distance for 2 strings of DNA. This edit distance is based on penalty values from a 2D array of integers that follows the following rules:

- 2pt penalty for gap
- 1pt penalty for different characters
- 0pt penalty for same character

This assignment is able to determine the edit distance for string of the ecoli genome of up to 28000 nucleotide pairs. The program runs until the array hits [0][0] where it checks the value in [0][0] which is the edit distance for the 2 strings.

Key Concepts and Algorithms:

The key algorithm for the assignment is the Needleman-Wunch algorithm. This algorithm follows the following procedure for an NxM matrix.

- Start with $[N][M] = 0$
- Add 2 to the previous element for the right column and bottom row.
- For the remaining columns and rows choose the smallest penalty:
 - $[N][M+1]$ or $[N+1][M] + 2$ to add a gap
 - $[N+1][M+1] + \text{penalty value}$

This algorithm produces the edit distance and the aligned strings.

What I learned in this assignment:

I learned how the Needleman-Wunch algorithm works and how to take a mathematical algorithm and adapt it into computer code. This assignment taught me how to take algorithms that I use in my math classes and figure out how to implement them in a program to compute the result automatically.