Faculty : Information and Communications Technology
Students : ICT1008 Data Structures and Algorithms – Group 18
Ang Zi Yan (1802851), Lai Wei Fang Ted(1802183),
Chua Woi Zhao (1801961), Yap Yong Sheng (1801233)

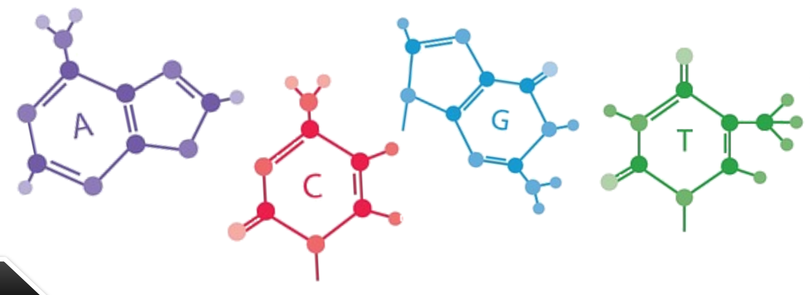
Sequence alignment for similarity in biological data.

Background

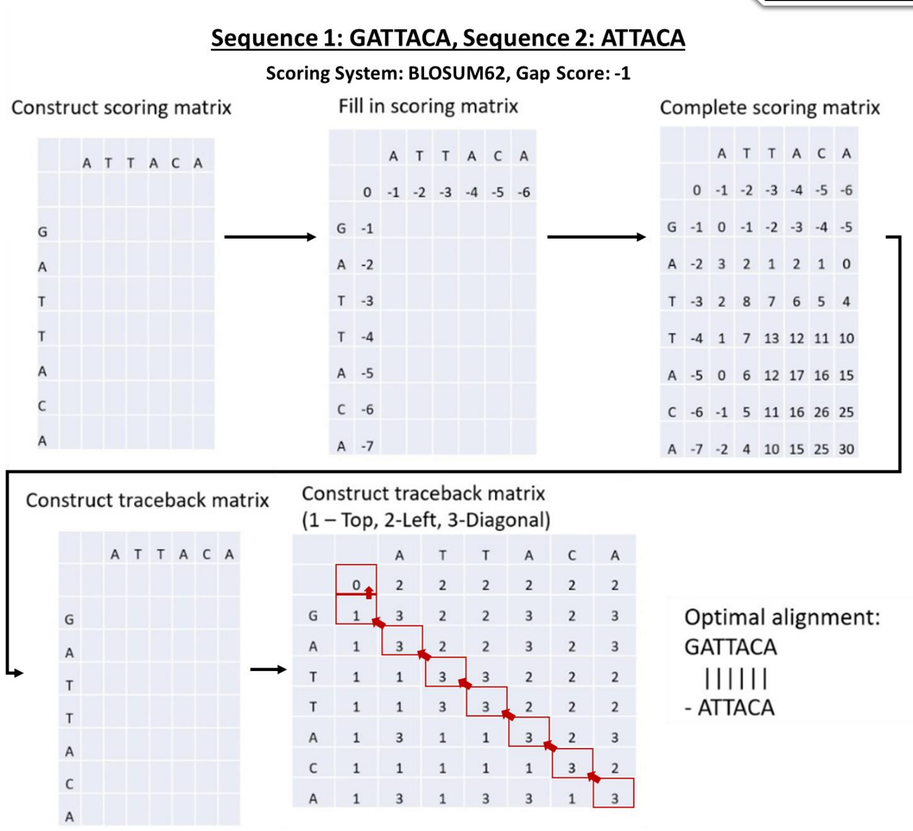
The basic concept involves pattern recognition or pattern matching for DNA, RNA and protein sequences. To put things into perspective, when a new gene is discovered, the biologists use information available from previously known genes to find some characteristics of the new gene. There are different algorithms which can be utilized for sequence alignment and our team have chosen the Needleman- Wunsch algorithm.

Abstract

Every living organism's cells have their own Deoxyribonucleic Acid (DNA) that carries genetic information. Using algorithms for pattern recognition and matching, biologists can learn more about new DNA, RNA and Protein sequences by comparing them to existing known sequences. An optimal algorithm will be used to provide the users with analysed information of the sequences, so as to aid in learning more about a new sequence and its characteristics.



Our Algorithm



Needleman-Wunsch algorithm steps.

Time Complexity: **O(mn)**.

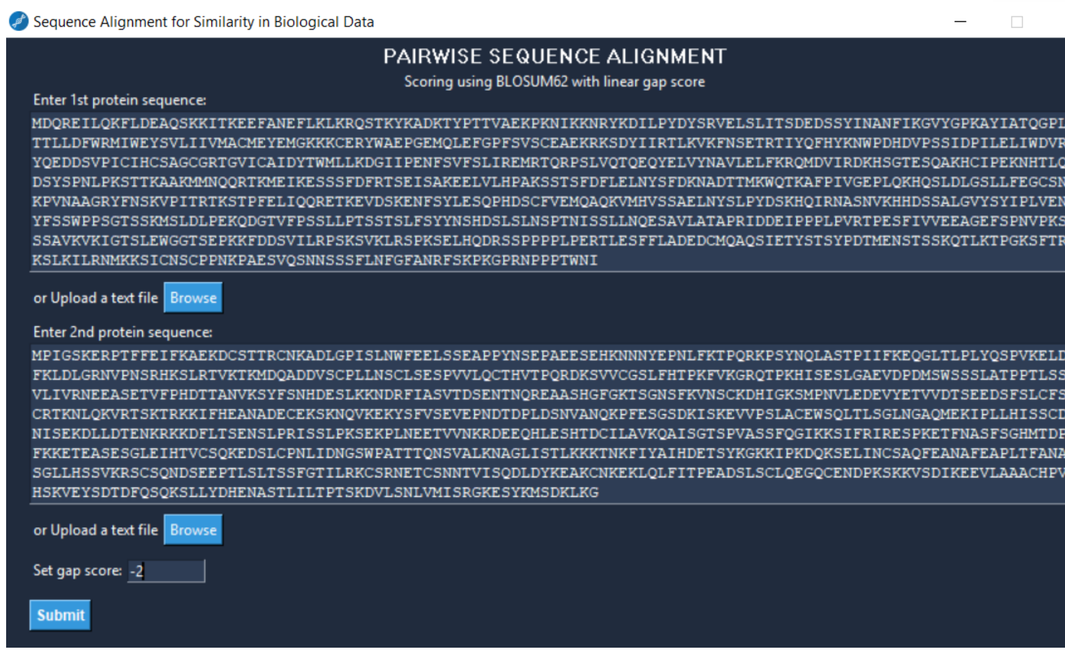
The Needleman–Wunsch algorithm is an algorithm developed by Saul B. Needleman and Christian D. Wunsch and published in 1970, commonly used in bioinformatics to globally align protein or nucleotide sequences with the highest quality of alignment. The algorithm uses dynamic programming to compare the sequences, essentially divides a large problem (e.g. the full sequence) into a series of smaller problems (one pair of amino acid) and uses the solutions to the smaller problems to reconstruct a solution to the larger problem.

Assessing the relationships between two sequences can be done by counting the number of identical and similar amino acids. The number of identical and similar amino acids may then be compared to the total number of amino acids in the protein, giving the percentage of sequence identity and similarity.

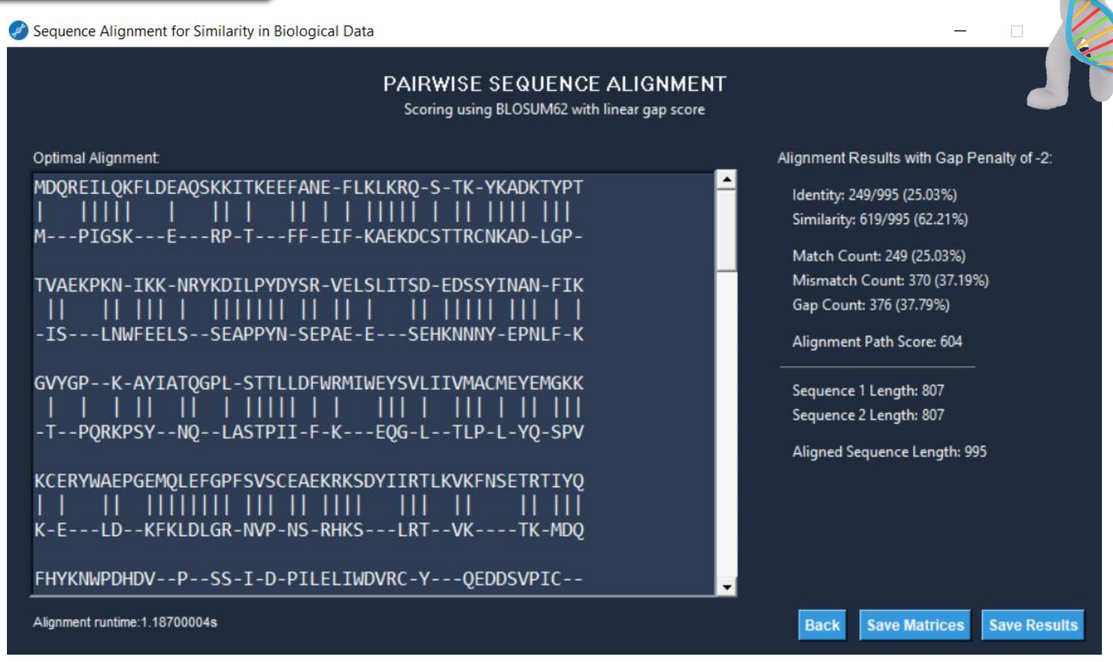
Algorithm Walkthrough

1. Build the first matrix for containing the scores.
2. For each pair of sequence, lookup the BLOSUM62 matrix to get the match/mismatch score.
3. Build the traceback matrix to get the optimal alignment
4. Based on the scoring allocated, decide if an indel (gap) or a match/mismatch would be optimal.
5. Build the optimal alignment based on the traceback matrix.

Our Application



Program GUI – Main Screen.



Program GUI – Results Screen.

www.tinyurl.com/seqaligner