WARSAW UNIVERSITY OF TECHNOLOGY
Faculty of Mathematics
and Information Science

Introduction to Bioinformatics

Project of
Needleman-Wunsch Algorithm
(Global sequence alignment)

Done by:

Amir Ali
317554

MSC Data Science

9 November 2022

Outline

1. Abstract

In this project, we implement the Needleman Wunch Algorithm(global sequence alignment0 from scratch. We took a dataset from the NCBI database in FASTA format. The dataset contains information on Homologous genes alignment and Insulin sequence alignment. And we experiment with this data into two scoring functions and get the optimal alignment and score, respectively.

2. Problem Statement

Global alignment, you should basically determine how one big sequence relates to another; you're comparing two giant sequences to each other, and it was developed the reason.

3. Introduction

Needleman Wunch Algorithm, called Global sequence alignment, was developed by South Needleman and Christian Wunch from Chicago and published in 1970 [1]. This algorithm can efficiently find the optimal alignment for any two sequences given a user-specific scoring function. It was widely used in the early research on protein sequence analysis.

However, as more and more sequences became known, limitations of aligning two sequences globally were noticed first, with more and more proteins sequenced. Researchers have found that those two functionally related proteins might differ significantly in their whole sequences but share similar critical functional domains. [2]

The sequence fragment of the functional part might be very conservative across different proteins in the same protein family and determine the biological function.

4. Methodology

In this part, we will explain our methodology. We implement Needleman Wunch Algorithm in python from scratch, and the following step follows in this implementation:

- Firstly, we initialize the Needleman Wunch Algorithm matrix with the score.
- Secondly, we fill the Needleman Wunch Algorithm matrix with maximum scores.

- At last, Retrace the residues to ensure proper alignment.

Building a Needleman Wunch algorithm from scratch is a tough and complex challenge. Therefore, we split our tasks into small parts and then sum up all the small jobs to solve them.

We build five functions which are the following:

1. **global_sequence_alignmen:** Our main function basically returns all ideal alignments and alignment scores for two sequences that need to be aligned.

   Parameters that need to pass in that function:

   sequence_1: The value of the very first sequence

   sequence_2: The value of the second sequence for comparison

   match: Scores are adjusted for sequence positions that match.

   Mismatch: Scores are adjusted for sequence positions that mismatch.

   Gap: scores are adjusted for gaps in the sequence

   Returns tuple containing a table of pointers for all grids and pointer table.

2. **build_score_table:** In this function, we build a result table and traceback pointer table and return both in the end.
3. **get_score_of_cell:** By implementing this function we get the max points & arrow list for a given grid.
4. **build_traceback:** This function help to find all optimal sequence alignments using the arrow grid generated by the scoring step.
5. **final_result:** This function prints the sequence's results.

5. Result and Discussion

We implement Needleman Wunch Algorithm for two alignments: Homologous genes and Insulin Sequence. For reference, you will find two folders in the source code for each implementation. And in the end, we compare both sequences' results with different scoring functions.

Below you will see the table of results:

	Alignment Score (Scoring function 1)	Alignment Score (Scoring function 2)
Homologous gene sequences	-4183	-3023
Insulin sequence alignment	-1050	-342

Table 5.1: Results

The sequence of Homologous and Insulin is so big. It took so much time to execute the program. And for reference, you can also find the optimal alignment of each sequence in the output text for each scoring function.

6. Conclusion

Needleman Wunch Algorithm is a very interesting technique to compare two sequences and find optimal alignment and purpose to discover all feasible alignments having the maximum points. And it also provides flexibility to do an experiment on a different sequence, and by changing the parameters like a mismatch, match, and gap.

Reference:

[1] Saul B. Needleman, Christian D. Wunsch,A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology, Volume 48, Issue 3, 1970, Pages 443-453, ISSN 0022-2836, https://doi.org/10.1016/0022-2836(70)90057-4.

[2] Jararweh, Y., Al-Ayyoub, M., Fakirah, M. et al. Improving the performance of the needleman-wunsch algorithm using parallelization and vectorization techniques. Multimed Tools Appl 78, 3961–3977 (2019). https://doi.org/10.1007/s11042-017-5092-0