Warsaw University of Technology

bioinformatics

Implentation of Needleman-Wunsch Algorithm

Authors:
Abba Umar

October 31, 2023

Contents

1 Project description **2**

2 Method **2**
2.1 Needleman-Wunsch algorithm 2
2.2 Dataset 3
2.3 Result 3
2.4 Discussion 4

List of Figures

1 Needleman-Wunsch. UNI (2023) 2
2 Resulting Alignment (Scoring Function 1) 3
3 Resulting Alignment (Scoring Function 2) 3
4 Resulting Alignment (Scoring Function 1) 4
5 Resulting Alignment (Scoring Function 2) 4

1 Project description

The project aims to implement Python 3.5 the Needleman-Wunsch global sequence alignment technique and apply it to two different scenarios: aligning the protein sequences of human and hamster insulin and aligning homologous genes with a variation of about 10%. Because the scoring functions of the Needleman-Wunsch algorithm can be customized, we employed two distinct sets of scoring parameters for our alignments.

2 Method

2.1 Needleman-Wunsch algorithm

To implement the Needleman-Wunsch algorithm, a score matrix is created, with each item in the matrix representing the best alignment score that can be achieved for the two prefixes of the sequences up to that point. Using the following principles, the score matrix is filled in recursively, beginning in the top-left corner:

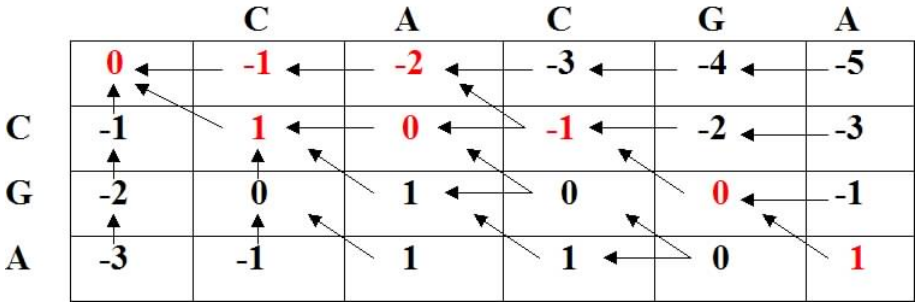

Figure 1: Needleman-Wunsch. UNI (2023)

- The score for aligning the empty prefix of one sequence with the empty prefix of the other sequence is 0.
- The score for aligning a character from one sequence with a character from the other sequence is equal to the match score if the characters are the same, or the mismatch score if the characters are different.
- The score for aligning a character from one sequence with a gap in the other sequence is equal to the gap penalty.

The optimal alignment is then found by tracing back the path through the score matrix that leads to the highest score.

2.2 Dataset

The sequences are downloaded from the NCBI database using Biopython, a popular bioinformatics package. The retrieved homologous gene sequences differ by around 10% and the protein sequences of human and hamster insulin. The retrieval of sequences required the NCBI API key, email for authentication, and the actual accession numbers.

I selected the gene "lysine acetyltransferase 5 (KAT5)," also referred to as "TIP60," from both Humans (Homo Sapiens) NCBI (2023a) and Sperm Whales (Physeter catodon) NCBI (2023b) because this particular gene exhibits a high degree of similarity between these two species, making it particularly intriguing.

2.3 Result

Homologous Gene Sequences: I downloaded two homologous gene sequences, which were stored as sequence1 and sequence2. The sequences were aligned using two distinct scoring functions.

Alignment with Scoring Function 1:

- Match Score: 4
- Mismatch Score: -2
- Gap Penalty: 0

```
Optimal Alignment:
AGTGGAGGGAGGGAAGATGGCGGAGGTGGTGAGTCCGGTGCCCGGGGCGGGGCGGAGGGAGCCAGGGGAGGTGGGTAGAGCCCGAGGCCCCCCAGTAGCCGACCCTGGCG-TCGCGCTGTCTCCCC
----------------ATGGCGGAGGTGGTGAGTCCGGTGCCCGGGGCGGGGCGGAGGGAGCCAGGGGAGGTGGGTAGAGCCCGAGGCCCCCCAGTAGCCGACCCTGGCGC-CGCGCTGTCTCCCC

Score: 7668
```

Figure 2: Resulting Alignment (Scoring Function 1)

Alignment with Scoring Function 2:

- Match Score: 2
- Mismatch Score: -1
- Gap Penalty: -2

```
Optimal Alignment:
AGTGGAGGGAGGGAAGATGGCGGAGGTGGTGAGTCCGGTGCCCGGGGCGGGGCGGAGGGAGCCAGGGGAGGTGGGTAGAGCCCGAGGCCCCCCAGTAGCCGACCCTGGCGTCGCGCTGTCTCCCCA
----------------ATGGCGGAGGTGGTGAGTCCGGTGCCCGGGGCGGGGCGGAGGGAGCCAGGGGAGGTGGGTAGAGCCCGAGGCCCCCCAGTAGCCGACCCTGGCGCCGCGCTGTCTCCCCA

Score: 3613
```

Figure 3: Resulting Alignment (Scoring Function 2)

Insulin Protein Sequences:

We downloaded the protein sequences of human and hamster insulin, which were aligned using the same two scoring functions.

Alignment with Scoring Function 1:

```
Optimal Alignment:
M-ALWMRLLPLLALL-ALW-GP-DPA-AAFVNQHLCGSHLVEALYLVCGERGFFYTPK-TRR--EAED-LQV-GQ-VELGGGPGA--GSLQ-PLALE---GSLQKRGIV-EQCCTSICSLYQLENYCN
MT-LWMRLLPLLALLV-LWE-PN-PAQ-AFVNQHLCGSHLVEALYLVCGERGFFYTPKS-RRGV--EDP-QVA-QL-ELGGGPGADD--LQT-LALEVAQ---QKRGIVD-QCCTSICSLYQLENYCN
Score: 368
```

Figure 4: Resulting Alignment (Scoring Function 1)

Alignment with Scoring Function 2:

```
Optimal Alignment:
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
MTLWMRLLPLLALLVLWEPNPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRRGVEDPQVAQLELGGGPGADDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN
Score: 166
```

Figure 5: Resulting Alignment (Scoring Function 2)

2.4 Discussion

The sequence alignments using the Needleman-Wunsch algorithm show how various scoring factors affect the alignment results. Notably, adjusting the score criteria may cause alignment variations that affect the final sequences that are aligned.

Through the alignment of homologous gene sequences, it was possible to determine that gaps were more preferred under Scoring Function 2, as longer gaps in the alignment were produced by a higher match score and a greater gap penalty. This is in line with the hypothesis that various scoring schemes may have an impact on how match/mismatch and gap penalties are distributed, which ultimately influences alignment.

However, the alignment of the insulin protein sequences revealed that altering the gap resulted in a distinct alignment while altering the match and mismatch scores yielded identical alignments. This could suggest that these extremely similar protein sequences could not have been aligned uniquely using the exact scoring parameters that were applied.

NOTE:

To visualize the result in your Jupyter Notebook you need to run:

”jupyter notebook –NotebookApp.iopub_data_rate_limit=1.0e10” in your terminal

References

NCBI (2023a), "Homo sapiens." URL https://www.ncbi.nlm.nih.gov/nuccore/NM_182710.3?report=fasta.

NCBI (2023b), "Physeter catodon." URL https://www.ncbi.nlm.nih.gov/nuccore/XM_007128188.3?report=fasta.

UNI (2023), "Cs: Needleman-wunsch." URL http://www.cs.uni.edu/~fienup/cs188s05/lectures/lec7_2-1-05.htm.