Implementation of Needleman-Wunsch algorithm

Paulina Kulczyk

November 8, 2023

Abstract

The Needleman-Wunsch algorithm is used in bioinformatics to find the optimal global alignment of two sequences (DNA, mRNA, and amino acids). In this article, we look at the problem of optimal alignment and discuss the mechanism of the Needleman-Wunsch algorithm. We also compare the alignment of the sequences of two exemplary homologous genes and the sequences of human insulin and hamster insulin with the use of this algorithm.

1 Problem description

The Needleman-Wunsch algorithm (Needleman and Wunsch (1970)) is used to find the optimal alignment of two sequences. The sequence alignment mechanism is used in bioinformatics to compare the sequences of nucleotides in nucleic acids and amino acids in proteins. This is done to identify regions showing structural similarity. This similarity can be a consequence of functional or evolutionary relationships. Therefore, the information obtained by the method of optimal alignment is essential for studying the evolution, function, and structure of proteins, organs, systems as well as whole organisms.

In the sequence alignment method, we can distinguish two approaches: global and local. The global approach involves matching entire sequences - from beginning to end. It is primarily applicable when we assume the overall similarity of the sequences being matched and when they are of similar length. In the local approach, only the best-matching subsequences of sequences are matched. This approach is more useful for sequences that do not have a high degree of similarity in general, which are assumed to contain similar sub-sequences. The Needleman-Wunsch algorithm makes it possible to find the optimal global alignment.

2 Methods

The Needleman-Wunsch algorithm is based on dynamic programming (it divides the problem into smaller subproblems and then uses the solutions of the smaller subproblems to solve the entire problem). As an input, this algorithm takes two sequences to be compared to each other. The letters (symbols) contained in the sequences form the alphabet of sequences. We consider finite sequences:

- $\sum$ - alphabet (ex. $\sum = \{A, G, C, T\}$)

- $X, Y \in \sum^*$ (ex. X =ACAAT)
- $|X| = m$ - length of X
- $|Y| = n$ - length of Y

In addition, we introduce a symbol that denotes a gap: "-".
Alignment is each pair (X', Y') where $X', Y' \in (\sum \cup "-")^*$ that meets: $|X'| = |Y'| = k$.
To find the best global match, we use a scoring function:

$$score(x, y) = \begin{cases} 1, & \text{if } x = y \text{ (match)} \\ -1, & \text{if } x \neq y \text{ (mismatch)} \\ -2, & \text{if } x = - \text{ or } y = - \text{ (gap)} \end{cases} \tag{1}$$

(Note: we can set the values for match, mismatch and gap ourselves).
The quality of fit (X', Y') we define as:

$$quality = \sum_{l=1}^{k} score(X'_l, Y'_l), \tag{2}$$

where X'_l is l-th element of sequence X', Y'_l analogously.

3 Results

Global similarity of the substring $X_1, .._i$, $Y_1, \ldots,_j$ $(v_{i,j})$ is computing dynamically:

$$v_{i,j} = max \begin{cases} 0, & \text{if } i = j = 0) \\ v_{i-1,j-1} + score(X_i, Y_j), & \text{if } i, j > 0 \\ v_{i-1,j} + score(X_i, -), & \text{if } i > 0 \\ v_{i,j-1} + score(-, Y_j), & \text{if } j > 0 \end{cases} \tag{3}$$

$v_{|X|,|Y|}$ is the quality of the optimal alignment, and enables its reconstruction - we can trace the path from which the maximum values comes from. To do that we start from last cell of matrix that was filled with values $v_{i,j}$ (where $v_{i,j}$ is cell on i-th row and j-th column). Symbols from two sequences are aligned if maximum value comes from neighboring diagonal cell, there is a gap in sequence 1 if maximum value comes frome above cell, similarly, there is a gap in sequence 2 if maximum value comes from left.

4 Alignment examples with the Needleman-Wunsch algorithm

In our work at first, we make an alignment of two homologous genes (differing by around 10%). To do this we choose the creatine kinase B gene from the NCBI database (which is located at chromosome 14, locus GRCh38.p14) and align it to the rat creatine kinase gene. With the use of classic values of match, mismatch, and gap (that is 2, -1, -2) we get the alignment with the quality score: 273. Then we change the values of mismatch to

-2 and the gap to -1 and we get the quality score: 317. So it seems that if the gaps don't bother us so much, it's better to choose a match with the changed values of the variables.

Next in our work we decide to compare the alignment of amino acids of human and hamster insulin proteins (insulin is a protein which aim is to decrease the level of sugar in blood). Now we write out the alignments since the amino acid sequences of these proteins are much shorter than those of the earlier genes.
With basic values of scoring function we get:
[0, 49] ----------FVNQHLCGSHLVEALYLVCGERGFFY
[0, 49] MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY
[50, 99] TPK----------S----GIVDQCCTSIC
[50, 99] TPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIC
[100, 110] SLYQLENYCN
[100, 110] SLYQLENYCN
with the quality score: -19.
Then we modify values like before to mismatch = -2, gap = -1 and we get:
[0, 49] ----------FVNQHLCGSHLVEALYLVCGERGFFY
[0, 49] MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY
[50, 99] TPK----------S----GIVDQCCTSIC
[50, 99] TPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIC
[100, 110] SLYQLENYCN
[100, 110] SLYQLENYCN
with quality score: 39.
In this case, we see that nothing has changed in the matching sequence, but the different values of the quality function with the selected variable values tell us that there are more gaps than mismatches in this sequence.
The results of this experiment prompt us to see what would happen if we didn't care about gaps, but we are very disturbed by mismatches (set gap = 0, mismatch = -4):
[0, 49] ----------FVNQHLCGSHLVEALYLVCGERGFFY
[0, 49] MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY
[50, 99] TPK----------S----GIV-DQCCTSI
[50, 99] TPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVE-QCCTSI
[100, 111] CSLYQLENYCN
[100, 111] CSLYQLENYCN
We can see that the mismatched letters have been replaced with gaps. Now the quality score is equal to 100.

5 Discussion

The Needleman-Wunsch algorithm is an easy-to-implement and fast tool for global matching of two amino acid or nucleotide sequences. The quality value allows us to determine how similar the sequences are to each other. With the ability to manipulate the values in the score function, we can dictate to the algorithm whether we prefer more mismatches or more gaps, and how much we expect a match. However, for those who do not deal with bioinformatics daily, selecting values of these arguments and interpreting the quality values can be a challenge. Hence, there is a need for a deeper discussion of these values and to issue some standards that will allow us to objectively answer whether sequences are similar or not.

It is worth mentioning that the Needleman-Wunsch algorithm has also found application in the computer stereo vision field. So it is reasonable to wonder whether it would not also be applicable in other domains (e.g. comparing two radio waves).

References

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.