An Approximation approach for Multiple Sequence Alignment Using Central Star Method with DNA Sequence Datasets

Gudipati Sai Krishna(M190241CS)
Nandi Dileep Kumar(M190437CS)

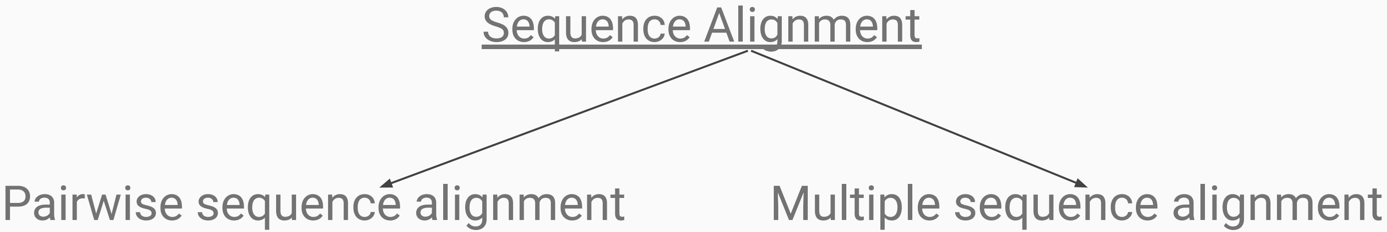
Contents

- Introduction
- Problem Definition
- Literature Survey
- Proposed Method
- Implementation
- Results
- Conclusion
- References

Introduction

Sequence Alignment:

- Arranging sequences of DNA,RNA or protein to identify regions of similarity.
- The similarity may indicate the functional.structural and evolutionary significance of the sequence.



Sequence Alignment

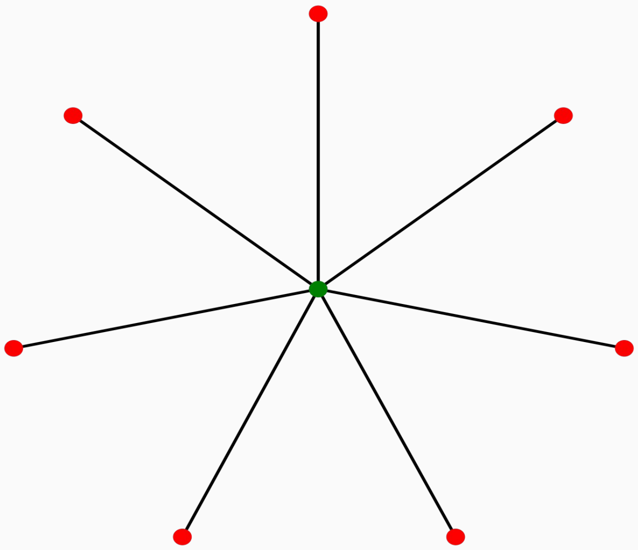
Pairwise sequence alignment

Multiple sequence alignment

Central Star Method

- It is an approximation solution for aligning multiple sequences.

- it uses heuristic methods rather than global optimization because identifying the optimal alignment between more than a few sequences of moderate length is computationally expensive and consumes more time.



Problem Definition

Problem Statement: Find a multiple sequence alignment of S that maximizes a similarity function or minimize a distance function.

Input: A set S of sequences.

Output: A new set of sequences of equal length formed after multiple sequence alignment

Literature Survey

Performing an **MSA** on a given set of DNA sequence and extracting maximum information from the alignment comprises a number of prominent steps:

- The selection of sequences.
- The choice of the scoring function used to compare sequences or sequence blocks.
- The application and optimization of this scoring function in compiling the alignment.

Literature Survey

Needleman-Wunsch algorithm:

- it is used to align protein or nucleotide sequences.
- It was one of the first applications of dynamic programming to compare biological sequences.
- The algorithm assigns a score to every possible alignment, and the purpose of the algorithm is to find all possible alignments having the highest score.It produces pairwise aligned sequences after calculating penalties and inserting gaps.

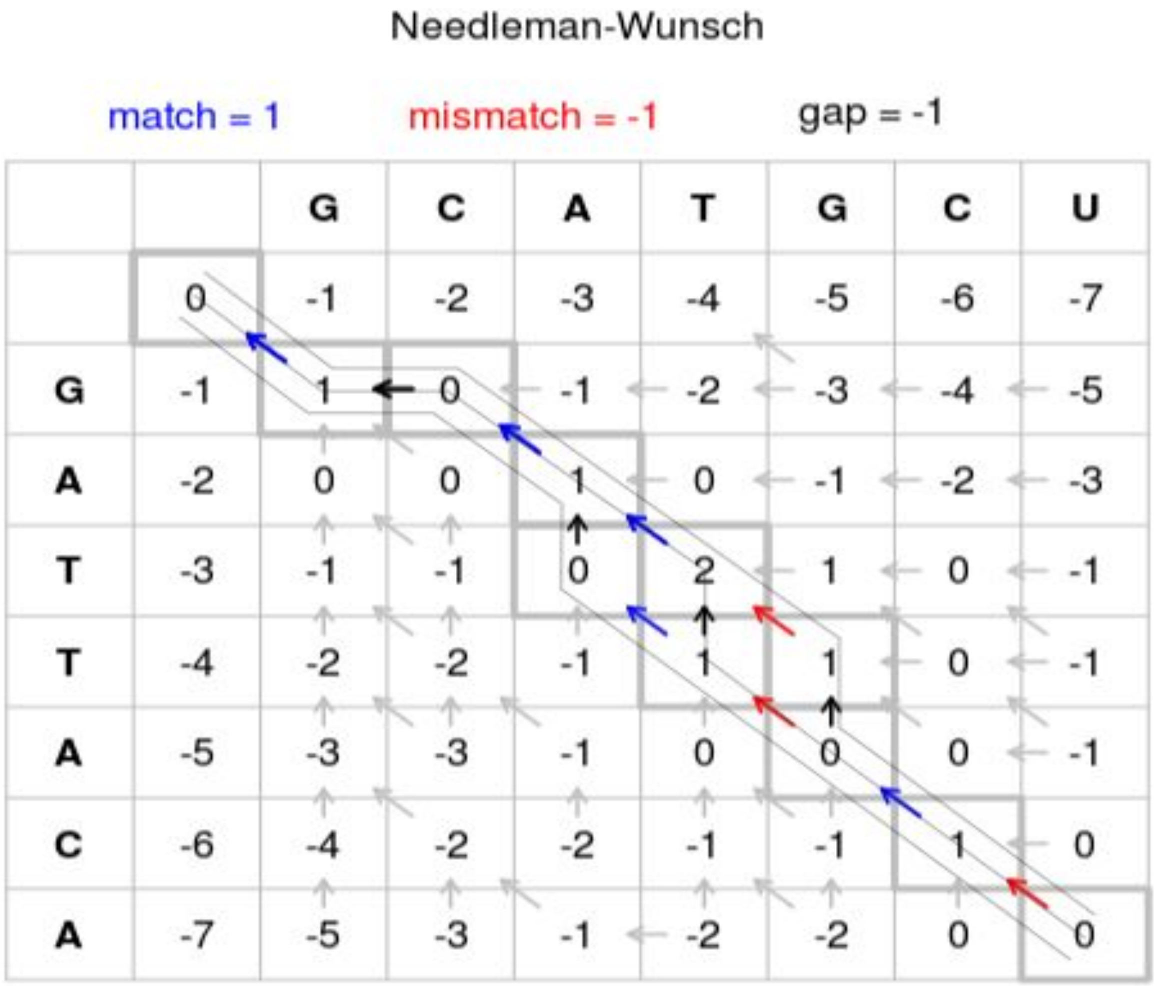
Example

Formula:

$$M_{i,j} = Maximum\left[M_{i-1,j-1} + S_{i,j},\ M_{i,j-1} + W,\ M_{i-1,j}+W\right]$$

Scoring System:

- Match: 1
- Mismatch: -1
- Gap: -1



Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Proposed Method

- Given a family {S1,S2,....Sk} of k sequences, such that the sequences are similar to each other, we would like to find out the common characteristics of this family.

- A multiple alignment is a new set of sequences {S’1 ,S’2 ,....S’k } such that:
  - ➔ All the strings in are of equal length. We denote this length by l.
  - ➔ Each S’i was generated from Si by inserting spaces.

Proposed Method

ALGORITHM:

A multiple alignment of M with sum of pair distances at most twice that of the optimal alignment of S.

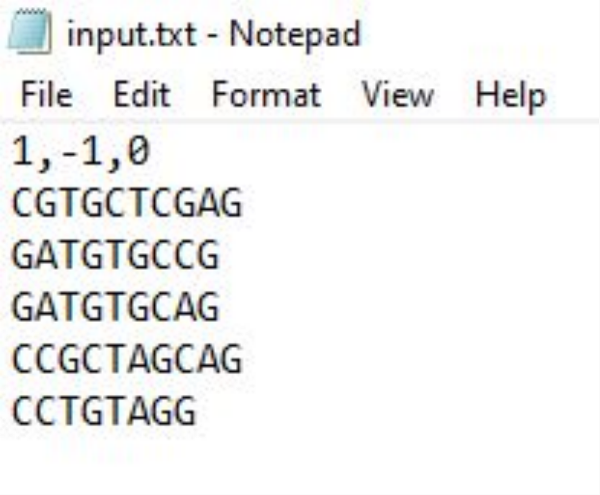
1. Find for all i,j
2. Find the center sequence Sc which minimizes
3. For every choose an optimal alignment Sc and Si.
4. Introduce spaces into Sc so that the multiple alignment M satisfies the alignment found in step 3.

Implementation

- Match,Mismatch and Penalty scores along with the list of k sequences are taken as input from the input file
- Calculate the pairwise alignment between the pairs of sequences using NW Algorithm and store its results along with the sequences in a 2D-matrix.
- Now,Compute $D(S_i,S_j)$for all i,j.
- Find the center sequence S_c which minimizes $\Sigma^k_{i=1}D(S_c,S_i)$.
- Now,Align the S_c with the remaining sequences.
- Merge all the pairwise alignments to make multiple alignment which is consistent and store it in a output file.

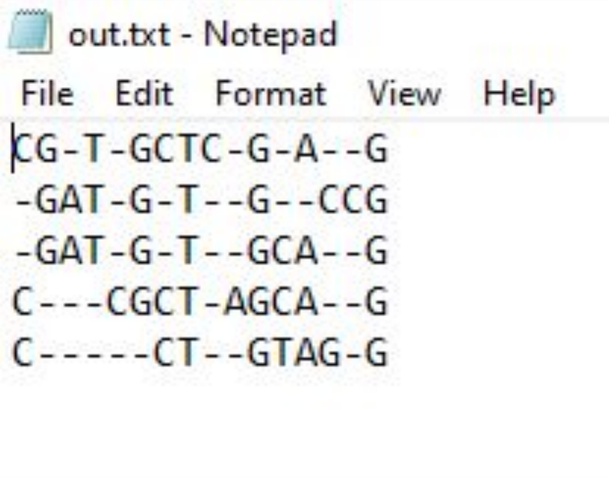
Results

INPUT



```
1,-1,0
CGTGCTCGAG
GATGTGCCG
GATGTGCAG
CCGCTAGCAG
CCTGTAGG
```

OUTPUT



```
CG-T-GCTC-G-A--G
-GAT-G-T--G--CCG
-GAT-G-T--GCA--G
C---CGCT-AGCA--G
C-----CT--GTAG-G
```

Conclusions

- An MSA can be observed as a representation that offers a unified picture of sequence similarity by averaging out matched residues that perhaps cannot be reliably matched over the entire lengths of the sequences.
- This is because of evolution, mutations, insertions, and deletions of sequence fragments.
- Given these difficulties, building a reliable MSA for a query set of sequences is an overwhelming task.
- In this unit it has been made strong that the increased attention to multiple sequence alignment methodology has ensued in recent developments regarding most of its facets.

References

- An algorithm of multiple sequence alignment based on consensus sequence searched by simulated annealing and star alignment.

- https://ieeexplore.ieee.org/document/7344909

- https://www.kaggle.com/thomasnelson/working-with-dna-sequencedata-for-ml

“Thank You For Your Attention”

- Gudipati Sai Krishna
- Nandi Dileep Kumar