

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

Gudipati Sai Krishna(M190241CS)¹ & Nandi Dileep Kumar(M190437CS)²

Abstract—Multiple sequence alignment is the assignment of recognizing evolutionarily or structurally related positions in a collection of different DNA sequences. That implies it can tell us approximately the evolution of the life forms, it can see which locales of a quality are prone to change and which can have one buildup substituted by another without modifying function. It is one of the foremost basic topics of bioinformatics research, which has noteworthy effects on analyzing quality and the relationship of species, evolution, comparing quality and conservative components in regulatory regions. Though the multiple sequence alignment issue has been considered for a few decades, numerous advanced ponders have demonstrated critical development in refining the precision or scalability of multiple sequence alignment algorithms.

I. INTRODUCTION

Sequence alignment could be a way of organizing sequences of DNA, RNA or protein to recognize regions of similarity and is made to align the whole sequence. The closeness may demonstrate the functional, structural and developmental noteworthiness of the sequence. The sequence alignment is made between a known sequence and unknown sequences. The known sequence is called the reference sequence and the unknown sequence is called the query sequence.

Pairwise sequence alignment

Pairwise sequence alignment strategies are utilized to find the best-matching piece-wise (local or global) alignments of two query sequences. Pairwise arrangements can be utilized between two sequences at a time, but they're effective to calculate and are frequently utilized for strategies that do not require extraordinary exactness. The three essential strategies of making pairwise arrangements are dot-matrix methods, dynamic programming, and word methods. In spite of the actual fact that every strategy has its individual strengths and shortcomings, all three pairwise methods have trouble with exceedingly repetitive sequences of low information content - particularly where the quantity of repetitions contrast within the 2 sequences to be aligned.

Multiple sequence alignment

Multiple sequence alignment is an extension of pairwise alignment to include over two sequences at a time. Multiple alignment methods try and align all of the sequences in a very given query set. Multiple alignments are often employed in identifying conserved sequence regions across a gaggle of sequences hypothesized to be evolutionary related[1]. Such conserved sequence motifs is employed in conjunction with structural and mechanistic information to locate the catalytic active sites of enzymes. Alignments are wont to aid in establishing evolutionary relationships by constructing phylogenetic trees. Multiple sequence alignments are computationally dif-

ficult to supply and most formulations of the matter result in NP-complete combinatorial optimization problems.

II. PROBLEM DEFINITION

Central Star Method 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[2].

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.

III. LITERATURE SURVEY

Many MSA techniques accomplish global alignment and match sequences over their complete lengths. Difficulties with this approach can arise when sequences that are only homologous over local regions are matched. In such situations, global alignment techniques might be unsuccessful to identify extremely similar internal regions because these may be dominated by divergent stretches and high gap penalties normally required to achieve proper global matching. Moreover, many biological sequences are modular and show shuffled domains, which can extract a global alignment of two complete sequences meaningless. The occurrence of varying numbers of internal sequence repeats can also strictly limit the applicability of global methods. Generally, when there is a large difference in the lengths of two sequences to be compared, it is desirable to take account of local alignment techniques in the analysis. To address these problems, Smith and Waterman developed a so-called local alignment technique in which the most similar regions in two sequences are carefully chosen and aligned. For multiple sequences, the main automatic methods include the Central Star Method Gibbs sampler. These local MSA programs often perform well when there is a clear block of ungapped alignment common to all of the sequences, but perform poorly under moderate gap requirements and show inferior results over general sets of test cases when compared with global methods.

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

1. The selection of sequences.
2. The choice of the scoring function used to compare sequences or sequence blocks.

3. The application and optimization of this scoring function in compiling the alignment.

A. Selecting the Sequences

An MSA can be equivocal when an arrangement set contains sequences that are not homologous. In a perfect world, all the sequences should be homologous, but practically, it is difficult to guarantee that. It should be stressed that most MSA schedules will create an arrangement indeed in the case of naturally irrelevant groupings, which can allow rise to wrong suggestions with respect to the proteins' structure or function[1]. The most common method to create set of sequences around a given query sequence is to utilizing homology searching technique to scour sequences in public sequence databases

B. The Scoring Function

The scoring function is the formalization of the organic information utilized in aligning the sequences. In reality, it should contain all available knowledge about structural, functional, and evolutionary aspects, so that the scoring function gauges the biological reality. In practice, however, this information is often not available or cannot be formalized numerically. Although each cross-comparison of a residue between two sequences should in reality be calculated individually based on its structural and functional context.

C. Applying the Scoring Function

Apart from being a fundamental genetics challenge, MSA is also a computationally intense problem, which means that for all but the smallest data sets of less than 10 sequences, an exact solution is not realistic. Algorithms that perform simultaneous alignment over a multidimensional search matrix, where each sequence in the MSA represents an extra dimension, come closest to an exact solution. The most populous class of algorithms is that of progressive MSA methods. The progressive strategy infers that an algorithm for pairwise sequence alignment is repetitively utilized in a step wise fashion until all the given sequences are aligned[1]. In the huge majority of progressive methods the Dynamic Programming (DP) strategy is approved. The DP procedure ensures that, given an amino acid exchange matrix and gap penalty values, the most noteworthy scoring or ideal pairwise arrangement is calculated. The dynamic arrangement technique reuses the pairwise DP calculations in a "greedy" way; i.e., alignments formed during the progression towards the ultimate MSA cannot be changed any longer. The most contrast between the accessible DP based strategies is the way in which the data of aligned blocks of sequences is represented. Whereas early strategies utilized agreement arrangements to speak to arrangement squares, current strategies all utilize a profile formalism to speak to the data in an MSA. Later improvements in multiple alignment strategies have primarily concentrated on delicate and ideal models to represent MSA information. A class of practises that are able to revisit and optimize is that of iterative multiple alignment techniques. Iterative techniques attempt to enhance

the alignment quality by gleaning information from a multiple alignment assembled in an earlier round, which is then useful in a next round to improve the alignment according to a given scoring scheme. Another classes of alignments is stochastic alignments, where probabilistic frameworks such as hidden Markov models and Bayesian networks have been attempted. Other techniques based on fast computational techniques such as suffix trees and fast Fourier transforms (FFT) and approximation algorithms like Central Star Method.

D. Needleman-Wunsch algorithm

The Needleman-Wunsch algorithm is an algorithm used in bioinformatics to align protein or nucleotide sequences. It was one of the first applications of dynamic programming to compare biological sequences. The algorithm essentially divides a large problem into a series of smaller problems, and it uses the solutions to the smaller problems to find an optimal solution to the larger problem. It is also sometimes referred to as the optimal matching algorithm and the global alignment technique. The Needleman-Wunsch algorithm is still widely used for optimal global alignment, particularly when the quality of the global alignment is of the utmost importance. 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.

E. Central Star Method

Central Star Method 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[2].

IV. PROPOSED METHOD

Given a family $\{S_1, S_2, \dots, S_k\}$ of k sequences, such that the sequences are similar to each other, we would like to find out the common characteristics of this family[2]. Aligning each pair of sequences from family separately, often does not reveal this common information. A multiple alignment is a new set of sequences $\{S'_1, S'_2, \dots, S'_k\}$ such that:

- All the strings in are of equal length. We denote this length by l .
- Each S'_i was generated from S_i by inserting spaces.

The brief idea of how the center star method of approximation works as follows:

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

1: Find $D(S_i, S_j)$ for all i, j

2: Find the center sequence S_c which minimizes $\sum_{i=1}^k D(S_c, S_i)$

3: For every $S_i \in S - \{S_c\}$ choose an optimal alignment S_c and S_i .

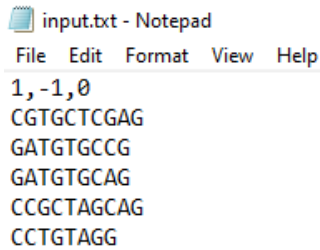
4: Introduce spaces into S_c so that the multiple alignment M satisfies the alignment found in step 3.

V. 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 $\sum_{i=1}^k D(S_c, S_i)$.
- Now, Align the S_c with the remaining sequences other than S_c .
- Merge all the pairwise alignments to make multiple alignment which is consistent and store it in a output file..

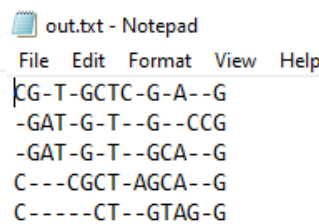
VI. RESULTS

The below images are the result obtained by implementing center star method of approximation:



```
input.txt - Notepad
File Edit Format View Help
1, -1, 0
CGTGCTCGAG
GATGTGCCG
GATGTGCAG
CCGCTAGCAG
CCTGTAGG
```

Fig. 1. Input



```
out.txt - Notepad
File Edit Format View Help
CG-T-GCTC-G-A--G
-GAT-G-T--G--CCG
-GAT-G-T--GCA--G
C---CGCT-AGCA--G
C-----CT--GTAG-G
```

Fig. 2. Output

VII. 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. So the sequence alignment inconsistencies can well arise under divergent evolution. 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. The increased focus has also led to the construction of new benchmark databases and novel evaluation protocols. More developments will be significantly dependent on the integration and representation of biological knowledge in new quality criteria. There are now a multitude of high-quality MSA techniques, each with particular strengths and weaknesses. Increased sensitivity could flourish as a result of new consensus protocols to utilize the combined power of the techniques, or new techniques to determine the kind of alignment problem at hand and then invoke the most appropriate method or combination of methods available. In the meantime, however, it remains important for the end user

to run a combination of different MSA methods to optimize the biological information derived from a set of sequences

REFERENCES

- [1] An algorithm of multiple sequence alignment based on consensus sequence searched by simulated annealing and star alignment
- [2] <https://ieeexplore.ieee.org/document/7344909>
- [3] <https://www.kaggle.com/thomasnelson/working-with-dna-sequence-data-for-ml>