Grenze International Journal of Engineering and Technology, Special issue

A Survey on Recent Research in Algorithms for Sequence Alignment

Chikku George[1], Sreelakshmi K.Sivadas[2], Sidharth V[3], Bincy Babu[4] and Dhanya Sudarsan[5]
[1-5]Dept. of Computer Science, Muthoot Institute of Technology & Science, Puthencruz
Email: chikkugeorge24@gmail.com, dhanyasudarsan127@gmail.com

***Abstract*—Bioinformatics is a field of research which uses computational technologies efficiently to analyze genomic data for genetic studies. The main applications of genomic studies come in similarity study, disease conformation, ancestral identification, genetic therapeutics, medicinal testing etc. All these applications need the comparison between the genetic sequences. So the algorithm development for sequence alignment is a major research area of interest for bioinformatics researchers. The paper conducts a survey on the recently proposed sequence alignment algorithms, focuses on identifying the pros and cons in each algorithm and also tries to suggest the best algorithm for each application.**

***Index Terms*— Sequence alignment; BLAST; FASTA; Tcoffee;Shifted Hamming Distance; MTRAP; MUSCLE; Dynamic Weight Assignment Algorithm.**

I. INTRODUCTION

Bioinformatics is an ever growing field of research that provides various software tools for studying the biological data particularly DNA, RNA and protein sequences. Bioinformatics combines Computer Science, Mathematics, Statistics and Engineering so that the biological data can be analyzed and interpreted in a creative manner. The comparison of genetic and genomic data, analysis of evolutionary aspects of molecular biology can be easily carried out by using the Bioinformatics tools.

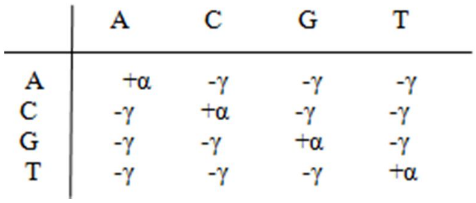
Sequence alignment algorithms are designed for either comparing a sequence with the database of sequences or for generating multiple sequence alignment, it can be divided into Pair wise Sequence Alignment and Multiple Sequence Alignment in which pair wise sequence alignment finds the similarity of two sequences by using the highest score whereas multiple sequence alignment deals with the similarity of multiple sequences. Pair wise sequence alignment falls into two major categories like Global and local alignment. The global alignment compares one whole sequence with other entire sequences while local method tries to align a subset of sequence to subset of other sequences.

It is possible to align smaller sequences by hand. But extremely large sequences can’t be aligned by the human. So algorithms are developed to produce the sequence alignment with high quality. For the alignment of two DNA sequences, a scoring matrix is used. The elements of scoring matrix are depicted in fig 1.

The scoring matrix will give a score α to all matches and a penalty score γ to all mismatches.

Since Sequence alignment is the most required process for most of the bioinformatics research a lot of work is being conducted in this field. Other than legacy sequence alignment algorithms like BLAST and FASTA many algorithms has been developed recently. Each one is having its own merits and demerits, so it is better to find the best algorithm suiting each application.

Grenze ID: 01.GIJET.3.2.26



	A	C	G	T
A	+α	-γ	-γ	-γ
C	-γ	+α	-γ	-γ
G	-γ	-γ	+α	-γ
T	-γ	-γ	-γ	+α

Fig 1: Scoring Matrix

Our paper gives an overview of the proven algorithms and conducts a survey on the recent research in sequence alignment algorithms, identifies the pros and cons in each algorithm and tries to find out the best algorithm suiting each application.

II. LEGACY SEQUENCE ALIGNMENT ALGORITHMS

There exist different proven methods for sequence alignment such as Needleman-Wunsch algorithm, Smith-Waterman algorithm, FASTA, and BLAST (Basic Local Alignment Search Tool). Needleman –Wunsch is a global pair wise sequence alignment that uses dynamic programming to align the sequences. This algorithm is expensive in terms of time and space and hence is not preferred for long sequences. Smith-Waterman algorithm is a local pair wise sequence alignment that uses highest scoring pairs to find a match between the two sequences. The sequence search performed by this algorithm is relatively slow and it produces more complicated local alignment matches.

FASTA is a heuristic approach to sequence alignment, which uses certain assumptions and approximations. It works in 3 steps namely identify very short exact matches, extend the best short hits and optimize the best hits with some form of dynamic programming. FASTA can be used to find the functional and evolutionary relationships between sequences as well as helps to identify the members of gene families.

BLAST provides only the high scoring pairs which score above the threshold value. This algorithm can be implemented in 3 steps namely compiling a list of high-scoring words, scanning the database for hits and extending hits. This algorithm is simple, robust and fast. BLAST is more time-efficient than FASTA. BLAST can be used for identifying species, locating domains, establishing phylogeny, DNA mapping, and comparison.

FASTA and BLAST are used for the database search where a quick alignment is necessary with lower accuracy.

III. RECENT RESEARCH ALGORITHMS

A. Shifted Hamming Distance

Shifted Hamming Distance is an edit-distance based filter that can quickly check whether the minimum number of edits including insertions, deletions, and substitutions between the two strings is smaller than a threshold value. This algorithm filters the string pairs with edit-distance greater than the threshold value and doesn't validate the string pairs with edit-distance smaller than the threshold value.

Shifted Hamming Distance algorithm provides high accuracy with high speed. But, as the edit-distance threshold increases the accuracy gets reduces. So this algorithm can be used for applications like disease identification where both accuracy and time is equally important.

B. MTRAP

MTRAP is a pair wise sequence alignment method by a new measure based on transition probability between two consecutive pairs of residues. Here we calculate the Q (quality) score which is defined as the ratio of the number of correctly aligned residue pairs in the test alignment to the total number of aligned residue pairs in the reference alignment. The Q score will take a maximum value of 1 when all pairs are aligned correctly and if none of the pairs are aligned then the Q score will have a minimum value of 0.

This algorithm provides higher accuracy compared to the existing alignment algorithms. So MTRAP will fit to applications like gene therapeutics which requires high accuracy but can compromise with time.

C. TCoffee

TCoffeee is a Tree-based Consistency Objective Function for alignment Evaluation. Tcoffee uses a set of local and global pair wise alignments between all of the sequences to be aligned. For each pair of the aligned residues in the library, a weight is assigned. By a simple process of addition the local and global alignment information is efficiently combined. If the two libraries contain similar pair, then it is merged in to a single entry with a weight equal to the sum of two weights. Otherwise, a new entry is created. Then the so called library is extended such that the two residues are aligned with residues from the rest of the sequences. A distance matrix is produced between all the sequences, which in turn creates a guide tree to group the sequences. The two closest sequences is aligned first using normal dynamic programming and so does for the next two closest sequences until all the sequences have been aligned.

Tcoffee is a simple, flexible and accurate algorithm for generating multiple alignments and is fast and relatively robust. This algorithm can have the disadvantage of weak scalability. A maximum of only 100 sequences can be aligned by Tcoffee without loss of accuracy. TCoffee can be used for the construction of evolutionary trees and protein engineering.

D. MUSCLE

MUSCLE is a Multiple Sequence Comparison by Log-Expectation. MUSCLE uses two distance methods; k-mer distance for unaligned pair of sequences and Kimura distance for aligned pair of sequences. Here a binary a tree is constructed by estimating the distance between the sequences using k-mer counting and clustered using UPGMA (Unweighted Pair Group Method with Arithmetic Mean).

An initial multiple sequence alignment is produced from the resulted binary tree. The initial guide tree is then reestimated using Kimura distance and reclustered using UPGMA to produce a second guide tree so that the second multiple sequence alignment is obtained.

Using the first and second multiple alignments, a new multiple alignment is produced by aligning the profile parameters such as residue frequencies and gap frequencies. If the second multiple alignment improves the sum of pairs, then the new alignment is kept and the old is discarded. Otherwise, the first alignment is used by deleting it.

MUSCLE creates alignments with average accuracy compared to the best current methods. This algorithm is used in applications like phylogenetic tree estimation, secondary structure prediction and critical residue identification.

E. Dynamic Weight Assignment Algorithms

The global alignment can be improved by having dynamic weight. Based on this, we can have two new methods for sequence alignment like Weighted Alignment and Diagonal Alignment

Weighted alignment

In weighted alignment, weights are given according to the position and direction of each cell.

A diagonal is drawn on the matrix which divides the matrix into two parts; upper and lower parts. In case of homogenous sequences, the same nucleotide will locate on the diagonal of the alignment matrix. Hence, low weight is assigned to the nucleotides that are neighbor to the diagonal. The horizontal and vertical movements are given lower weight respectively in the upper and lower part of the diagonal.

Diagonal alignment

In diagonal alignment, it will give different costs according to the position in the matrix alignment. The weights are calculated according to the location of the cells in relation to the diagonal. Two diagonals are drawn on the matrix which divides the matrix into three parts; upper part, lower part and between the diagonals. If the movement is closer to the diagonal then it has no penalty. No penalty is given in between the diagonals. Also, the penalty is not given to the horizontal and vertical movements respectively in the upper and lower part of the diagonal.

These methods suit for similarity study, ancestral identification kind of applications where high scalability is needed.

F. MAFFT

MAFFT is a multiple sequence alignment algorithm which includes the identification of Homologous regions using Fourier transform in which an amino acid sequence is converted to a sequence composed of volume and polarity values of each amino acid residue. A simplified scoring system is used to reduce the CPU time and increasing the accuracy of alignments even for sequences having large insertions or extensions as well as distantly related sequences of similar length. Two different heuristics, the progressive method (FFT-NS-2) and the iterative refinement method (FFT-NS-i), are implemented in MAFFT.

MAFFT algorithm overall CPU utilization time is very less compared to CLUSTALW and Tcofee.It also ensures a comparable accuracy. So the algorithm is most suitable for researchers who process limited computational capability.

G. DIALIGN

DIALIGN is an algorithm which combines global and local alignment features for pair-wise and multiple alignments of nucleic acid and protein sequences. Here all respective optimal pair-wise alignments are carried out. i.e., for each pair of input sequences, a continuous local fragment alignments with maximum total weight score is identified. A fragment is an un-gapped local pair-wise alignment, and the weight score of such a fragment is calculated based on a P-value where p is the probability of its random occurrence.
DIALIGN is proven to be highly accurate but takes consumes more time compared to other algorithms. To overcome this drawback the concept of parallelism is also incorporated with the algorithm. This parallel version of DIALIGN can reduce the program running time up to 97%.
Applications include the alignment distantly related sequences which is having some isolated local homologies, detection of regulatory elements by multiple alignment, phylogenetic studies and identification of signature sequences to detect pathogenic viruses.

IV. CONCLUSION

The paper provides an insight on the basic traditional algorithm for sequence alignment. A survey has been conducted on the seven recent innovations in sequence alignment algorithms, identified the pros and cons of each algorithm and explored the best suited algorithm for each application.

REFERENCES

[1] Robert C. Edgar, “MUSCLE: Multiple sequence alignment with improved accuracy and speed”, Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004

[2] Daniel Saad Nogueira Nunes, Mauricio Ayala-Rinc´on, “A Practical Semi-External Memory Method for Approximate Pattern Matching”, Published by Elsevier B.V

[3] Mohad. Faizal Omar, Rosalina Abdul Salam, Nura Abdullah. Multiple Sequence Alignment using Genetic Algorithm and Simulated Annealing, ieeexplore.ieee.org/ie15/9145/29024/01307828.pdf.IEEE.2004

[4] Dr. Mamta C. Padole, “Search Algorithm Used in FASTA" Conference Paper, May 2005

[5] Altschul, S.F. et al (1990), “Basic local alignment search tool", Journal of Molecular Biology, vol. 215,pp 403-410

[6] Hongyi Xin , John Greth , John Emmons , Gennady Pekhimenko, Carl Kingsford , Can Alkan and Onur Mutlu, “Shifted Hamming Distance: A Fast and Accurate SIMD -Friendly Filter to Accelerate Alignment Veri_cation in Read Mapping", Bioinformatics Advance Access published January 10, 2015

[7] Leila Alimehr, “The Performance of Sequence Alignment Algorithms”

[8] Toshihide Hara, Keiko Sato, Masanori Ohya, “MTRAP: Pair wise sequence alignment algorithm by a new measure based on transition probability between two consecutive pairs of residues”

[9] CeÂdric Notredame, Desmond G. Higgins and Jaap Heringa “T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment”

[10] Robert C. Edgar, “MUSCLE: Multiple sequence alignment with improved accuracy and speed”

[11] William J. Pearson and David J. Lipman, “Improved tools for biological sequence comparison”, *Proc. Natl. Acad. Sci. USA* Vol. 85, pp. 2444-2448, April 1988 Biochemistry

[12] Martin Schmollinger,Kay Nieselt,Michael Kaufmann and Burkhard Morgenstern ,”DIALIGN P: Fast pair-wise and multiple sequence alignment using parallel processors”, BMC Bioinformatics,2004.