Invention Journal of Research Technology in Engineering & Management (IJRTEM)
ISSN: 2455-3689
www.ijrtem.com Volume 2 Issue 5 ǁ May 2018 ǁ PP 13-21

Performance Improvement of BLAST with Use of MSA Techniques to Search Ancestor Relationship among Bioinformatics Sequences

[1,] Dr. Kamal Shah, [2,] Mr. Nilesh N. Bane
[1,] Professor & Dean- R & D TCET
[2]ME PART-II TCET

ABSTRACT: *BLAST is most popular sequence alignment tool used to align bioinformatics patterns. It uses local alignment process in which instead comparing whole query sequence with database sequence it breaks query sequence into small words and these words are used to align patterns. it uses heuristic method which make it faster than earlier smith-waterman algorithm. But due small query sequence used for align in case of very large database with complex queries it may perform poor. To remove this draw back we suggest by using MSA tools which can filter database in by removing unnecessary sequences from data. This sorted data set then applies to BLAST which can then indentify relationship among them i.e. HOMOLOGS, ORTHOLOGS, PARALOGS. The proposed system can be further use to find relation among two persons or used to create family tree. Ortholog is interesting for a wide range of bioinformatics analyses, including functional annotation, phylogenetic inference, or genome evolution. This system describes and motivates the algorithm for predicting orthologous relationships among complete genomes. The algorithm takes a pairwise approach, thus neither requiring tree reconstruction nor reconciliation*

KEYWORDS: *BLAST, MSA, CLUSTALW, ORTHOLOGS, PARALOGS, RSD, CLIQUE*

I. INTRODUCTION

Background: Bioinformatics is a field which is related to developing and improving methods for storing, organizing. Retrieving and analyzing biological data. An important work in bioinformatics is to create a tool to capture useful biological knowledge. Bioinformatics is a field which is related to developing and improving methods for storing, organizing. Retrieving and analyzing biological data. An important work in bioinformatics is to create tools to capture useful biological knowledge. It can be used in many areas of science and technology to process biological data. High configuration machines are used to process bioinformatics data at a highest rate. . Knowledge generation from biological data may involve study of areas like data mining, artificial intelligence, simulation, Genetic Algorithms, Fuzzy Logic and image processing. DNA patterns, RNA patterns, protein Patterns are some of the examples of Bioinformatics pattern. DNA patterns are of DNA sequences. Various functional structures such as promoters and genes, or larger structures like bacterial or viral genomes, can be analyzed using DNA patterns. Protein patterns are sequence of amino acid molecule connected to each other. It contains almost 16 type of amino acid molecule.
Protein patterns example:-

"MQKSPLMEKASFISKLFFSWTTPILRKGYRHHLELSDIYQAPSADSADHLSEKLEREWDREQASKKNP
QLIHALRRCFFWRFLFYGILLYLGEVTKAVQPVLLGRIIASYDPENKVERSIAIYLGIGLCLLFIVRTLLLH
PAIFGLHRIGMQMRTAMFSLIYKKTLKLSSRVLDKISIGQLVS..."

It is the large sequence of characters which represent their respective amino acid molecule. To extract information from these large patterns is necessary to match them with another pattern. Sequence alignment method has been using to capture similarity between these patterns. In bioinformatics, a sequence alignment is a way of matching the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences [1]. The sequences are represented in matrix form. Each row in matrix represents aligned sequences of amino acid residues. Spaces are introduced between the residues so that identical characters are come under same column. The sequence alignment process has two methods: based on completeness and based on number of sequences in used. Sequence alignment based on completeness of sequences deals with way of matching sequence i.e. whether whole sequences are matched to achieve the alignment of sequences or will it be mapped partially. Depending upon nature it can be categorized as GLOBAL and LOCAL mapping. Here are examples of global and local alignments. The global alignment looks for comparison over the entire range of the two sequences involved.

GCATTACTAATATATTAGTAAATCAGAGTAGTA
| | | | |
AAGCGAATAATATATTTATACTCAGATTATTGCGCG

As you can see only a portion of this two sequences can be aligned. By contrast, when a local alignment is performed, a small seed is uncovered that can be used to quickly extend the alignment. The initial seed for the alignment:

TAT

| | |

AAGCGAATAATATATTTATACTCAGATTATTGCGCG

And now the extended alignment:

TATA T ATTAGTA

| | | | | |

AAGCGAATAATATATTTATACTCAGATTATTGCGCG

Sequence alignment based on number of sequences in used deals with total number sequences involving alignment processive. it checks whether two sequences are aligned at a time or 'n' sequences are used to for sequence alignment. Depending upon No of sequences it can be categorized as 'Pairwise sequence alignment' and 'multiple sequence alignment 'These aligned sequences are matched based their match score which can calculate using scoring scheme. The scoring scheme is consisting on character substitution score (i.e. score for each possible character replacement) plus penalties for gaps. The alignment score is sum of substitution scores and gap penalties. The alignment score reflects "GOODNESS OF ALIGNMENT"

Protein substitution matrices are significantly more complex than DNA scoring matrices. PAM, i.e. "Point Accepted Mutation" family AND BLOSUM, i.e. "Blocks Substitution Matrix" family (BLOSUM62, BLOSUM50, etc.) are most commonly used substitution matrix for protein substitutions. BLOSUM62 is the matrix calculated by using the observed substitutions between proteins which have at most 62% sequence identity. Various algorithms have been used to align sequence but most popularly known algorithm is BLAST (Basic Local Alignment Search Tool). The BLAST accepts query sequence and weighted matrix as input and give output in various format such as HTML, Plain text etc. It follows local paired matching of sequences. It breaks input query sequence into small words (e.g. default 3 characters/per word). It will scan whole database for word matches. Extends all matches to seek high scoring alignments due to which it is highly sensitive to speed.

The classification of genes according to evolutionary relations is essential for many aspects of comparative and functional genomics. Evolutionary relations are often described as pairwise relations. Two genes that share a common ancestor are defined as homologs, while genes that are similar in sequence without a common origin are termed analogs. Homologs can be divided into several classes [1]:

orthologs, which originate from a speciation event

paralogs, which originate from gene duplication

xenologs, which originate from horizontal gene transfer.

Orthologs are valuable in numerous analyses, including reconstruction of species phylogenies, protein function inference, database annotation, and genomic context analysis.

Motivation: Suppose organization contains bioinformatics database containing 'n' number of sequences where 'n' is very large number. And they want to create ortholog clusters which highly random in nature using BLAST. Even though BLAST is most popular algorithm in terms of and even though it is fastest than Smith–Waterman algorithm it is less accurate in nature as well as it uses word by word local alignment matching process it may take unnecessary extra time which can be minimized. Using BLAST bioinformatics sequences can categorize into ORTHOLOGS AND PARALOGS which can be used to construct inheritance tree in given context.

Scope of the system: This system can be used to identify related sequences in bioinformatics database which can be used to identify homology among them. Along with that system aims towards increasing speed and accuracy of sequence alignment process. It can be used to predict relationship among two persons.

II. RESEARCH GAPS AND PROBLEM DEFINITION

Research Gaps: A number of techniques have been developed for Bioinformatics Sequence Alignment such as BLAST, MSA, smith-waterman, Needleman-Wunsch algorithm.

Needleman-Wunsch algorithm : This algorithm first take the two sequences and create a 2-dimensional array with the length of Multiply of the two sequence's length, each cell can be evaluated from the maximum of the three cells around it and at the same time keep a pointer of the maximum value to make the trace-back to get optimal solution the problem of these dynamic algorithms is that it take more time to fill all the matrix of the two sequences, although there are unused data but it must be found to help in filling all the cells in the matrix, so these algorithm take many times to make this computation

Smith-Waterman algorithm: The Smith-Waterman algorithm is a method of database similarity searching which considers the best local alignment between a query sequence and sequences in the database being searched. The Smith-Waterman algorithm allows consideration of indels (insertions/deletions) and compares fragments of arbitrary lengths between two sequences and this way the optimal local alignments are identified. Sequence similarity searches performed using the Smith-Waterman algorithm guarantees you the optimal local alignments between query and database sequences. Thus, you are ensured the best performance on accuracy and the most precise results - aspects of significant importance when you cannot afford to miss any information gained from the similarity search as e.g. when searching for remote homology. The Smith-Waterman algorithm being the most sensitive algorithm for detection of sequence similarity has however some costs. Time is a considerable disadvantage and performing a Smith-Waterman search is both time consuming and computer power intensive.

Basic Local Alignment Search Tool: BLAST (Basic Local Alignment Search Tool) also identifies homologous sequences by database searching. BLAST identifies the local alignments between sequences by finding short matches and from these initial matches (local) alignments are created. The BLAST algorithm is a development of the Smith-Waterman algorithm suggesting a time-optimized model contrary to the more accurate but time-consuming calculations of the Smith-Waterman algorithm BLAST (Basic Local Alignment Search Tool) also identifies homologous sequences by database searching. BLAST identifies the local alignments between sequences by finding short matches and from these initial matches (local) alignments are created. The BLAST algorithm is a development of the Smith-Waterman algorithm suggesting a time-optimized model contrary to the more accurate but time-consuming calculations of the Smith-Waterman algorithm

Problem Definition: BLAST is very popular tool to align protein sequence in Local Pair wise alignment method. The problem with BLAST is it can work best only for pair of single queries. IT may fail with complex queries as well as in large database while using isolated.
"if 'S' is candidate bioinformatics sequence set contains 'n' bioinformatics sequences find the optimal way of creating clusters of orthologs sequences and find parlogs among them" .

Research Objectives

- To find the efficient method or techniques for sequence alignment
- Methods should be such that it improves performance and accuracy of BLAST
- System can be able to identify orthologs and paralogs given query sequence from available datasets

III. PROPOSED METHODOLOGY

Materials and methods Technology
Software etc.
BLOSUM62: In bioinformatics, for sequence alignment of proteins BLOSUM matrix is used as substitution matrix. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments. BLOSUM matrices were first introduced in a paper by Henikoff S. and Henikoff J.G. They scanned the BLOCKS database for very conserved regions of protein families and then counted the relative frequencies of them and their substitution probabilities. Then, they calculated a log-odds score for each of the 210 possible substitution pairs of the 20 standard amino acids.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Multiple sequence alignment tools: Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied. Used to generate a concise, information-rich summary of sequence data. Sometimes used to illustrate the dissimilarity between groups of sequences. Alignments can be treated as models that can be used to test hypotheses. It uses two method dynamic and progressive alignment methods. Crustal has become the most popular algorithm for multiple sequence alignment. This program implements a progressive method for multiple sequence alignment. As a progressive algorithm, Crustal adds sequences one by one to the existing alignment to build a new alignment. The order of the sequences to be added to the new alignment is indicated by a precomputed phylogenetic tree, which is called a guide tree. The guide tree is constructed using the similarity of all possible pairs of sequences. The algorithm consists of 3 phases that are described below:
Stage 1 —Distance Matrix: All pairs of sequences are aligned separately in order to calculate a distance matrix based on the percentage of mismatches of each pair of sequences.

Stage 2 —Neighbor joining: The guide tree is calculated from the distance matrix using a neighbor joining algorithm. The guide tree defines the order which the sequences are aligned in the next stage.

Stage 3 —Progressive alignment: The sequences are progressively aligned following the guide tree.

Now we discuss the complexity for all stages. Give N sequences and sequence length L, calculating the distance matrix in stage 1 takes O(N2L2) time. A neighbor joining algorithm is O(N4) time for constructing the guide tree in stage 2. And for the last stage, progressive alignment is O(N3+NL2) time. The summary is shown in Table 1 [17].

Table 1. Complexity of the sequential Crustal. The big-O asymptotic complexity of the elements of Crustal as a function of L, the sequence length, and N, the number of sequences, retaining the highest-order terms in N with L fixed and vice versa.

Stage	**Time Complexity**
Distance Matrix	$O(N^2L^2)$
Neighbor joining	$O(N^4)$
Progressive alignment	$O(N^3+NL^2)$
Total	$O(N^4+L^2)$

Basic Local Alignment Search Tool
(BLASTP)
BLAST is very popular tool to align protein sequence in Local Pair wise alignment method. The problem with BLAST is it can work best only for pair of single queries. IT may fail with complex queries as well as in large database

Working of Basic Local Alignment Search Tool:

- A locally maximal segment pair (LMSP) is any segment pair (s, t) whose score cannot be improved by shortening or extending the segment pair. Align all words with sequence in database
- A maximum segment pair (MSP) is any segment pair (s, t) of maximal alignment score σ(s, t).
- Given a cutoff score S, a segment pair (s, t) is called a high-scoring segment pair (HSP), if it is locally maximal and $\sigma(s, t) \geq S$.

Finally, a word is simply a short substring of fixed length w.

- **Localization of the hits:** The database sequence d is scanned for all hits t of w-mers s in the list, and the position of the hit is saved.
- **Detection of hits:** First all pairs of hits are searched that have a distance of at most A (think of them lying on the same diagonal in the matrix of the SW-algorithm).
- **Extension to HSPs:** Each such seed (s, t) is extended in both directions until its score σ(s, t) cannot be enlarged (LMSP). Then all best extensions are reported that have score ≥ S, these are the HSPs. Originally the extension did not include gaps, the modern BLAST2 algorithm allows insertion of gaps.
- The list L of all words of length w that have similarity > T to some word in the query sequence q can be produced in O(|L|) time.
- These are placed in a "keyword tree" and then, for each word in the tree, all exact locations of the word in the database d are detected in time linear to the length of d.
- As an alternative to storing the words in a tree, a finite-state machine can be used.found to have the faster implementation.

The reciprocal smallest distance algorithm: One such computationally intensive comparative genomics method, the reciprocal smallest distance algorithm (RSD), is particularly representative of the scaling problems faced by comparative genomics applications. RSD is a whole-genomic comparative method designed to detect orthologous sequences between pairs of genomes. The algorithm employs BLAST [2] as a first step, starting with a subject genome, J, and a protein query sequence, i, belonging to genome I. A set of hits, H, exceeding a predefined significance threshold (e.g., E < 10-10, though this is adjustable) is obtained. Then, using crustal, each protein sequence in H is aligned separately with the original query sequence i. If the alienable region of the two sequences exceeds a threshold fraction of the alignment's total length (e.g., 0.8, although this is also adjustable), the BLOSUM62 is used to obtain a maximum likelihood estimate of the number of amino acid substitutions separating the two protein sequences, given an empirical amino acid substitution rate matrix. The model under which a maximum likelihood estimate is obtained in RSD may include variation in evolutionary rate among protein sites, by assuming a gamma distribution of rate across sites and setting the shape parameter of this distribution, α, to a level appropriate for the phylogenetic distance of the species being compared. Of all sequences in H for which an evolutionary distance is estimated, only j, the sequence yielding the shortest distance, is retained. This sequence j is then used for a reciprocal BLAST against genome I, retrieving a set of high scoring hits, L. If any hit from L is the original query sequence, i, the distance between i and j is retrieved from the set of smallest distances calculated previously. The remaining hits from L are then separately aligned with j and maximum likelihood distance estimates are calculated for these pairs as described above. If the protein sequence from L producing the shortest distance to j is the original query sequence, i, it is assumed that a true orthologous pair has been found and their evolutionary distance is retained.

Clique Algorithm: To search for maximal, completely connected subgraphs in a graph, where the vertices are genes and the edges are verified pairs. To compute cliques, algorithms exist to maximize either the size of the clique (number of vertices) or the total weight of cliques (sum of edge weights). Figure A shows a graph with edges between all vertices except (z1, z2) and (z1, y2), which are paralogous relations. The highest scoring partition is {w1, x1, z1}, {y2, z2}, with the total sum of edge weights of 700 + 800 + 900 + 1000 = 3600. The score is higher than the highest scoring maximum size clique {w1, x1, y2, z2}, {z1}, where the sum of the scores is 200+300+400+500+800+1000 = 3200. Hence, a smaller clique is chosen due to higher edge weights, which correctly assigns orthologs according to the hypothetical evolutionary scenario in Figure B where the duplication gives subscripts that correspond to functionality. Finding cliques is a NP-complete problem and the implementations used here are based on an approximation of the vertex cover problem [19]. Each clique constitutes an orthologous group, where the sequence pairs in an orthologous group are denoted group pairs (GP), corresponding to close orthologs.

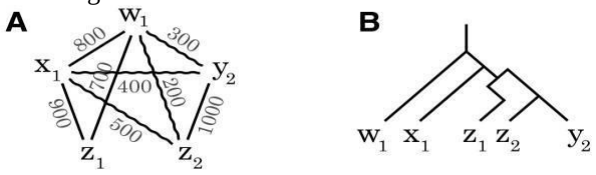

Figure 1:A.An example graph containing one 4-clique, four 3-cliques, and eight 2-cliques is provided. The highest scoring partition of the graph is {w1, x1, z1}, {y2, z2}. B. A possible evolutionary scenario corresponding to the graph.

Proposed model: Core objective of this system is to combine method of multiple sequence alignment and Basic linear alignment search so as to increase speed and efficiency of target system.

Ortholog Cluster Creator: It accept a set of complete genomes and gives pairs of orthologous genes that will be clustered into orthologous groups. The algorithm follows four steps

ALL X ALL Comparison: - To find homology, system will compute pairwise alignments between all pairs of sequences for all genes in all genomes. Pairs with good alignment scores are selected as candidate pairs. The goal of the first step is homology detection. This includes following steps

a. Given dataset has been applied to the multiple sequence alignment tool which will introduce gap for alignment which will allow short sequence to compare with long one or allow partial sequence to compare with the whole. With help of BLOSUM62, all pairs of sequences are aligned separately in order to calculate a distance matrix based on the percentage of mismatches of each pair of sequences and calculate score. Depending upon score filter the database. proposed system used CLUSTALW algorithm for this process. Crustal has become the most popular algorithm for multiple sequence alignment. This program implements a progressive method for multiple sequence alignment. As a progressive algorithm, ClustalW adds sequences one by one to the existing alignment to build a new alignment.
b. This selected dataset has been applied to Basic Linear Alignment Search Tool which is very popular tool to align protein sequence in Local Pair wise alignment method.
c. Since consider entire proteins as the basic evolutionary unit, why then not use global alignments? Protein ends are often variable, and thus, it is reasonable to ignore them by using local alignments. To guarantee that a significant fraction of a sequence is aligned, use a length tolerance criterion. The length of the shorter aligned sequence must be at least the fraction ℓ of the longest sequence. That is
$\min(|a1|, |a2|) > \ell \cdot \max(|s1|, |s2|)$

where a1 and a2 are the lengths of the aligned subsequences of s1 and s2. a "plateau" is observed for $0.6 < \ell < 0.9$. Alignments that pass both the length and score criteria are upgraded to candidate pairs (CP).

The all-against-all step is computationally expensive, and the run time increases quadratically with the size of a protein sequence. The use of a heuristic-based algorithm such as BLAST could potentially increase the speed of the search

Formation of Stable Pairs: - Orthologs are usually the nearest genes in two genomes, because they started diverging at speciation, whereas paralogs started diverging at a duplication prior to speciation. Genes across genomes that are mutually the most closely related sequences, taking into account inference uncertainty, are upgraded to stable pairs. In the second step of the algorithm, potential orthologs are detected by finding sequences in two genomes that are near to each other than to any other sequence in the other sequence. These sequences called as stable pairs (SP). This name was chosen due to its close association with the stable marriage problem in computer science. The rsd algorithm is used to find candidate pairs.

Verification of Stable pair: - Although the construction of stable pairs is likely to identify the corresponding ortholog of each sequence, at least one special case exists in which systematic failure will occur: differential gene loss. This problem affects all pairwise approaches, and is shown in Fig A. An ancient duplication event is followed by two speciation events resulting in three species X, Y, and Z. In two of these species, each of the duplicates is lost (e.g. x2 and y1), and as a result, when comparing species X and Y, x1 and y2 are the highest scoring match. In such a case, (x1, y2), although paralogs, form a stable pair.

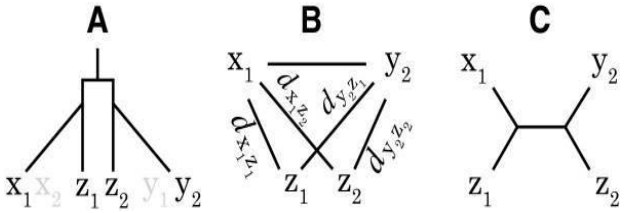

Figure 2:Identification of potential paralogs

The purpose of the third step is to detect such stable pairs corresponding to non-ortholog. The presence of a third genome Z, which has retained both copies z1 and z2 of the duplication event, acts as a witness of non-ortholog. We previously described the details of this procedure [21], and the idea is illustrated in Figure B. If dx1z1 is significantly shorter than dx1z2 and dy2z2 is significantly shorter than dy2z1, there is evidence that x1 and y2 may not be orthologs. Figure C depicts the most likely quartet predicted from the data provided in Figure A. This approach can also be viewed as a tree-reconciliation that is based on quartets without assuming any species tree topology.

Each stable pair is verified by comparison to all other genomes. Stable pairs for which no witness of non-ortholog could be found are termed verified pairs (VP) and are likely to be orthologs. Furthermore, stable pairs that are not verified were defined as broken pairs (BP) and are likely to correspond to paralogs.

Clustering of orthologs: - The final step of the algorithm creates groups of orthologs. Such grouping is non-trivial, because orthology is defined over pairs of sequences and is not necessarily a transitive relation. For instance, a sequence in one genome may form several verified pairs with sequences in another genomes, corresponding to several orthologous relations (co-orthologs). These in turn cannot be orthologous to each other. This problem is addressed by making available both pairwise orthologous relations (the verified pairs) and groups of genes in which all pairs are orthologs. A clique algorithm is used to search for maximal, completely connected subgraphs in a graph, where the vertices are genes and the edges are verified pairs. Each clique constitutes an orthologous group, where the sequence pairs in an orthologous group are denoted group pairs (GP), corresponding to close orthologs.

Table 1:Sequence pairs and their corresponding evolutionary relationships

Pairs	**Evolutionary Relation**
All pairs (AP)	Any
Candidate pairs (CP)	Homologs
Stable pairs (SP)	Orthologs, Pseudo-orthologs
Broken pairs (BP)	Paralogs
Verified pairs (VP)	Orthologs
Group pairs (GP)	Close orthologs

IV. EXPECTED OUTCOMES/ DELIVERABLES

Expected Outcome Using MSA: The MSA tools will provide efficient way to filter out given database with ‘n’ sequences. which can be reduce size of dataset. which is help to remove unmatched sequences reference to the given query string. Which in turns reduce load on BLAST tool

Expected Outcome Using BLAST: The BLAST tool is used to find out relevant sequence from filtered dataset outputted from MSA tools. It can be use to find out their relationship score so as to we can able to categorized them.

Deliverables: - An application which accepts bioinformatics textual sequence and give HOMOLOGS sequences available in Database. this application can further predict relationship among two humans.

V. CONCLUSION

We have proposed new hybrid method for aligning bioinformatics sequences to understand the nature of relationship among them. The proposed method will use MSA tools to deduct irrelevant data from available database and decrease size of data. then BLAST tool will be used to identify matches among them and try to

identify ancestor relationship among query string and clusters of ortholog from available data in bioinformatics database. The core intention of this proposed method is to decrease unnecessary time utilized by BLAST in comparing word by word and increase accuracy of clustering. Orthology is useful for a wide range of bioinformatics study, including functional annotation, phylogenetic inference, or genome evolution. This system describes and motivates the algorithm for predicting orthologous relationships among complete genomes. The algorithm takes a pairwise approach, thus neither requiring tree reconstruction nor reconciliation, and offers the following improvements over the standard bidirectional best hit approach: i) the use of evolutionary distance, ii) a tolerance that allows the inclusion of one-to-many and many-to-many orthologs, iii) consideration of uncertainty in distance estimations, iv) detection of potential differential gene losses.

REFERENCES

[1] Min-Sung Kim, Choong-Hyun Sun, Jin-Ki Kim, Gwan-Su Yi (2006) Whole Genome Alignment with BLAST on Grid Environment, Proceedings of The Sixth IEEE International Conference on Computer and Information Technology (CIT'06)

[2] Sing-Hui To, Hoon-Jae Lee, Kyeong-Hoon D0 (2009), Basic Sequence Search by Hashing Algorithm in DNA Sequence Databases, ISBN 978-89-5519-139-4, Feb 15-18,2009, ICACT

[3] Rodolfo Bezerra Batista and Alba Cristina Magalhaes Alves de Melo (2006), Z-align: An Exact and Parallel Strategy for Local Biological Sequence Alignment in User-Restricted Memory Space

[4] Sangha Mitra Bandyopadhyay, Senior Member, IEEE, and Ramakrishna Mitra (2009), A Parallel Pair Wise Local Sequence Alignment Algorithm

[5] Amlan Kundu, Suvasini Panigrahi, Shamika Sural and Arun K. Majumdar in (2009)BLAST-SSAHA Hybridization for Credit Card Fraud Detection, Dependable and Secure Computing, IEEE Transactions on Volume: 6 , Issue: 4

[6] Stephen Pellicer∗, Guihai Chen, Member, IEEE, Keith C. C. Chan, and Yi Pan, (2008) Distributed Sequence Alignment Applications for the Public Computing Architecture, IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 7, NO. 1, MARCH 2008

[7] Ken D. Nguyen, Member, IEEE, and Yi Pan, Senior Member, IEEE in (2011) An Improved Scoring Method for Protein Residue Conservation and Multiple Sequence Alignment

[8] Farhana Naznin, Member, IEEE, Ruhul Sarker, Member, IEEE, and Daryl Essam (2012) Progressive Alignment Method Using Genetic Algorithm for Multiple Sequence Alignment, IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 16, NO. 5, OCTOBER 2012

[9] Ching Zhang and Andrew K. C. Wong. (1998). A Technique of Genetic Algorithm and Sequence Synthesis for Multiple Molecular Sequence Alignment 0-7803-4778-1 198 @ 1998 IEEE

[10] A. Agarwal and s.k. khaitan (2008) A new heuristic for multiple sequence alignment, Electro/Information Technology, 2008. EIT 2008. IEEE International Conference, 978-1-4244-2029-2

[11] Jagadamba, P.V.S.L. ; DST Project, JNTUK, Kakinada, India ; Babu, M.S.P. ; Rao, A.A.; Rao, P.K. (2011).An improved algorithm for Multiple Sequence Alignment using Particle Swarm Optimization, I EEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 10, NO. 4, DECEMBER 2011

[12] Z. Liu, J. Bornean, and T. Jiang (2004) A Software System for Gene Sequence Database Construction Proceedings of the 26th Annual International Conference of the IEEE EMBS San Francisco, CA, USA • September 1-5, 2004

[13] Ken D. Nguyen, Member, IEEE, Yi Pan, Member, IEEE (2013) A Knowledge-based Multiple sequence Alignment Algorithm, IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 1545-5963/13 © 2013 IEEE

[14] Jun Wang and Young Sun (2012) A Sliding Window and Keyword Tree Based Algorithm for Multiple Sequence Alignment, 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012),978-1-4673—0024-7/10, IEEE

[15] Bugra Ozer, Gizem Gezici, Cem Meydan, Ugur Sezerman (2009). Multiple Sequence Alignment Based on Structural Properties IEEE-2009

[16] Joanne Bai, Siamak Rezaei (2005). Multithreaded Multiple Sequence Alignments. Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1-4, 2005

[17] R.C. Edgar, "Muscle: a multiple sequence alignment method with reduced time and space complexity," BMC Bioinformatics. 2004, vol. 5, 2004.

[18] Distinguishing homologous from analogous proteins. Fitch WM Syst Zool. 1970 Jun; 19(2):99-113.

[19] Balasubramanian R, Fellows M, Raman V. An improved fixed-parameter algorithm for vertex cover. Information Processing Letters. 1998;65:163–168. doi: 10.1016/S0020-0190(97)00213-5.
[20] A dimensionless fit measure for phylogenetic distance trees.[J Bio inform Comput Biol. 2005]
[21] Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. Decimos C, Beckmann B, Roth AC, Gannet GH, Nucleic Acids Res. 2006; 34(11):3309-16.
[22] Needleman, Saul B. & Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology
[23] Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes(https://www.ncbi.nlm.nih.gov)
[24] An Algorithm to Cluster Orthologous Proteins across Multiple Genomes, Sunshin Kim, Chung Sei Rhee, Jung-Do Choi
[25] Automatic protein function annotation through candidate ortholog clusters from incomplete genomes A. Vashist; C. Kulikowski; I. Muchnik 2005 IEEE Computational Systems Bioinformatics Conference - Workshops (CSBW'05)
[26] Clustering orthologs based on sequence and domain similarities Fa Zhang; Sheng Zhong Feng; H. Ozer; Bo Yuan Eighth International Conference on High-Performance Computing in Asia-Pacific Region (HPCASIA'05)
[27] Ortholog Clustering on a Multipartite Graph Akshay Vashist; Casimir A. Kulikowski; Ilya Muchnik IEEE/ACM Transactions on Computational Biology and Bioinformatics

Dr. Kamal Shah, and Mr. Nilesh N. Bane. "Performance Improvement of BLAST with Use of MSA Techniques to Search Ancestor Relationship among Bioinformatics Sequences." Invention Journal of Research Technology in Engineering & Management