IMPLEMENTING HIERARCHICAL CLUSTERING METHOD FOR MULTIPLE SEQUENCE ALIGNMENT AND PHYLOGENETIC TREE CONSTRUCTION

Harmandeep Singh[1], Er. Rajbir Singh Associate Prof.[2] , Navjot Kaur[3]

[1]Lala Lajpat Rai Institute of Engineering and Tech., Moga
Punjab, INDIA
har_pannu@yahoo.co.in

ABSTRACT

In the field of proteomics because of more data is added, the computational methods need to be more efficient. The part of molecular sequences is functionally more important to the molecule which is more resistant to change. To ensure the reliability of sequence alignment, comparative approaches are used. The problem of multiple sequence alignment is a proposition of evolutionary history. For each column in the alignment, the explicit homologous correspondence of each individual sequence position is established. The different pair-wise sequence alignment methods are elaborated in the present work. But these methods are only used for aligning the limited number of sequences having small sequence length. For aligning sequences based on the local alignment with consensus sequences, a new method is introduced. From NCBI databank triticum wheat varieties are loaded. Phylogenetic trees are constructed for divided parts of dataset. A single new tree is constructed from previous generated trees using advanced pruning technique. Then, the closely related sequences are extracted by applying threshold conditions and by using shift operations in the both directions optimal sequence alignment is obtained.

General Terms

Bioinformatics, Sequence Alignment

KEYWORDS

Local Alignment, Multiple Sequence Alignment, Phylogenetic Tree, NCBI Data Bank

1. INTRODUCTION

Bioinformatics is the application of computer technology to the management of biological information. It is the analysis of biological information using computers and statistical techniques; the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. Bioinformatics is more of a tool than a discipline, the tools for analysis of Biological Data.

In bioinformatics, a multiple sequence alignment is a way of arranging the primary 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. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that residues with identical or similar characters are aligned in successive

DOI : 10.5121/ijcseit.2013.3101

columns. Mismatches can be interpreted, if two sequences in an alignment share a common ancestor and gaps as indels i.e. insertion or deletion mutations introduced in lineages. In DNA sequence alignment, the degree of similarity between amino acids occupying a particular position in the sequence can be presented as a rough measure of how conserved a particular region is among lineages. The substitutions of nucleotides whose side chains have similar biochemical properties suggest that this region has structural or functional importance.

2. DATA MINING

Data mining is the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform this data into information. As part of the larger process known as knowledge discovery, data mining is the process of extracting information from large volumes of data. This is achieved through the identification and analysis of relationships and trends within commercial databases. Data mining is used in areas as diverse as space exploration and medical research. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The knowledge discovery process includes:

- Data Cleaning
- Data integration
- Data selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

2.1 Data Mining Techniques

To design the data mining model the choice has to be made from various data mining techniques [6][8], which are as follows:

a. Cluster Analysis
- Hierarchical Analysis
- Non- Hierarchical Analysis

b. Outlier Analysis
c. Induction
d. Online Analytical Processing (OLAP)
e. Neural Networks
f. Genetic Algorithms
g. Deviation Detection
h. Support Vector Machines
i. Data Visualization
j. Sequence Mining

In the present work we have adopted Hierarchical Cluster Analysis as a Data Mining approach, as it is most suitable to work for a common group of protein sequences.

2.2 Data Mining Methods

Prediction and description are two important goals of data mining. Prediction uses the current database to predict unknown or probable values of other variables of interest, and description extracts the human-interpretable patterns describing the data.

These goals can be achieved using following data-mining methods:

a. Classification
b. Regression
c. Clustering
d. Summarization
e. Dependency Modeling
f. Change and Deviation Detection

In the present work, we have picked Classification Method for the sequence alignment problem.

3. ALIGNMENT METHODS

The sequences can be aligned manually that are short or similar. However, lengthy, highly variable or extremely numerous sequences cannot be aligned solely by human effort. Computational approaches are used for the alignment of these sequences. Computational approaches divided into two categories:

- Global
- Local

3.1 Global and Local Alignments

When the sequences in the query set are similar and of roughly equal size then, Global alignments are useful which attempt to align every residue in every sequence. (This does not mean global alignments cannot end in gaps.) The Needleman-Wunsch algorithm is a global alignment technique, which is mainly based on dynamic programming. For dissimilar sequences that are suspected to contain regions of similarity within their larger sequence context local alignments are more useful. A general local alignment technique is Smith-Waterman algorithm which is also based on dynamic programming. There is no difference between local and global alignments with sufficiently similar sequences.

3.2 Pair wise Alignment

To find the best-matching local or global alignments of two query sequences a Pair wise sequence alignment methods are used. At a time only two sequences can be align in pair wise alignment methods, which are efficient to calculate and are often used for methods that do not require extreme searching a database for sequences to a query. For producing Pair wise alignments, there are basically three methods:

- dot-matrix methods
- dynamic programming
- word methods

However; to align pairs of sequences multiple sequence alignment techniques can also be used. Although each and every individual method has its own pros and cons, with highly repetitive sequences all these Pair wise methods have difficulty because of low information content - especially where the numbers of repetitions differ in the two sequences to be aligned. the 'maximum unique match', or the longest subsequence one way of quantifying the utility of a given Pair wise alignment is that occurs in both query sequence. Longer MUM sequences typically reflect closer relatedness.

3.3 Dot Matrix Method

The dot-matrix approach is qualitative and simple to produce a family of alignments for individual sequence regions but it is time-consuming to analyze on a large scale. From a dot-matrix plot certain sequence features (such as insertions, deletions, repeats, or inverted repeats) can easily be identified. The dot-matrix plot is created by designating one sequence to be the subject and placing it on the horizontal axis and designating the second sequence to be the query and placing it on the vertical axis of the matrix. Some implementations vary the size or intensity of the dot which depends on the degree of similarity between two characters, to accommodate conservative substitutions. The dot plots of very closely related sequences will make a single diagonal line along the matrix. To assess repetitiveness in a single sequence, dot plots can also be used. A sequence can be plotted against itself and regions that share significant similarities will appear as diagonal lines along the matrix. This effect can occur when a DNA/RNA consists of multiple similar structural domains. A dot matrix plot is a method of aligning two sequences to provide a picture of the homology between them. The main diagonal represents the sequence's alignment with itself; lines off the main diagonal represent similar or repetitive patterns within the sequence.

3.4 Progressive Method

A Progressive also known as hierarchical or tree methods, generate a multiple sequence alignment by first aligning the most similar sequences and then adding successively less related sequences or groups to the alignment until the entire query set has been incorporated into the solution. Sequence relatedness is describing by the initial tree that is based on Pair wise alignments which may include heuristic Pair wise alignment methods. Results of progressive alignment are dependent on the choice of most related sequences and thus can be sensitive to inaccuracies in the initial Pair wise alignments. In most progressive multiple sequence alignment methods, the sequences are weighted in the query set according to their relatedness, which reduces the likelihood of making a poor choice of initial sequences and thus improves alignment accuracy.

4. INTRODUCTION TO PHYLOGENETICS ANALYSIS

Phylogenetics is the study of evolutionary relationships. Phylogenetic analysis is useful for inferring or estimating these relationships. Phylogenetic analysis is usually depicted as branching, tree like diagrams that represent an estimated pedegree of the inherited relationships among gene trees, organisms, or both. In phylogenetics because the word clade used which is a set of descendants from a single ancestor, is derived from the Greek word for branch, that’s why it is sometimes called as cladistics. However, for hypothesizing about evolutionary relationships, cladistics method is used.

In cladistics, members of a group or clade share a common evolutionary history and are more related to each other than to members of another group that particular group is recognized by sharing unique features that were not present in distant ancestors. These shared, derived characteristics can be anything that can be observed and described from two organisms. Usually, cladistic analysis is performed by comparing multiple characteristics at once, either multiple phenotypic characters or multiple base pairs or nucleotides in a sequence.

1) There are three basic assumptions in cladistics: Organisms of any group is related by descent from a common ancestor.
2) There is a bifurcating pattern of cladogenesis. This assumption is controversial.
3) A phylogenetic tree represents the resulting relationship from cladistic analysis.

5. METHODOLOGY

The methodology for this work involves the uses the cluster analysis techniques to compute the alignment scores between the multiple sequences. Based on the alignment the phylogenetic tree is constructed signifying the relationship between different entered sequences. The data is taken from the databank of NCBI.

5.1 SNAP (synonymous non-synonymous analysis program)

In SNAP, based on a set of codon aligned nucleotide sequences, synonymous and non-synonymous substitution rates are calculates.
They should be provided in table format:

Seq1ACTGCCTTTGGC...
Seq2ACTGCCTATGGG...

The first field is the sequence name and the second field is the sequence, and then returns to a new

line for the second sequence. A single insertion could be treated as a way to keep the codons ACT and GGC intact:

ACTTGCC ==> ACTT--GCC

ACTGGC ==> ACT---GCC

To represent ambiguous bases, the letter N should be used.

A dash, '-' for insertions, Only A C G T N - are allowed.

Table 4.1: Summary of the Synonymous and Non-Synonymous Information

Compare		Sequences names		Sd	Sn	S	N	ps	pn	ds	dn	ds/ds	ps/ps
0	1	GU259621	GU068562	2832.8333	6615.1667	1245	6670	2.1062	0.7123	NA	0	NA	2.9527
0	2	GU259621	DQ233207	2127.5	6609.5	1282	6633	2.6673	0.7056	NA	0	NA	3.6686
0	3	GU259621	AY716362	2848	6596	1266.6667	6650.3333	2.087	0.7122	NA	0	NA	2.9302
0	4	GU259621	GU068564	2848.1667	6616.8333	1236.5	6678.5	2.1311	0.7123	NA	0	NA	2.9917
0	5	GU259621	GQ870049	2949.5	6607.5	1215.5	6699.5	2.2621	0.7089	NA	0	NA	3.1628
0	6	GU259621	EU870721	3038	6605	1200.5	6714.5	2.336	0.7069	NA	0	NA	3.3047
0	7	GU259621	GU068521	2849.1667	6612.8333	1237.1667	6677.8333	2.1307	0.7121	NA	0	NA	2.9922
0	8	GU259621	GU068553	2832.8333	6615.1667	1245	6670	2.1062	0.7123	NA	0	NA	2.9527
0	9	GU259621	GU068556	2847.3333	6617.6667	1241.8333	6673.1667	2.122	0.7124	NA	0	NA	2.9766
0	10	GU259621	DQ619977	2176.5	6616.5	1272.6667	6642.3333	2.6959	0.7056	NA	0	NA	3.8372
0	11	GU259621	AJ289678	3203.3333	6607.6667	1266.5	6648.5	2.5293	0.7086	NA	0	NA	3.5967
1	2	GU068562	DQ233207	569.8333	798.1667	1270.6667	9210.3333	0.4485	0.0867	0.6826	0.091	7.5096	5.221
1	3	GU068562	AY716362	1182	2077	1496.6667	9096.3333	0.7961	0.2283	NA	0	NA	3.4867
1	4	GU068562	GU068564	220.1667	758.8333	1235.1667	9255.8333	0.1661	0.082	0.1878	0.0868	2.1632	2.0265
1	5	GU068562	GQ870049	359.6667	792.3333	1206.1667	9276.8333	0.2758	0.0855	0.3428	0.0908	3.7865	3.2269
1	6	GU068562	EU870721	475.3333	792.6667	1289.1667	9291.8333	0.3687	0.0853	0.5076	0.0906	5.6027	4.3222
1	7	GU068562	GU068521	220.1667	761.8333	1235.8333	9255.1667	0.1661	0.0823	0.1877	0.0872	2.1528	2.0176
1	8	GU068562	GU068553	0	0	1333.6667	9267.3333	0	0	0	0	NA	NA
1	9	GU068562	GU068556	56.8333	171.1667	1330.5	9250.5	0.0612	0.0185	0.0626	0.0187	3.3626	3.2272
1	10	GU068562	DQ619977	606.3333	798.6667	1261.3333	9219.6667	0.4907	0.0867	0.7682	0.091	8.4618	5.6096
1	11	GU068562	AJ289678	663.1667	789.8333	1255.1667	9225.8333	0.5126	0.0847	0.8622	0.0899	9.5936	6.0503
2	3	DQ233207	AY716362	1426.5	2096.5	1421.6667	9159.3333	1.0036	0.2287	NA	0	NA	4.3879
2	4	DQ233207	GU068564	568.3333	789.6667	1262.1667	9218.8333	0.4503	0.0847	0.6879	0.0899	7.6506	5.3138
2	5	DQ233207	GQ870049	309.5	666.5	1241.1667	9239.8333	0.2496	0.069	0.3031	0.0726	4.1877	3.6137
2	6	DQ233207	EU870721	209	220	1226.1667	9254.8333	0.1706	0.0353	0.1934	0.0361	5.3516	4.8319
2	7	DQ233207	GU068521	567.3333	791.6667	1262.8333	9218.1667	0.4493	0.085	0.6856	0.0902	7.6008	5.2879
2	8	DQ233207	GU068553	569.8333	798.1667	1270.6667	9210.3333	0.4485	0.0857	0.6826	0.091	7.5096	5.221
2	9	DQ233207	GU068556	569.1667	776.8333	1267.5	9213.5	0.4333	0.0832	0.6665	0.0882	7.3216	5.2079
2	10	DQ233207	DQ619977	79.5	170.5	1198.3333	9282.6667	0.0663	0.0182	0.0695	0.0184	3.776	3.6508
2	11	DQ233207	AJ289678	126.1667	156.8333	1192.1667	9288.8333	0.1058	0.0165	0.1141	0.0167	6.8616	6.4172
3	4	AY716362	GU068564	1165.5	2102.5	1476.1667	9104.8333	0.776	0.2309	NA	0	NA	3.3604
3	5	AY716362	GQ870049	1268	2096	1455.1667	9125.8333	0.8716	0.2296	NA	0	NA	3.8158
3	6	AY716362	EU870721	1250.5	2092.5	1440.1667	9140.8333	0.9377	0.2289	NA	0	NA	4.0966
3	7	AY716362	GU068521	1166.5	2102.5	1476.8333	9104.1667	0.7762	0.2309	NA	0	NA	3.3616
3	8	AY716362	GU068553	1182	2077	1496.6667	9096.3333	0.7961	0.2283	NA	0	NA	3.4867
3	9	AY716362	GU068556	1192.3333	2075.6667	1481.5	9099.5	0.8048	0.2281	NA	0	NA	3.5282
3	10	AY716362	DQ619977	1665	2116	1412.3333	9168.6667	1.0221	0.2308	NA	0	NA	4.4232
3	11	AY716362	AJ289678	1672.3333	2118.6667	1406.1667	9174.8333	1.0678	0.2309	NA	0	NA	4.6273
4	5	GU068564	GQ870049	253.3333	832.6667	1295.6667	9285.3333	0.2727	0.0897	0.339	0.0956	3.549	3.041
4	6	GU068564	EU870721	457.3333	810.6667	1280.6667	9300.3333	0.3571	0.0872	0.4949	0.0927	5.3351	4.0969
4	7	GU068564	GU068521	0	2	1317.3333	9263.6667	0	0.0002	0	0	NA	NA

4	8	GU048564	GU048553	220.1667	758.8333	1325.1667	9255.8333	0.1661	0.082	0.1878	0.0868	2.1632	2.0265
4	9	GU048564	GU048554	218.1667	746.8333	1322	9259	0.165	0.0807	0.1864	0.0853	2.1841	2.046
4	10	GU048564	DQ419977	616.6667	778.3333	1252.8333	9328.1667	0.4922	0.0834	0.801	0.0885	9.055	5.8991
4	11	GU048564	AJ389678	654.5	767.5	1246.6667	9334.3333	0.525	0.0822	0.903	0.0871	10.3684	6.385
5	6	GQ870249	EU670731	250.3333	629.6667	1259.6667	9321.3333	0.1987	0.0676	0.2309	0.0708	3.2616	2.9419
5	7	GQ870249	GU048531	353.1667	829.8333	1296.3333	9284.6667	0.2724	0.0894	0.3385	0.0952	3.5572	3.0482
5	8	GQ870249	GU048553	359.6667	793.3333	1304.1667	9276.8333	0.2758	0.0855	0.3438	0.0908	3.7865	3.2249
5	9	GQ870249	GU048554	339.1667	767.8333	1301	9280	0.2607	0.0827	0.3203	0.0877	3.6536	3.1508
5	10	GQ870249	DQ419977	333.5	669.5	1231.8333	9349.1667	0.2707	0.0716	0.3359	0.0753	4.4625	3.7806
5	11	GQ870249	AJ389678	356.8333	679.1667	1225.6667	9355.3333	0.2911	0.0726	0.3685	0.0764	4.826	4.0103
6	7	EU670731	GU048531	456.8333	811.1667	1281.3333	9299.6667	0.3565	0.0872	0.4838	0.0927	5.2174	4.0875
6	8	EU670731	GU048553	475.3333	792.6667	1289.1667	9291.8333	0.3687	0.0853	0.5074	0.0906	5.6027	4.3222
6	9	EU670731	GU048554	454.5	768.5	1286	9295	0.3534	0.0827	0.4779	0.0876	5.4554	4.2746
6	10	EU670731	DQ419977	259	338	1216.8333	9364.1667	0.2128	0.0361	0.2503	0.037	6.7674	5.8969
6	11	EU670731	AJ389678	285	329	1210.6667	9370.3333	0.2354	0.0351	0.2825	0.036	7.8568	6.7047
7	8	GU048531	GU048553	220.1667	761.8333	1325.8333	9255.1667	0.1661	0.0823	0.1877	0.0872	2.1528	2.0174
7	9	GU048531	GU048554	218.1667	749.8333	1322.6667	9258.3333	0.1649	0.081	0.1863	0.0857	2.1734	2.0366
7	10	GU048531	DQ419977	616.1667	777.8333	1253.5	9327.5	0.4916	0.0834	0.799	0.0884	9.0387	5.8946
7	11	GU048531	AJ389678	654.5	767.5	1247.3333	9333.6667	0.5247	0.0822	0.902	0.0871	10.3569	6.3812
8	9	GU048553	GU048554	54.8333	171.1667	1330.5	9250.5	0.0412	0.0185	0.0424	0.0187	2.2624	2.2273
8	10	GU048553	DQ419977	606.3333	798.6667	1261.3333	9319.6667	0.4807	0.0857	0.7682	0.091	8.4418	5.6094
8	11	GU048553	AJ389678	643.1667	789.8333	1255.1667	9325.8333	0.5124	0.0847	0.8622	0.0899	9.5936	6.0503
9	10	GU048554	DQ419977	585.6667	775.3333	1258.1667	9322.8333	0.4655	0.0832	0.727	0.0881	8.2473	5.5972
9	11	GU048554	AJ389678	620.5	766.5	1252	9329	0.4956	0.0822	0.8109	0.087	9.3183	6.032
10	11	DQ419977	AJ389678	48.1667	69.8333	1182.8333	9398.1667	0.0407	0.0074	0.0419	0.0075	5.6067	5.4803

5.2 Jukes cantor Method

The Jukes and Cantor model is a model which computes probability of substitution from one state to another. From this model we can also derive a formula for computing the distance between 2 sequences. This model was for nucleotides, but this can easily be substituted by codons or amino acids. In this model the probability of changing from one state to a different state is always equal as well as the different sites are independent. The evolutionary distance between two species is given by the following formula.

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{N_d}{N}\right)$$

Where Nd is the number of mutations (or different nucleotides) between the two sequences and N is the nucleotide length.

5.3 Algorithm

The algorithm finds the optimal alignment of the most closely related species from the set of species. The work includes the construction of phylogenetic trees for the triticum wheat varies. The sequences for the varieties are loaded from the NCBI database. The phylogenetic distances are calculated based on the jukes cantor method and the trees are constructed based on nearest-neighbor method. The final tree is obtained by using tree pruning. The closely related species are selected based on the threshold condition. As the sequences are very lengthy and the alignment is tedious work for these sequences. To obtain the multiple sequence alignment, the consensus sequence (fixed sequence for eukaryotes) is aligned with the available sequence. This helps in locating the positions near to the optimal alignment. The sequences are aligned based on local alignment and shift right and shift left operations are performed five times to obtain the optimal multiple sequence alignment. The algorithm steps are written below:

- Load the m wheat sequences from NCBI database
- Calculate the distances from jukes cantor method.
- Create the matrix1 for different species based on JC distance.

- Find the smallest distance by examine the original matrix. Half that number is the branch length to each of those two species from the first Node.
- Create a new, reduced matrix with entries for the new Node and the unpicked species. The distance from each of the unpicked will be the average of the distance from the unpicked to each of the two species in Node 1.
- Find the smallest distance by examining the reduced matrix. If these are two unpicked species, then they Continue these distances calculation and matrix reduction steps until all species have been picked.

- Construct the tree 1.
- Load the n more wheat sequences from NCBI database.
- Calculate the distances from jukes cantor method.
- Create the matrix2 for different species based on JC distance.
- Repeat step iv to vi
- Construct tree 2.
- Apply tree pruning to join the tree1 and tree2.
- Consider the p species above threshold value 0.7.
- Consider the consensus sequence (TATA box) for wheat varieties.
- Load the p sequences (S1, S2, ……..Sp)
- Align TATA consensus sequence with S1 to Sp considering local alignment.
- Align the sequences using local alignment with TATA consensus sequence.
- Calculate the alignment with shift five shift right and five shift left operations.
- Consider the alignment with minimum score.

6. RESULTS AND DISCUSSIONS

The triticum wheat varieties shown in table 1 are used as input for the present research work. The data is loaded from National Center for Biotechnology Information advances science and health (ncbi.nlm.nih.gov). The five wheat varieties are chosen. The evolutionary distance is calculated with the help of Jukes Cantor Method. The phylogenetic tree is created using nearest neighbor technique. The tree obtain is shown in Fig. 2.Then the different seven varieties are chosen. The tree is constructed by the same method and is combined with the first tree by using tree pruning techniques. The final tree for twelve wheat varieties is shown in Fig. 3.

The Jukes Cantor values for these varieties are shown below.

Columns 1 through 15

34.2415 2.2956 2.2840 2.3334 3.5762 4.5120 34.2415 1.3255 2.2956 0.7981

Columns 16 through 30

0.9966 1.7686 1.0740 0.0161 1.0029 0.7850 0.7981 0.7830 0.5973 0.8810 1.3949

Columns 31 through 45

1.0590 0.8596 0.6849 0.8790 2.2364 1.0842 0.0183 0.0522 0.9209 0.9863 1.7589

TRITICUM WHEAT VARIETIES

- Compactum Cultivar (GU259621)
- Compactum Isolate (GU048562)
- Compactum Glutenin (DQ233207)
- Macha Floral (AY714342)
- Macha Isolate (GU048564)
- Macha Glutenin (GQ870249)

- Sphaerococcum (EU670731)
- Tibeticum Isolate (GU048531)
- Yunnanense Isolate (GUO48553)
- Vavilovii Isolate (GU048554)
- Vavilovii Caroxylase (DQ419977)
- Gamma Gladin (AJ389678)

Columns 46 through 60

0.9966 1.7686 1.0740 0.0161 0.9092 0.9751 1.7275 1.0985 0.0500 0.8219 1.1671

Columns 61 through 66

1.9545 0.9863 3.3140 1.7589

The threshold condition is applied to obtain the most closely related varieties from the set of twelve varieties. The most commonly related varieties are shown in Fig. 4. These varieties are:

- 'yunnanense_isolate'
- 'vavilovii_isolate'
- 'compactum_isolate'
- 'macha_isolate'
- 'tibeticum_isolate'

The multiple sequence alignment for these varieties is obtained. The MSA is shown in Fig. 5.

7. CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

To simplify the complexity of the problem of Multiple Sequence Alignment several different heuristics have been employed. The solution to the neighborhood of only two sequences at a time is restricted by pair-wise progressive dynamic programming. Using pair-wise progressive dynamic programming method, all sequences are compared pair-wise and then each is aligned to its most similar partner or group of partners. Each group of partners is then aligned to finish the complete multiple sequence alignment. This is a quite tedious job and is suited for limited number of sequences and also of small length. Also, it doesn't guarantee the optimal alignment. Complexity is reduced by restricting the search space to only the most conserved local portions of all the sequences involved. The model is constructed for aligning the DNA sequences of different wheat varieties. There are different sequence formats available from which plain text format is utilized. The two phylogenetic trees are constructed for different datasets.

The closely related sequences are extracted based on the threshold condition. Then the consensus sequence is used for to obtain the local alignment with each sequence. Based on the local alignment, the sequences are arranged at the corresponding positions detected by the consensus sequence. Shift right and shift left operations are performed on each sequence to obtain the optimal alignment.

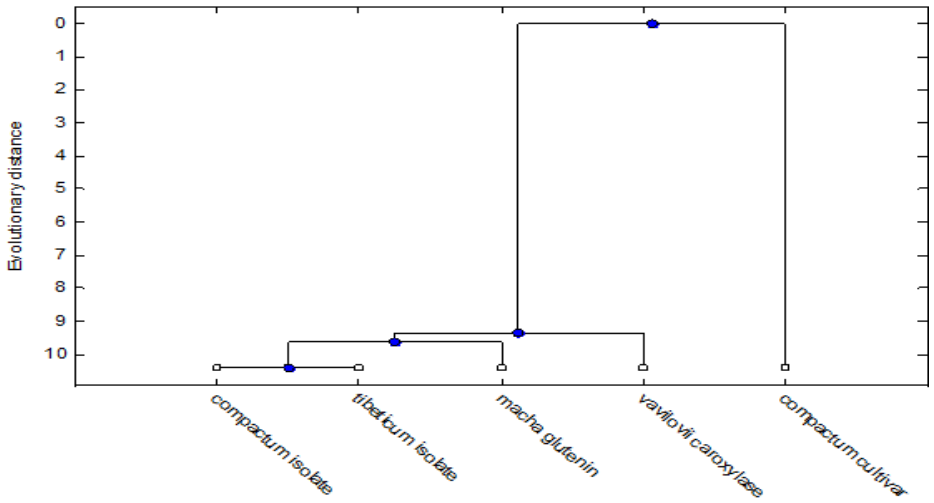

Figure1. Evolutionary distance between for five wheat varieties

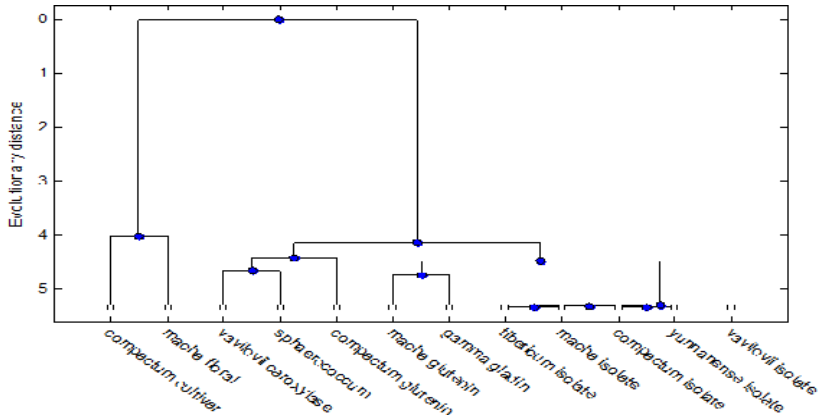

Figure2. Evolutionary distance between for 12 wheat varieties

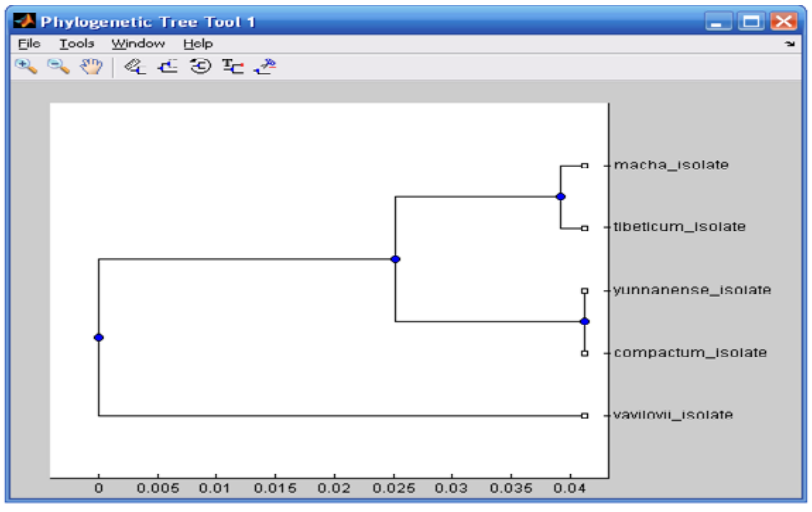

Figure3. Phylogenetic tree for five closely related varieties

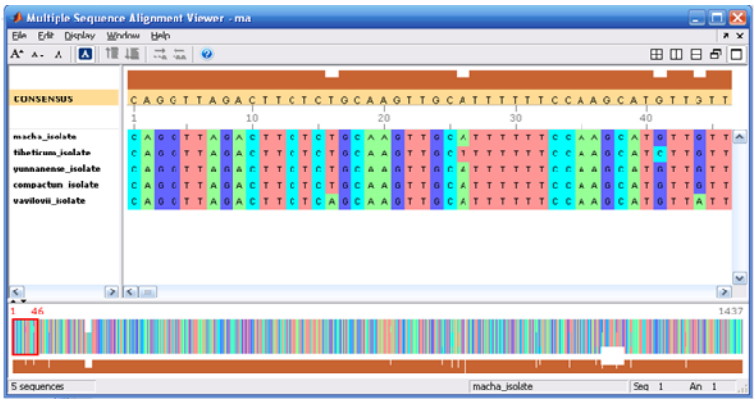

Figure4. Multiple sequence alignment for five closely related species

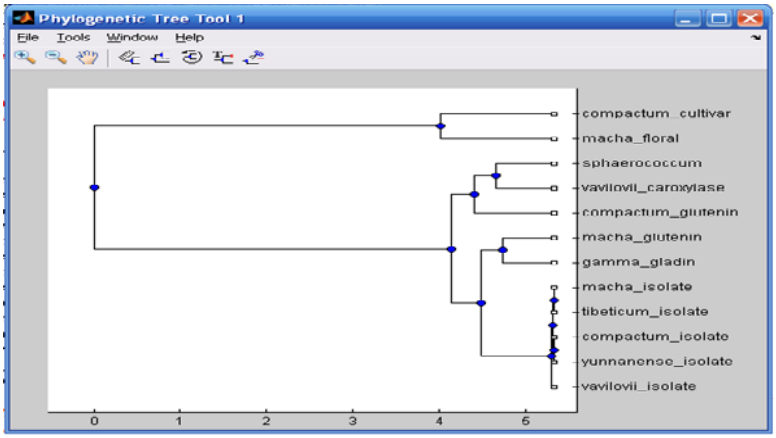

Figure5. Phylogenetic tree for twelve closely related varieties

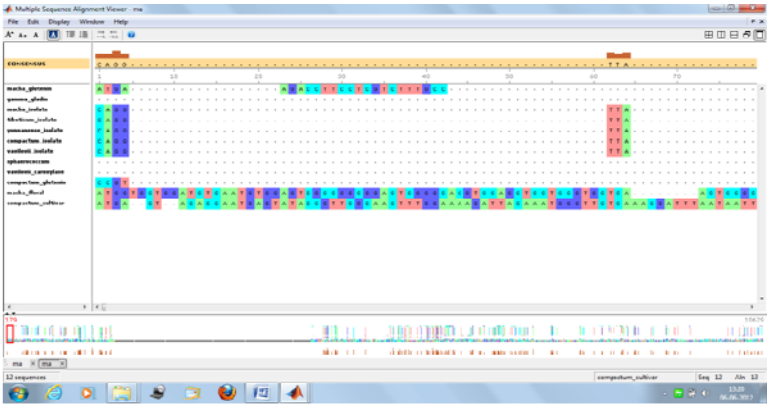

Figure 6. Multiple sequence alignment for twelve closely related species

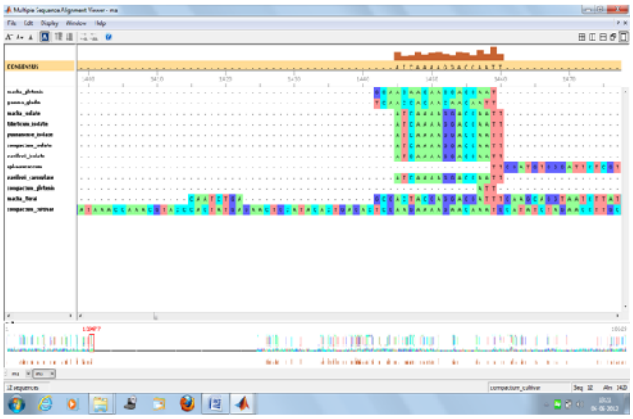

Figure 7. Contd. Multiple sequence alignment for twelve closely related species

Figure 8. Contd. Multiple sequence alignment for twelve closely related species

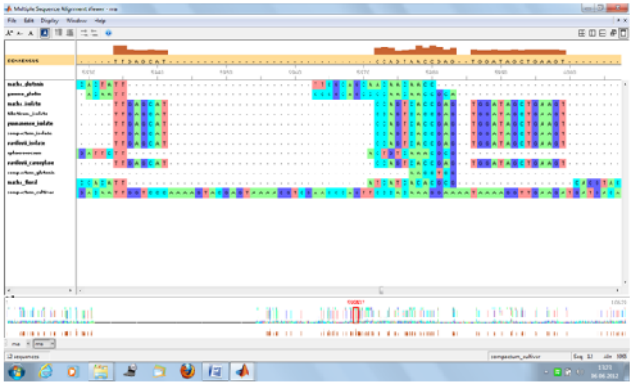

Figure 9. Contd. Multiple sequence alignment for twelve closely related species

7.2 Future Scope

Following improvements regarding the developed model of bioinformatics can be made:

- The model can be extended for protein sequence alignment.
- Various Scoring matrices like PAM and BLOSUM can be incorporated for computing the evolutionary distances.

8. REFERENCES

[1] B.Bergeron, Bioinformatics Computing, Pearson Education, 2003, pp. 110- 160.
[2] Clare, A. Machine learning and data mining for yeast functional genomics Ph.D. thesis, University of Wales, 2003.
[3] Han, J. and Kamber M., Data Mining: Concepts and Techniques, Morga Kaufmann Publishers, 2004, pp. 19-25.
[4] Jiang, D. Tang, C. and Zhang, A., "Cluster Analysis for Gene Expression Data", IEEE Transactions on knowledge and data engineering, vol. 11, 2004, pp. 1370-1386.
[5] Kai, L. and Li-Juan, C., "Study of Clustering Algorithm Based on Model Data", International Conference on Machine Learning and Cybernetics, Honkong, 2007, Volume 7, No. 2., 3961-3964
[6] Kantardzic, M., Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, 2000, pp. 112-129.
[7] Baxevanis, A.D and Ouellette, B.(2001) " Sequence alignment and Database Searching" in Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. John Wiley & Sons, New York, pp. 187-212.
[8] Tzeng YH. Et al. (2004) "Comparison of three methods for estimating rates of synonymous and nonsynonymous nucleotide substitutions", Molecular. Biology Evol. , pp. 2290-8.

AUTHOR

Harmandeep Singh has received B-Tech degree from Punjab Technical University, Jalandhar in 2009 and his M-Tech Degree from Punjab Technical University, Jalandhar in 2012.

Er. Rajbir Singh Cheema is working as a Head of Department in Lala Lajpat Rai Institute of Engineering & Technology, Moga, Punjab. His research interests are in the fields of Data Mining, Open Reading Frame in Bioinformatics, Gene Expression Omnibus, Cloud Computing, Routing Algorithms, Routing Protocols Load Balancing and Network Security. He has published many national and international papers.

Navjot Kaur has received B-Tech degree from Punjab Technical University, Jalandhar in 2010 and pursuing her M-Tech Degree from Punjab Technical University, Jalandhar. She is working as an Assitant Professor in Lala Lajpat Rai Institute of Engineering & Technology, Moga, Punjab.