ABSTRACT
Phylogenetic tree is a branched structure which represents the evolutionary relationships among genes and organisms. Multiple sequence alignment is an initial step in constructing a phylogenetic tree. The most widely used tools for phylogenetic analysis i.e. PHYLIP (Phylogeny Inference Package) and PAUP (Phylogenetic analysis using parsimony) have so far been used for inferring phylogenies. However, the above referred packages intern had to rely on other tools for input. In this context, many open source MSA tools are available for generating both multiple sequence alignment and phylogenetic tree. The purpose of the present paper is to highlight various open source MSA tools for constructing phylogenetic trees using distance based methods after generating the alignment. A comparative study of five MSA tools Geneious, ClustalX, DNAMAN, STRAP and MUSCLE is presented here with a motive of creating awareness among bioinformaticians about MSA tools that helps in constructing phylogenetic trees.

General Terms
Bioinformatics, Software

Keywords
Multiple sequence alignment, Neighbor-joining, UPGMA, Distance matrix

1. INTRODUCTION
Trees can be used to graphically depict the relationship among sequences within the alignment. Once an alignment has been generated and an appropriate model of sequence evolution has been selected a phylogenetic tree can be inferred. A rooted phylogenetic tree is a directed tree with a unique node corresponding to the (usually imputed) most recent common ancestor of all the entities at the leaves of the tree. Unrooted trees illustrate the relatedness of the leaf nodes without making assumptions about common ancestry. While unrooted trees can always be generated from rooted ones by simply omitting the root, a root cannot be inferred from an unrooted tree without some means of identifying ancestry.

There are various methods of building and analyzing phylogenetic trees. Distance methods are based on creating a distance matrix. From the obtained distance matrix a phylogenetic tree is calculated with clustering algorithm. The most commonly used clustering algorithm are UPGMA (unweighted pair group method with arithmetic mean) clustering [1] produces rooted trees and requires a constant rate assumption - that is, it assumes an ultra metric tree in which the distances from the root to every branch tip are equal and Neighbour-Joining clustering [2,3] apply general data clustering techniques to sequence analysis using genetic distance as a clustering metric produces unrooted trees, but it does not assume a constant rate of evolution. These algorithms which calculate genetic distance from multiple sequence alignments are simplest to implement, but do not invoke an evolutionary model.

Table 1. Few requisite properties of MSA tools

<table>
<thead>
<tr>
<th>Name</th>
<th>Link</th>
<th>Year</th>
<th>Author</th>
<th>Operating System</th>
<th>Programmed in</th>
<th>Newer Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNAMAN</td>
<td><a href="http://www.lynnon.com">www.lynnon.com</a></td>
<td>2005</td>
<td>Huang &amp; Zhang</td>
<td>Windows, Macintosh &amp; Linux</td>
<td>JAVA</td>
<td>Same</td>
</tr>
<tr>
<td>STRAP</td>
<td><a href="http://www.charite.de/bioinf/strap">www.charite.de/bioinf/strap</a></td>
<td>2004</td>
<td>Christoph Gille</td>
<td>Windows, Macintosh &amp; Linux</td>
<td>JAVA</td>
<td>Same</td>
</tr>
<tr>
<td>ClustalX</td>
<td><a href="http://www.clustal.org">www.clustal.org</a></td>
<td>1994</td>
<td>Julie Thompson &amp; Toby Gibson</td>
<td>Windows, Macintosh &amp; Linux</td>
<td>C++</td>
<td>Clustal Omega</td>
</tr>
</tbody>
</table>

2. INTERNATIONAL SCENARIO
Many comparative studies were made in the past of MSA tools [4,5,6] but did not highlight the phylogenetic analysis. Essoussi Nadia [7] compared several MSA tools such as ClustalX, Align-m, T-Coffee, SAGA, ProbCons, MAFFT,
MUSCLE and DIALIGN to illustrate comparative phylogenetic trees analysis for two datasets. MUSCLE outperforms the different alignment methods in producing more identical test trees to the reference ones on all datasets used in this analysis.

There have been many algorithms and software programs implemented for the inference of phylogeny. But, still biologists are still dependent on the patent bioinformatics tool like Phylip and PAUP to construct phylogenetic trees which requires the input of a multiple sequence alignment file. Hence, they had to depend on two tools for constructing a phylogenetic tree.

This study helps the bioinformaticians to divert their mind to the upcoming MSA bioinformatical tools that perform the two in one task of performing both the multiple sequence alignment and then constructing phylogenetic tree from that alignment.

3. GENESIS OF THE PRESENT STUDY
Following MSA tools (open source) have been systematically compared to construct a phylogenetic tree by using the distance based methods.

- Geneious (Drummond)
- DNAMAN (Huang and Zhang)
- STRAP (Christoph and Cornelius [8]).
- MUSCLE (Edger [9])
- ClustalX(Thompson [10]).

The reason for choosing the above mentioned tools are as most of them are upcoming tools and can perform both multiple sequence alignment and construct a phylogenetic tree. Few requisite properties of the tools are shown in Table I.

In this study, the Multiple Sequence Alignment (MSA) tools are compared based on the features used for constructing a phylogenetic tree. The features include the algorithms, output format, etc. The study also highlights the best tool which gives a good outcome revealing the genetic relationship by calculating the degree of fit.

4. EXPERIMENTAL SETUP
Five open source multiple sequence alignment tools have been downloaded. Table 1 shows information related to the five tools. Geneious, STRAP and DNAMAN are available as executable files and run on windows as well as linux platform. Secondly, these three tools requires java runtime environment to execute without which it will fail to run. Only MUSCLE is a command line program, hence it requires a lot of manual intervention. But since it is a command line argument program it works faster as it do not require to download any graphics. MUSCLE is the software that has been executed on Linux platform with fedora4 operating environment.

Fifteen protein sequences of Medicago sativa L. plant which is biologically also known as alfa-alfa have been downloaded from Protein Databank (PDB) database (www.pdb.org). The PID’s of these protein sequences are: >1BQ6, >1FM8, >2GAS, >1CGK, >1EYQ, >1FP1, >1FPX, >1H6, >1JX0, >1KYZ, >1SUL, >1U0V, >1YMU, >1J2S AND >1D6H. These protein sequences are saved in fasta format. These five tools were executed on a laptop with Intel Core i3 processor, 2GB ram and 320GB hard disk with a dual boot operating system i.e. Windows7 and Fedora5.

<table>
<thead>
<tr>
<th>Name</th>
<th>Method</th>
<th>Distance Model</th>
<th>Viewer</th>
<th>Export format Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geneious</td>
<td>N-J &amp; UPGMA</td>
<td>Jukes-cantor</td>
<td>Tree &amp; Chromatogram</td>
<td>Phylip, Fasta, Nexus, Geneious</td>
</tr>
<tr>
<td>DNAMAN</td>
<td>N-J &amp; UPGMA</td>
<td>Jukes-cantor, kimura, poisson correction</td>
<td>DNAMAN viewer</td>
<td>Nexus</td>
</tr>
<tr>
<td>STRAP</td>
<td>N-J</td>
<td>Jukes-cantor</td>
<td>ATV, JalView &amp; GeneBe e</td>
<td>Phylip &amp; Nexus</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>N-J</td>
<td>Kimura</td>
<td>NA</td>
<td>ClustalX, msf, Phylip, html</td>
</tr>
<tr>
<td>ClustalX</td>
<td>N-J &amp; UPGMA</td>
<td>Kimura</td>
<td>NA</td>
<td>Clustal, Phylip &amp; Nexus</td>
</tr>
</tbody>
</table>

NJ: Neighbor joining
UPGMA - unweighted pair group method with arithmetic mean

5. RESULTS AND DISCUSSION
In this study five most upcoming and widely used MSA tools have been tested under the windows and/or linux platform to construct the phylogenetic trees. The procedure to construct a phylogenetic tree from any of the five MSA tools is shown in Fig.1. The study also finds degree of fit (R²) by using the TreeView[11] tool.

Fig 1: Procedure to construct Phylogenetic Tree

The procedure is as follows:
- Downloaded fifteen protein sequences from the public database i.e. PDB (Protein Data Bank, url: www.pdb.org).
- The fifteen sequences are given as input to the software which first finds the sequence alignment.
- Set the parameters required to construct the tree like the distance method and the distance model used.
- The multiple sequence alignment from the above step is given as input to the same software and displays the phylogenetic tree.

Table 3. Degree of fit in terms of $R^2$

<table>
<thead>
<tr>
<th>Tools</th>
<th>Geneious</th>
<th>DNAMA</th>
<th>STRAPER</th>
<th>MUSCLE</th>
<th>ClustalX</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.2</td>
<td>0.05</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Among the selected tools in this study tools like ClustalX and MUSCLE constructed the phylogenetic tree (Newick Format [12]) by using the above steps but did not view the tree. The degree of fit of a tree to a matrix of genetic distances can be quantified with $R^2$, the proportion of variation in the genetic distance matrix that is explained by the tree. If $R^2$ is near 1.0, the tree represents a good summary of the genetic relationships shown in the distance matrix. If $R^2$ is not near 1.0, the tree does not represent a good summary of the genetic relationships among populations [13]. The degree of fit $R^2$ is calculated by using the tool TreeView. TreeView only takes nexus and phylip format of the phylogenetic trees as input file. Table 3 shows the $R^2$ values generated from TreeView which accepted the nexus format phylogenetic tree as input from the five tools.

6. CONCLUSION

The comparative analysis accomplished in the present paper evidences that the phylogenetic tree generated from strap tool describes a better genetic relationship. The same is also evident from the $R^2$ value as the same is greater as compared to the other $R^2$ values as shown in Table 3. All the tools perform equally well in producing reliable phylogenetic trees, whereas tools such as MUSCLE and ClustalX lack the feature of displaying phylogenetic tree.

The authors are in a process to extend the research work by taking into account more prevailing tools so as to get better degree of fit values. This will empower the Bioinformaticians to undertake the calculations of $R^2$ values pertaining to phylogenetic tree irrespective of other tools.

7. REFERENCES


