International Journal of Knowledge Engineering
ISSN: 0976-5816 & E-ISSN: 0976-5824, Volume 3, Issue 2, 2012, pp.-184-187.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000230

BOOTSTRAPPING ANALYSIS FOR A SET OF SEQUENCES FROM VARIOUS PLANTS CONTAINING CHALCONE SYNTHASE

SRIMANI P.K.[1]* AND KUMUDAVALLI M.V.[2]

[1]R&D, Bangalore University, Bangalore-560078, Karnataka, India.
[2]SCSVMV University, Kanchipuram- 631561, TN, India.
*Corresponding Author: Email- profsrimanipk@gmail.com

Received: October 25, 2012; Accepted: November 06, 2012

Abstract- Enzymes are biological catalysts. They speed up the chemical reactions that take place within living cells, without having any overall change. Chalcone synthase is one such enzyme which is present in the set of sequences under study. A set of species if related, then the relationship is called a Phylogeny. Generally these relationships are represented by a phylogenetic tree. The task of phylogenetics is to infer this tree from observations upon the existing organism. One of the tree building algorithms used in the present study is Maximum Parsimony, which works by finding the tree which can explain the observed sequence with a minimal number of substitutions. For the data set the Bootstrapping method is applied to assess the reliability of trees.

Keywords- chalcone synthase, sequences, bootstrapping, parsimony, phylogenetics, reliability, random, consensus.

Citation: Srimani P.K. and Kumudavalli M.V. (2012) Bootstrapping Analysis for a Set of Sequences from Various Plants Containing Chalcone Synthase. International Journal of Knowledge Engineering, ISSN: 0976-5816 & E-ISSN: 0976-5824, Volume 3, Issue 2, pp.-184-187.

Introduction

A phylogenetic tree or an evolutionary tree is a branching diagram which indicates the ancestral relationships among the species. The phylogenetic tree of a group of sequences does not necessarily reflect the phylogenetic tree of their host species, because gene duplication is another mechanism, in addition to speciation, by which two sequences can be separated and diverge from a common ancestor. Genes which diverged because of speciation are called orthologues. Genes which diverge by gene duplication are called Paralogues. A set of orthologues gene sequences are selected for the current study.

The main method for phylogenetic analysis is predicting the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequence from common ancestral sequences is Maximum Parsimony method. For the same reason, it is also sometimes referred to as the minimum evolutionary method. This method is best suited for sequences that are quite similar and is limited to small number of sequences.

Parsimony

It is a tree building algorithm which works by finding the tree which can explain the observed sequences with a minimal number of substitutions. Unlike in distance-based algorithms it uses a different general strategy, instead of building a tree, it assigns a cost to a given tree, and it is necessary to search through all topologies or to pursue a more efficient search strategy that achieves this effect, in order to identify the 'best tree'. The two components of the algorithm are:

I. The computation of a cost for a given tree T and
II. A search through all trees, to find the overall minimum of this cost.

Bootstrap

The tree building algorithm i.e Maximum Parsimony method gives us a tree but with no measure of how much they should be trusted. As suggested by [1], usage of bootstrap [2] as a method of assessing the significance of some phylogenetic feature, such as the segregation of a particular set of species on their own branch or a 'clade' is considered. Bootstrapping is a method of assessing the reliability of trees. Bootstrapping method differs by deliberately constructs sequence data sets that differ by some small random fluctuations from the real sequence. The method is then repeated on the randomized sequences in order to see whether the same tree topology is obtained. The randomized sequences are constructed by sampling columns from the original sequence alignment.

To carry out bootstrapping analysis, many sets of randomized sequences are constructed (usually 100 or 1000). The tree construction method is repeated on each set of sequences to produce a set of possible trees, some of which will be different. We then look at

each group of species in the original tree and determine what percentage of the randomized trees contains this same group. This percentage gives us a measure of confidence that those species do really form a related group.

This paper is organized as follows: Section II deals with related works, Section II discusses the need and importance of the problem, Section IV Deals with the methodology, Section V gives the data set Description, Section VI deals with the experiments and results and Section VII gives the conclusion.

Related Works

A thorough survey of the literature pertaining to the subject reveals that very sparse literature is available in this direction. Some recent works include [3-8]. Absolutely no work is available with regard to the present work. Hence, the present investigation is carried out.

Need and Importance of the Problem

Phylogenetic tree or Evolutionary tree construction and analysis are a major phase in Bioinformatics. As discussed earlier parsimony algorithm builds a best tree with a minimum number of substitutions. But there is no measure of how much these trees are to be trusted. For the same reason the Bootstrap analysis is to be made for assessing by how much percentage the components of the trees are related. Various tools and computer programs are available in this regard. Usage of some available tools for tree building and bootstrapping the same with input data and the analysis of its various outputs are the main objective of this research work**.** Since not much work in this area has been done, the present investigation is carried out to throw some light on the qualitative as well as quantitative aspects of the problem.

Methodology

The following steps were used in the analysis process,

Step 1: The data set was used in CLUSTALW for alignment and to get the resulting file in PHYLIP format.

Step 2: The output file was then used to get the maximum parsimony tree. The algorithm followed in Maximum Parsimony method is guaranteed to find the best tree, because all possible trees relating a group of sequences are examined. For this reason, the method is quite time-consuming and is not useful for data that include a large number of sequences with a large amount of variation. The main program for maximum parsimony analysis in the PHYLIP package [9], DNAPARS- which treats gaps as a fifth nucleotide state was used for the same.

Step 3: The set of sequences were used in CLUSTALX to create a '.phb' file.

Step 4: Later the .phb file was used in N-J plot to create the Neighbor -joining tree which gives the bootstrap percentages.

Bootstrap works as follows: Given a data set consisting of an alignment of sequences, an artificial dataset of the same size is generated by picking columns from the alignment at random with replacement. A given column in the original dataset therefore appears several times in the artificial dataset. The tree building algorithm is then applied to this new dataset, and the whole selection and tree building procedure is repeated the same number of times, typically of the order of 1000 times. The frequency with which a chosen phylogenetic feature appears is taken to be a measure of the confidence we can have in this feature. For certain probabilistic models, the bootstrap frequency of a phylogenetic feature F can be shown to approximate the posterior distribution P (F | data). When the bootstrap is applied to a non-probabilistically formulated model, such as parsimony it can be interpreted in terms of statistical hypothesis testing [10].

Data Set Description

All enzymes are proteins. One such enzyme is 'chalcone synthase'. A set of 23 sequences belonging to plants kingdom which contain the protein coded gene 'chalcone synthase' was being selected from NCBI's Nucleotide Database using BLAST as in [Table-1]. Then the data set was used in DNAPARS program of PHYLIP package to find the most parsimony tree. The set of sequences under study was used in CLUSTALX to get the resulting file with '.phb' extension, which was later used for bootstrapping using N-J Plot.

Table 1- Contains The Sequence Details

S. No.	Locus/Accession	Organism
1	JN830647.1	*Pyrus pyrifolia*
2	HQ853494.1	*Malus toringoides*
3	DQ286037.1	*Sorbus aucuparia*
4	AF400567.1	*Rubus idaeus*
5	HQ423171.1	*chalcone synthase*
6	AB201756.1	*Fragaria x ananassa*
7	JQ247184.1	*Camellia sinensis*
8	AM263200.1	*Humulus lupulus*
9	AJ413277.1	*Rhododendron simsii*
10	X94706.1	*Juglans nigra*
11	JN654702.1	*Vaccinium corymbosum*
12	JQ627646.1	*Lonicera japonica*
13	AB009350.1	*Citrus sinensis*
14	JF795272.1	*Gossypium hirsutum*
15	HQ127337.1	*Phlox drummondii*
16	EU430077.1	*Senna tora*
17	AY237728.1	*Glycine max*
18	L24517.1	*Trifolium subterraneum*
19	FJ705842.1	*Capsicum annuum*
20	AY170347.1	*Arachis hypogaea*
21	DQ208973.1	*Cardamine maritima*
22	AF112108.1	*Barbarea vulgaris*
23	AF144530.1	*Rorippa amphibia*

Experiments and Results

In this section, the results of the present investigation are presented in [Fig-1], [Fig-2], [Fig-3], [Fig-4], [Fig-5], [Fig-6]. The experiments are conducted as per the steps mentioned in section IV.

[Fig-2] shows the best possible tree for the given data set consisting of 23 sequences. During the process it generates various trees which differ in the order in which the sequences have been assigned to the ancestral nodes so as to minimize the number of changes needed in the whole tree.

[Fig-3] is the most parsimonious UNROOTED Tree for the given Dataset

[Fig-4] shows the most parsimonious tree for the given set of se-

quences as a PHYLOGRAM. A Phylogram which is also known as an Additive tree has additional information where edge lengths are drawn proportional to some attribute.

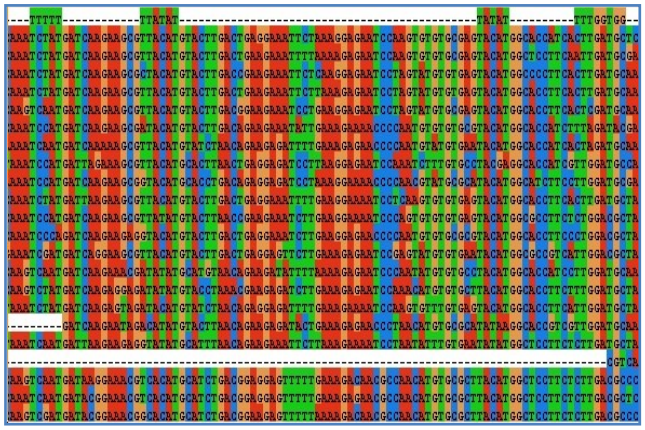

Fig. 1- A Partial view of the sequence alignment using CLUSTALX

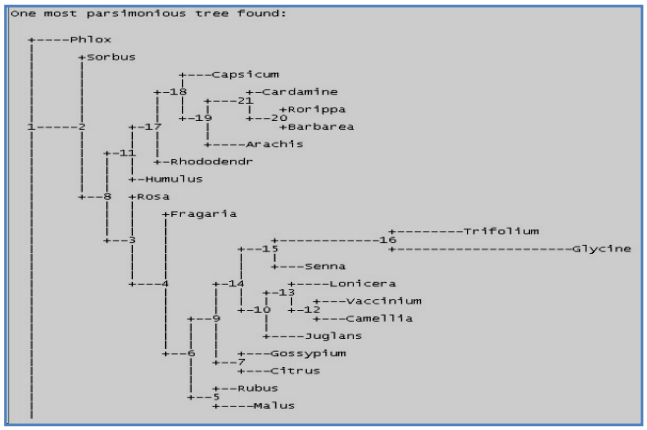

Fig. 2- Most Parsimonious Tree

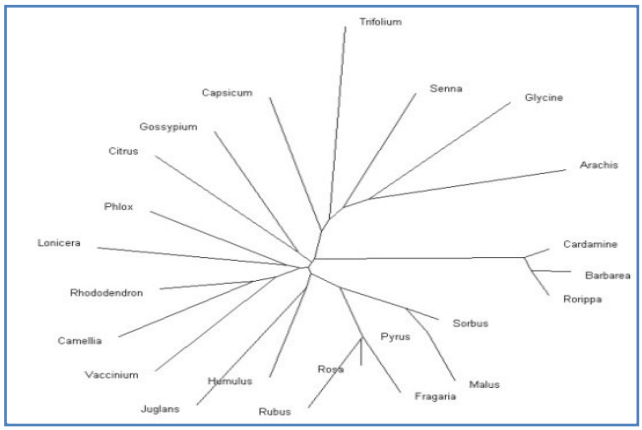

Fig. 3- Parsimonious Tree (Unrooted).

The numbers that appear on the internal nodes of the tree are called Bootstrap Percentages. From the figure5 it is clear that Malus and Pyrus have 100% bootstrap support. Similarly Sorbu and Fragaria and also their connecting ancestral node have 100% bootstrap support. Whereas, Rorippa and Barbarea have 90% support. The weakest point in this phylogenetic tree is the very low value of 37.7% at the node linking the Vaccinum, Phlox, Lonicera, Rhododendron, Camellia and Juglans, Humulus.

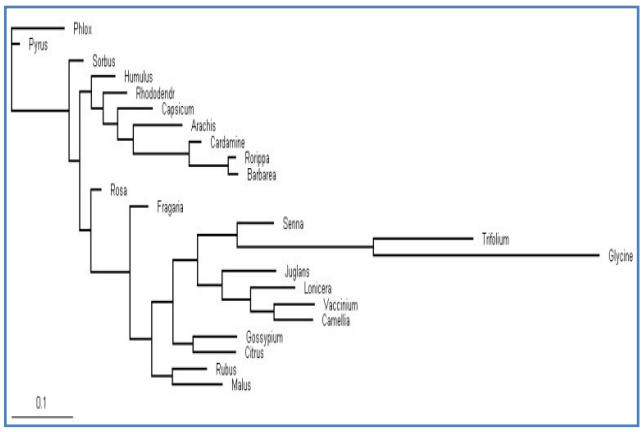

Fig. 4- Phylogram

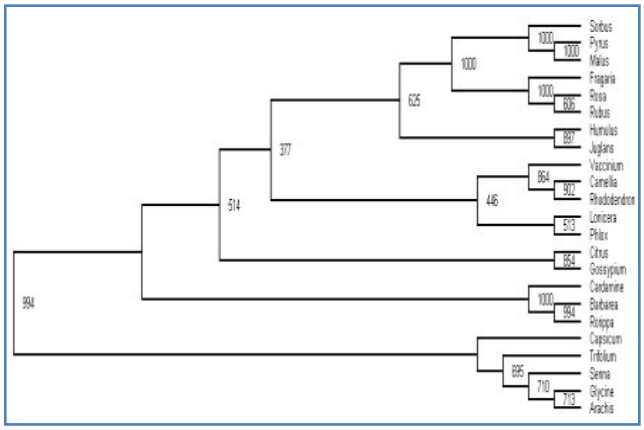

Fig. 5- Bootstrap N-J Tree.

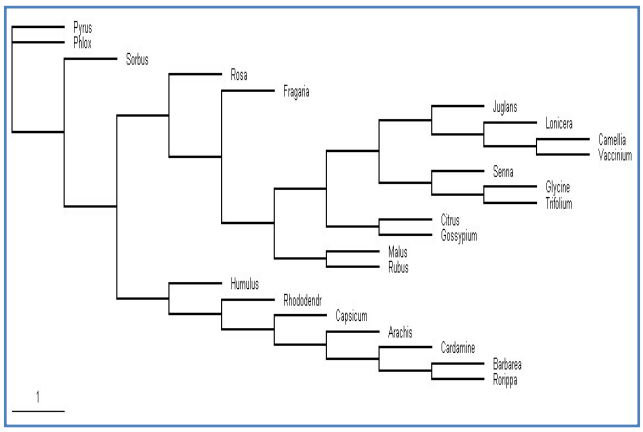

Fig. 6- Consensus Tree

Conclusion

There is no precise rule to say how high a bootstrap percentage has to be before we can be sure that the group of species forms a 'true clade'. Generally it is often thought to be greater than 70%. But sometimes may not be an accurate percentage to decide the reliability of the output trees. Therefore Bootstrap results are often presented in the form of consensus trees as in Figure 6. The frequency of occurrence of each possible clade in the set of bootstrap trees is determined and clades are ranked in descending order of frequency. The consensus tree is constructed by adding clades one at a time working from the top of the ranking list. Each clade

added is, the one with the highest frequency that is consistent with all the clades already added. The final consensus topology may differ slightly from the tree obtained with the original full set of data. It is then a matter of choice whether to present the original tree labeled with bootstrap percentage for the clades obtained in that tree, or to present the consensus tree, which will tend to contain clades with slightly higher bootstrap values that were not present in the original tree. Well-determined clades, with highest bootstrap values, will almost always occur in both the consensus and the original tree, so this issue affects only the way the results are presented for the least well-determined parts of the tree. The results are encouraging.

Acknowledgement

One of the authors Mrs. Kumudavalli M.V acknowledges Dayananda Sagar Institutions, Bangalore, Karnataka and SCSVMV University, Kanchipuram, Tamilnadu, India for providing the facilities for carrying out the research work.

References

[1] Felsenstein J. (1985) *Evolution,* 39, 783-791.

[2] Efron B. and Tibshirani R.J. (1993) *An Introduction to the Bootstrap*.

[3] Scott M. Lanyon (1985) *Systematic Zoology*, 34(4), 397-403.

[4] Naruya Saitou (1989) *Systematic Zoology*, 38(1), 1-6.

[5] James W. Archie (1989) *Systematic Zoology*, 38(3), 239-252.

[6] Teresa Przytycka (2007) *Journal of Computational Biology*, 14 (5), 539-549.

[7] Thorpe and Dickinson W.J. (1988) *Systematic Zoology*, 37(2), 97-105.

[8] Srimani P.K., Kumudavalli M.V. (2012) *International Journal of Current Research*, 4(5), 206-210.

[9] Felsenstein J. (1996) *Methods in Enzymology*, 266, 418-427.

[10] Efron B., Halloran E. and Holmes S. (1996) *Proceedings of the National Academy of Sciences of the USA,* 93, 13429-13434.