A Survey on Methods of Phylogenetic Analysis for Constructing Phylogenetic Tree

B.MADHAV RAO
Assistant Professor, Department of CSE
SIR C.R.REDDY COLLEGE OF ENGINEERING

Abstract- Protein function prediction plays the major role in designing drug and identifying the cause for the disease. For this purpose we are using many methods like sequence, genomic, structure, phylogenitics etc. Among these phylogentic analysis plays a major role for detecting function of the Protein by constructing pylogenitic tree. These trees are analyzed based on two techniques optimality criterion and choice of search strategy.

Keywords- Phylogenetic Analysis; Protein Sequence; Maximum likelihood

I. INTRODUCTION

Phylogenetic analysis [1] is a process of finding relation among the molecules phenotypes, and organisms. The Phenomena we consider for comparing identified characteristics of the species, natural assumption that species with similar characteristics are closer. Phylogeny analyze the relationship between species, generally we represent this as phylogenetic trees. Traditional phylogenitics consider physical features like color, size, etc.Modern phylogenetics uses information extracted from genetic materials like protein sequence, and DNA and RNA.We consider sequences for constructing and analysis of phylogentic trees. The relationships between species are then reduced from well conserved blocks in the alignment of several sequences, one from each examined species.

This Analysis consists of two major decisions.

1. Optimality criterion
2. Choice of *search strategy.*

Problems in choice of *search strategy* for consideration of tree space we can't determine the best tree among all possible trees for a set of protein sequences, for below reasons.

i) If number of sequences increases then the possible number of trees are increases exponentially. So that even though small increase in sequence it generates more phylogenetic trees.

ii) Because *complete* evaluation of such large numbers of trees is impracticable for data sets that contain a dozen sequences or more.

Analyzing such large trees was difficult for sequence dataset contain a dozen or more sequences.

Phylogeneticists developed many search strategies that effectively finding the best phylogenetic tree. All these search mechanisms are faster and accurate then exhaustive searching techniques, but does not guarantee to generate best tree. so that they used different optimality criteria techniques to analyze the phylogenetics.

THE MAXIMUM LIKELIHOOD OPTIMALITY CRITERIA:

Maximum likelihood is a mathematical method[4][5] to determine the unknown characteristics (parameters) of a probability model. A parameter is some characteristic feature of a model. In this generally we use parameters as mean and variance. In phylogenetic analysis we consider parameters as differential transformation cost,rates,size and tree characteristics .It was defined as quantity is directly proportional to the probability of given dataset of model, P(D|M).we can find the probability of the dataset of a model. Then we analyze the likelihood function and find maximum value of the parameters generally the tree length.

Advantages of maximum likelihood methods are most accurate and suitable for analyzing the DNA sequences, low variance that will reduce the sampling error, this method was statistically well structured, and by using the sequence information it analysis different tree topologies.

Disadvantages of this method was slow to evaluate, result generated in this method was dependent on the model and data. Questionably applicable to complex data like morphology given the difficulty of modeling the numerous processes, philosophically this model was less well established.

PHYLIP

PHYLIP[8](PHYLogeny Inference Package) contains more than thirty five programs. PHYLIP program source code will be written on C programming language. Many precompiled executables are available all platforms. It takes input as ASCII or plain text format. many protein databases provide PHYLIP supported datasets for download. It generates a tree as output. This network format used to analysis the tree using another programs.

Advantages of PHYLIP

1. PHYLIP is open source program

2. It has the cross platform compatibility
3. Well Structured documentation for user understanding
4. Easy to Use
5. It used to explain phylogenetics easily when compared to other methods.
6. Execution of this can be automated by using command files and input redirection
7. Many programs supports the PHYLIP format

Limitations of PHYLIP

1. Tree search is less thorough when compared PAUP*.
2. Very slow When Compared to packages like PAUP* and RAxML
3.It supports only command interface
4. Manual steps such as renaming file names can be to long
5. Still no: codon model, Bayesian inference.
6. Only few basic options available to end user
7. It does not support some files like NEXUS standard files.

RAXML

Randomized Accelerated Maximum Likelihood (RAXML)[6] was designed to perform phylogenetic analysis for both parallel and sequential using maximum likelihood optimality criteria. It was designed to analysis large number of datasets. It was developed to process any data set irrespective of their size. RAXML can use a variety of different character sets, including nucleotide, amino acid, binary, and multi-state character state data.

RAXML working phenomena based on optimality criteria, In this first step RAXML construct the basic tree or starting tree with zero sequences, then after one by one sequence is added to starting tree in random order, and determine the optimality location on the tree under the parsimony optimality criterion. For taking sequences in random order each time we run this program we get new starting trees. RAXML is compared to all most all real world biological

PAUP*

PAUP*[11] defined as **P**hylogenetic **A**nalysis **U**sing **P**arsimony and star (*) represents other methods. This was developed by David L. Swofford.before version 4.0 PAUP uses only parsimony to generate phylogenetic trees. After version 4.0 this supports other methods like likelihood and distance matrix.PAUP supports rich graphical user interface. With the help of MacClade program it shares the NEXUS data format.PAUP* has a limitation that does not perform some features in MEGA. It is not available for free of cost.PAUP uses two searching strategies for optimal trees; heuristic and exact. Heuristic search methods do not guarantee for optimality,but it takes less time to perform search. Exact search method guarantee to discover the optimal trees, but it consumes more time to execute.

BEAST

BEAST[7] is a package for hypothesis testing and evolutionary parameter estimation .In this large number of different evolutionary models possible ,therefore summarizing this is difficult. The features published by BEAST software results the high throughput. Next version of this software includes faster and more flexible codon-based substitution models, models of continuous character evolution and new relaxed clock models based on random local molecular clocks. The overall architecture of BEAST package contains several components like BEAUti,Tracer. BEAUti is a software developed in JAVA programming language and distributed with BEAST that provides a GUI for generating BEAST XML input files for a number of simple model combinations.Tracer proves graphical tool for BEAST output analysis. It analysis output of common MCMC packages such as MrBayes and Bali-Phy. BEAST is a cross-platform program for Bayesian analysis of molecular sequences using MCMC. BEAST uses MCMC to average over tree space, so that each tree is weighted proportional to its posterior probability. It uses relaxed or strict molecular time clocks. BEAST package contains more than 35 different program modules.

BOOTSTRAP

This Method starts by creating B replicate datasets.Each Dataset is generated by taking repeatedly sampling n adjustment sites with replacement. After getting these B replicate datasets we apply maximum likelihood estimation to these B bootstrapped[2] sequence alignments. By summarizing the similarities between bootstrapped phylogenetic data we assessed uncertainity.Figure1 shows Some bootstrap datasets generated from an observed sequence alignment. In this columns are random for each data set sequence alignment.Bootstrap sampling is concept is easy, but summarizing bootstrap trees is very difficult to implement. If we repeatedly execute this method for same phylogentic datasets to obtain replicated molecular sequence alignments. Then based on this alignments we estimating the phylogenies. Maximum likelihood method

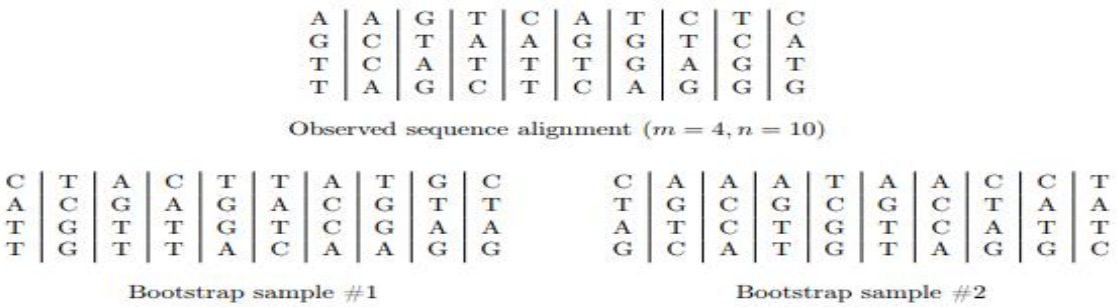

Figure :1 An example of bootstrap sampling for sequence alignment data. We obtain bootstrap datasets by randomly sampling $n = 10$ sites (i.e. columns) with replacement from the observed sequence alignment (top). We provide two examples of possible bootstrap datasets (bottom).

used to estimate the sampling variability of our estimation.

II. CONCLUSION

The Trees generated from above methods like BEAST,RAXML,PAUP* and etc are different from each other. In order to compare and analysis we need to assume some evolutionary model so that the trees may be tested. The data we consider for constructing the phylogenetic tree may changed from one method to another menthod,because in phylogenitic analysis we consider all or some parameters of the molecules or protein sequences.

III. REFERENCES

[1]. Antonis Rokas, "Phylogenetic Analysis of Protein Sequence Data Using the Randomized Accelerated Maximum Likelihood (RAXML) Program", Current Protocols in Molecular Biology 19.11.1-19.11.14, October 2011.

[2]. Abascal, F., Zardoya, R., and Posada, D. 2005. Prottest: Selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105.

[3]. Whelan, S., Lio, P., and Goldman, N. 2001. Molecular phylogenetics: State-of-the-art methods for looking into the past. Trends Genet. 17:262-272.

[4]. Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367-372.

[5]. Zwickl, D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Doctoral thesis. The University of Texas at Austin.

[6]. Buckley, T., C. Simon, and G. K. Chambers, "Exploring Among-Site Rate Variation Models in a Maximum Likelihood Framework Using Empirical Data: Effects of Model Assumptions on Estimates of Topology, Branch Lengths, and Bootstrap Support". Syst. Biol. 50(1):6786, 2001.

[7]. Alexei J Drummond*1,2 and Andrew Rambaut3," BEAST: Bayesian evolutionary analysis by sampling trees", Published: 8 November 2007 BMC Evolutionary Biology 2007, 7:214 doi:10.1186/1471-2148-7-214

[8]. Amrit Dhar and Vladimir N. Minin," Maximum Likelihood Methods for Phylogenetic Inference",1-13,

[9]. http://evolution.genetics.washington.edu/phylip/general.html

[10]. http://evolution.gs.washington.edu/sisg/2014/2014_SISG_12_7.pdf

[11]. 11.http://phylosolutions.com/paup-documentation/paupmanual.pdf