Well- Established Bootstrap Algorithms

Asmaa Abdel-Kader[1], Ahmad Moustafa[2], Ibrahim Elhenawy[3]
[1]Teaching assistant at faculty of computer and informatics, Zagazig University.
[2] Lecturer at faculty of computer and informatics, Zagazig University.
[3] Professor at faculty of computer and informatics, Zagazig University.
Zagazig, Egypt

Abstract

Maximum-Likelihood is considered to be a powerful statistical method for the process of phylogenetic inference from sequence data (DNA & RNA).Unfortunately, the computation power need for carrying out this method is very high. Although there is a continuous advance in the computational tools and the tools for high-performance computing, the process of finding the phylogenetic tree using ML is still problematic due to the burden of bootstrapping process used in the ML method, which requires very high computations. In this paper, we will go through the different methods developed to optimize the process of bootstrapping.

Keywords: *Bootstrap; phylogenetic analysis; phylogenetic tree.*

I. INTRODUCTION

Phylogenetic analysis is the study of the evolution of a group of organisms or spices - called “Taxonomies” in the phylogenetic language – to find the evolutionary relationships between these organisms through analyzing changes occurred in the organism’s genetic information through their physical characteristics. Phylogenetic analysis also seeks to estimate divergence time between groups of organisms sharing a common ancestor. Phylogenetic analysis is used in many areas. It can be used in the discovery of new drugs and can be also used in many biological applications such as conservation biology, epidemiology and much other application as Stamatakis mentioned in his survey [1]. The output of the phylogenetic analysis is a phylogenetic tree that models the relationship between organisms in a branching diagram of a tree.

Bootstrap method have been developed at first by Eron[1].The basic idea behind bootstrapping is, given a datapoints$x_1,x_2,\ldots,x_n$ which aredrawn form a certain distribution and an estimator

$$t = T(x_1,x_2,\ldots,x_n)$$

Bootstrapping is used to create new resampled (fictional) dataset,

$$t^* = T(x_1^*,x_2^*,\ldots,x_n^*)$$

In the original version of bootstrapping made by Efron the new resampled dataset is a copy of the original dataset with some data points omitted and others duplicated. The process is repeated many times, each time a new estimator *t* is produced. We end up with many estimators that almost follow the same distribution. It will help to predict the original one. So bootstrapping is useful when the distribution from which the original data points are drawn form is unknown or when the statistical function T is very complex and untraceable. New methods of bootstrapping are then designed such as Bayesian bootstrap,Smooth bootstrap, parametric bootstrap. In phylogenetic analysis, a part of the genetic information for each organism is used to construct a phylogenetictree then bootstrapping is used to test uncertainty about the constructed tree by resampling the dataset and building a phylogenetic tree for each of the resampled data item, as a result of bootstrapping processa new tree is created with annotation on the branches that show branch support values based on the number of time a given branch occurred in all of the newly created trees.

II. BACKGROWND

Phylogenetic tree is a diagram or a tree that is constructed using computational methods for a set of organisms to show the evolutionary relationships between them. The tree consists of nodes (internal and external) and branched. External nodesrepresents the organisms under investigation (taxonomic units), theinternal nodes represents the relationship between the taxonomic units and branch lengths shows the number of changes happened to the sequence over the time period under investigation.There are two types of trees Rooted and Unrooted. In the rooted tree, the root represents the common ancestor of all the organisms for which the phylogenetic tree is constructed. In the unrooted tree, all the organisms are connected but It is not known which of them is common ancestor. There are two main categories of methods for building a phylogenetic tree. The first one is the Distances based methods such as Neighbor joining and UPGAMA methods. The second one is the Character based methods such as Maximum Parsimony and Neighbor joining methods.The first step in building any phylogenetic tree is finding the multiple sequence alignment for the sequences to be used in the tree construction process. In Distance based methods, The sequences is transformed into pairwise distances, and then use the matrix during tree building. In character based methods the alignment is used directry.For Maximum Parsimony (MP), It works on minimizing the total number of evolutionary steps that can explain the dataset used for tree construction. All possible trees are constructed then the algorithm searched for the optimal one which has the minimum number of changes (i.e. the tree with the minimum tree length). This method is suitable for closely related sequences. For Maximum Likelihood method (ML), It uses a statistical model for building the tree. Given a model of nucleotide substitution, ML method try to find the most likely tree that obeys the evolutionary model. This method is best used with distantly related sequences.

III. STATEOFTHE ART

A. Standard Bootstrapping (SBS)[2]

Is the first bootstrapping technique developed by Felsensteinwho applied the statistical model made by Efron[1]in phylogenetic analysis.In this technique a set of pseudo-replicates are created as a result of a sampling process. Each replica is a copy of the original data file with some characters duplicated and other dropped out. The phylogenetic tree for each replica is then constructed and the newly created trees are used to build a consensus tree. This method was very successful and replaced the confidence intervals in phylogenetic trees but the main disadvantage is that it requires very high computationalresources to build a tree for only few dozens of taxa.

B. Resampling Estimated Log Likelihoods(RELL)[1][3]

In 1990 Kishano Developed two different techniques that find enhance the performance of the bootstrap process. He was the first one to find the bootstrap value without inferring the full phylogenetic tree for each replica of data. The two developed techniques was Multivariate Normal Distribution (MND) and RELL. The bootstrap probabilities estimated from RELL and MND was almost the same. RELL is simpler but MND is favorable when the sequence is too long because MND’s running time is independent from the sequence size. RELL method was much more promising technique which was further used in many enhanced techniques.

D. Local Bootstrap Probabilities(LBP)[4]

The LBP method is inferred mainly from the RELL method. In LBP the tree is traversed branch by branch, for each one a local rearrangement of the outgroups of the branch called Nearest-Neighbor Interchanges (NNI) is constructed and the RELL method is the used to calculate the bootstrap probability for each of the rearrangements and the one with the highest probability is chosen .The process continues until the tree is traversed without finding any further improvements. The main disadvantage of this technique is that LBS may be deceptive if the relationship between the outgroups in the branch’s subtree is incorrect. In addition the algorithm may be trapped in local maxima very often.

E.Subtree pruning and regrafting (SPR)[5]

The SPR algorithm is an efficient way to search the tree space. It works on two stages. First, it works globally to discard the SPR moves which tend to be less promising based on a distance-based method. Second, it workslocally by estimating the likelihood change for the remaining SPRs. The likelihood values resulting from this technique was as good as the other phylogenetic inference method but the computation time was greatly reduced. In addition, the algorithm works well even if the starting tree was poor.

F. approximate Likelihood-Ratio Test (alRT)[6]

In 2006, Anisimova has modified of the standard LRT method to produce ApproximateLikelihood-Ratio Test (aLRT).Instead of computing the LRT value as 2(l1-l0) where l0 is the maximum-likelihood of the tree whichassume that the branch of interest is collapsed (null hypothesis) ,the value of aLRT is computed as the difference between the branch length in the best tree and the second best ML Tree between the alternatives around a certain branch multiplied by 2 i.e. $2(l_1-l_2)$.aLRT results in a more conservative test than LRT. The algorithm was very fast because the log-likelihood value l_2 is computed and optimized only over the branch of interest and its four adjacent branches where other parameters are fixed to the optimal value.

G.Shimodaira - Hasegawa approximate Likelihood Ratio Test (SH-aLRT)[7]

In 2010, Guindon has suggested two modifications to improve bootstrappingof the algorithms used in PHYML software. He used the SPR with a maximum parsimony method instead of distance-based method. This small change has enhanced the performance greatly because parsimony and likelihood are theoretically interconnected with an inverse relationship (i.e. minimizing parsimony is maximizing Likelihood).The second one was to use the SH test with aLRT. Due to the difference between simulated and real data, the aLRT is expected to violate the substitution model assumption. To correct the violation the SH test is used to check whether the log-likelihood obtained with each method and each dataset was significantly smaller than the likelihood of the most likely tree found for the corresponding alignment.

H.Ultrafast Boot[8]

This algorithm makes use of the ML-tree used in the process of the ML-tree construction in the likelihood evaluation process to evaluate the likelihood of the bootstrap MSA.It collects the trees with log-likelihood values larger than a predefined threshold value (l) in the tree space resulting from the NNI method. Then the RELL bootstrapping is applied to the collected trees and the tree with higher RELL value is chosen to be the current bootstrap tree for the corresponding bootstrap alignment. The consensus tree is computed from the set of collected bootstrap trees, but this method may lead to the problem of overoptimisticbootstrap supports especially for short branches because it saves only one optimal tree for each bootstrap.

In 2017, a new version of UFboot is released[9] to overcome the problem of overoptimistic.Instead of maximizing the RELL value for the selected tree, It selects randomly one of the trees with a RELL score less than some threshold value so it will never give overoptimistic branch support anymore.

H. Maximum parsimony Boot(MPBoot)[10]

The MP boot is a method that inferred from UFboot[9] to carry out bootstrap for trees constructed using maximum parsimony instead of ML-Based trees. It employs an extra tree search algorithm and a parsimony rachet.The algorithm was compared with the TNT[11] and PAUP[12] programs and MPBoot was the fastest between them and results in better scores.

Table I - Bootstrap-based tree construction methods summary

Year	Method	Inferred from	Software	Ref.
1985	SBS	Statistical Model	-	[2]
1990	MND & RELL	Statistical model	-	[1] [3]
1996	LBP	RELL method	MOLPHY	[4]
2005	SPR	Tree pruning method	PhyML	[5]
2006	aLRT	LRT statistical Model	PhyML	[6]
2010	SH-aLRT	alRT	PhyML	[7]
2017	UFBoot	Quarter Puzzling	iqTree	[9]
2018	MPBoot	UFBoot	MPBoot	[10]

IV.CONCLUSION

There is a wide variety of techniques and algorithms that were made to solve the problem of bootstrapping in order to enhance the process of phylogenetic tree construction and remove the computation burdens of this process. But the bootstrapping is still the bottleneck of the phylogenetic tree reconstruction.

REFERENCES

[1] B.Efron, “Bootstrap Methods: Another Look at the Jackknife BT - Breakthroughs in Statistics: Methodology and Distribution,” S. Kotz and N. L. Johnson, Eds. New York, NY: Springer New York, 1992, pp. 569–593.

[2] J.Felsenstein, “Confidence limits on phylogenies: an approach using the bootstrap,” Evolution (N. Y)., vol. 39, no. 4, pp. 783–791, 1985.

[3] M.Hasegawa and H. Kishino, “Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree,” 1994.

[4] J.Adachi and M. Hasegawa, “MOLPHY Version 2.3:Programs for Molecular Phylogenetics Based on Maximum Likelihood Jun Adachi and Masami Hasegawa,” Comput. Sci. Monogr., vol. 28, pp. 1–150, 1996.

[5] W.Hordijk and O. Gascuel, “Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood,” Bioinformatics, vol. 21, no. 24, pp. 4338–4347, 2005.

[6] M.Anisimova and O. Gascuel, “Approximate Likelihood-Ratio Test for Branches : A Fast , Accurate , and Powerful Alternative,” vol. 55, no. 4, pp. 539–552, 2006.

[7] S.Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, “New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0,” Syst. Biol., vol. 59, no. 3, pp. 307–321, 2010.

[8] B.Q.Minh, M. Anh, T. Nguyen, and A. Von Haeseler, “Ultrafast Approximation for Phylogenetic Bootstrap,” no. March, 2013.

[9] D.T.Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, and S. V. Le, “UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution.,” Mol. Biol. Evol., vol. 35, no. 2, p. msx281, 2017.

[10] D.T.Hoang, L. S. Vinh, T. Flouri, A. Stamatakis, and A. Von Haeseler, “MPBoot : fast phylogenetic maximum parsimony tree inference and bootstrap approximation,” pp. 1–11, 2018.

[11] P.A.Goloboff, J. S. Farris, and K. C. Nixon, “TNT, a free program for phylogenetic analysis,” Cladistics, vol. 24, no. 5, pp. 774–786, 2008.

[12] D.L.Swofford, “PAUP*: phylogenetic analysis using parsimony (* and other methods). Sunderland, MA.” Sinauer Associates, 2002.