Phylogenetic Reconstruction with Disk-Covering and Bayesian Approaches

Yan Guo, Fei Ye and Jijun Tang

Abstract

The DCM approach is commonly used to divide the dataset into smaller subproblems, analyze each subproblem using a base method to obtain sub-trees, then recombine these subtrees to build the final phylogeny over the whole dataset. In recent years, the new and improved method MrBayes, a Bayesian Markov Chain Monte Carlo (MCMC) approach is widely used for phylogeny analysis. In this paper, a new method for large scale Bayesian phylogeny analysis is proposed. This new method (DCM3-MrBayes) is an improved version of Rec-I-DCM3 (Recursive Iterative Disk-Covering Method), which uses a divide-and-conquer approach and is designed for large dataset analysis.

To integrate MrBayes with Rec-I-DCM3, we have to deal with some unique problems and proposed several methods to tackle these problems. Our improvements include a cache system that can avoid unnecessary computations and a method to eliminate weak branches indicated by the Bayesian analysis to filter out potential bad branches. Our experiments on simulated datasets shows promising improvement over the original DCM. One of the most important advantages of using Bayesian method for phylogeny reconstruction is being able to calculate the posterior probabilities. A divide-and-conquer Bayesian method looses its ability to calculate the posterior probabilities due to the fact that each subproblem generates its own posterior probabilities, which posts some difficulties for obtaining the posterior probability for the whole problem. In order to preserve the advantage of Bayesian approach, we also introduce an algorithm that calculates the posterior probabilities of the whole phylogeny from the subproblems' posterior probabilities.

I. Introduction

Phylogenetic reconstruction is a procedure to infer the evolutionary history among organisms and is one of the most fundamental problems in biological research. To date, DNA sequence data is still the most used data type for phylogenetic reconstruction, and Maximum Parsimony (MP) and Maximum Likelihood (ML) are commonly used as the optimization criteria for reconstructing phylogenies.

In recent years, Bayesian-based inference of phylogeny found its way into the phylogeny field despite its long tenure in statistics. Even though Bayesian based methods still use the same models of evolution as many other reconstruction methods, it represents a powerful tool that can address many of the long-standing and complex questions in evolutionary biology.

One of the most difficult problems facing any phylogenetic reconstruction method is how to produce accurate phylogeny trees for large datasets (with thousands of species) within reasonable time constraints. Since the number of all possible trees grows exponentially with the number of taxa [1], enumerating all trees for several dozens taxa may take more than several centuries.

One way to remedy this problem is to use a type of divide-and-conquer approach, in which the set of taxa is decomposed into a collection of overlapping subsets, each of which optimizes some criterion designed to make reconstruction on the subset as accurate and efficient as possible.

The best divide-and-conquer approach to date is the family of Disk-Covering Methods (DCM), introduced by Warnow and her group [6], [5] and since shown in several experimental studies to produce better results on sequence-based data than any other distance- or parsimony-based method [8], [9]. Generally speaking, all DCMs proceed in four major phases: (1) decomposing the dataset into smaller and overlapping sub-problems, (2) solving the subproblems using some base methods, (3) using a supertree method to merge the results from subproblems and form a tree on the whole dataset, and (4) refining the resulted tree.

Although DCMs are designed and able to work with any base method, such as the maximum parsimony method PAUP* [16] and the distance based method Neighbor-joining [13], integrating them with Bayesian methods poses some unique problems. For example, during the recombination step, DCMs do not check if the sub-tree produced from the subproblem has weak branches, although by applying the Bayesian approach, each branch will be assigned a credibility value and branches with weak credibility values should be filtered out.

On the other hand, in a Bayesian analysis, inferences of phylogeny is based upon the posterior probabilities of phylogenetic trees [11]. Such posterior probabilities of the model can not be obtained by analysis even though it is easy to formulate, because it involves a summation over all trees,

This work was supported by NIH grant R01 GM078991.

Yan Guo and Jijun Tang are with Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA guo27,jtang@cse.sc.edu.

Fei Ye is with Cancer Biostatistics Center/Division of Biostatistics Department of Biostatistics, Vanderbilt University, 571 Preston Research Building, Nashville TN 37232, USA Fei.Ye@vanderbilt.edu.

The first two authors should be regarded as joint first authors.

[1] Given N taxa, the number of all possible trees is $(2N-5)!! = (2N-5) \times (2N-7) \times \cdots \times 3$.

and for each tree, integration over all possible combinations of branch length and substitution model parameter values. Current DCM methods do not provide a way to obtain the posterior probability of the whole phylogeny based on the probabilities of all sub-trees.

Since Bayesian method can produce accurate phylogeny trees in a fast fashion [11], using a Bayesian approach as base method for the Disk-Covering methods is an ideal approach for building a more robust statistical model.

In this paper, we present our new improvements on Rec-I-DCM3 (the latest DCM method) to address the above problems. The paper is structured as following. First we give necessary background on the Bayesian approach and the Disk-Covering methods. We then introduce the various techniques we implemented in the new method (called DCM3-MrBayes). Our experimental results on simulated datasets show significant improvements on topology correctness, and the results are shown in section IV.

II. Background

A. Bayesian Approach for Phylogenetic Reconstruction

Huelsenbeck and Ronquist introduced MrBayes, a program using Bayesian approach to estimate phylogenies from sequence data [3], [11]. The current version of MrBayes is v3.1, which is a completely rewritten and restructured version. The hallmark of the new MrBayes is a powerful framework for phylogenetic inference under mixed models accommodating data heterogeneity. This framework will help the user to specify mixed models and exploit the computational efficiency of Bayesian MCMC analysis in dealing with composite data sets.

Unlike MP and ML based methods, using Bayesian approach of phylogeny reconstruction combines the prior probability of a phylogeny with the tree likelihood to produce a posterior probability distribution on trees [11]. In ML or MP methods, topologies and branch lengths are not treated as parameters but random variables. In Bayesian analysis, the best estimate of the phylogeny can be obtained by selecting the tree with the highest posterior probability, in a way the posterior probability of a tree can be interpreted as the probability of that tree being the true tree. Usually all trees are considered a priori equally probable.

Since the posterior probabilities of the model can not be obtained by analysis, they have to be approximated by the Markov Chain Monte Carlo (MCMC) approach [1]. The principle of the MCMC approach is to build a succession of states, and once convergence is reached, the consecutive states are assumed to be drawn from the target probability distribution. The objective of MCMC when associated with Bayesian methods is to compute the global optimum of some posterior probability. Markov chains are used to explore the posterior probability surface by integrating over the space of model parameters. Usually the trees are sampled at a fixed frequency and through those samples, the posterior probability is approximated.

In general for phylogeny, the MCMC algorithm involves two steps: first, a new tree is proposed by stochastically perturbing the current tree. Second, this tree is either accepted or rejected with an acceptance probability. Upon acceptance, the new tree is subjected to further perturbation [2].

The acceptance probability is defined as the minimum of one or the likelihood ratio times the prior ratio times the proposal ratio, where the likelihood ratio is the ratio of the likelihoods of the new state to the old state, the prior ratio is the ratio of the prior probability of new state to the old state, and the proposal ratio is the ratio of the probability of proposing the old state to the probability of proposing the new state [3].

Even though many of the analysis of difficult model are made possible by MCMC algorithm, it is not a silver bullet, as Markov Chains can fail to converge to the stationary distribution for various reasons.

A Bayesian approach on phylogeny can be generalized in the following formula:

$$Tree|Data] = \frac{P[Data|Tree] \times P[Tree]}{P[Data]}$$

the new MrBayes applies the general Bayesian approach using the following rule:

$$f(\tau,\upsilon,\theta|X) = \frac{f(\tau,\upsilon,\theta)f(X_{|\tau|},\upsilon,\theta)}{f(X)}$$

where X is the data matrix, τ is the topology of the tree, υ is a vector of edge lengths on the tree and q is a vector of substitution model parameters. $f(\tau,\upsilon,\theta)$ is the prior, which specifies the prior probability of different parameter values. $f(X_{|\tau|},\upsilon,\theta)$, is the likelihood function, which describes the probability of the data under different parameter values. $f(X)$ is the total probability of the data summed and integrated over the parameter space. $f(\tau,\upsilon,\theta|X)$ is the posterior distribution. MrBayes uses a Metropolis-Hasting Sampler to update single parameter or blocks of related parameters in each step.

In general, Bayesian based algorithms avoid the standard approach of specifying only one hypothesis as the null hypothesis then asking if the data are strong enough to reject it. Since the output of a Bayesian analysis is the posterior probability of any solution, standard probability rules can still be used as measurement to select the most reasonable and strong hypothesis. For example, if one hypothesis is consistent with k different trees from the tree space and the alternative is consistent with all other tree topologies, the probability that the first hypothesis is correct is simply the sum of the posterior probabilities of the k trees. The Bayesian approach is intuitive and is particularly useful when the number of alternative hypotheses is huge.

B. Disk-Covering Methods

To date, there are three major DCMs: DCM1 [4], DCM2 [5] and Rec-I-DCM3 [12]. All DCMs proceed in the four divide-and-conquer phases described above, but variants of DCMs come from different decomposition methods: both DCM1 and DCM2 operate solely from the pairwise distance matrix of the taxa, whereas DCM3 uses a dynamically updated guide tree (in practice, the current estimate of the phylogeny) to direct the decomposition. The last three phases are identical for all DCMs.

As the latest variant of DCM approaches, Rec-I-DCM3 uses iterations to escape local optima, a divide-and conquer approach to reduce problem size, and recursions to enable further localization and reduction in problem size. One of the improvements is that it tries to avoid dividing the data into very large subset by applying DCM3 recursively on subsets that are too large. Another improvement is to use a dynamically updated guide tree to direct the decomposition, so that it will produce different decompositions for different guide trees. In other words, Rec-I-DCM3 iteratively refines the guide tree and produces better decomposition as the iteration proceeds. Experiments showed that Rec-I-DCM3 not only reduces the size of the explored tree space, but also finds a larger fraction of MP trees with better scores than other methods, and provides more accurate phylogeny trees [12].

So far, Rec-I-DCM3 has not been offered to work with any Bayesian method. Although the CIPRES portal[2] has a plan to add such capability, the current release still only provide Neighbor-joining, RAxML [14] (for likelihood) and PAUP [16] (for parsimony and likelihood) as base methods.

Our integration of Rec-I-DCM3 and MrBayes (DCM3-MrBayes) provides an alternative to the traditional ML and MP approaches. It should be a very powerful tool for inferring phylogeny, evaluating clade probabilities, detecting selection, detecting sample substitutions, and counting number of synonymous and non-synonymous changes.

III. DCM3-MrBayes

In this section, we introduce the various improvements in our new program DCM3-MrBayes, including a procedure to eliminate weak branches, a method to compute the posterior probability of the whole phylogeny based on information from sub-trees, and a cache system to avoid unnecessary computations.

A. Eliminating Weak Branches

Our first improvement is to eliminate the weak branches from the sub-trees MrBayes returns. MrBayes reports two trees in newick format, one contains the topology, branch length and probability of the partition indicated by the branch, and the other one contains information only on the topology and branch length. Fig. 1 shows the output tree from MrBayes which displays the topology, branch length and probability.

The probability, also known as the clade credibility value is in fact the posterior probability of how likely the given branch is a true branch. We assume a branch is weak indicates that it is unlikely to be part of the true tree. Since DCMs use a majority consensus rule when they integrate sub-trees to form the final phylogeny, errors introduced in sub-trees will make the consensus hard to achieve and sometime problematic, thus these weak branches should be eliminated from sub-trees.

Our elimination procedure works as following. Based on the branch percentage a cut off point is set, any branch that has less than the cut off percentage will be deleted, its child node will be connect to the nearest branch as a star. In other words, this branch is treated as unresolved in this sub-tree. Fig. 2 shows an example. Since a weak branch indicates that such branch may be wrong, removing it from the sub-tree prevents potential conflicts in the recombination stage of Rec-I-DCM3 and allows stronger branches to dominate the recombination.

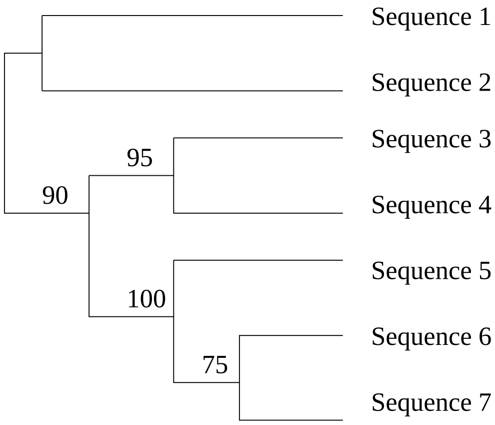

Fig. 1. MrBayes output tree which displays clade credibility values.

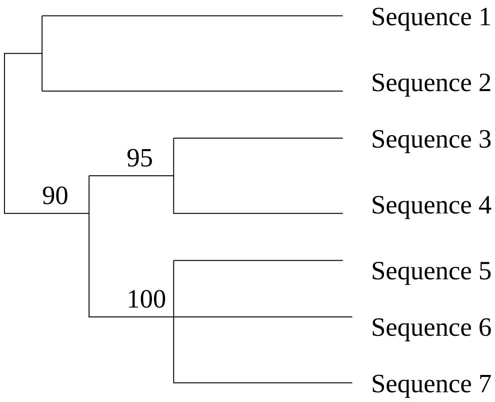

Fig. 2. Eliminating the weak branch (the one with 75% credibility value) from the tree shown in Fig. 1.

B. Obtaining Posterior Probabilities

Providing posterior probabilities is one of the advantages of Bayesian method since posterior probabilities can be used as easily interpretable alternatives to *p* values. The original Rec-I-DCM3 might work reasonably well with a Bayesian method as base method even without eliminating weak branches, however, there are no previously known algorithm to generate posterior probabilities for the final phylogeny of the whole dataset.

[2] under development with funding from the National Science Foundation to support large-scale sequence-based phylogenetic analysis

In this study, the posterior probabilities are equal to the clade probabilities. A clade is defined in Cladistics, which is the hierarchical classification of species based on their evolutionary ancestry, and the diagrams generated by cladistics are called cladograms. Fig. 3 shows the clades in the tree shown in Fig. 1. In a cladogram, a clade is defined as a taxonomic group comprising of a single common ancestor and all the descendants of that ancestor. In cladistics, a clade that is located within another more inclusive clade is said to be *nested* within that clade.

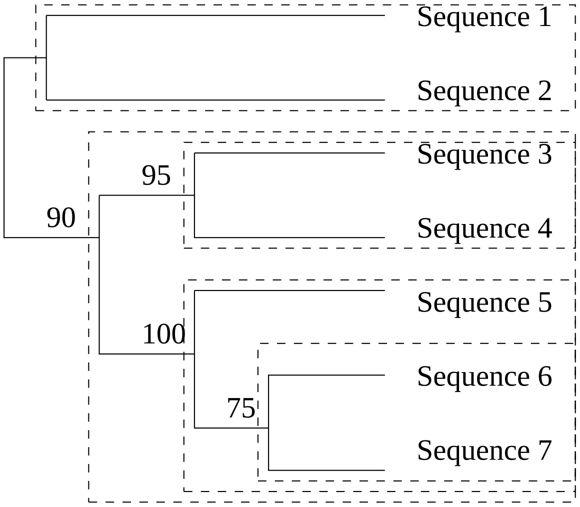

Fig. 3. Example of clades, which are bounded by dashed boxes.

Consider an unrooted phylogeny tree which contains many clades. A clade can be formed in several ways. The simplest clade is formed directly by several (minimum two) leaves (species). A more complicated clade can be formed by one or more smaller clades with one or more single leaves. For example, in Fig. 3, the clade containing sequences $5 \sim 7$ is an example of a complicated clade, which is composed of two smaller clades.

The difficulty associated with calculating posterior clade probabilities for the combined final tree lies in the area of how to find an algorithm that is statistically meaningful to combine the posterior clade probabilities of sub-trees.

We made several assumptions here:

- For all of the sub-trees generated by DCM3 during the decomposition step, they are independent of each other, even though they must overlap slightly.
- For all clades in one tree, some may contain others, they are independent to each other, which means, if a bigger clade has clade probability of x, and it contains a smaller clade whose clade probability is y, x and y are independent of each other. In other words, the probability of the contained clade does not depend on its parent clade's probability, or vice versa.

With this knowledge in hand, an algorithm can be constructed by the following steps. Since a clade may appear in several sub-trees, it is assigned multiple posterior probabilities, one from each sub-tree. We can also assign a weight to each clade, which determines its importance. If a clade has more weight, its probability is more important than its copies that appear in other sub-trees. In this paper, the weight is set to be 1, so each clade has equal weight in each sub-tree.

Let L be a list of clades, L_i denote the ith clade in the list; let LW be an array of weight, and LW_{im} denotes the mth weight of the clade L_i; let LP be an array of probabilities, and LP_{im} denotes the mth probability of the clade L_i. Also, for each clade L_i, define n_i, the number of sub-trees containing clade L_i.

All the above lists (L, LW and LP) can be easily obtained by traversing all sub-trees before the recombination stage of DCM3. If the ith clade L_i appears in the final whole phylogeny, its posterior probability can be calculated as

$$P_i = \frac{\sum_{m=0}^{n_i}(LP_{im} \times LW_{im})}{n_i}.$$

By checking all clades appear in the final tree, we can easily get the posterior probability of the whole phylogeny using the above equation.

C. Avoiding Unnecessary Computation

During some preliminary test runs on DCM3-PAUP, one interesting problem surfaced is that depending on the size of the datasets (number of sequences), the subproblems obtained during the iterative decomposing step can not only overlap each other, but also sometimes identical to each other. That's to say, since DCM3 proceeds in an iterative way, one iteration may contain subproblems that have appeared in a previous iteration. Rec-I-DCM3 will recognize the identical subproblems as different ones, thus applies the base method on the same subproblems every single time it appears. This unnecessary and undesirable step will hinder the run speed by doing avoidable computation.

As an effort to trim the potential running time, a computation saving technique for Rec-I-DCM3 is developed and applied. The saving technique saves the resulted sub-tree for each subproblem by creating a cache file and a cache folder, as detailed in the following steps:

- Each dataset has its own cache directory, which name is based on the input data file name.
- Each subset will have a unique cache file name based on the genomes in the subset. To save the checking time, the genomes are sorted by name when it is passed to MrBayes or the recursive procedure of DCM3.
- When MrBayes and the recursive DCM3 get the subset, it will check whether or not the cache file for this subset exists in the cache directory. If such a file exists, it copies the file back as the result for DCM; otherwise, it moves on to compute and cache the result.

With this improvement, the number of subproblem is reduced by up to 60 percent. Because Bayesian analysis is very expensive, the computation of subproblems dominates the whole analysis. As a result, the cache system can provide considerable speedup, and the time spent on writing and reading files is negligible. The cache system will be more effective when the maximum size of subproblems is set to

TABLE I

The Robinson-Foulds error rates from the true trees to the inferred trees. An error rate of less than 5% is considered very accurate for a method.

number of taxa	100	200	400	800	1000
DCM3-Paup error	0	0.9%	1.75%	2.75%	15.3%
DCM3-MrBayes error	0	0.9%	1.5%	2.75%	8.1%

TABLE II

The Branch Score Distances of DCM3-Paup and DCM3-MrBayes.

number of taxa	100	200	400	800	1000
DCM3-Paup Distance	7	97	289	578	654
DCM3-MrBayes Distance	6	90	200	489	611

TABLE III

Parsimony scores of the trees returned by DCM3-MrBayes and DCM3-PAUP

number of taxa	100	200	400	800	1000
DCM3-Paup score	740	1185	11000	12343	16835
DCM3-MrBayes score	731	1185	10930	12300	16432

be small (for example, fewer than 20 taxa per subproblem). Since the duplicated subproblem is created during the decomposition steps, this technique can be used as a general fix on DCMs with any base method except fast distance-based methods.

IV. Experimental Results

We set out to examine the accuracy of our new DCM3-MrBayes method. We concentrated our experiments on simulated datasets because topological accuracy can be easily assessed when the true trees are known.

We used two criterion to measure the topological accuracy of a method: the Robinson-Foulds error rate and the Branch Score Distance.

If the true tree has an edge defining a bipartition with no equivalent in the reconstructed tree, that edge is a *false negative (FN)*; conversely, if the reconstructed tree has an edge defining a bipartition with no equivalent in the true tree, that edge is a *false positive (FP)*. The goal of all phylogeny methods is to obtain both lower false negative and false positive. The Robinson-Foulds error rate [10] is defined as the number of false (FP and FN) edges divided by the number of internal edges of the true tree ($N-2$ for N taxa).

Another popular measurement for the accuracy of a phylogeny program is the Branch Score Distance [7] which uses branch lengths and can be calculated when the trees have lengths on all branches. A method returns lower Robinson-Foulds error rate and smaller Branch Score Distance is generally considered more accurate.

The ROSE (Random Model of Sequence Evolution) [15] software package is a widely used simulator for sequence evolution, which implements the HKY85 model of DNA sequence evolution and allows for insertions and deletions. In this experiment, we first create random trees and use ROSE to generate sequences on all internal and leaf nodes. All datasets are tested using both DCM3-MrBayes and DCM3-PAUP to compare the results.

Tables I and II show the Robinson-Foulds error rates and the Branch Score Distances of these two methods. These tables clearly show that for larger dataset size, DCM3-MrBayes infers more accurate phylogenies than the commonly used DCM3-PAUP method.

The quality of the inferred trees is measured by computing the maximum parsimony scores of these trees. Table III shows these parsimony scores obtained by these two methods. Although DCM3-MrBayes does not explicitly seek the most parsimony trees, this table suggested that the resulted trees do require fewer number of events than DCM3-PAUP, which uses the criterion of maximum parsimony to select the phylogenies.

We also test Paup with its built-in Maximum Likelihood method. Even DCM3 tries to decompose very small subproblems, the ML method of Paup is just too slow and we many of the above datasets cannot be finished after several days of computation. The newly improved RAxML may be fast enough to work with DCMs, and we will test its performance against our DCM3-MrBayes in the future.

V. Summary and Conclusions

In this paper, we present our new method to handle phylogeny reconstruction on large dataset and report experimental results on simulated datasets. Our testing confirms that the new DCM3-MrBayes is more accurate than the current DCM3 methods. This method also enable us to reconstruct the posterior probability for divide-and-conquer based Bayesian approach, which will make the Bayesian approach more useful in large scale phylogeny analysis.

VI. Acknowledgments

The authors were supported by US National Institutes of Health (grant number R01 GM078991).

References

[1] Gilks, W.R., S. Richardson and D. Spiegelhalter (Eds) (1996). Markov Chain Monte Carlo in Practice. Chapman and Hall, London.

[2] Huelsenbeck, J.P., F. Ronquist, R. Nielsen and J.P. Bollback (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294: 2310-2314.

[3] Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES. Bayesian inference of phylogeny. *Bioinformatics* 17: 754-755.

[4] Huson, D., S. Nettles, L. Parida, T. Warnow, and S. Yooseph (1998). The Disk-Covering Method for Tree Reconstruction. *Proc. Algorithms and Experiments (ALEX'98)*, 62-75.

[5] Huson, D., L. Vawter, and T. Warnow (1999a). Solving large scale phylogenetic problems using DCM-2. *Proc. 7th Conf. on Intelligent Systems for Mol. Biol. (ISMB'99)*, 118–129.

[6] Huson, D., S. Nettles, and T. Warnow (1999b). Disk-covering, a fast-converging method for phylogenetic tree construction. *Journal of Computational Biology*, 6: 369-386.

[7] Kuhner, M.K. and J. Felsenstein (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11(3): 459-68.

[8] Moret, B.M., U. Roshan, and T. Warnow (2002). Sequence length requirements for phylogenetic methods. *Proc. 2nd Workshop on Algorithms in Bioinformatics (WABI'02)*, Volume 2452 of *Lecture Notes in Computer Science*, 343-356.

[9] Nakhleh, L., B.M. Moret, U. Roshan, K. St John, and T. Warnow (2002). The accuracy of fast phylogenetic methods for large datasets. *Proc. 7th Pacific Symp. on Biocomputing (PSB'02)*, 211–222.

[10] Robinson, D.F. and L.R. Foulds (1981) Comparison of phylogenetic trees. *Mathematical Biosciences* 53, 131–147.

[11] Ronquist, F. and J.P. Huelsenbeck. (2003). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572-1574.

[12] Roshan, U., B.M. Moret, T. Warnow and T.L. Williams (2004). Rec-I-DCM3: a fast algorithmic technique for reconstructing large phylogenetic trees *Proc. IEEE Comput. Syst. Bioinform. Conf. (CSB'04)* 98-109.

[13] Saitou, N. and M. Nei (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Bio. and Evo.* 4: 406-425.

[14] Stamatakis, A., T. Ludwig and H. Meier (2005). RAxML-III: A Fast Program for Maximum Likelihood-based Inference of Large Phylogenetic Trees. *Bioinformatics* 21: 456-463.

[15] Stoye, J., D. Evers and F. Meyer (1998). Rose: generating sequence families. *Bioinformatics* 14: 157-163.

[16] Swofford, D.L. (1996). PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods). Sinauer Associates, Sunderland, Massachusetts, Ver 4.0.