Pak. J. Statist.
2019 Vol. 35(3), 217-227

MAXIMUM LIKELIHOOD INFERENCE OF ROOTED SPECIES TREES FROM ROOTED VS UNROOTED GENE TREES

Ayed R.A. Alanzi
Department of Mathematics, College of Science and Human Studies
Hotat Sudair Majmaah University, Majmaah 11952, KSA
Email: a.alanzi@mu.edu.sa

ABSTRACT

This work investigates that how the maximum likelihood estimate (MLE) of the phylogenetic tree works. MLE is one of the most standard method of estimation among all the methods used in estimation. It is obligatory to find ML before finding the core value of the parameters which proceed to figure out the likelihood function. Scholars used the MLE frequently to prove the notion of species trees from gene trees inspired by the incomplete lineage sorting. They have traditionally used rooted genes tree to understand rooted species tree or unrooted genes tree to realize unrooted species tree. Depending upon the knowledge of the topology of unrooted gene tree, Allman *et al.* (2011) gave a notion to relate the rooted species tree from unrooted genes trees. This theory was helpful to acknowledge the features of the rooted specie tree. Allman *et al.* (2011) discovered for the first time that the roots of species tree are using unrooted specie tree as the input data without assuming an outgroup. When an appropriate outgroup is challenging and gene trees do not go with a molecular clock, this approach will be valuable. This study follows the MLE method to compute the maximum value of the correct tree. In addition to that, this study will compare between using rooted gene trees and unrooted gene trees, both with and without DNA sequences, for different number taxa.

KEYWORDS

Multispecies coalescent, outgroup, unrooted gene trees, MLE, DNA sequences.

1. INTRODUCTION

Edwards and Cavalli-Sforza (1964) started the likelihood method studying phylogenies to understand the data of gene frequency. After that, Felsenstein (2004) revised the history of likelihood methods. Felsenstein also elaborate the approach of Neyman (1971), who was basically serious in the use of likelihood methods, was the most appropriate to apply them to molecular sequences. Felsensten including Kashyap and Subas (1974) as well as his own work, concludes by mentioning those who built on this that applies to nucleotide orders.

Knowles and Kubatko (2011) proved the facts that the estimations for ML species trees can be gathered in a couple of successive stages. Earlier in these stages, the calculation of the likelihood function must be applicable when working with a multi-

locus data set for any available species tree. Second stage depends upon the first stages so it is quite essential to develop a process for locating a likelihood maximizing tree by exploring through the range of possible species trees in the existed space. They enhanced their knowledge in the application of a likelihood function to both species trees and gene trees. The likelihood of species trees is computed in a variety of ways depending on whether the sample merely involves gene tree topologies, a gene tree topology with branch lengths, or multi-locus sequence data. On the other hand, the likelihood of gene-tree data includes gene-specific DNA series data.

Maximum parsimony (MP) technique (Eck and Dayhoff, 1966) and the maximum likelihood (ML) technique (Felsenstein, 1981) are the most popular and result oriented methods. According to the parsimony method (Eck and Dayhoff, 1966), trees resulting from fewer variations are more probable. This method can be utilized to account for radical changes that prefers to hypothesize the lowest number of variations. (Felsenstein, 1981) in the maximum likelihood method highest probability mutation got more preference as compare to smallest number of mutations. According to stochastic models of nucleotide sequences. The best maximum likelihood estimate (MLE) for a tree can be resolute by taking the highest value of the MLE. It commonly happens by estimating the ML of the branch lengths for specific tree topology and DNA substitution model. This process is repeated many times with other topologies (Felsenstein, 1981). The MLE method has an established statistical foundation (Felsenstein, 1981; Goldman, 1990) and is widely accepted at reestablishing the true topology of trees by using a computer simulation study (Fukami-Kobayashi and Tateno, 1991; Hasegawa *et al.*, 1991). Bouckaert *et al.* (2014); Ronquist *et al.* (2012), and Alfaro *et al.* (2003) use the frequentist parametric approach to estimate tree phylogeny. They try to estimate trees and the substitutes of the Bayesian and nonparametric approaches. Their argument about its value is part of a long tradition in the biology history. The best features of the MLE of asymptotic properties are developed by Rogers (1997) and Yang (1994). Consistency was the issue that they encountered during their study. To assess the consistency of ML trees, Yang (1994) has developed a method that reflect the complication of the difficulties.

One more central feature of the ML method is that it can compute the evolutionary trees variable models in a statistical framework. To infer species trees many remarkable plans have been developed in recent times. In accordance to maximum likelihood, Wu (2012) brought up a new algorithm for ST inference. The likelihood method is based on probabilities of rooted gene tree topologies. He compares his algorithm to the algorithm of Degnan and Salter (2005), and proves his algorithm faster. He named his algorithm STELLS (stands for Species Tree Inference with Likelihood for Lineage Sorting). The approaches used by Yu *et al.* (2011) and Yu *et al.* (2013) depend on hybridization as well as incomplete lineage sorting (ILS). Yu *et al.* (2013) tries to resolve the inference issues by discovering the space of phylogenetic networks that they undertake by using search heuristics in the software PhyloNet (Than *et al.*, 2008). PhyloNet can infer species trees and networks using probabilities of rooted gene tree topologies.

Their opportunity was prolonged in Yu *et al.* (2014) to inspect the best phylogenetic network based on multi-locus data (see also Nguyen and Roos (2015) for a various likelihood structure, where sites instead of genes are treated as independent and ILS is overlooked). The likelihood-based technique in Yu *et al.* (2014), used in PhyloNet, gives

a concrete hypothetical context to calculate the maximum likelihood phylogenetic network from a set of gene tree (GT). It has several profits: it integrates uncertainty on the GT assessed from sequence data, estimates for a contextual level of GT discordance owing to ILS, and regulates the trouble of network with a cross authentication step. However, its likelihood calculation is significant and becomes insistent when increasing sum of taxa or the number of hybridizations, making this technique realistic for small scenarios up to 10 species and 4 hybridizations in the network.

To estimate phylogenetic networks from multi-locus data, Solís-Lemus and Ané (2016) specify a persistent arithmetical method. pseudo-likelihood of a network model was presented by Solís-Lemus and Ané (2016). They did it by emerging the amount of the genome with 4-taxon tree each (quartet concordance factors) as projected in the coalescent model prolonged by hybridization measures, and they confirm generic identifiability of the model. The experimental quartet concordance features as collected from the multi-locus data was used by Solís-Lemus and Ané (2016) to estimate the species network.

Another method for inferring the root is recommended by Allman *et al.* (2011) to use maximum likelihood of the unrooted gene trees. This technique has not been applied yet; however, likelihood estimation for rooted gene tree topologies scale exponentially in the number of taxa (Degnan and Salter (2005); Wu (2012); Rosenberg (2007); Disanto and Rosenberg (2016); Disanto and Rosenberg (2017)), recommend that this technique will not measure well in the number of taxa. Furthermore, the only existing method to estimate the unrooted gene tree probabilities is to sum gene tree possibilities over all given root locations. To calculate the MLE, this research, from the beginning, needs to calculate the likelihood function. This method follows for unrooted trees and this is the formula that the PhyloNet program (Than *et al.*, 2008) practices to estimate the MLE for unrooted trees:

$$\prod_{i=1}^{(2m-5)!!} P_i^{n_i} = \prod_{i=1}^{(2m-5)!!} \left(\sum_{j=1}^{2m-3} P_{ij} \right)^{n_i}$$

where i is the index to the topology, P_i = Probability of the i^{th} unrooted topology, n_i is the number of trees observed with topology i, j indexes the root location within the i^{th} unrooted topology, and m is the number of taxa. From the left hand side of the equation, the likelihood is multinomial, where the number of categories are the number of unrooted tree topologies. However, I also use the PhyloNet program to compute the likelihood function for rooted trees, and also the PhyloNet program needs to use this formula to compute the MLE:

$$\prod_{ij} P_{ij}^{n_{ij}} = \prod_{k=1}^{(2m-3)!!} P_k^{n_k}$$

This research for the first time compute the Maximum likelihood by using unrooted gene trees, which is implemented by the PhyloNet program (Than *et al.*, 2008).

2. SIMULATION METHODS

The species trees are inferred under four conditions. First, it consists of rooted, known gene trees. Secondly, it consists of rooted gene trees estimated from DNA sequences. The third consists of unrooted, known gene trees. Finally, it consists of unrooted gene trees estimated from DNA sequences.

This study requires seven stages to proceed. The first step uses the library from the R-package, that is TreeSim (Stadler, 2014). This library simulates the species tree. Now it is very important to have the following information to simulate the species tree by using the library TreeSim. First thing is to know the number of taxa which has two dissimilar varieties in this case as 8 taxa and 12 taxa. Then, it is essential to determine how many trees to simulate. It is also required to specify the value of λ (birth rate). In this research, λ has five difference values, which are 0.1, 0.25, 0.5, 0.75, and 1.0. The value of μ/λ (turnover) is the final information required to simulate the species tree. There are four values of μ/λ, which are 0.0, 0.25, 0.5, and 0.75.

The second stage is to enlarge the outgroup to the species tree. This study utilizes the R code to do it. Dual versions of the species tree are kept, which makes one with the outgroup to compare with the result from the PhyloNet program (Than *et al.*, 2008) for rooted gene trees also the other form without the outgroup to compare with the results from the PhyloNet program for unrooted gene trees. The third stage is to run the ultrametric program, that makes the simulated tree from the primary step into a molecular clock tree with the outgroup. The fourth step is the usage of the molecular clock species tree as an input to the Hybrid-Lambda program (Zhu *et al.*, 2015) to simulate the gene trees. The fifth stage is the lengthiest step since this study required to run the PhyloNet program.

The sixth step is to simulate DNA sequences by using the Seq-Gen program (Rambaut and Grass, 1997) and the PhyML program (Guindon and Gascuel, 2003). The Seq-Gen program practices the gene tree that is adapted from Hybrid-Lambda as input to the program. The PhyML program uses the output data from Seq-Gen as input. The research work runs an R code to mark the output data from PhyML as a rooted tree and keep it as data simulated from DNA sequences. The final step, which is the seventh step, is all alike to the fourth step in this study.

The outgroup is detached when the gene trees are obtained from the hybrid-lambda program to calculate the MLE from the unrooted gene trees without DNA by using the PhyloNet program. Furthermore, the outgroup required to delete from the DNA sequences of the PhyML program to compute the MLE of the unrooted gene trees with DNA by using the PhyloNet program. Figure 1 shows the steps for building the simulation code and how it is compared with the results of this work.

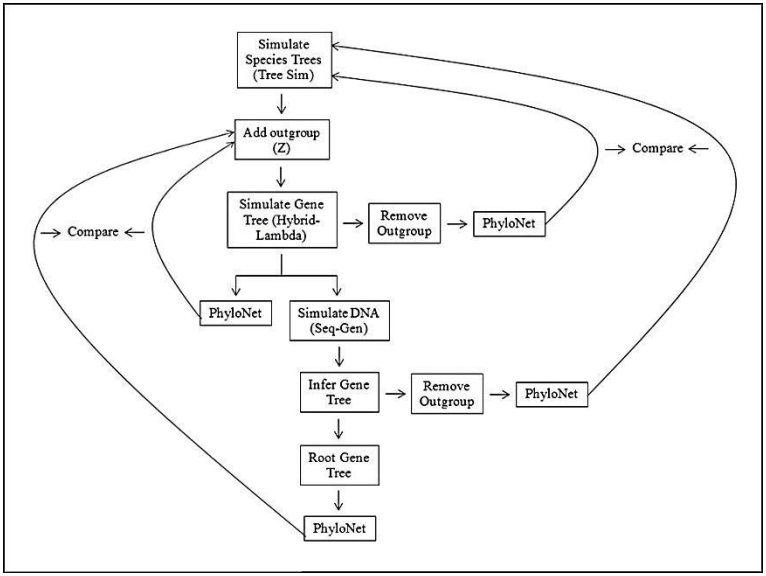

Figure 1: Diagram of the Simulation Method

3. SIMULATION RESULTS

Figure 2 and Figure 3 demonstrates the comparison between rooted gene trees and unrooted gene trees for cases that have the gene trees with and without DNA for both 8 taxa and 12 taxa with 100 genes. On the horizontal axis depict the values of λ, and the vertical axis show the average of the tree distance to the species tree from both rooted gene trees and unrooted gene trees with and without DNA, where the μ is constant in each subfigure. Those figures describe how the λ changes when μ is risen. Figure 2 and Figure 3 present that there are not many variations between the rooted gene trees and unrooted gene trees in both cases with and without DNA. In contrast, Figure 2 and Figure 3 display that when the λ grows up the variation of differences rise with all values of μ. Figure 2 demonstrate that unrooted gene trees become more accurate in all values of μ when λ is big. For 8-taxa, the rooted gene tree lacking DNA makes a better estimation with small values of λ, but when λ becomes greater, the unrooted gene tree without DNA does better, as seen in Figure 2. Moreover, when I rise the number of taxa from 8 taxa to 12 taxa, the rooted gene tree without DNA makes the estimates well for 12 taxa in all values of λ and μ, as seen in Figure 3. Figure 3 expresses the worst case for both rooted genes trees and unrooted gene trees with and without DNA when both λ and μ are large.

Figure 4 and Figure 5 illustrates the association between rooted gene trees and unrooted gene trees for cases which have the gene trees with and without DNA for both 8 taxa and 12 taxa with 100 genes versus 500 genes, where the μ is equal to zero. Figure 4 indicates that 500 loci do slightly better than the 100 loci. The unrooted gene trees without DNA for 500 loci are more accurate than others as shown in Figure 4. However, Figure 5 shows the comparison between 100 loci and 500 loci, which doesn't depict much variation between them. According to this result I can finalize that no big difference occurs with increasing the sample size of the genes.

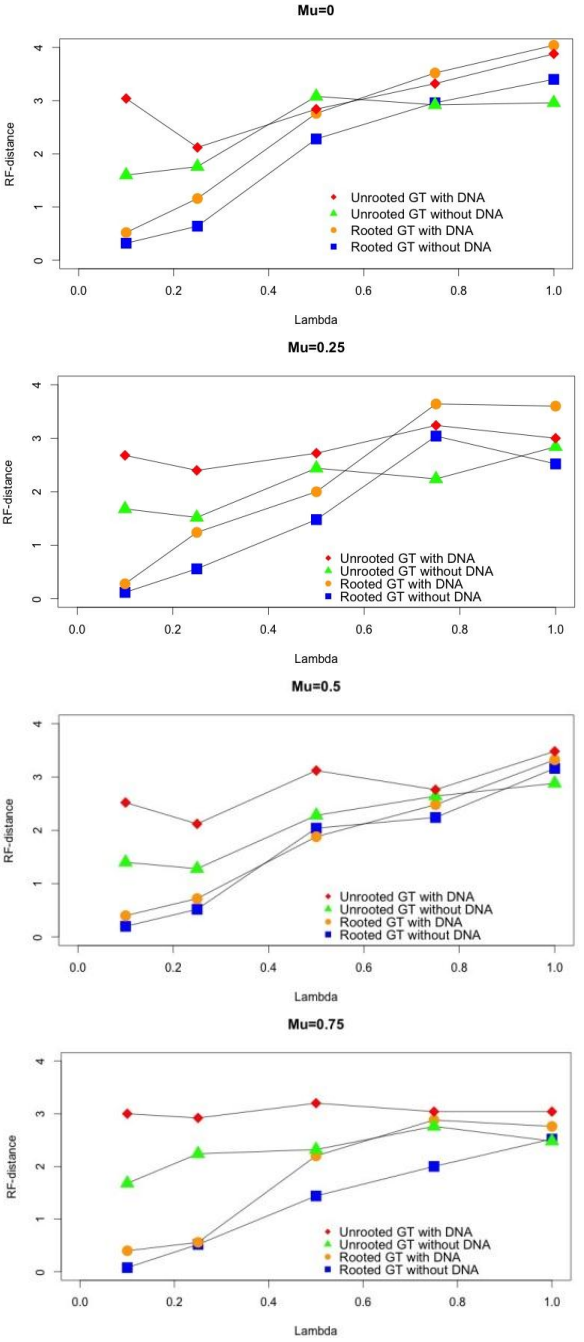

Figure 2: Rooted GT vs Unrooted GT for 8 Taxa

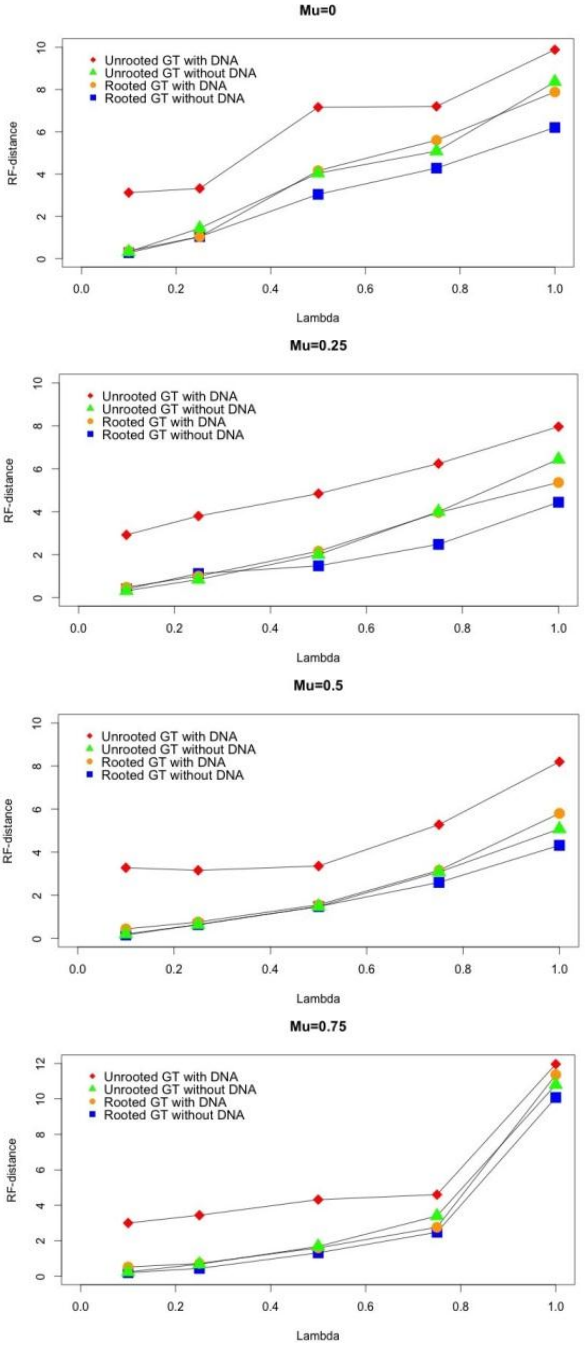

Figure 3: Rooted GT vs Unrooted GT for 12 Taxa

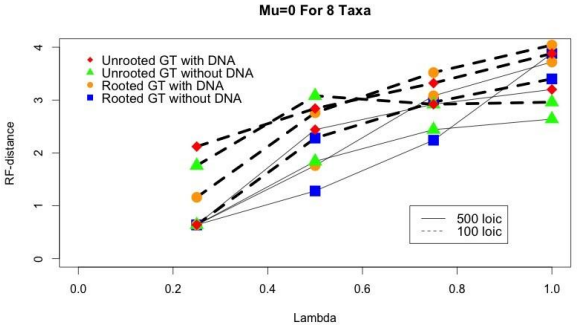

Figure 4: Rooted GT vs Unrooted GT for 8 Taxa For 100 loci vs 500 loci

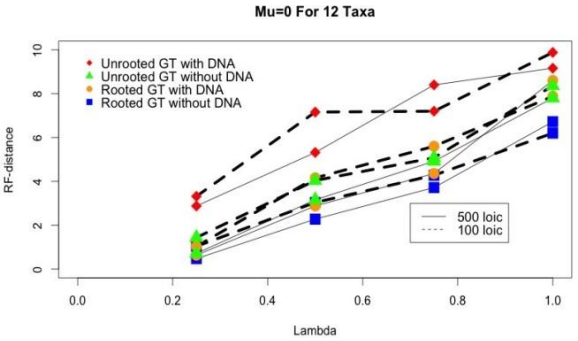

Figure 5: Rooted GT vs Unrooted GT for 12 Taxa for 100 loci vs 500 loci

4. DISCUSSION

Recent available technique for computing unrooted gene tree probabilities is to sum gene tree probabilities over all possible root locations. ML species trees can be attained in two successive steps. In the first step, when working with a multi-locus data set for any available species tree the evaluation of the likelihood estimation is necessarily to be applicable. The second stage depends upon the accomplishment of the first, it is compulsory to develop a technique for locating a likelihood maximizing tree by combing through the selection of possible species trees within a given space.

The highest maximum likelihood estimate (MLE) for a tree can be determined by selecting the highest value of the MLE, that mostly happens by estimating the ML of the branch lengths for specific tree topology and DNA substitution model.

This research first time uses the phyloNet program to compute the MLE for rooted species tree by using the unrooted genet tree. After that, the researchers modified the PhyloNet program to be able to use the unrooted gene trees as input data. PhyloyNet was calculating the MLE for rooted species tree by using rooted gene trees only before this, which is discus in Yu and Nakhleh (2015). In this study PhyloNet is available to compute MLE for rooted species tree by using unrooted gene tree without assuming the outgroup by Allman *et al.* (2011). When they discuss that can estimate the rooted species tree

without assuming the outgroup by using the unrooted gene tree. Alanzi and Degnan (2017) discussed the estimation rooted species tree by using the unrooted gene tree by the Approximate Bayesian computation (ABC), which was the first work to estimate rooted species tree by this way of computation. In addition to this the MLE is also the first work to compute the rooted species tree by this way.

This approach will be useful in the cases where an appropriate outgroup is problematic and gene trees do not follow a molecular clock. The researcher also uses the MLE method to compute the maximum value of the correct tree.

REFERENCES

1. Alanzi, A.R. and Degnan, J.H. (2017). Inferring rooted species trees from unrooted gene trees using approximate bayesian computation. *Molecular phylogenetics and evolution*, 116, 13-24.
2. Alfaro, M. E., Zoller, S. and Lutzoni, F. (2003). Bayes or bootstrap? A simulation study comparing the performance of bayesian markov chain monte carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution*, 20(2), 255-266.
3. Allman, E.S., Degnan, J.H. and Rhodes, J.A. (2011). Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology*, 62(6), 833-862.
4. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A. and Drummond, A.J. (2014). Beast 2: A software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.*, 10(4), e1003537.
5. Degnan, J.H. and Salter, L.A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59(1), 24-37.
6. Disanto, F. and Rosenberg, N.A. (2016). Asymptotic properties of the number of matching coalescent histories for caterpillar-like families of species trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(5), 913-925.
7. Disanto, F. and Rosenberg, N.A. (2017). Enumeration of ancestral configurations for matching gene trees and species trees. *Journal of Computational Biology*, 24(9), 831-850.
8. Eck, R.V. and Dayhoff, M.O. (1966). *Atlas of protein sequence and structure*, National Biomedical Research Foundation. Maryland: Silver Springs.
9. Edwards, A.F. and Cavalli-Sforza, L.L. (1964). Reconstruction of Evolutionary Trees. In *Phenetic and Phylogenetic Classification* (Heywood, W.H. and McNeill, J. Eds.) *Syst. Assoc. Pub.* No. 6, London, 67-76.
10. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6), 368-376.
11. Felsenstein, J. (2004). *Inferring phylogenies*, Volume 2. Sinauer Associates Sunderland.
12. Fukami-Kobayashi, K. and Tateno, Y. (1991). Robustness of maximum likelihood tree estimation against different patterns of base substitutions. *Journal of Molecular Evolution*, 32(1), 79-91.

13. Goldman, N. (1990). Maximum likelihood inference of phylogenetic trees, with special reference to a poisson process model of DNA substitution and to parsimony analyses. *Systematic Biology*, 39(4), 345-361.
14. Guindon, S. and Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5), 696-704.
15. Hasegawa, M., Kishino, H. and Saitou, N. (1991). On the maximum likelihood method in molecular phylogenetics. *Journal of Molecular Evolution*, 32(5), 443-445.
16. Kashyap, R. and Subas, S. (1974). Statistical estimation of parameters in a phylogenetic tree using a dynamic model of the substitutional process. *Journal of Theoretical Biology*, 47(1), 75-101.
17. Knowles, L.L. and Kubatko, L.S. (2011). *Estimating species trees: practical and theoretical aspects*, John Wiley and Sons.
18. Neyman, J. (1971). Molecular studies of evolution: A source of novel statistical problems. In Statistical decision theory and related topics (pp. 1-27). Academic Press.
19. Nguyen, Q. and Roos, T. (2015). Likelihood-based inference of phylogenetic networks from sequence data by phylodag. In *International Conference on Algorithms for Computational Biology*, pages 126-140. Springer.
20. Rambaut, A. and Grass, N.C. (1997). Seq-gen: an application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences: CABIOS*, 13(3), 235-238.
21. Rogers, J.S. (1997). On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Systematic Biology*, 46(2), 354-357.
22. Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A. and Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3), 539-542.
23. Rosenberg, N.A. (2007). Counting coalescent histories. *Journal of Computational Biology*, 14(3), 360-377.
24. Solís-Lemus, C. and Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS genetics*, 12(3), e1005896.
25. Stadler, T. (2014). Treesim: Simulating trees under the birth-death model. R Package version 2.0.
26. Than, C., Ruths, D. and Nakhleh, L. (2008). Phylonet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9, Article 322.
27. Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, 66(3), 763-775.
28. Yang, Z. (1994). Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic biology*, 43(3), 329-342.
29. Yu, Y. and Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(10), Article S10.
30. Yu, Y., Than, C., Degnan, J.H. and Nakhleh, L. (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, 60(2), 138-149.

31. Yu, Y., Barnett, R.M. and Nakhleh, L. (2013). Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic Biology*, 62(5), 738-751.
32. Yu, Y., Dong, J., Liu, K.J. and Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46), 16448-16453.
33. Zhu, S., Degnan, J.H., Goldstien, S.J. and Eldon, B. (2015). Hybrid-lambda: simulation of multiple merger and kingman gene genealogies in species networks and species trees. *BMC Bioinformatics*, 16(1), Article 292.