GMR

The Original by FUNPEC-RP

Large number of taxa used to estimate a rooted species tree with the ABC method from an unrooted gene tree

A.R.A. Alanzi

Department of Mathematics, College of Science and Human Studies at Hotat Sudair, Majmaah University, Saudi Arabia

Corresponding author: A.R.A. Alanzi
E-mail: a.alanzi@mu.edu.sa

Genet. Mol. Res. 19 (1): gmr18433
Received July 10, 2019
Accepted January 11, 2020
Published January 31, 2020
DOI http://dx.doi.org/10.4238/gmr18433

ABSTRACT. Various approaches use gene trees to infer species trees produced from incomplete lineage sorting. Generally, one of these approaches is used to deduce the rooted species tree from a rooted gene tree, or another method can be used to determine the unrooted species tree from an unrooted gene tree. Typically, this unrooted species is then rooted through at least one outgroup. However, in theory, the unrooted gene tree can be used consistently and directly to infer the rooted species tree without using an outgroup. We used an unrooted gene tree with the assumption of a multispecies coalescent model to infer the rooted species tree by using the approximate Bayesian computation (ABC). In certain cases, this could be useful, especially when it is hard to locate a fitting outgroup neglected by gene trees. To address the challenges of increasing the taxa number, an ABC was used to gauge the rooted species tree of a large number of taxa, using an unrooted gene tree to develop the rooted species tree. This is the first ABC application that can handle very large numbers of taxa. Based on the results, the Robinson-Foulds (RF) distance is generally equal to 2 when the species tree is in imbalance. When the shape of the species tree is balanced, the RF distance is normally equal to 0. Out of all shapes of species trees, the most recent one is the most appropriate for every clade.

Key words: Number of taxa; Unrooted gene tree; Outgroup; ABC method; Coalescent model

INTRODUCTION

Over the last decade, methods that use gene trees for inferring rooted species tree have experienced rapid development. According to Degnan (2018), an essential issue for any species tree inference is how to connect the time scale, used in the multispecies coalescent model in a species tree, to other evolutionary models sequentially used on gene trees. These methods include, for example, rooted triple consensus (Ewing et al., 2008), STAR (Liu et al., 2009), triplet MRP (Wang and Degnan, 2011), ASTRAL (Mirarab et al., 2014), and approximate Bayesian computation (ABC) strategy (Fan and Kubatko, 2011; Alanzi and Degnan (2017). A clear shift in methodology has been observed, from methods utilizing rooted gene trees as input (most procedures used from 2005 to 2010–2011) to strategies that utilize unrooted gene trees as inputs (most procedures used from 2010 to present). In certain circumstances, one can view the latter methods as unrooted varieties of prior rooted techniques. Some examples include pseudo likelihood utilizing quartets rather than rooted triples (Solis-Lemus and Ane, 2016) and an unrooted copy of the standard required for minimizing deep coalescence (MDC) (Yu et al., 2011). One can also view the NJst (Liu and Yu, 2011) procedure as an unrooted group from the prior STAR procedure (Liu et al., 2009); Allman et al., 2018a). It was observed whether the unrooted copy of MDC, which was applied in PhyloNet (Than et al., 2008), can use the unrooted gene tree as an input to produce a rooted species tree. This strategy is compared to ABC in Alanzi and Degnan (2017).

In part, strategies that utilize unrooted trees are advantageous whenever fast likelihood programs such as RAxML (Stamatakis, 2006) and PHYML (Guindon et al., 2010) can only give estimations of unrooted trees in the absence of a clock. Furthermore, rooting gene trees may result in systematic errors that go beyond the usual errors as a result of the estimation error for gene tree arising from short alignments or improper models for substitution. Specifically, even if one can consider a taxon as an outgroup of the level for the species tree, it does not generally follow that as it is an outgroup for any gene tree. For instance, Huang and Knowles (2009) conducted simulations that revealed that given six coalescent units (i.e., 6N generations), and after separating the root of the species tree including the outgroup and the root of the ingroup taxa, only a 95% probability of monophyly existed for the ingroup taxa. This signifies that 5% of gene trees that were rooted through an outgroup would be rooted improperly.

Allman *et al.* (2011) obtained a theoretical result that demonstrated that when five or more taxa are included, the rooted species tree can be deduced utilizing the true distribution of unrooted gene tree topologies. The basis of this result is to know about the probabilities of every topology of the unrooted gene tree. It does not use information that is in the sequence data. In practice, a limited number of loci for sequence data are used to estimate the unrooted gene trees that take place from those sequence data. Thus, it can only estimate topology probabilities. As a result, the simplicity range of loci remains unclear due to the limited numbers it uses for estimating rooted species tree and utilizing evaluated unrooted gene trees. Part of the reasoning for determining the rooted species tree given unrooted gene trees is the fact that they do not have equalities, which they can have in the gene tree probabilities.

The result that Allman *et al.* (2011) obtained does not automatically lead to an approach to infer the rooted species tree. Moreover, the ABC-inspired algorithm designed by Alanzi and Degnan (2017) has been applied in the current study to estimate the shape of root from the unrooted species tree under the assumption of unknown-unrooted species tree. However, the type of species was inferred using the simulated method by R-package.

Fan and Kubatko (2011) formulated the ABC algorithm and used the input data from a collection of gene trees. It inferred species trees through the simulation of data sets that bear similarities to the input data set. The prior distribution determined the species trees, which are used as a set of input data. After that, it measured how the observation data is different compared to the simulated species tree. It saves the species trees that showed the least difference. These species trees are then used for estimation of the species trees' posterior distribution. The aim for species trees is that they are moderately nearer to the genuine species and are randomly chosen to be produced by data set of simulation. In this way, a greater similarity must be existed for the observation data set so that when compared with known species trees, a difference from the genuine species tree is shown.

The method that Fan and Kubatko (2011) used is similar to ABC. However, in ABC strategies, simulation of an informational collection depends on the haphazardly selected parameters by determined by prior researchers, such as Alanzi and Degnan (2017), and Fan and Kubatko (2011). Furthermore, it computes differences between both data sets, which are simulated and observed. Fan and Kubatko (2011) determined the normal quantities of quality tree topologies utilizing the dispersion of gene trees from the multispecies coalescent through COAL programming as opposed to recreating gene trees from the prior chosen species tree (Degnan and Salter, 2005). Buzbas (2012) criticized this method as it is not of genuine ABC strategy since it does not have the ability to simulate data sets.

Additionally, Alanzi and Degnan (2017) made modifications to the method of Fan and Kubatko (2011). They utilized the standard ABC strategy for data set simulation for each priorly chosen parameter, which is a species tree. Moreover, the researchers utilized a prior result, which was made up of rooted species trees having the same unrooted topology. They also utilized unrooted topologies instead of the rooted ones. In spite of choosing the established species trees from the earlier ones, in each species tree, the recreated quality trees are dealt with as unrooted, taking into account the estimated separation between the estimated and observed sets of quality trees.

The method we used here is similar to that of Alanzi and Degnan (2017). However, the ABC is applied with large amounts of taxa after randomly selected species are simulated at each iteration of simulation. The same step of algorithm was then used for the computation of the eight taxa with varying speciation and extinction rates.

The difference between observation dataset and simulation dataset was used to compute the summary statistic for the ABC method. Priorly simulated species trees were related to small separations and after that filled in as the reason for the parameter's calculable posterior distribution. In Bayesian statistics, the objective is often to work out the parameter's posterior distribution supported by the data. The posterior distribution in ABC is calculable providing the outline data point closer to the information. The aim is to approximate the posterior distribution based on the data (Sisson *et al.* (2007); Joyce and Marjoram (2008)). Thus, using adequate statistics is desirable for the summary statistics (information outlines that keep all useful information for the inference) (Casella and Berger, 2002).

Computation of the difference between datasets is done by employing a summary statistic of each simulated and observed data sets. Priorly simulated species trees that compared to little separations filled in as reason for the estimation of the parameter's posterior distribution. In Bayesian statistics, the general objective is to identify the parameter's posterior distribution based on the data. The ABC approach is vital in circumstances at the molecular clock but it is not applicable and a suitable outgroup, which is hard to discover. An example of this is the study conducted by Boykin *et al.* (2010) about rooting methods for Orcuttiaea. It is likely that this method will be inefficient except if a small number of competitor species trees is present.

The objective of this study is to apply the ABC method to estimate the rooted species tree by using the unrooted gene tree, which is considered as the first ABC application for large numbers of taxa. According to Allman *et al.* (2011), estimating rooted species trees from unrooted gene tree is easier to compute than an rooted gene tree. Moreover, Allman *et al.* (2018b) argues that the distribution of branch lengths in the gene trees gives information regarding times when the species had split and then merged before the final divergence. For this reason, this paper is as the first demonstration for a large number of taxa that uses an unrooted gene tree to estimate a rooted species tree using ABC method. Therefore, Mirarab and Warnow (2015) affirmed that setting X is considered is a problem when large numbers of taxa, few gene trees or high levels of discordance exist.

The methods section below provides details of the ABC algorithm. Simulations that were performed using 8-, 12- and 16-taxon trees were used to determine the best strategy for various branch lengths and species tree topologies.

MATERIAL AND METHODS

The ABC Approach

For ABC methods, the normal method involves a first simulation from the prior parameter distribution (in this situation, a species tree having branch lengths). Consequently, data are simulated from the parameter, which indicates that species trees contain gene trees. The distance between the real data set and simulated data set is then recorded. For this project, the observed and simulated data were made up of the topologies for the unrooted gene tree. The calculated difference between the real and simulated data is affected by the summary statistic used. Some of those variations were then utilized. The approach to trees was then applied with 8, 12, and 16 taxa.

Fan and Kubatko (2011) managed this issue for eight taxa just noting rooted gene tree counts in the real data. After that, the number of topologies for gene tree was counted within the simulated data corresponding to one amongst the input trees. Nevertheless, the measure of correct tree topologies could still be too high, with each gene tree sometimes having one kind of topology (Salichos and Rokas, 2013). The researchers discovered that their method was precisely accurate for trees having eight branches instead of four branches. It is speculated that this could be a result of the dimensionality problem. Alanzi and Degnan (2017) utilized splits for eight taxa. Rather than recording the counts every split allowed, the symmetric difference is recorded between the simulated data set of splits versus real data set of splits. Splits are only for 8, 12, and16 taxa. The symmetric distinction between the real data set of splits versus simulated data set of splits is then recorded. More explanation for this algorithm provided below (Figure 1).

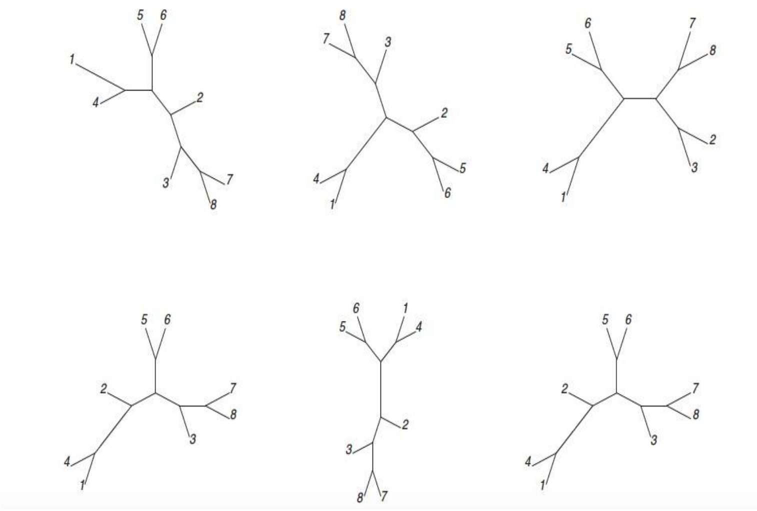

Figure 1. An example of how to explain computing distance using the multiset. The first row shows the observed data set and the second row shows the simulated data (Alanzi and Degnan, 2017). The multisets are sets to keep the track of the multiplicity of each element - the number of times needed for each element of the set occurs.

Splits for trees having 8, 12, and16 taxa are registered as multisets. Multisets refer to sets that monitor the multiplicity of each element. Multiplicity is that the quantity of times for each part of the observed set. For instance, the application of multiset thought, Figure 1 in Alanzi and Degnan (2017), gives a test case that outlines two arrangements of three trees. After that, the distance got from the multisets of splits was computed. The distance D = |Sobs \ Ssim| + |Ssim \Sobs| refers to the total sizes of the set variations. This algorithm below was adapted from the research.

Algorithm

By Simulation the observed gene tree that used to extract the splits to multiset S_{obs}.

It Begins with $J = 1$.

By simulation the species tree from the prior distribution to have the rooted species tree.

By simulation the gene trees obtained a species tree from the prior distribution from the third step, which used the Hybrid-Lambda program to simulate the gene tree and extracted all splits to multiset $S_{sim.}$

Compute the symmetric difference for splits by using this formula
D = | $S_{obs}\backslash S_{sim}$ | + | $S_{sim}\backslash S_{obs}$|, for both sets.

Redo all steps from step 2 to step 5 by increasing the value of j by 1 until reach
J times.

From αJ choose the small values obtained from stage 5. After that the size of species tree, which corresponds to the small value is then retained. The trees are capable of estimating the posterior distribution.

In stage 5, the observation of the symmetric distance refers to the distance for both multisets. Accordingly, it could be noted that if the basis of the sets S_{obs} and S_{sim} are the observation (real) and one simulation tree, then, the Robinson-Foulds (RF) distance would be reduced by D (Robinson and Foulds, 1981). In the algorithm, the calculated value of D generalizes the RF distance when two sets of trees are present instead of two individual trees. This work presents a summary of the posterior distribution of the species tree splits as a way of inferring the rooted species tree through the majority-rule consensus. This study implements all the calculations by using a number of scripts, which includes code in R (Ihaka and Gentleman, 1996).

Besides, this research bears similarity to Alanzi and Degnan (2017) in that species trees are estimated based on the steps of the algorithm. However, expected counts of small amounts of taxa were computed by Alanzi and Degnan (2017) under the assumption that the species tree is known. On the other hand, this study assumes that the species tree is unknown when it calculates expected counts for large numbers of taxa. This work used unrooted gene trees to estimate rooted species trees. The priors utilized were restricted to finding the root of the species tree given an unrooted tree.

Simulation

Fifty species trees were simulated using TreeSim (Stadler, 2011). A pure birth model was used with birth parameters of λ= 0.25, 0.5, 0.75, and 1.0. Four values of μ/λ were used -0.0, 0.25, 0.5, and 0.75. This approach is similar to other species tree inference papers that used a pure birth model to summarize results over species trees (Huang et al., 2010).

The species tree serves as the parameter and the data is made up of the topologies of the unrooted gene tree. The researcher started with the step that involves data simulation and recording. The second step is the utilization of a prior to simulate species trees (the parameters). It then uses the simulated species trees for data simulation in the form of unrooted gene trees. It utilized a uniform prior for the species tree topologies. The prior is made up of $2n - 3$ possible rooted topologies for the unrooted tree topology. According to Furnas (1984), exactly $2n - 3$ times full rooted binary trees with n leaves exist for unrooted binary trees having n leaves. The study used an exponential distribution having a 1.0 coalescent unit's rate for branch lengths of the species trees. This includes the lengths of the external branches. The study used the exponential prior in order to simulate a broad range of coalescent data. Once data simulation with those priors happens, recording of the split counts take place.

Simulation of gene trees was performed from species trees through Hybrid-Lambda (Zhu et al., 2015). $J = 50,000$ was used for the simulation. For every repetition of J, it used a sample size equal to 100 genes. The study sets the value of α to 0.002 to retain the 100 best species trees. It used the formula αJ to identify the desired number of species trees, with the smallest D as the basis. The posterior probability was computed using the algorithm when 8-,12, and16-taxon trees were inferred. It also used the sets library (Hornik and Meyer, 2009) from the R-package (Ihaka and Gentleman, 1996) and the tape library (Paradis et al., 2004) to compute for the symmetric difference among the multisets of splits. This is for the distance between the observed and simulated data. In this process, splits in the observed data are found, but not in the simulated data. On the contrary, it can likewise discover the splits in the simulated data, however not discover splits in the observed data. Figure 2 illustrates the 8-, 12, 16- taxon trees that were utilized for simulation, as well as their equivalent unrooted topologies. The consensus tree was used to summarize a posterior distribution of trees, a practice that is fairly typical in Bayesian phylogenetic (Holder et al., 2008).

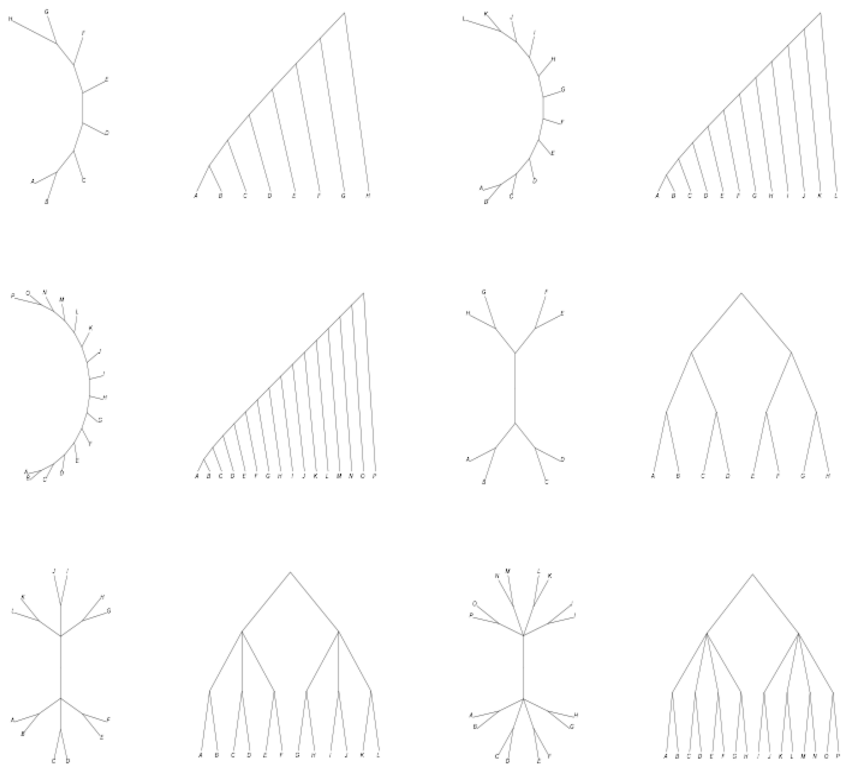

Figure 2. Topology of the species used in the simulation for each taxon: 8- 12- 16- taxa with two types of shape, which are the unbalanced and balanced shapes. The first six shapes show the unbalanced taxa with rooted shape versus unrooted shape and the rest of the figure shows the balanced shape with rooted versus unrooted shape.

RESULTS

The generation of these species trees took place under a pure birth model that had varying values of λ. Different results were achieved with different μ values. In this study, branch lengths had a tendency to be longer when the values of λ were smaller. For 8-taxon trees, the consensus trees obtained for the 100 best species trees possessed an average RF distance that is shown in Table 1. As seen from the second to the fifth column, each row represents different λ values while each column represents different μ values. Columns six to nine show information about accurate (RF = 0). For the rest of the Table 1, the percentage of trees having an RF value of 2 or lower (i.e. at most one clade is wrong) is given. Based on Table 1, it was revealed that the smallest RF distance was observed with the smallest λ value, which is 0.25, and μ value of 0.5. However, the highest proportion of recovering is completely accurate (RF = 0) for all λ and μ values, when λ = 1 and μ = 0.5. The next highest proportion is when λ is 0.75 and μ is 0.25. It was also observed that 74% of trees possessed an RF value of 2 or lower when λ = 1 and μ = 0.5 and when λ = 0.25 and μ = 0.25.

Table 1. Summary for eight taxa.

	Average RF				**Proportion of accurately RF = 0**				**Proportion RF of 2 or Lower**			
	μ				μ				μ			
λ	0.00	0.25	0.50	0.75	0.00	0.25	0.50	0.75	0.00	0.25	0.50	0.75
1.00	2.4	2.36	2.44	2.72	0.22	0.10	0.26	0.18	0.79	0.64	0.74	0.64
0.75	2.64	2.4	2.48	2.84	0.20	0.18	0.22	0.20	0.50	0.68	0.68	0.64
0.50	2.76	2.32	2.28	2.52	0.12	0.18	0.14	0.08	0.62	0.70	0.62	0.56
0.25	2.08	2.64	2.00	2.44	0.22	0.12	0.18	0.12	0.62	0.74	0.72	0.62

Based on Table 2, the smallest RF distance was observed to have taken place with the smallest λ value, which is 0.25, and μ value of 0.5. Furthermore, the highest proportion is recovered completely and accurately (RF= 0) among all λ and μ values, when λ = 0.5 and μ = 0.75. Around 70% of trees were observed to have had an RF of 2 or lower when λ = 0.5 and μ = 0.5 and λ = 0.25 and μ = 0.5.

Table 2. Summary for 12 -taxa.

	Average RF				**Proportion of accurately RF = 0**				**Proportion RF of 2 or Lower**			
	μ				μ				μ			
λ	0.00	0.25	0.50	0.75	0.00	0.25	0.50	0.75	0.00	0.25	0.50	0.75
1.00	3.32	2.92	3.16	3.08	0.10	0.16	0.10	0.20	0.50	0.58	0.56	0.56
0.75	3.52	3.16	3.16	3.12	0.20	0.10	0.12	0.06	0.64	0.52	0.52	0.58
0.50	3.00	2.92	2.44	3.00	0.14	0.10	0.16	0.20	0.52	0.62	0.70	0.54
0.25	3.20	2.96	2.44	2.72	0.12	0.12	0.12	0.16	0.54	0.56	0.68	0.64

Table 3 shows that the smallest RF distance was observed with the smallest λ values, which are 1 and 0.5, both with a μ value of 0.75. However, all values of λ render the smallest RF distance with μ = 0.75. On the other hand, proportion of completely recovered RF shows the highest accurately (RF = 0) with the different values of λ that is greater than 0.25 and the value of μ = 0.75 while other exception was λ = 0.25 and μ = 0. Moreover, 62% and 66% of trees possessed an RF of 2 or lower with all λ values when μ = 0.75, except for when the value of λ = 0.5 and μ = 0.25. In the latter's case, 58% of trees possessed an RF of 2 or lower. Based on Table 2 and 3, it is observed that μ had an effect on the estimation of the result, given an increase in the number of taxa and a decrease in the values of λ.

Imbalanced species trees have a tendency to exhibit higher gene tree discordance compared to balanced species trees even if they have similar branch lengths (Degnan and Salter, 2005). This is a possible evidence for why the caterpillar is preferred in the posterior when there are absolutely arbitrary gene trees, which is the star species tree, and why there is underestimation of caterpillars when no variation is observed in the gene trees, which is the species tree has long branches. In line with this expectation, the ratio of times when there is an inference for the balanced species tree is lower compared to the expected value for an uninformative prior for the star tree.

Table 3. Summary for 16 taxa.

	Average RF				**Proportion of accurately RF = 0**				**Proportion RF of 2 or Lower**			
	μ				μ				μ			
λ	0.00	0.25	0.50	0.75	0.00	0.25	0.50	0.75	0.00	0.25	0.50	0.75
1.00	3.84	4.52	3.12	2.84	0.10	0.06	0.16	0.20	0.46	0.40	0.54	0.62
0.75	3.60	3.24	4.00	2.96	0.08	0.14	0.10	0.16	0.44	0.52	0.42	0.62
0.50	3.32	3.16	3.24	2.84	0.12	0.14	0.14	0.18	0.48	0.58	0.46	0.50
0.25	2.92	3.04	2.92	2.88	0.20	0.16	0.10	0.12	0.54	0.54	0.56	0.66

Furthermore, a bias exists in favor of balanced topologies when the gene tree topologies fail to exhibit variations (Alanzi and Degnan, 2017). This seems as dissimilar to the instance of deducing unrooted species trees from unrooted gene trees or rooted species trees from rooted gene trees, longer internal branches don't naturally facilitate the inferences. Variation in the gene trees is needed to infer the rooted species tree from unrooted gene trees. This argument is in agreement with results obtained for all numbers of taxa in this study. The above figures only illustrate one iteration for every posterior probability on clade since every iteration provides a different species tree topology. This results into difficulties in computing the average of posterior probability for every iteration, because there is a change in the correct clade with every iteration. For the computation of the posterior probabilities when 8, 12, and16 taxa were possessed by the species tree, only split counts were utilized in the gene trees. The highly posterior probability tree could not be matched with the species tree. Thus, the RF distance was always equivalent to 2.0. In addition, the species trees are divided to two equal leaves or almost equally as seen in Figures 3, 4, and 5. Most of species trees that RF = 0 turned to balanced shaped, even though it is not balanced shape. However, all the shapes illustrated the highest posterior probability for every clade.

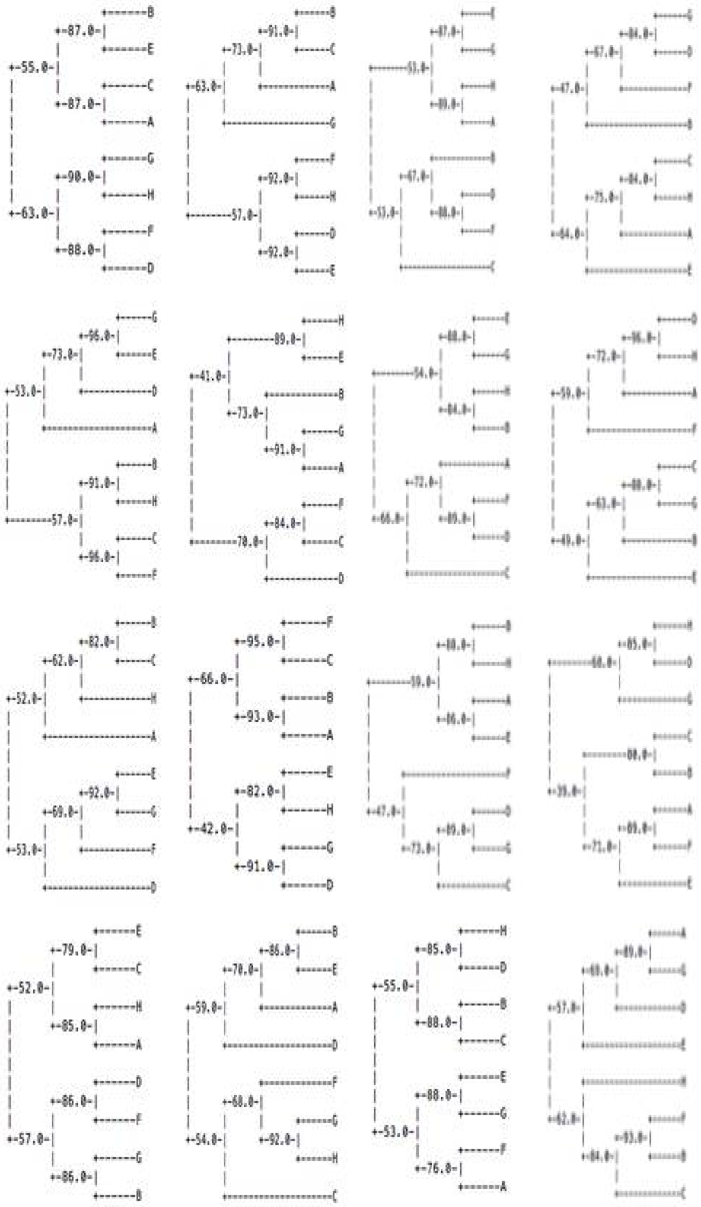

Figure 3. One shape of posterior probability for eight taxa with all values of lambda and all values of Mu where the RF distance is equal to zero. Note that every row of this figure shows different values of μ. The first row shows that $\mu = 0$, the second row shows that $\mu = 0.25$, the third row shows that $\mu = 0.50$, and the fourth row shows that μ = 0.75. Also the columns show different value of λ from column one to column four, ranging from 0.25 to 1.

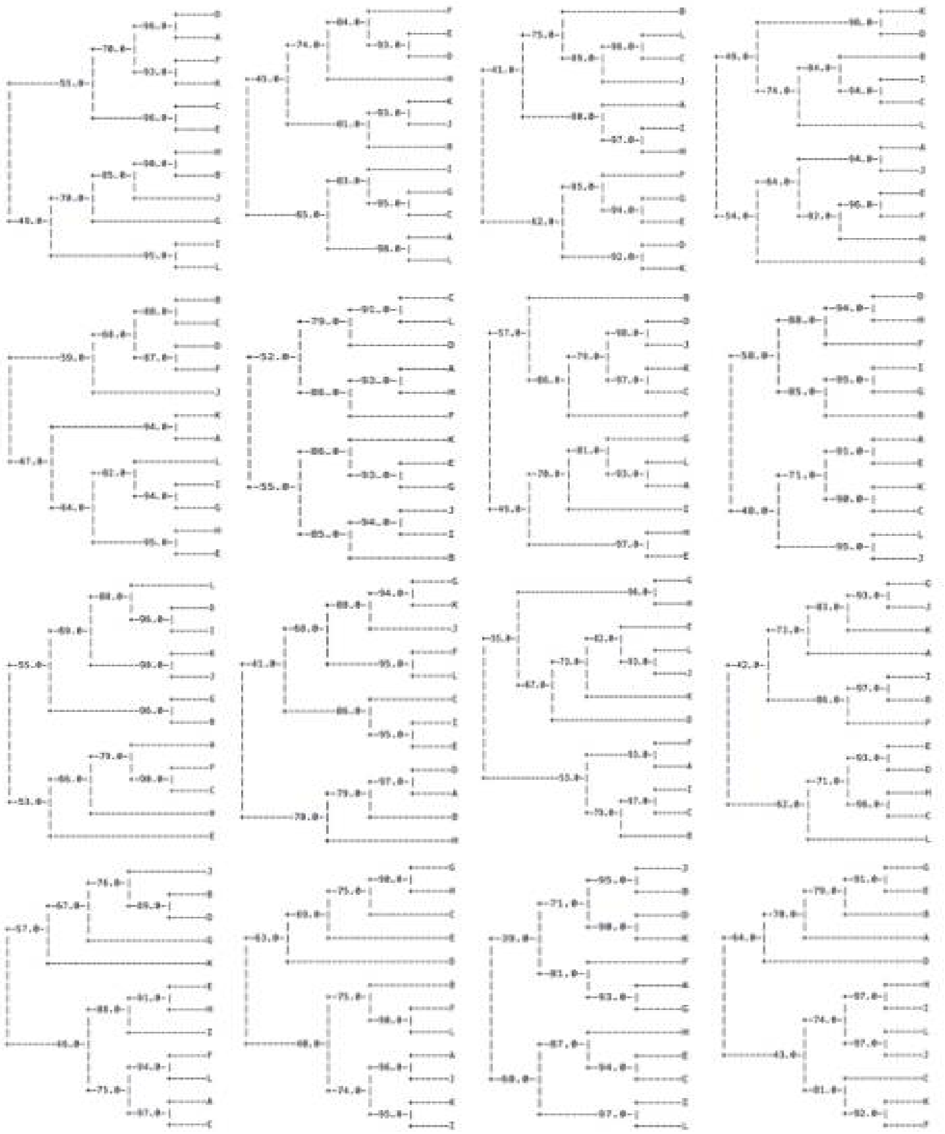

Figure 4. One shape of posterior probability for 12 taxa with all values of lambda and all values of Mu where the RF distance is equal to zero. Note that every row of this figure show different value of μ. The first row shows that μ = 0, the second row shows that μ = 0.25, the third row shows that μ = 0.50, and the fourth row shows that μ = 0.75. Also each column shows different value of λ, ranging from 0.25 to 1.

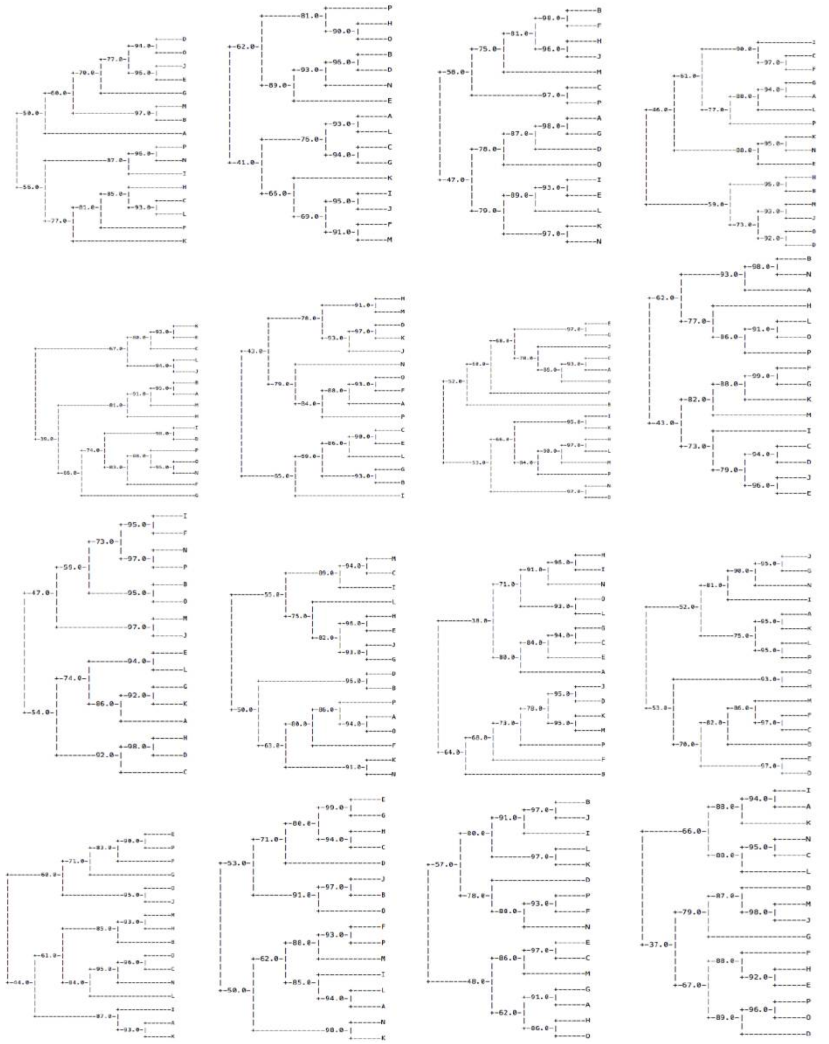

Figure 5. One shape of posterior probability for 16 taxa with all values of lambda and all values of Mu where the RF distance is equal to zero. Note that every row of this figure shows different value of μ. The first row shows that μ = 0, the second row shows that μ = 0.25, the third row shows that μ = 0.50, and the fourth row shows that μ = 0.75. Also the columns have λ values ranging from 0.25 to 1.

DISCUSSION AND CONCLUSIONS

This study uses the coalescent model with unrooted gene trees to estimate rooted species trees. It is different from the author's previous research (Alanzi and Degnan (2017).in using a method that allows large numbers of taxa. However, there was also an improvement of a version of the MDC (minimize deep coalescence method) by using the unrooted gene trees to infer the rooted species trees (Liu and Yu, 2011). When there is failure for two or more lineages to coalesce in a population (i.e. in the case of incomplete lineage sorting), extra lineages could have arisen. These extra lineages will be held during speciation events that go back in time to compute the MDC score tallies for the lowest base measure of these additional lineages that are required for taking place in a specific species tree from a gene tree. When combining the species tree with the gene trees, the development of MDC was initially in light of the concept to infer the rooted species tree that needed the least number of extra lineages. This eventually included unrooted gene trees through minimization for the replacement of rooted for gene trees. Hence, one would then be able to use it to restore a rooted species tree even when unrooted gene trees serve as input (Yu et al., 2011). Thus, the MDC score that a gene tree contributes for a candidate species tree will be calculated, because of the minimum amounts of additional lineages. Moreover, the ABC approach in this study bears similarities to the method that Alanzi and Degnan (2017) developed. However, they did not make any inferences regarding the root of the gene trees.

The MDC criterion, even though it was one of the principal criteria to be executed for gene trees to infer species trees (Maddison and Knowles, 2006), was determined as statistically conflicting when used within the rooted setting, which rooted gene trees used as inputs. This took place around the same time of publishing unrooted extension (Than and Rosenberg, 2011). This implies that it has a more prominent inclination to restore an inaccurate species tree the closer the quantity of info gene trees near interminability. It does not have frequent usage since methods with more accuracy have been developed. Furthermore, Alanzi and Degnan (2017) noted that PhyloNet is capable of utilizing unrooted MDC to restore a rooted species tree even at the four- taxa case. However, four-taxon gene tree topologies are not capable of identifying the rooted species tree that is present underneath the multispecies coalescent (Allman et al., 2011).

The ABC method for root location estimation is relatively slow because of the numerous computations. However, it scales sensibly well with the number of taxa. This study agrees with Alanzi and Degnan (2017) regarding the simulation time for various amounts of taxa. However, the simulation in this study is slower than that by Alanzi and Degnan (2017) because the species tree is not known, giving a rise to the need for calculating the priors for every new species tree. Alanzi and Degnan (2017) already owned a known species tree. Furthermore, the prior of this species and constant for all iterations were known as well.

It could be concluded that the root location for 8-, 12-, 16-taxa species trees is hard to infer in practice within 100 loci, however, it is possible in a few cases. It could also be stated here that there is a need to do more work in order to observe the effect of adding more loci. The capacity of ABC for inferring the root additionally appears to be highly sensitive to the split and branch length combinations.

One can use the ABC method with a flat or an informative prior. A flat prior was used in this study, with the assumption that a specific unrooted species tree was not known. If there is also uncertainty in the unrooted species tree, it could be reflected by including more rooted trees in the prior. Also, given characteristic birth-death processes, certain unrooted trees can be more likely than others in cases of six or more taxa (Steel, 2012). Thus, this could be the basis of the prior instead of making every labelled topology that has equal likelihood in the prior. This study concludes that the ABC method will not function well with very large amounts of taxa. When this method uses evaluated gene trees rather than known gene trees, the ABC method needs to estimate $G??J$ gene trees. The G refers to the quantity of gene tree loci and J refers to the amount of data sets in simulation. This is a more laborious approach since gene trees have to be estimated. However, it is also scaled linearly with the number of loci.

ACKNOWLEDGMENTS

The author thanks Prof. James H. Degnan for introducing this interesting topic. Thanks is also due to the Deanship of Scientific Research at Majmaah University for support and encouragement.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Alanzi AR and Degnan JH (2017). Inferring rooted species trees from unrooted gene trees using approximate Bayesian computation. *Mol. Phylogenetics Evol.* 116: 13-24.

Allman ES, Degnan JH and Rhodes JA (2011). Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math Biol.* 62: 833-862.

Allman ES, Degnan JH and Rhodes JA (2018a). Species tree inference from gene splits by unrooted STAR methods. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 15: 337-342.

Allman ES, Degnan JH and Rhodes JA (2018b). Split probabilities and species tree inference under the multispecies coalescent model. *BullMat. Biol.* 80: 64-103.

Boykin LM, Kubatko LS and Lowrey TK (2010). Comparison of methods for rooting phylogenetic trees: A case study using Orcuttieae (Poaceae: Chloridoideae). *Mol. Phylogenetics Evol.* 54: 687-700.

Buzbas EO (2012). On the article titled "estimating species trees using Approximate Bayesian Computation (fan and kubatko, molecular phylogenetics and evolution 59: 354-363). *Mol. Phylogenetics Evol.* 65: 1014-1016.

Casella G and Berger RL (2002). Statistical inference, Volume 2. Duxbury Pacific Grove, CA.

Degnan JH (2018). Modeling hybridization under the network multispecies coalescent. *Syst. Biol.* 67: 786-799.

Degnan JH and Salter LA (2005). Gene tree distributions under the coalescent process. *Evolution.* 59: 24-37.

Ewing GB, Ebersberger I, Schmidt HA and Von Haeseler A (2008). Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol.* 8: 118.

Fan HH and Kubatko LS (2011). Estimating species trees using approximate Bayesian computation. *Mol. Phylogenetics Evol.* 59: 354-363.

Furnas GW (1984). The generation of random, binary unordered trees. *J. Classif* . 187-233.

Guindon S, Dufayard J-F, Lefort,V, Anisimova M, et al. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59: 307-321.

Holder MT, Sukumaran J and Lewis PO (2008). A justification for reporting the majority-rule consensus tree in Bayesian phylogenetics. *Syst. Biol* .57: 814-821.

Hornik K and Meyer D (2009). Generalized and customizable sets in R. *J. Stat. Softw.* 31: 2.

Huang H and Knowles LL (2009). What is the danger of the anomaly zone for empirical phylogenetics? *Syst. Biol.* 58: 527-536.

Huang H, He Q, Kubatko LS and Knowles LL (2010). Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst. Biol.* 59: 573-583.

Ihaka R and Gentleman R (1996). R: a language for data analysis and graphics. *J. Comput. Graph Stat.* 5: 299-314.

Joyce P and Marjoram P (2008). Approximately sufficient statistics and bayesian computation. *Stat. Appl. Genet. Mol. Biol.* 7: 26.

Liu L and Yu L (2011). Estimating species trees from unrooted gene trees. *Syst. Biol.* 60: 661-667.

Liu L, Yu L, Pearl DK and Edwards SV (2009). Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 58: 468-477.

Maddison WP and Knowles LL (2006). Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55: 21-30.

Mirarab S and Warnow T (2015). ASTRAL II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics.* 31: i44-i52.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, et al. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics.* 30: i541-i548.

Paradis E, Claude J and Strimmer K (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics.* 20: 289-290.

Robinson DF and Foulds LR (1981). Comparison of phylogenetic trees. *Math Biosci.* 53: 131-147.

Salichos L and Rokas A (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature.* 497: 327.

Sisson SA, Fan Y and Tanaka MM (2007). Sequential monte carlo without likelihoods. *Proc. Natl. Acad. Sci. USA.* 104: 1760-1765.

Solis-Lemus C and Ane C (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics.* 12: e1005896.

Stadler T (2011). Simulating trees with a fixed number of extant species. *Syst. Biol.* 60: 676-684.

Stamatakis A (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 22: 2688-2690.

Steel M (2012). Root location in random trees: A polarity property of all sampling consistent phylogenetic models except one. *Mol. Phylogenetics Evol.* 65: 345-348.

Than C, Ruths D and Nakhleh L (2008). Phylonet: a software package for analyz- ing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics.* 9: 322.

Than CV and Rosenberg NA (2011). Consistency properties of species tree inference by minimizing deep coalescences. *J. Comput. Biol.* 18: 1-15.

Wang Y and Degnan JH (2011). Performance of matrix representation with parsimony for inferring species from gene trees. *Stat. Appl. Genet. Mol.* 10: 21.

Yu Y, Warnow T and Nakhleh L (2011). Algorithms for mdc-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. *J .Comput. Biol.* 18: 1543-1559.

Zhu S, Degnan JH, Goldstien SJ and Eldon B (2015). Hybrid-lambda: simulation of multiple merger and kingman gene genealogies in species networks and species trees. *BMC Bioinformatics.* 16: 292.