Molecular Phylogenetics and Evolution 65 (2012) 1014–1016

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Letter to the Editor

On the article titled "Estimating species trees using approximate Bayesian computation" (Fan and Kubatko, Molecular Phylogenetics and Evolution 59:354–363)

In their article titled "Estimating species trees using approximate Bayesian computation" Fan and Kubatko present an algorithm called ST-ABC to sample the posterior distribution of species trees (Molecular Phylogenetics and Evolution 59: 354–363). The authors claim that ST-ABC is an approximate Bayesian computation (ABC) algorithm. Here, I argue that one of the steps in their algorithm differs from the approach that would be taken by a proper ABC algorithm. Therefore, the distribution sampled by ST-ABC might not approximate the true posterior distribution as in proper ABC algorithms. As a consequence, the estimates based on posterior samples obtained by ST-ABC might not recover the true species tree.

ABC algorithms sample an approximate form of the posterior distribution of a parameter of interest without directly evaluating the likelihood of the parameter given the data (Tavaré et al., 1997; Beaumont et al., 2002; Marin et al., 2011). Given the observed data set D, a prior distribution $\pi(\theta)$ for parameter θ, and a probability model $P(x|\theta)$, an ideal version of ABC algorithm where there is no approximation error proceeds as follows:

Algorithm. ABC (ideal)

(1) Simulate a parameter value θ^* from $\pi(\theta)$.
(2) Simulate a data set x^* from $P(x|\theta^*)$.
(3) Accept the parameter value θ^* if $x^* = D$.

Iterated many times, Algorithm–ABC samples the posterior distribution $\pi(\theta|D)$ of the parameter θ by the following argument. For a given iteration, a value θ^* is accepted with probability proportional to $\pi(\theta^*)P(x^*|\theta^*)\mathbf{I}_{\{x^*=D\}}$, where $\mathbf{I}_{\{S\}}$ is the indicator function taking a value of 1 on set S and 0 otherwise (Marin et al., 2011). Summing over all iterations, a Monte Carlo approximation proportional to the sampling probability of θ^* obtained by Algorithm–ABC is given by

$$h(\theta^*) \propto \sum_{x^*} \pi(\theta^*)P(x^*|\theta^*)\mathbf{I}_{\{x^*=D\}} \quad (1)$$

$$= \pi(\theta^*)P(D|\theta^*) \quad (2)$$

$$\propto \pi(\theta^*|D),$$

where the proportionality follows by Bayes Theorem. This equation verifies that the distribution sampled by Algorithm–ABC indeed has the correct posterior probability $\pi(\theta^*|D)$ for each θ^*.

In practice, with high-dimensional datasets, the equality $x^* = D$ to accept a parameter value in the third step of Algorithm–ABC is hard to satisfy. Therefore, in a typical *approximate* Bayesian computation algorithm, the third step is replaced by the approximation $x^* \approx D$. The premise of ABC is that the correct posterior distribution is sampled in the ideal version of the algorithm (e.g., Algorithm–ABC) and an approximate form of the posterior distribution is sampled when using $x^* \approx D$ in place of $x^* = D$. Below, we first provide a theoretical explanation of why the Algorithm–ST-ABC does not always sample the correct posterior distribution, even when the ideal version of ABC is used, implying that it also does not sample the correct distribution when an approximate version is used. We then work out a three-taxon example to show that Algorithm–ST-ABC does not sample the correct posterior distribution of the species tree in a specific case.

To apply the ABC argument to ST-ABC, the parameter of interest denoted by θ is the species tree (topology together with the branch lengths). Let the probability distribution of gene tree topologies induced by the species tree θ^* be $P(\beta|\theta^*)$. The ST-ABC algorithm is as follows (modified from Fan and Kubatko (2011)):

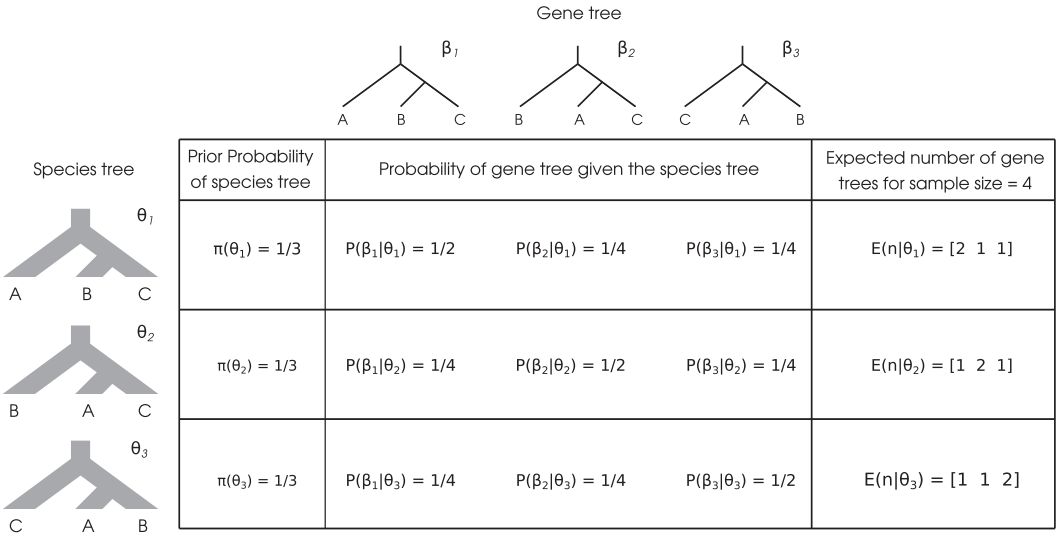
Algorithm. ST-ABC

(1) Set $j = 1$.
(2) Sample a parameter value θ^* from a specified prior distribution $\pi(\theta)$.
(3) Using θ^*, analytically compute the probability distribution of gene trees $P(\beta|\theta^*)$ and obtain the expected frequency of each gene tree topology, $\mathbf{n}_{\text{exp}} = (n_{\text{exp},1}, n_{\text{exp},2}, \ldots, n_{\text{exp},G})$, by multiplying each probability by the sample size N.
(4) Compute[a] $D_j = \sum_{i=1}^{G} \frac{(n_{\text{obs},i} - n_{\text{exp},i})^2}{n_{\text{exp},i}}$.
(5) Increment j by 1 and repeat steps (2)–(4) J times.
(6) Retain the αJ sampled species trees with the smallest values of D_j.

[a] The denominator in Step 4 of the original algorithm in Section 2.1 of Fan and Kubatko reads $n_{\text{exp,j}}$ due to a typographical error which we replaced by $n_{\text{exp,i}}$. Another typographical error kindly pointed out by the Associate Editor appears in Section 2.3.1, where the tree (H (C (G,O)) should read as (((H,C),G),O).

For illustrative purposes, consider an exact version of Algorithm–ST-ABC, in which for all retained values of θ, the statistics D_j computed in Step 4 of Algorithm–ST-ABC are exactly zero, so that $\mathbf{n}_{\text{obs}} = \mathbf{n}_{\text{exp}}$. The arguments carry over straightforwardly for the genuine ABC algorithm where $D_j > 0$.

1055-7903/$ - see front matter
http://dx.doi.org/10.1016/j.ympev.2012.08.014



Prior Probability of species tree	Probability of gene tree given the species tree			Expected number of gene trees for sample size = 4
$\pi(\theta_1) = 1/3$	$P(\beta_1\|\theta_1) = 1/2$	$P(\beta_2\|\theta_1) = 1/4$	$P(\beta_3\|\theta_1) = 1/4$	$E(n\|\theta_1) = [2\ 1\ 1]$
$\pi(\theta_2) = 1/3$	$P(\beta_1\|\theta_2) = 1/4$	$P(\beta_2\|\theta_2) = 1/2$	$P(\beta_3\|\theta_2) = 1/4$	$E(n\|\theta_2) = [1\ 2\ 1]$
$\pi(\theta_3) = 1/3$	$P(\beta_1\|\theta_3) = 1/4$	$P(\beta_2\|\theta_3) = 1/4$	$P(\beta_3\|\theta_3) = 1/2$	$E(n\|\theta_3) = [1\ 1\ 2]$

Fig. 1. Species trees and gene trees for the counterexample with three taxa. The probability of each gene tree is calculated for each species tree assuming that the time between speciation events is $t = -\log(3/4)$.

A Monte Carlo approximation proportional to the sampling probability of a species tree θ^* obtained by Algorithm–ST-ABC is given by

$$\begin{aligned} h'(\theta^*) &\propto \pi(\theta^*)P(\mathbf{n}_{\exp}|\theta^*, N)\mathbf{I}_{\{\mathbf{n_{obs}}=\mathbf{n}_{\exp}\}} \\ &= \pi(\theta^*)\mathbf{I}_{\{\mathbf{n_{obs}}=NP(\beta|\theta^*)\}} \\ &= \pi(\theta^*)\mathbf{I}_{\{(\mathbf{n_{obs}}/N)=P(\beta|\theta^*)\}}. \end{aligned} \tag{3}$$

The equality on the second line follows from the fact that given a species tree θ^* and the sample size N, the expected values for the numbers of observations of the various gene tree topologies given θ^* are known (Degnan and Salter, 2005).

If Algorithm–ST-ABC is a proper ABC algorithm, then $h'(\theta^*)$ must be proportional to $h(\theta^*)$ given in expression (1). Substituting the generic data D in expression (2) with the specific data $\mathbf{n_{obs}}$ for the ST-ABC case and setting the right-hand sides of expressions (2) and (3) proportional to each other, we get

$$\pi(\theta^*)P(\mathbf{n_{obs}}|\theta^*) \propto \pi(\theta^*)\mathbf{I}_{\{(\mathbf{n_{obs}}/N)=P(\beta|\theta^*)\}}.$$

Hence, Algorithm–ST-ABC correctly samples the posterior distribution of species tree *only if* $\mathbf{I}_{\{(\mathbf{n_{obs}}/N)=P(\beta|\theta^*)\}}$ is proportional to the likelihood of the observed data under θ^*:

$$P(\mathbf{n_{obs}}|\theta^*) \propto \mathbf{I}_{\{(\mathbf{n_{obs}}/N)=P(\beta|\theta^*)\}}.$$

Since the indicator function $\mathbf{I}_{\{(\mathbf{n_{obs}}/N)=P(\beta|\theta^*)\}}$ takes a value of 1 if $\mathbf{n_{obs}}/N = P(\beta|\theta^*)$ and is 0 otherwise, the last proportionality implies that in Algorithm–ST-ABC, the probability of the observed data $\mathbf{n_{obs}}$ given the species tree θ^* is taken to be 1 if the observed frequencies $\mathbf{n_{obs}}/N$ match the probabilities $P(\beta|\theta^*)$ in the distribution of gene tree topologies and is 0 otherwise. However, the Degnan–Salter distribution of gene tree topologies $P(\beta|\theta^*)$ assigns a positive value to $P(\mathbf{n_{obs}}|\theta^*)$ for *any* θ^* because gene trees in the observed sample can always be embedded in any species tree θ^*, implying that $\mathbf{n_{obs}}$ is a possible outcome under any θ^*. Such probabilities are not taken into account in sampling the species trees by Algorithm–ST-ABC. In general, $P(\mathbf{n_{obs}}|\theta^*)$ is not proportional to $\mathbf{I}_{\{(\mathbf{n_{obs}}/N)=P(\beta|\theta^*)\}}$.

The following counterexample shows that the Algorithm–ST-ABC does not always sample the correct posterior distribution. We consider a three-taxon case with one lineage in each taxon. In this case, there are three labeled species tree topologies, which we denote by $\theta_1, \theta_2, \theta_3$, and three labeled gene tree topologies, which we denote by $\beta_1, \beta_2, \beta_3$. Given species tree θ_i, the probability of each gene tree β_j whose labels at the leaves do not match the species tree labels is given by the well-known formula $P(\beta_j|\theta_i) = (1/3)e^{-t}$, $j \neq i$, where t is time between the two speciation events measured in coalescent units. The probability of the gene tree whose labels do match the labels of the species tree is then given by $P(\beta_j|\theta_i) = 1 - (2/3)e^{-t}$, $j = i$. For computational convenience we fix $t = -\log(3/4)$ so that $P(\beta_j|\theta_i) = 1/4$, $j \neq i$ and $P(\beta_j|\theta_i) = 2/4$, $j = i$. Further, we assume that each species tree has a prior probability of 1/3. We consider a sample of size $N = 4$ gene trees and given the probabilities for each gene tree β_j under each species tree θ_i, we compute the expected value of the distribution of gene trees under each species tree θ_i by $\mathbf{n}_{\exp_i} = [NP(\beta_1|\theta_i)\ NP(\beta_2|\theta_i)\ NP(\beta_3|\theta_i)]$ (Fig. 1 shows the probabilities of each gene tree given each species tree and provides expected values of gene trees for a sample of size 4).

Let us now assume that the observed sample satisfies $\mathbf{n_{obs}} = \mathbf{n}_{\exp_1}$. That is, using the last column for species tree θ_1 in Fig. 1, we have $\mathbf{n_{obs}} = [2\ 1\ 1]$. We can now compute the statistic D_i for each species tree θ_i by Step 4 of Algorithm–ST-ABC. The statistics are as follows:

$$\begin{aligned} D_1 &= \frac{(2-2)^2}{2} + \frac{(1-1)^2}{1} + \frac{(1-1)^2}{1} = 0, \\ D_2 &= \frac{(2-1)^2}{1} + \frac{(1-2)^2}{2} + \frac{(1-1)^2}{1} = \frac{3}{2}, \\ D_3 &= \frac{(2-1)^2}{1} + \frac{(1-1)^2}{1} + \frac{(1-2)^2}{2} = \frac{3}{2}. \end{aligned}$$

Note that we have chosen the observed counts to be exactly equal to the expected counts $\mathbf{n}_{\exp_1}$, such that only D_1 is zero. The posterior sample from Algorithm–ST-ABC for our example will consist of θ_1 with probability 1, because only the species trees with $D_i = 0$ are retained in the sample, and this condition is satisfied only for $i = 1$. However, we have a positive prior probability of 1/3 for each species tree θ_i and there exists positive probability of observing the sample $P(\mathbf{n_{obs}}|\theta_i)$ under each species tree, since all the gene tree probabilities $P(\beta_j|\theta_i)$ in Fig. 1 are positive. Therefore, $P(\mathbf{n_{obs}}|\theta_i)\pi(\theta_i)$ must be positive for all species trees, implying that the posterior probability of each species tree θ_i –which is

proportional to $P(\mathbf{n_{obs}}|\theta_i)\pi(\theta_i)$– must be positive as well. Consequently, the true posterior distribution of species trees has more than one value of the species tree (i.e, θ_1) in its support. This true posterior distribution is not recovered by Algorithm–ST-ABC, because Algorithm–ST-ABC samples only the one value of the species tree, θ_1, with probability 1 and assigns zero posterior probability to the other species trees θ_2 and θ_3.

The key difference between Algorithm–ST-ABC and Algorithm–ABC is as follows. In standard ABC, a *new random data set* is generated under the model conditional on each parameter value drawn from its prior distribution (Step 2 of Algorithm–ABC), whereas in Algorithm–ST-ABC *the fixed quantity* $\mathbf{n}_{\text{exp}}$ (the expected value calculated from the distribution of gene tree topologies under the parameter value for a sample of size N) is used instead of such a random data set. Using the fixed quantity $\mathbf{n}_{\text{exp}}$ in Algorithm–ST-ABC eliminates the sampling variability associated with the likelihood of a species tree given the observed data, or $P(\mathbf{n_{obs}}|\theta)$. In ST-ABC, it is *as if* the random data set that would have been simulated under Algorithm–ABC is always substituted by $\mathbf{n}_{\text{exp}}$.

Fan and Kubatko found that Algorithm–ST-ABC performs well in practical species tree estimation problems. Algorithm–ST-ABC can be converted into a proper ABC algorithm if Step 3 of Algorithm–ST-ABC is substituted by a step in which a data set of size N is randomly drawn from the probability distribution of gene tree topologies $P(\beta|\theta^*)$ and the data set generated is compared with the observed data $\mathbf{n_{obs}}$. It would then be interesting to examine how a proper ABC algorithm compares with Algorithm–ST-ABC in its ability to estimate species trees.

Acknowledgements

The author thanks Cuong Than and Ethan Jewett for helpful discussions on gene trees and species trees. Support was provided by NSF grant DBI-1146722.

References

Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. Genetics 162, 2025–2035.

Degnan, J., Salter, L.A., 2005. Gene tree distributions under the coalescent process. Evolution 59, 24–37.

Fan, H.H., Kubatko, L.S., 2011. Estimating species trees using approximate Bayesian computation. Molecular Phylogenetics and Evolution 59, 354–363.

Marin, J.M., Pudlo, P., Robert, C.P., Ryder, R. 2011. Approximate Bayesian Computational Methods. arXiv:1101.0955v2.

Tavaré, S., Balding, D.J., Griffiths, R.C., Donelly, P., 1997. Inferring coalescence times from DNA sequence data. Genetics 145, 505–518.

Erkan Ozge Buzbas *
Department of Biology, Stanford University,
371 Serra Mall, Stanford, CA 94305-5020, USA
* Tel.: +1 650 724 5122.
E-mail address: buzbas@stanford.edu

Available online 12 September 2012