TECHNICAL COMMENT

AVIAN GENOMICS

Comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree"

Liang Liu[1] and Scott V. Edwards[2]*

Mirarab *et al*. (Research Article, 12 December 2014, p. 1250463) introduced statistical binning to improve the signal in phylogenetic methods using the multispecies coalescent model. We show that all forms of binning—naïve, statistical, and weighted statistical—display poor performance and are statistically inconsistent in large regions of parameter space, unlike unbinned sequence data used with species tree methods.

Mirarab *et al.* introduced statistical binning as a method for improving the signal in species tree phylogenetic methods using the multispecies coalescent model and claimed that it can improve the accuracy of coalescent-based estimation of species trees (*1*). Statistical binning is a method for signal augmentation in multilocus species tree reconstruction, designed to reduce gene tree estimation error by estimating supergene trees from DNA sequences concatenated across the genes that do not conflict above an arbitrary bootstrap threshold. Mirarab *et al.* show a number of examples in which statistical binning appears to outperform unbinned species tree analysis as measured by the frequency of achieving accurate estimates of known phylogenies and species tree branch lengths. However, they explore a limited region of species tree parameter space that is favorable to binning analyses. We have recently shown (*2*) that naïve binning (NB), in which sequences are binned at random to create longer supergenes, without regard to their chromosomal location, exhibits poor performance in some regions of parameter space, because the method produces incorrect species trees with increasing certainty as the number of genes increases. Here, we show that statistical binning (SB), as well as its recent update, weighted statistical binning (WSB) (*3*), also exhibit inconsistent behavior, tending to distort the distribution of gene trees and converging

[1]Department of Statistics, University of Georgia, Athens, GA 30602, USA. [2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA.
***Corresponding author. E-mail: sedwards@fas.harvard.edu**

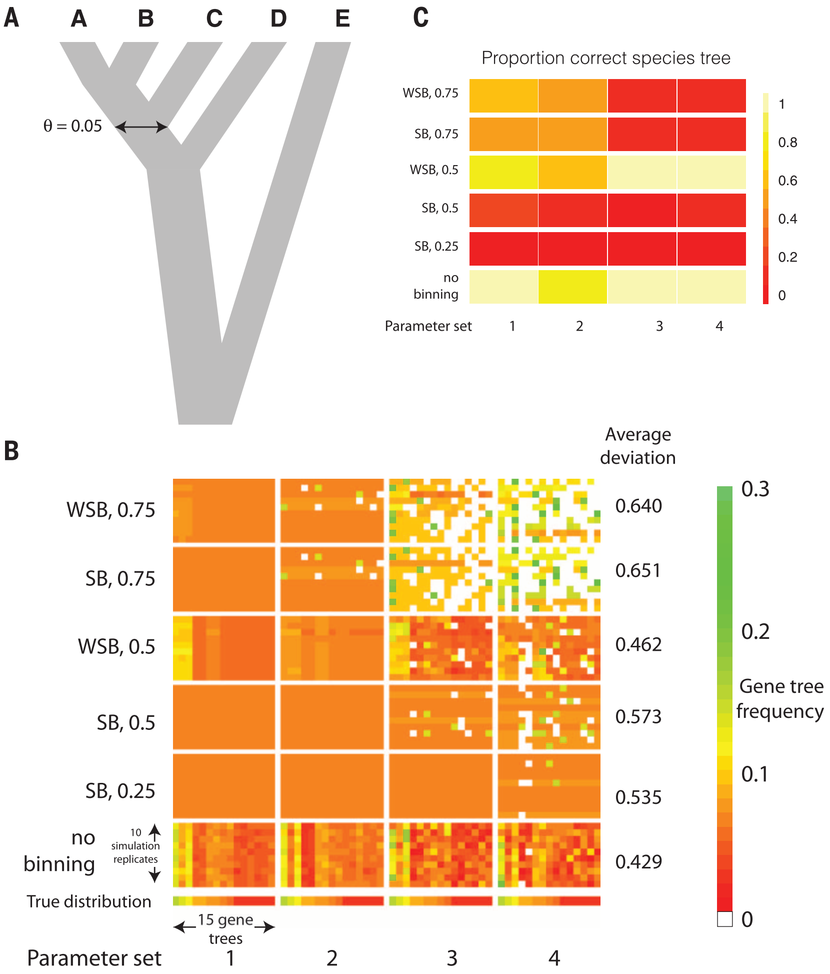

Fig. 1. Binning simulation. Gene trees were simulated from a five-taxon species tree and then used to simulate DNA sequences using Seq-Gen (*9*) with the general time-reversible (GTR) + gamma model. We adopted the same GTR + gamma parameters used in (*1*) to simulate sequence data. We considered the following situations: number of genes = 100 or 1000, sequence length = 100 or 1000 base pairs (bp), and bootstrap threshold = 0.25, 0.50, or 0.75. The maximum likelihood (ML) and bootstrap gene trees were built for each gene using RAxML with the correct model. The estimated gene trees with bootstrap percentages were used without binning, or as input for statistical and weighted binning algorithms with varying thresholds, and the sequences within each bin were concatenated as a supergene according to the binning algorithm. Supergene trees were built using RAxML (*10*) with the correct model. Each simulation was repeated 10 times, but all the trends in our results were upheld with 100 replicates for each simulation. (**A**) The species tree used in the simulation is ((((A:0.005, B:0.005):0.005,C:0.01):0.005,D:0.015):0.5,E:0.515) (branch lengths in substitutions per site), with the population size parameter θ = 0.05. Binning is expected to perform worse for species trees in the anomaly zone (*5*, *6*). (**B**) The true and estimated distributions of gene trees across simulations for six binning protocols, including no binning. Bootstrap thresholds are indicated. From left to right, four parameter sets, as follows: set 1: number of genes (n.g.) = 1000, sequence length (s.l.) = 1000 bp; set 2: n.g. = 1000, s.l. = 100 bp; set 3: n.g. = 100, s.l. = 1000 bp; set 4: n.g. = 100, s.l. = 100 bp. The 15 possible gene trees are represented along the *x* axis for each block of simulations. The *y* axis in each block represents 10 replicate simulations. Colors represent the values of probabilities, with white for gene trees not produced by the simulation. Flat distributions of gene trees are indicated by rows of the same color within blocks. The average deviation across the four parameter sets of the observed, reconstructed distribution of gene trees from the true distribution is indicated at left. For 100 replicates, including WSB0.25, these deviations are: no binning, 0.448; SB0.25, 0.543; WSB0.25, 0.449; SB0.50, 0.568; WSB0.50, 0.470; SB0.75, 0.678; WSB0.75, 0.667. These deviations were calculated as the sum of the absolute values of the differences between observed and true frequencies of gene trees. (**C**) A heat map for the proportion of the true species tree estimated among 10 replicates for each of four parameter sets, as in (B). Colors represent proportions. On the *y* axis are five different thresholds of statistical and weighted binning. Results for simulations with 100 replicates are similar, with the values for WSB0.25 varying from 1 (n.g. = 1000, s.l. = 100 bp) to 0.67 (n.g. = 100, s.l. = 100 bp).

to the wrong result under some parameter sets that otherwise fulfill the assumptions of the neutral multispecies coalescent model. Such convergence toward an incorrect result with increasing data set size is one prediction of an inconsistent method, yet species tree methods such as maximum pseudo-likelihood estimation of species trees (MP-EST) (*4*) have not exhibited this inconsistency under any parameter sets, including the anomaly zone, for analyses using unbinned loci.

SB yields a series of bins of roughly equal sizes, each of which includes a set of sequences consistent with a gene tree. Because the algorithm ensures that all supergene trees have frequency 1 in the binned data set, SB flattens the distribution of gene trees (*3*), thereby removing the coalescent signal maintained in individual gene trees and misleading downstream estimation of species trees for many parameter sets. The authors state that there is a high chance of binning genes with different histories (*1*), especially when the threshold is high and the bootstrap percentages on estimated gene trees are low. Such an outcome is likely when the internal branches in the species tree are short, a situation that generates short branches in gene trees. SB can produce highly biased distributions of supergene trees under some conditions (Fig. 1), and, just as full concatenation of alignments from genes with different histories can positively mislead species tree estimation (*5*, *6*), so can SB.

We used simulation to evaluate the performance of SB under a five-taxon (A to E) species tree that is close to but outside of the anomaly zone (Fig. 1A). As the root species E is fixed, there are 15 possible rooted gene trees. When θ = 0.05 [high levels of incomplete lineage sorting (ILS)] and the bootstrap threshold is low (0.25), there is a high probability (>0.8) that SB will create a perfectly flat distribution of gene trees (Fig. 1B). When the threshold is high (0.75), bootstrap percentages on most gene trees are less than the threshold in our simulation, and most gene trees will be randomly distributed to different bins by the binning algorithm, similar to NB (*7*) (Fig. 1B). Under all sampling scenarios in our simulation, the gene tree distribution produced by SB, when combined with MP-EST, resulted in higher rates of estimating an incorrect species tree compared with no binning (Fig. 1C). This behavior is predicted by our sketch of the inconsistency of species tree methods under SB (Fig. 2).

WSB, in which each bin is assigned a weight equal to its size, has been proposed as a fix for the tendency of SB to flatten the distribution of gene trees (*3*). However, when WSB is applied to estimated gene trees, it may not be able to correct the flat distribution produced by SB. If the bootstrap percentages on most estimated gene trees are less than the threshold, the binning algorithm will assign those gene trees at random to different bins, again resulting in flat distributions of gene trees under many parameter sets (Fig. 1B). Consistent with this tendency, we observe a much lower rate of correct species tree estimation across all simulations than without binning (Fig. 1C).

A

Theorem 1: If the sequence length l is finite, there exists a species tree S with branch lengths L and a sufficiently small mutation rate μ such that $P(\mathrm{BS}_{max} > t) < \varepsilon$ for every $\varepsilon > 0$ and threshold t ($\mathrm{ML}_{LB} < t < 1$).

Sketch of proof: Let x_i be the number of mutations occurring on branch i of an unrooted gene tree g. Under the substitution model, x_i has a Poisson distribution with mean = b_i (the length of branch i). The probability that no mutation occurs on branch i is e^{-b_i} . Because the number of mutations on different branches (x_i) are independent of one another, the probability that no mutation occurs on all branches is $\prod_{i=1}^{2N-3} e^{-b_i}$. This probability converges to 1 as $b_i \to 0$ for $i = 1,...,2N\text{-}3$. Because $b_i \to 0$ as the mutation rate $\mu \to 0$, the probability that no mutation occurs on all branches of the gene tree converges to 1 as $\mu \to 0$. When the alignments have no mutation, the EBPs on the ML gene tree equal the lower bounds LB of the EBPs. Because $t > \mathrm{ML}_{LB}$, it follows that $P(\mathrm{BS}_{max} < t) > 1 - \varepsilon$ for every $\varepsilon > 0$. Thus, $P(\mathrm{BS}_{max} > t) < \varepsilon$ for every $\varepsilon > 0$.

B

Theorem 2: If the sequence length l is finite, there exists a species tree S with branch length L and population size θ, for which all supergene trees generated from unweighted or weighted binning converge in probability to the wrong trees as the number of genes goes to infinity.

Sketch of proof: The number of bins identified by unweighted or weighted binning equals the number of statistically different gene trees. As the number of taxa N is fixed, the number of bins is bounded. By Theorem 1, there exists a species tree S with branch lengths L and a sufficiently small mutation rate μ such that there is $(1 - \varepsilon)$ proportion of gene trees, on which BS_{max} is less than the threshold t. These gene trees are randomly assigned to the bins identified by unweighted or weighted binning, and the probability distribution of the gene trees within each bin is identical to the probability distribution of gene trees under the multispecies coalescent model. Since ε can be arbitrarily small, we can find an $\varepsilon > 0$ such that the effect of the remaining (ε) genes in estimating the supergene trees is negligible. In addition, because the number of bins is bounded, the number of genes within a bin goes to infinity as the number of genes increases. It follows from Theorem 1 in [5] that supergene tree SG_i built from the concatenated alignments converges to the wrong tree in probability as the number of genes goes to infinity.

Fig. 2. Inconsistency of binning. Let *S* be an *N*-taxon species tree with topology *T* and branch lengths *L*. Let *SG* be the ML supergene trees estimated from the sequences concatenated across genes within bins. Let *t* be the threshold defined in the binning algorithm for identifying statistically identical gene trees. BS_{max} denotes the maximum expected bootstrap percentage (EBP) on a ML gene tree. Let M_{LB} be the maximum of the lower bounds *LB* of EBP. The lower bounds *LB* are achieved when no mutations are observed among all sequences. It is assumed that one allele is sampled from each species, so that the number of taxa in gene trees is equal to the number of taxa in the species tree. (**A**) Theorem 1 shows that the majority of the estimated gene trees generated from an anomalous species tree are poorly supported. Thus, given a threshold *t*, there exists an anomalous species tree such that the EBPs on the majority of gene trees are less than *t*. (**B**) Theorem 2 further shows that binning gene trees generated from an anomalous species tree positively misleads the estimation of supergene trees.

The empirical trees on which Mirarab *et al.* tested SB have many taxa, and the probability of generating two identical gene trees is very low, resulting in a true flat distribution of gene trees, leaving little opportunity for inconsistency of SB on species tree estimation. Mirarab *et al.* claim that binned trees are better estimated and more congruent with other analyses, but using concatenated trees as a benchmark is questionable. Nearly all of the species tree branches in unbinned analyses that Mirarab *et al.* claim are incorrect [figure 5 in (*1*)] differ nonsignificantly [as measured by bootstrap support (BS) less than 0.90] from binned analyses.

Outside of collecting more data, methods for signal augmentation in phylogenetics are extremely rare, with most methods instead focusing on improving model fit. We question the motivation behind SB: to improve the signal in gene trees and hence species trees. Rather, we suggest that the low signal often found in species trees is a real result that calls for more data collection—feasible even in analyses that purport to analyze whole genomes—or improved coalescent models, which binning is not. Binning (concatenation) might be used most profitably while taking genomic location into account, such as concatenating adjacent exons as frequently occurs in transcriptome data, which minimizes intralocus recombination, even though recombination is not a severe problem (*8*). Our demonstration that SB exhibits inconsistent behavior not observed in unbinned analyses and frequently distorts the distribution of estimated gene trees compels us to discourage its use. When support for a species tree is deemed too low, we suggest collecting more data and improving model fit rather than binning.

REFERENCES AND NOTES

1. S. Mirarab, M. S. Bayzid, B. Boussau, T. Warnow, *Science* **346**, 1250463 (2014).
2. L. Liu, Z. Xi, S. Wu, C. C. Davis, S. V. Edwards, *Ann. N.Y. Acad. Sci.* (2015).
3. M. S. Bayzid, S. Mirarab, B. Boussau, T. Warnow, *PLOS ONE* **10**, e0129183 (2015).
4. L. Liu, L. Yu, S. V. Edwards, *BMC Evol. Biol.* **10**, 302 (2010).
5. S. Roch, M. Steel, *Theor. Popul. Biol.* **100**, 56–62 (2015).
6. L. S. Kubatko, J. H. Degnan, *Syst. Biol.* **56**, 17–24 (2007).
7. M. S. Bayzid, T. Warnow, *Bioinformatics* **29**, 2277–2284 (2013).
8. H. C. Lanier, L. L. Knowles, *Syst. Biol.* **61**, 691–701 (2012).
9. A. Rambaut, N. C. Grassly, *Comput. Appl. Biosci.* **13**, 235–238 (1997).
10. A. Stamatakis, *Bioinformatics* **22**, 2688–2690 (2006).

ACKNOWLEDGMENTS

We thank T. Warnow and L. Kubatko for discussion and M. Steel, J. DaCosta, D. Mindell, and an anonymous reviewer for comments. The R scripts for running the simulations in Fig. 1 are available at https://github.com/lliu1871/R-code-for-simulation-on-binning.git. The complete distribution of gene trees for the simulations with 100 replicates can be found at http://odyssey.bioinformatics.uga.edu/~lliu/data/binning.zip. Author contributions: L.L. and S.V.E. designed the study, L.L. performed analysis, and L. L and S.V.E. wrote the paper.

20 January 2015; accepted 28 July 2015
10.1126/science.aaa7343

Comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree"
Liang Liu and Scott V. Edwards (October 8, 2015)
Science **350** (6257), 171. [doi: 10.1126/science.aaa7343]