Statistical binning leads to profound model violation due to gene tree error incurred by trying to avoid gene tree error

Richard H. Adams, Todd A. Castoe

Department of Biology, The University of Texas at Arlington, Arlington, TX 76019, USA

ARTICLE INFO

Keywords:
Phylogenetic inference
Phylogenomics
Species trees
Concatenation
Coalescent
Supergene

ABSTRACT

Fundamental to all phylogenomic studies is the notion that increasing the amount of data to entire genomes when possible will increase the accuracy of phylogenetic inference. Simply adding more data does not, however, guarantee phylogenomic inferences will be more accurate. Even genome-scale reconstructions of species histories can suffer the effects of both incomplete lineage sorting (ILS) and gene tree estimation error (GTEE). Weighted statistical binning was originally proposed as a technique to assist the avian phylogenomics project in solving the bird tree of life, which has long eluded resolution as a result of both ILS and GTEE. These so-called “statistical binning procedures” seek to overcome GTEE by concatenating loci into longer multi-locus “supergenes” that are used to reconstruct a species tree under the assumption that the supergene tree set is an accurate estimate of the true underlying gene tree distribution. Here we evaluate the performance of the method using the original avian phylogenomics dataset. Our results suggest that statistical binning constructs false supergenes that concatenate loci with different coalescent histories more often than not, and the e ects on the accuracy of species tree estimates, and it is not immediately clear whether one should prioritize either.

The avian Tree of Life is a prime example of an important vertebrate phylogeny that has long eluded resolution because of both ILS and GTEE (Jarvis et al., 2015; Mirarab et al., 2014). Researchers thus face a gauntlet of challenges when analyzing phylogenomic data: concatenate loci and suffer the consequences of ILS, or do not concatenate loci and suffer the consequences of GTEE. Both sources of conflict can have major debilitating effects on the accuracy of species tree estimates, and it is not immediately clear whether one should prioritize either.

In light of the challenges facing phylogenomic analyses, a new method (“weighted statistical binning”; referred to as “statistical binning” hereafter) was originally developed to enable the avian phylogenomics project in resolving the relationships of modern birds (Bayzid et al., 2015; Jarvis et al., 2015; Mirarab et al., 2014). The method has since been used to infer the evolutionary relationships of placental mammals (Tarver et al., 2016), teleost fishes (Malmstrøm et al., 2016), and many other major radiations (i.e., Blaimer et al., 2016; Branstetter et al.,

1. Introduction

Much of our understanding and practice of evolutionary biology relies on knowledge of the species-level relationships of organisms (i.e., species trees). Two major sources of phylogenetic conflict can pose serious challenges for species tree reconstruction: incomplete lineage sorting (ILS) and gene tree estimation error (GTEE). Standard phylogenetic analysis of concatenated loci, for example, will be statistically inconsistent in the presence of ILS and yield highly-supported but incorrect species trees (Edwards et al., 2007; Kubatko and Degnan, 2007). To address this, coalescent-based methods have been developed that are statistically consistent under ILS and will return the true species-level phylogeny with high confidence given sufficient information (Degnan and Rosenberg, 2009; Heled and Drummond, 2010; Knowles, 2009; Liu, 2008; Liu et al., 2010, 2015b). While ILS is an inherent property of the demographic processes of speciation and divergence, GTEE is a fundamentally different source of conflict that represents statistical sampling error and variation between the true tree and one estimated from a dataset of finite size and information content. Although modern phylogenomic datasets often consist of millions to billions of base pairs (bp), any one aligned locus is often limited to < 3 kbp of aligned orthologous sequence data, and thus individual gene trees may entail substantial error that can permeate to the level of species tree inference (Jarvis et al., 2014; Mirarab et al., 2014). Researchers thus face a gauntlet of challenges when analyzing phylogenomic data: concatenate loci and suffer the consequences of ILS, or do not concatenate loci and suffer the consequences of GTEE.

The avian Tree of Life is a prime example of an important vertebrate phylogeny that has long eluded resolution because of both ILS and GTEE (Jarvis et al., 2015; Mirarab et al., 2014; Prum et al., 2015). In light of the challenges facing phylogenomic analyses, a new method (“weighted statistical binning”; referred to as “statistical binning” hereafter) was originally developed to enable the avian phylogenomics project in resolving the relationships of modern birds (Bayzid et al., 2015; Jarvis et al., 2015; Mirarab et al., 2014). The method has since been used to infer the evolutionary relationships of placental mammals (Tarver et al., 2016), teleost fishes (Malmstrøm et al., 2016), and many other major radiations (i.e., Blaimer et al., 2016; Branstetter et al.,...
Supergenes represent profound phylogenetic model misspecification, because standard ML-analysis assumes that all sites within an alignment evolved under the same gene tree, and thus, only one tree will be estimated when there should be three (right example). Regardless of whether this ML topology is the blue, red, purple, or some other topology, the answer is the same: ML-analysis cannot be statistically consistent because it cannot estimate three unique trees. False supergene trees are likely to reflect an amalgamation of conflicting phylogenetic signal (here three distinct trees), such that the gene tree with the most support (i.e., highest number of informative sites) may have disproportionate influence (see Fig. 4). The relevant questions is thus whether statistical binning tends to infer true supergenes (left) or false supergenes (right), and although the method does not directly estimate a species tree, clearly supergene accuracy is likely to influence downstream species tree accuracy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Following publication of the avian phylogenomics project, substantial debate and contention has arisen over the use of statistical binning and similar methods (Bayzid et al., 2015; Liu and Edwards, 2015; Mirarab et al., 2015; Roch and Warnow, 2015). Authors have continued to argue both for and against these methods, and disagree over the statistical consistency (or lack of) of these approaches in the context of species tree estimation (Liu and Edwards, 2015; Mirarab et al., 2015; Roch and Warnow, 2015; Warnow, 2015). A follow-up study revealed that statistical binning distorted supergene tree distributions and likely biased species tree estimates (Liu and Edwards, 2015). Further studies corroborated this assertion: species trees reconstructed using supergenes obtained via statistical binning were likely to be highly inaccurate yet highly supported (Streicher et al., 2018). Subsequent response papers rejected the assertion that the method was statistically inconsistent, and instead argued for statistically consistency when the number of loci and the length of loci are both infinite (Bayzid et al., 2015; Mirarab et al., 2015). However, recent theoretical work has demonstrated the inconsistency of species tree methods that use supergenes inferred via statistical binning when the number of loci is unbounded but the length of each locus is bounded to a constant (Roch et al., 2018). These findings raise important questions about the nature of species tree inference under best-case scenarios (i.e., when the number and/or length of loci is infinite), and yet, we currently have relatively little understanding of the empirical performance of the statistical binning pipeline itself when both the number and length of loci are bounded.

When considering the properties of the method, it is imperative to acknowledge that the statistical binning pipeline itself only infers a set of supergene alignments, not a species tree. Statistical binning is therefore not a species tree estimation method per se, it is a supergene estimation method that uses gene tree estimates to infer topology congruency among loci. Distinguishing between species tree estimation and supergene estimation is critical, because both are fundamentally different statistical problems: species tree estimation seeks a single species-level topology and set of parameters (i.e., divergence times, effective population sizes), while supergene inference involves deciding whether individual loci share the same gene tree or not. In this sense, statistical binning represents the first “cog in the wheel” of the phylogenomic analysis pipeline, which is followed by supergene tree estimation using standard phylogenetic techniques, such as maximum likelihood (ML) analysis, and species tree estimation using coalescent-based summary methods. Understanding whether the statistical binning pipeline provides reliable supergene alignments is therefore paramount to assessing the performance of the method. At the end of a statistical binning analysis, ML-analysis of each supergene is conducted under the assumptions of the standard phylogenetic model. While different supergenes can have different topologies, ML-analysis of the individual supergene alignments assumes that each gene placed within a supergene shares the same coalescent history. Under these conditions (i.e., a “true supergene” containing only congruent genes), standard ML-analysis – which assumes all sites share the same tree (Felsenstein, 1981) – will converge with increasing probability to the single, true gene tree as the length of each congruent locus in the supergene increase (Fig. 1,
left).

In contrast, if a supergene incorrectly concatenates genes from multiple distinct topologies, standard ML-analysis of this “false supergene” will not converge to the true gene tree set (i.e., one tree for each distinct gene) as the length of each discordant gene increases, because it is restricted to inferring a single best-fit tree. In the right example shown in Fig. 1, a false supergene has been constructed by concatenating three genes with conflicting genealogies (red, purple, green). Even if the length of each of the three genes is infinite, standard ML-analysis will infer only a single supergene tree – instead of the “true” gene tree set comprised of three distinct topologies. Violation of this fundamental assumption of the phylogenetic model (i.e., all sites share the same tree) is of major consequence because it is the underlying cause of the failure of ML-analysis in the presence of ILS (Mendes and Hahn, 2017), and can also cause other modeling pathologies and biases, such as SPILS (“substitutions produced by ILS”; Mendes and Hahn, 2016). False supergene trees inferred using standard ML-analysis are likely to reflect an amalgamation of phylogenetic signal, such that the gene tree with the most support (i.e., highest number of informative sites) may have disproportionate influence. The overall supergene tree distribution will also likely be distorted as distinct gene trees are effectively “hidden” within false supergenes and may be poorly represented or absent in the set of supergene trees. False supergenes therefore represent profound phylogenetic model misspecification, and the hope is that methods such as statistical binning are able to avoid such sources of systematic bias by inferring accurate supergenes (i.e., Fig. 1 left vs. right).

A critical question therefore remains: how well does statistical binning infer topological congruency (or lack of) from gene tree estimates when attempting to construct true supergenes? Here we evaluate the performance of the method at this core function, and while previous studies have primarily focused on the theoretical properties of the method for species tree inference when aspects of the data are informative, we are interested in characterizing the degree to which supergene trees reflect the topologies of their constituent genes. A critical concern of concatenating genes into a single supergene is that, if genes do not share the same tree, the gene with the most informative sites will dominate and overwhelm gene tree signals from shorter or less informative genes. In such cases, the supergene tree may only reflect the relationships supported by the dominant genealogy, while conflicting topologies of shorter loci will be effectively “hidden” and likely absent from the supergene tree structure. To examine whether supergenes tend to be biased towards their longest constituent gene (and therefore capable of masking hidden gene trees from shorter gene constituents), we computed normalized Robinson-Foulds distances between each of the 14,446 gene trees and their associated supergene topology using the R package phangorn (Schliep, 2011).

2. Methods

2.1. Avian phylogenomic data

We downloaded the 14,446 alignments (8,251 exons, 2,516 introns, and 3,679 UCEs), the inferred supergene assignments for the 14,446 loci (i.e., assignment of each locus to a respective supergene), and the 2,021 ML supergene trees inferred via statistical binning for the avian phylogenomic analyses (Jarvis et al., 2015, 2014). For our simulation-based assessments of statistical binning accuracy, we downloaded the simulated gene tree sets and their associated inferred supergene assignments that were used in the original avian phylogenomic studies and were based on the estimated avian species tree (Jarvis et al., 2014; Mirarab et al., 2014).

2.2. Likelihood-based tests of statistical binning accuracy

We evaluated the accuracy of each inferred supergene using likelihood ratio tests (LRTs) implemented in ConcatPillar (Leigh et al., 2008) and SH-tests (Shimodaira and Hasegawa, 1999) implemented in RAxML v8.0.0 (Stamatakis, 2014). First, we used ConcatPillar to conduct LRTs to test whether a model consisting of a single topology or a model of multiple distinct topologies was better supported by the sequence data of each supergene based on the difference in log-likelihood scores between models (Fig. 2, top box). This approach effectively tests how many distinct topologies are supported by the data of each supergene and corrects for multiple comparisons throughout the process. If only a single topology best fits the data, this provides evidence that the supergene is likely to be accurate (i.e., Fig. 1, left). Conversely, if the data support multiple topologies, then the supergene likely violates the phylogenetic model because it exhibits evidence of incorrectly concatenated loci originating from distinct topologies (i.e., Fig. 1, right).

We used SH-tests in a similar fashion to test whether the difference in log-likelihood scores between the ML topology of each individual gene placed within a supergene and the overall ML supergene tree was statistically significant (Fig. 2, lower box). In other words, for each gene placed within an inferred supergene, we used SH-tests to compare the likelihood of the individual gene-specific ML topology with the overall supergene ML topology (Fig. 2, colored vs. gray trees in lower box). If the individual gene-specific ML tree was a statistically significant better fit than the supergene topology (i.e., P < 0.05), then that supergene was likely falsely constructed by statistical binning (i.e., concatenated loci with different phylogenetic histories, i.e., Fig. 1 right). The number of genes that reject the overall supergene tree in favor of a locus-specific tree provide an indication of the number of discordant genealogies present within a supergene alignment. SH-tests were conducted in RAxML 8.0.0 (Stamatakis, 2014) using the default GTR + I + Γ nucleotide substitution model independently for each locus.
false supergene occurs when multiple trees are supported by the data (i.e., Fig. 1 right). For both methods, we quantified the number and fraction of true supergenes (blue bar in right histogram) and false supergenes that incorrectly concatenate multiple trees (2-8 in this case, black bars and red area). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.3. Simulation-based assessment of statistical binning accuracy

We also evaluated the accuracy of statistical binning on the simulated gene tree sets provided in the original study (Jarvis et al., 2015; Mirarab et al., 2014), by testing whether supergenes inferred via the method included only simulated genes that share a common gene tree. For each inferred supergene, we computed pairwise Robinson-Foulds distances (Robinson and Foulds, 1979) between each simulated gene tree that statistical binning inferred to share a single supergene tree; all of the individual gene trees should be identical if statistical binning provided a correct supergene. An RF-distance of 0 between two trees means that the topologies are identical and an RF-distance > 0 means the topologies are different. If all gene trees placed within a supergene have an RF-distance of 0, then the supergene was accurately inferred (i.e., Fig. 1, left). If there is at least one RF-distance that is greater than 0, the supergene was inaccurate because it incorrectly concatenated loci that evolved along distinct, conflicting gene trees (i.e., Fig. 1, right). We computed unrooted RF-distances using the “multiRF” function provided in the phytools (Revell, 2012) package in R, and used these values to compute the mean RF-distance among gene trees across all inferred supergenes in each replicate simulation analysis (rightmost column of Supplementary Table 1). For reference, these supergenes were inferred in the original study using a bootstrap threshold of 75% (Jarvis et al., 2014).

2.4. Quantifying the impacts of statistical binning on gene tree distributions and species tree support

Considering evidence for spurious supergenes, we explored the impacts of statistical binning on both gene tree distributions and species tree support. To visualize differences in the underlying topological distributions due to statistical binning, we generated Densitree (Bouckaert, 2010) plots and summary consensus trees using TreeAnnotator (Rambaut and Drummond, 2016) of the unbinned gene tree and binned supergene tree distributions. We also quantified shifts in species tree support by measuring the difference in multispecies coalescent likelihoods of the unbinned gene trees and binned supergene trees using (1) the “unbinned” species tree (UST) estimated using the unbinned gene trees and (2) the “binned” species tree that was estimated using the binned supergene trees. For each of the 14,667 unbinned gene trees for the avian dataset, we measured the difference between the multispecies coalescent likelihood given the “binned” species tree and separately, the likelihood of the gene tree given the “unbinned” species tree: $\Delta \text{GeneTreeLnL} = \text{LnL(GeneTree|Binned Species Tree)} - \text{LnL(GeneTree|Unbinned Species Tree)}$. We also conducted this same analysis for the 2,021 supergenes inferred via statistical binning: $\Delta \text{SupergeneTreeLnL} = \text{LnL(SupergeneTree|Binned Species Tree)} - \text{LnL(SupergeneTree|Unbinned Species Tree)}$. To visualize the impacts of
Models-based evaluation of the performance of statistical binning on the avian phylogenomic data indicate that it does not provide reliable supergenes because it is highly prone to constructing “false supergenes” from loci with different coalescent histories – leading to profound and widespread phylogenetic model violation (Fig. 4). Both likelihood-based methods we employed indicate widespread error: 96.0% (1,940/2,021) of supergenes concatenated multiple, conflicting topologies using the LRTs and SH-tests, respectively (Fig. 3a and b). Our results therefore indicate that the vast majority (> 92%) of inferred supergenes represent false positives. We further evaluated the accuracy of statistical binning on the simulated datasets provided in the original avian study (Jarvis et al., 2015). Surprisingly, we found that 100% of multilocus supergenes (i.e., supergenes with at least 2 loci) across all simulation models and replicates were falsely constructed by statistical binning on species tree support, we compared the distributions of the 14,667ΔGeneTreeLnls and the 2,021 ΔSupergeneTreeLnls.

3. Results and discussion

3.1. Evidence of widespread model misspecification due to statistical binning

Although we have primarily presented the problem of “false supergenes” as a dichotomous phenomenon (i.e., either all genes are congruent or not), their impacts on species tree estimation may be more complex depending on the particular evolutionary parameters (i.e., species tree shape, divergence times, population sizes), and/or experimental conditions (i.e., number and length of loci). For example, “false supergenes” comprised of only two distinct trees may be less problematic than if they contain loci from three distinct trees. It also seems possible that particular branches and subclades may be more or less accurately estimated than others. This could occur, for example, if most genes within a false supergene agree on the placement of a particular
Fig. 5. The impacts of statistical binning on gene tree distributions and species tree support. Densitree plots showing the gene tree topology distribution for (a) the individual gene trees ("unbinned") and (b) the supergene trees. Plot of consensus trees with bipartition frequencies estimated using the individual, unbinned gene trees (c) and (d) the supergene trees constructed with statistical binning (d). Node circles are labeled and colored by the bipartition frequencies observed in their respective gene tree distributions. Histograms showing the distributions of multispecies coalescent likelihoods for the unbinned gene trees (ΔGeneTreeLnLs; e) and binned supergene trees (ΔSupergeneTreeLnLs f).
clade. Deeper nodes may be more accurately estimated than more recent species splits – perhaps because individual genes may exhibit little conflict in the placement of more ancient lineages (i.e., most ancient lineages are completely sorted). Nonetheless, ML-analysis of false supergenes will be a forced comprise of the conflicting signal exhibited across incongruent loci and thus, will likely suffer large-scale systematic error in topology, branch length estimates, and other parameters.

4. Conclusions

Perhaps surprisingly, genome-scale datasets do not yet equate to straightforward and robust resolution of phylogeny. Instead, both biology and methodology continue to pose serious challenges for phylogenomic analyses. There is certainly logical merit in approaches that straightforward and robust resolution of phylogeny. Instead, both

References


Jarvis, E.D., Mirarab, S., Boussau, B., Warnow, T., 2015. Weighted statistical binning: a forced comprise of the con-

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2019.02.012.


Molecular Phylogenetics and Evolution 134 (2019) 164–171

