PUBLISHED: 13 JANUARY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0056

PHYLOGENOMICS

Constrained gene tree inference

Data from many genes across the genome are now being routinely used in the hope of reconstructing challenging parts of the tree of life, and a new method provides a practical way of resolving the phylogenetic trees suggested by different genes.

Siavash Mirarab

A major challenge of using genome-scale data for constructing phylogenies is accounting for differences between gene trees and the species tree[1]. Several approaches have been proposed[2] that model gene trees as a function of the species tree and gene alignments as a function of gene trees (Fig. 1). Because of its computational efficiency, the most widely used approach first estimates gene trees separately from gene alignments and then summarizes them using a summary method to obtain the species tree. While scalable, a major shortcoming of this 'summary' approach is that short uninformative genes can lead to noisy estimates of gene trees, which then result in poor estimates of the species tree[3]. Consequently, the best practical method of analysing genomic data remains a topic of heated debate[4,5].

Writing in *Nature Ecology & Evolution*, Arcila *et al.*[6] address gene tree error using a practical approach, called gene genealogy interrogation (GGI), and use it to resolve a long-standing question in freshwater fish (Otophysi) phylogeny. GGI is based on a simple but elegant idea. The analyst provides a small set of hypotheses regarding relationships between major clades of undisputed monophyly; for Otophysi, the authors consider all 15 possible relationships among 5 undisputed otophysan groups. To find the hypothesis best supported by each gene, GGI enforces the monophyly of the major clades and performs a constrained maximum likelihood (ML) search for each hypothesis to resolve the rest of the tree. The resulting trees are ranked based on their likelihood scores and the statistical significance of support for all hypotheses is assessed via the established approximately unbiased test[7]. If a hypothesis finds overwhelming support among genes, GGI selects it as a resolution of the species tree. Alternatively, the set of top ranked constrained trees can be used as input to a summary method, such as ASTRAL[8].

Remarkably, for Otophysi, nearly half the genes preferred one hypothesis, twice as many as the second hypothesis. Not surprisingly, applying a summary method to GGI gene trees consistently recovered the same hypothesis. Interestingly, the selected hypothesis was consistent with traditional morphological analyses, but not with concatenation and some of the species tree analyses performed in this study. The authors argue convincingly in favour of the GGI resolution of Otophysi, but there are broader implications of the GGI method.

Enforcing monophyly of certain groups in gene trees raises many questions. How should we judge whether undisputed clades are really as certain as a user of the GGI approach would claim? What if there are too many major undisputed clades (for example, 10) and not all hypotheses (>2 million) could be easily tested? Beyond these practicalities, what makes constrained searches preferable to the traditional unconstrained search?

One reasonable claim is that when individual genes are phylogenetically uninformative, many alternative trees may have statistically indistinguishable support, and those that conform to our prior knowledge of undisputed clades should be preferred. However, if unconstrained trees have significantly better likelihood than constrained trees, shouldn't we prefer the unconstrained results even when they conflict with our prior beliefs? The answer, arguably, is that we should, and this observation leads to a potentially better variation of the GGI method that the authors only briefly consider. For each gene, we can perform both unconstrained and constrained ML searches and choose the top ranked constrained tree only if it is not

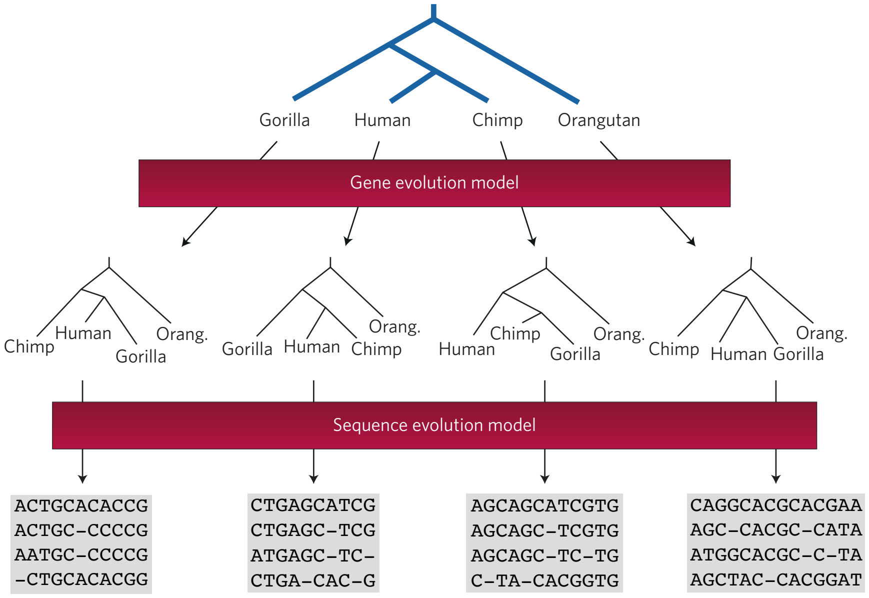

Figure 1 | Species tree estimation despite gene tree discordance. Schematics of a generative model adopted by the community. The species tree (top) generates gene trees (middle) using a model of gene tree evolution (for example, multi-species coalescence[9] or birth–death models of gene birth), and then each gene separately generates sequence data (bottom) using models of sequence evolution. Inference of the species tree starts from the data and follows the opposite directions of the generative model, either in two stages (summary methods), all at once (co-estimation), or skipping the middle layer (site-based methods).

significantly worse than the unconstrained tree. The resulting gene trees, a mix of unconstrained and constrained, can then be used as input to a summary method. The authors tried this approach and obtained the same results as the unconstrained-only GGI. I conjecture that as the number of genes and the number of sites in each gene increases to infinity, GGI based on a mix of constrained and unconstrained searches can be mathematically proved to be statistically consistent under the multi-species coalescence model, while the unconstrained-only GGI can be proved inconsistent and perhaps positively misleading.

The authors find that for a majority of genes in their Otophysi dataset, unconstrained ML trees are significantly better than the best constrained tree. Why would a gene tree strongly conflict with undisputed clades? One answer is that even if a clade is easy to recover in the species tree, incomplete lineage sorting (ILS) can still break its monophyly in many gene trees. The authors acknowledge this possibility and point out that for Otophysi, constrained branches are at least 20 million years long, making it unlikely that ILS would have a confounding effect. It is less clear whether similar arguments can be made for other datasets reanalysed in their paper. In general, applying constraints to gene trees is safe only when we can argue that ILS is very unlikely to break the monophyly of a branch in gene trees, and not whenever the monophyly of a clade is undisputed.

Besides ILS, the authors point out other possible explanations for strong conflict between estimated gene trees and undisputed clades, such as model specification and lateral transfer, but wisely leave the question to future research. Another question is whether GGI or any of the existing approaches remain valid when one of these confounding factors (and not noise) causes the violation of undisputed clades. For example, if the gene is laterally transferred or is a paralogue, what happens when we constrain its position in the gene tree? When the model is misspecified, should we trust constrained searches more than unconstrained ones?

The answers to these and similar questions are not clear. Nevertheless, GGI, especially the variant that mixes constrained and unconstrained gene trees, can provide an attractive practical alternative for analysing recalcitrant parts of the tree of life using phylogenomic data. ❐

Siavash Mirarab is at Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive, MC 0407 La Jolla, California 92093-0407, USA.
e-mail: smirarabbaygi@eng.ucsd.edu

References

1. Maddison, W. P. *Syst. Biol.* **46,** 523–536 (1997).
2. Szöllõsi, G. J., Tannier, E., Daubin, V. & Boussau, B. *Syst. Biol.* **64,** e42–e62 (2014).
3. Mirarab, S., Bayzid, S. M., Boussau, B. & Warnow, T. *Science* **346,** 1250463 (2014).
4. Springer, M. S. & Gatesy, J. *Mol. Phylogenet. Evol.* **94,** 1–33 (2016).
5. Edwards, S. V. *et al. Mol. Phylogenet. Evol.* **94,** 447–462 (2016).
6. Arcila, D. *et al. Nat. Ecol. Evol.* **1,** 0020 (2017).
7. Shimodaira, H. *Syst. Biol.* **51,** 492–508 (2002).
8. Mirarab, S. & Warnow, T. *Bioinformatics* **31,** i44–i52 (2015).
9. Pamilo, P. & Nei, M. *Mol. Biol. Evol.* **5,** 568–583 (1988).