New methods to calculate concordance factors for phylogenomic datasets

Bui Quang Minh[1,*], Matthew W. Hahn[2,3], and Robert Lanfear[1]

[1] Department of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, ACT, Australia

[2] Department of Biology, Indiana University, Bloomington, Indiana, USA

[3] Department of Computer Science, Indiana University, Bloomington, Indiana, USA

* To whom correspondence should be addressed.

Abstract

Summary: We introduce and implement two measures for quantifying genealogical concordance in phylogenomic datasets: the gene concordance factor (gCF) and the site concordance factor (sCF). For every branch of a reference tree, gCF is defined as the percentage of "decisive" gene trees containing that branch. This measure is already in wide usage, but here we introduce a package that calculates it while accounting for variable taxon coverage among gene trees. sCF is a new measure defined as the percentage of decisive sites supporting a branch in the reference tree. gCF and sCF complement classical measures of branch support in phylogenetics by providing a full description of underlying disagreement among loci and sites.

Availability and Implementation: An easy to use implementation and tutorial is freely available in the IQ-TREE software (http://www.iqtree.org).

Supplementary information:

Data are available at https://doi.org/10.5281/zenodo.1949290

Issue Section: Phylogenetics

Contact: m.bui@anu.edu.au

1. Introduction

Measures of branch support such as the bootstrap (Felsenstein, 1985; Minh et al., 2013) and Bayesian posterior probability (Ronquist and Huelsenbeck, 2003) are important for making robust inferences from phylogenetic trees. However, while measures such as these provide useful information on the statistical support for a given branch, neither captures the variation present in the underlying data (Kumar et al., 2012).

A complementary approach involves calculating the fraction of loci supporting a particular branch, thus capturing underlying disagreement in the data. Various approaches have been suggested, including the gene concordance factor (gCF) (Ane et al., 2007; Gadagkar et al., 2005) and internode certainty (Salichos and Rokas, 2013), with the former being the most widely used. The gCF describes for each branch in a reference tree the proportion of inferred single-locus trees that contain that branch. While intuitive, the gCF suffers from three limitations: there are few software implementations (Ane et al., 2007); there are unresolved issues with incomplete taxon sampling of gene trees; and low gCF values are hard to interpret because they may result from strongly supported discordance or weak phylogenetic signal.

Here, we resolve these issues by implementing the calculation of gCF while accounting for unequal taxon sampling in the popular IQ-TREE package (Nguyen et al., 2015), and by introducing the site concordance factor (sCF), which estimates the concordance for individual sites.

2. Gene Concordance Factor (gCF)

For a given reference tree, T (e.g. an estimate of the species tree), and a set of input trees, $S = \{T_1, \ldots, T_n\}$ (e.g. individual gene trees), we calculate the gCF for every internal

branch t in T. T_i must contain a subset of the taxa in T, but need not share the same taxon sets.

Each branch t is associated with four clades of T representing four taxon subsets, $A, B, C,$ and D, such that the bipartition $\{A \cup B\}|\{C \cup D\}$ corresponds to the unrooted topology of t. Let $A_i, B_i, C_i,$ and D_i be the set of taxa in $A, B, C,$ and D that are also present in T_i. We say T_i is *decisive* for t if A_i, B_i, C_i, D_i are non-empty (Sanderson et al., 2011), and we say T_i is *concordant* with t if T_i is *decisive* for t and the bipartition $\{A_i \cup B_i\}|\{C_i \cup D_i\}$ corresponds to a branch in T_i. The gene concordance factor for t is now defined as:

$$\mathrm{gCF}(t) = \frac{|\{i: T_i \text{ is concordant with t}\}|}{|\{i: T_i \text{ is decisive for } t\}|}$$

In other words, gCF is the proportion of input trees decisive for t that are concordant with t.

3. Site Concordance Factor (sCF)

To calculate the sCF for branch *t*, we randomly sample m quartets of taxa $q = \{a, b, c, d\}$, where $a, b, c,$ and d are in $A, B, C,$ and D, respectively. For each quartet q, we examine the sub-alignment of taxa a, b, c, d. For every site j in this alignment, we call j *decisive* for t if the characters $a_j, b_j, c_j,$ and d_j are all present and j is parsimony informative. Decisive sites can be concordant or discordant with *t.* We say that site j is *concordant* with t if $a_j = b_j \neq c_j = d_j$ (i.e., j supports the bipartition $\{a, b\}|\{c, d\}$). The concordance factor for q is defined as:

$$\mathrm{CF}_q(t) = \frac{|\{j: j \text{ is concordant with } t\}|}{|\{j: j \text{ is decisive for } t\}|}$$

Since many such quartets exist around branch t (when sampling individual tips from within *A*, *B*, *C*, and *D*), we define $\mathrm{sCF}(t)$ as the mean $\mathrm{CF}_q(t)$ over *m* random quartets:

$$\mathrm{sCF}(t) = \frac{1}{m}\left(\sum_{q} \mathrm{CF}_q(t)\right)$$

Thus, the sCF is the average proportion of sites decisive for *t* that are concordant with *t*. In effect, the sCF is a measure of concordance for sites that is directly comparable to the measure of concordance for single-locus trees provided by the gCF. sCF differs from related approaches (Pease et al., 2018; Strimmer and von Haeseler, 1997) because it describes variation at the level of the individual site, rather than the level of the whole alignment.

4. Implementation in IQ-TREE

We provide two new options, --gcf and --scf, in IQ-TREE (Nguyen et al., 2015) to compute gCF and sCF respectively. Both options can be combined in a single run, which will annotate each branch with the two concordance factors. While sCF can be calculated on a concatenated alignment, gCF requires individual gene trees. We have therefore implemented a convenient option, -S, to specify a partition file or a directory of single-locus alignments in which IQ-TREE will infer separate trees for each partition or alignment. A tutorial for how to use these options is provided at http://www.iqtree.org/doc/Concordance-Factor.

5. Application to a lizard dataset

To demonstrate the use of gCF and sCF, we analysed a dataset containing 3,220 ultra-conserved elements (UCEs) from lizards, totalling 1,301,107 bp for 43 species (Rodriguez et al., 2018). We reconstructed a concatenated maximum likelihood tree, 3,220 UCE trees, and the gCF and sCF values with IQ-TREE. Although all but one bootstrap value in the tree are 100%, the smallest gCF and sCF values are 4.5% and 33.2% (median: 21.9% and 57%),

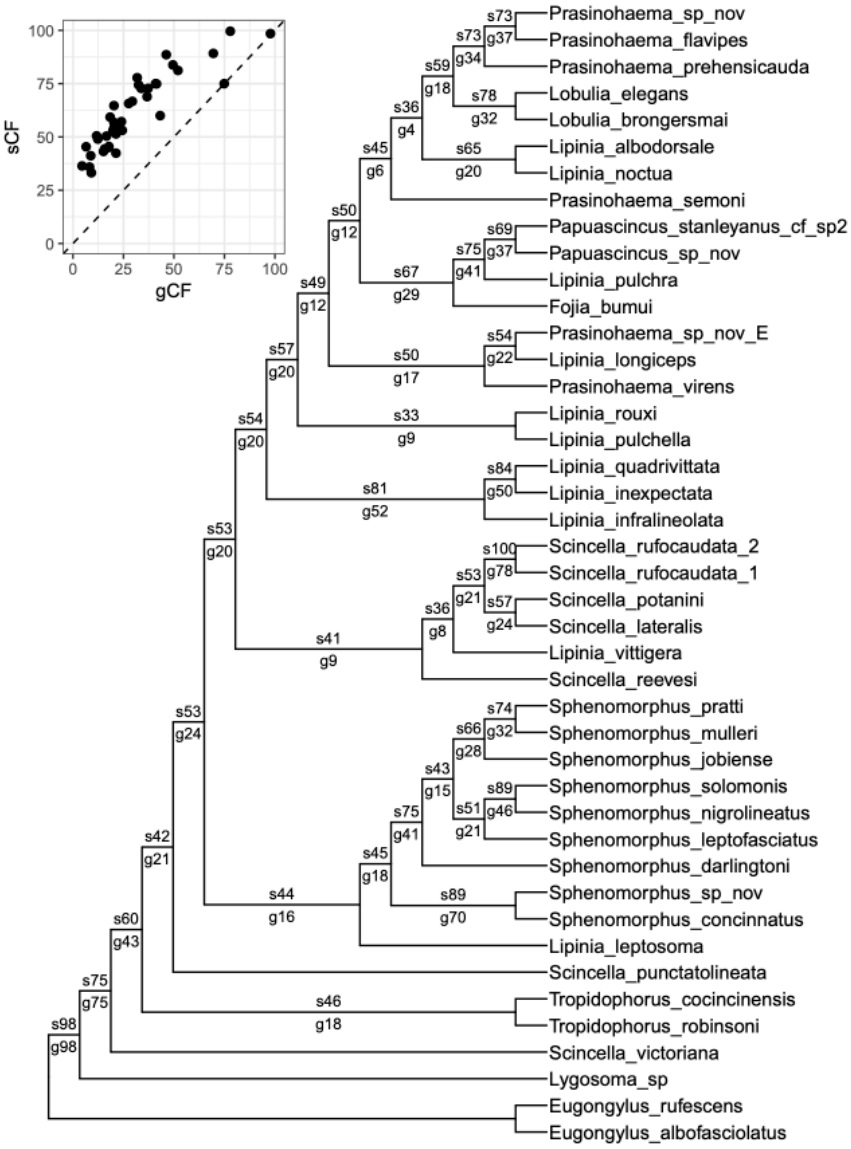

Figure 1. Tree for the lizard dataset (Rodriguez et al., 2018). Numbers on each branch show the site concordance factor (sCF) above the branch (e.g. s73) and the gene concordance factor (gCF) below the branch (e.g. g37). Bootstrap values are 100% on every branch, except for the branch leading to *Lipinia rouxi* and *L. pulchella*, which has a bootstrap value of 62%. The inset shows a scatter plot of gCF values against sCF values for all branches, revealing the large range of gCF and sCF values as well as the fact that for this dataset sCF values are always at least as large as gCF values.

respectively (Fig. 1). As expected, sCF values are always at least as large as gCF values (Fig. 1; top-left panel), likely due to the low phylogenetic signal in individual UCE alignments.

Acknowledgements

We thank Zachary Rodriguez for providing the lizard dataset.

Funding

BQM was supported by the Australian National University Futures grant.

Conflict of Interest: none declared.

References

Ane, C.*, et al.* (2007) Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.*, 24, 412-426.

Felsenstein, J. (1985) Confidence limits on phylogenies - an approach using the bootstrap. *Evolution*, 39, 783-791.

Gadagkar, S.R., Rosenberg, M.S. and Kumar, S. (2005) Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *J Exp Zool Part B*, 304b, 64-74.

Kumar, S.*, et al.* (2012) Statistics and truth in phylogenomics. *Mol. Biol. Evol.*, 29, 457-472.

Minh, B.Q., Nguyen, M.A.T. and von Haeseler, A. (2013) Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.*, 30, 1188-1195.

Nguyen, L.T.*, et al.* (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32, 268-274.

Pease, J.B.*, et al.* (2018) Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *Am. J. Bot.*, 105, 385-403.

Rodriguez, Z.B., Perkins, S.L. and Austin, C.C. (2018) Multiple origins of green blood in New Guinea lizards. *Sci Adv*, 4.

Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19, 1572-1574.

Salichos, L. and Rokas, A. (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497, 327-331.

Sanderson, M.J., McMahon, M.M. and Steel, M. (2011) Terraces in phylogenetic tree space. *Science*, 333, 448-450.

Strimmer, K. and von Haeseler, A. (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci. U. S. A.*, 94, 6815-6819.