Running head: COMPARING PHYLOGENETIC TREES

Quantifying and visualising divergence between pairs of phylogenetic trees:
implications for phylogenetic reconstruction

Martin R. Smith

Department of Earth Sciences, Lower Mount Joy, Durham University, Durham, DH1 3LE, UK;
+44 191 334 2320; martin.smith@durham.ac.uk

ABSTRACT

Tree topologies are a primary output of phylogenetic analysis. It is important to understand how such outputs are influenced by the choice of phylogenetic method, and the nature and quality of input data. This often entails the measurement of how closely an output corresponds to an ideal tree topology. The symmetric difference (partition) metric is the only widely-used measure that is defined for trees that contain polytomies, but expresses undesirable behaviour in certain situations.

Here I propose a divergence metric for unweighted tree topologies based on quartet statements, which measures the topological information common to two trees. This metric reflects the contributions of both accuracy and precision to tree quality; these components can be decomposed through the use of ternary diagrams. This approach is implemented in a new package for the R statistical environment, and applied to a recent controversy, where it provides

a new perspective on the relative merits of Bayesian and parsimony approaches to morphological phylogenetics.

Keywords.—phylogenetic inference; parsimony analysis; equal weights; implied weights; simulation analysis; information content

INTRODUCTION

Because phylogenetic trees are not vectors, there is no natural scale against which they can be compared (Penny & Hendy, 1985). Whilst it is possible to measure distances between edge-weighted trees in multi-dimensional space (Billera, Holmes, & Vogtmann, 2001; Owen & Provan, 2011), not all analyses generate meaningful edge lengths, meaning that their output can be difficult to compare.

This is unfortunate, as tree comparison underpins many fundamental considerations in phylogenetic analysis. Tree similarity has been used to determine the effect of missing data on phylogenetic analysis (Guillerme & Cooper, 2016), the impact of fossilization on phylogenetic results (Pattinson, Thompson, Piotrowski, & Asher, 2015; Sansom & Wills, 2013), the reliability of morphological data from different anatomical regions (Clarke & Middleton, 2008; Sansom, Wills, & Williams, 2017), and the most suitable method for morphological analysis (Congreve & Lamsdell, 2016; O'Reilly et al., 2016).

All existing measures of tree similarity have limitations.

The SPR distance measures the number of subtree pruning and regrafting rearrangements necessary to transform one of a pair of trees into the other (Penny & Hendy, 1985). Its limited range of integer values do not allow it to distinguish similar trees (one tip moved one edge) from

more dissimilar trees (a large clade moved a long distance), and it is only defined for pairs of bifurcating trees.

The path difference (Steel & Penny, 1993) counts differences in the number of edges that must be traversed to travel between each of the pairs of tips in a tree. The square of the difference in path length between the two trees is calculated for each pair of tips, and the square root of the sum of these quantities gives the path difference. The metric has high resolution, and behaves as expected in all situations, though the interpretation of its value is not straightforward, and it is only defined for pairs of bifurcating trees.

The symmetric difference measure (Penny & Hendy, 1985), a refinement of the Robinson-Foulds partition metric (Robinson & Foulds, 1981), is not restricted to bifurcating trees. It counts the proportion of edges (i.e. bipartitions or splits) present in each tree that are absent in the other. The measure has limited resolution, and can reach its maximum value when trees differ only in the position of a single tip (Steel & Penny, 1993). Worse, moving a single tip to a particular location can generate a higher distance metric than moving both that tip and its immediate neighbour to the same location. These counterintuitive properties mean that the metric offers a biased measure of tree similarity.

Finally, the quartet metric counts the number of four-taxon trees that are present as subtrees of each tree under consideration (Estabrook, McMorris, & Meacham, 1985; Steel & Penny, 1993). This method has a high resolution, allowing it to distinguish between relatively similar differences; and it does not exhibit the anomalous behaviour of the symmetric difference measure.

The method can be adapted in various ways to address pairs of trees that contain polytomies. If the trees each contain n tips, each of the $Q = {}_nC_4$ combinations of four taxa (i.e.

quartets) will satisfy one of five conditions (Estabrook et al., 1985): the quartet may be resolved in the same way (s) or a different way (d) on each tree; resolved in tree 1 only (r_1); resolved in tree 2 only (r_2); or unresolved in both trees (u). The proportions of quartets satisfying each of these conditions can be combined in various ways in order to capture different aspects of how trees agree, disagree, or differ in resolution (Estabrook et al., 1985).

One approach to combining these variables is to emphasize accuracy, measured by the number of incorrect quartets (i.e. “type II error over type I error” in Congreve & Lamsdell, 2016). Such a measure is unsatisfactory, as no tree can score better than a single polytomy that resolves no quartets incorrectly). The other extreme is to emphasize precision (Carpenter, 1988), which represents a preference for inaccuracy over ignorance, as better-resolved trees will typically contain a higher proportion of erroneous nodes.

The compromise position recognises accuracy and precision as two different but fungible components of the information held by a tree (MacKay, 1953). We desire a measure that allocates higher values to trees that more informatively describe a comparison tree, and whose value will not be affected by an information-neutral barter between precision and accuracy, for example by collapsing or resolving weakly supported nodes.

METHODS

If each quartet is taken to represent a single unit of information, then such a metric can be defined by calculating the number of quartet operations necessary to change from one tree into the other: each of the $d + r_1$ quartets that are present in the first tree must be unpicked, and each

of the $d + r_2$ quartets unique to the second tree must be forged. $(2d + r_1 + r_2) / 2Q$ provides an information-based measure of tree dissimilarity that reflects both accuracy and precision.

This measure is mathematically equivalent to the normalised symmetric difference metric, though it counts quartets in place of partitions as the unit of information. I call the measure the quartet divergence, by analogy with the Kullback-Leibler divergence (Kullback & Leibler, 1951) – though quartets are not independent so do not satisfy the statistical properties of Shannon-Weiner information.

A high value of this metric may represent high resolution (but some misinformation) or high accuracy (tempered by low resolution). The distinction between precision and accuracy can be instructive (e.g. Brown, Parins-Fukuchi, Stull, Vargas, & Smith, 2017), but is difficult to visualise. Because precision affects a tree's resolution and similarity score, plots that use these terms as their axes (e.g. O'Reilly et al., 2016; Puttick et al., 2017), will be dominated by autocorrelation, making them difficult to interpret. I instead employ a ternary plot with corners corresponding to the proportion of quartets (or partitions) that are the same in both trees (s), different in both trees (d), and unresolved in at least one tree ($r_1 + r_2 + u$).

Implementation

The new R package `slowQuartet` uses explicit enumeration to calculate the condition of each quartet, an $O(n^4)$ approach that is only practical for trees with up to a few dozen tips. A quicker $O(n \log n)$ solution (Sand et al., 2014) is applied to pairs of bifurcating trees. Ternary plots are produced using the `Ternary` package (Smith, 2017).

103 *Case study*

104 I used these approaches to visualise the results of a simulation study that evaluates different
 105 methods of phylogenetic reconstruction (Congreve & Lamsdell, 2016). These authors simulated
 106 55-character matrices from a 22-tip bifurcating tree using a Markov k -state 1 parameter model
 107 (Mk1, Lewis, 2001) with a gamma parameter (Wright & Hillis, 2014). I used TNT v1.1
 108 (Goloboff, Farris, & Nixon, 2008) to conduct parsimony search with equal and implied weights,
 109 using the parsimony ratchet (Nixon, 1999) and sectorial search (Goloboff, 1999) heuristics
 110 (search options: xmult:hits 20 level 4 chklevel 5 rat10 drift10). I took a strict
 111 consensus of all optimal trees obtained under equal weights, and under implied weights
 112 (Goloboff, 1993) at concavity constants of $k = 1, 2, 3, 5$ and 10. For each dataset I generated a
 113 further strict consensus of all trees that were optimal under any of the concavity constants 2, 3, 5
 114 and 10, but not $k = 1$ – which represents the extreme philosophy (Smith & Caron, 2015) that each
 115 step beyond the first makes a negligible contribution to tree score.

116 I also generated majority-rule consensus trees in MrBayes 3.2.2 (Huelsenbeck &
 117 Ronquist, 2001) using an Mk model (Lewis, 2001), with rates distributed according to a gamma
 118 parameter. I conducted four runs of four chains for 10 000 000 generations, sampling every
 119 10 000 generations and discarding the first 40% of samples as burn-in (Topology parameter:
 120 $0.999 < \text{PSRF} < 1.001$; $\text{ESS} > 400$). Scripts are provided in the Supplementary Information.

121 For each analytical configuration, I generated 20 further trees by progressively lowering
 122 the resolution of the most resolved tree. Under the Mk model, I collapsed clades with a posterior
 123 probability of $> x\%$, with x varying uniformly from 50–100. In parsimony analyses, strict
 124 consensus trees were produced including trees that were suboptimal by x steps (TNT command

subopt x ; bbreak;), with x corresponding to the integers 1..20 for implied weights, and drawn from a logarithmic distribution ($0.73^{0...19}$, $2.5 \times 10^{-3} \rightarrow 1 \times 10^0$) for implied weights.

To generate summary statistics, s , d , and r_2 were calculated for each tree relative to the bifurcating generative tree ($r_1 = u = 0$), and the mean of each of parameter was calculated for each analysis at each value of x .

RESULTS

There is no significant difference (at $p = 0.01$) between the quartet divergence of the best trees generated by the Mk model or implied weights ($k \in \{2, 3, 5, 10\}$), but the best trees generated by equal weights, or implied weights with $k = 1$, are significantly worse than those produced by the other methods (Figure 1).

Implied weights generated the highest precision, but not the highest accuracy (consistent with O'Reilly et al., 2016). Collapsing the least-supported nodes initially increases the accuracy (i.e. proportion of correctly to incorrectly resolved nodes), leading to a trivial increase in the overall informativeness of the tree. After a point, however, the gain in accuracy no longer offsets the information lost by collapsing nodes, and the tree diverges increasingly from the generative tree. The lower resolution of the equal weights and Bayesian results means that they do not experience this initial increase in tree quality: collapsing nodes immediately increases divergence from the generative tree.

Similar results occur if partitions are used instead of quartets to calculate tree divergence (Figure 2). The results of implied weights ($k \in \{2, 3, 5, 10\}$) and Bayesian analysis are statistically indistinguishable (at $p = 0.01$), whilst equal weights and implied weights with $k = 1$

generate trees that are significantly worse. Relative to the quartet measure, however, collapsing nodes more readily improves the partition score of a tree; optimal trees are obtained once relatively many nodes have been collapsed – indicating that the partition measure emphasizes accuracy over precision. These overall patterns and relationships continue to hold if datasets with a low consistency index are excluded.

DISCUSSION

The strict preference for accuracy (i.e. number of incorrect nodes) over resolution espoused by Congreve and Lamsdell (2016) can be visualised as a preference for trees that are closest to the upper right edge of the ternary space (Figure 3). This philosophy illustrates the danger of examining only the most resolved tree that a given method can produce (as discussed by Brown et al., 2017): even if the best-resolved tree of equal weights falls closer to this edge than the best-resolved trees under equal weights, collapsing nodes to polytomies causes all methods to approach the optimal score of zero incorrect nodes.

Divergence metrics offer a more meaningful comparison of methods, and demonstrates that at any given level of precision, equal weights and the extreme implementation of implied weights with $k = 1$ are significantly less accurate than other methods.

In general, the increase in accuracy attainable by intelligently reducing resolution (Goloboff, 1995; Salisbury, 1999) more than compensates for any reduction in precision, up to a point: though the partition metric differs from the quartet divergence metric in where this optimal trade-off lies. As a rule of thumb, collapsing weakly-supported nodes until trees reach an equal resolution to Bayesian results allows implied weights parsimony to generate trees that have an

equivalent accuracy to Mk trees, removing a reason to prefer Bayesian analysis to parsimony (cf. O'Reilly et al., 2016). Even more accuracy might be gained by collapsing nodes if other measures of node support (e.g. Giribet, 2003; Goloboff et al., 2003) are employed.

In conclusion, the quartet divergence metric produces notably different results to the partition metric of tree distance, with different implications for best practice. Ternary plots provide a means to visualise the contributions of accuracy and precision to tree quality, illustrating that perceived inaccuracy of implied weighting trees is simply a function of the greater precision that this method can, but need not, attain.

ACKNOWLEDGEMENTS

The TNT software is supported by the Willi Hennig Society.

DATA AVAILABILITY

The slowQuartet package is available at <https://github.com/ms609/SlowQuartet>. Its vignettes depict analytical results for each individual Congreve & Lamsdell dataset.

Supplementary data files have been uploaded to FigShare, 10.6084/m9.figshare.5659195.

REFERENCES

Billera, L. J., Holmes, S. P., & Vogtmann, K. (2001). Geometry of the space of phylogenetic

- 183 trees. *Advances in Applied Mathematics*, 27(4), 733–767. doi:10.1006/aama.2001.0759
- 184 Brown, J. W., Parins-Fukuchi, C., Stull, G. W., Vargas, O. M., & Smith, S. A. (2017). Bayesian
- 185 and likelihood phylogenetic reconstructions of morphological traits are not discordant when
- 186 taking uncertainty into consideration: a comment on Puttick *et al. Proceedings of the Royal*
- 187 *Society B: Biological Sciences*, 284(1864), 20170986. doi:10.1098/rspb.2017.0986
- 188 Carpenter, J. M. (1988). Choosing among multiple equally parsimonious cladograms. *Cladistics*,
- 189 4(3), 291–296. doi:10.1111/j.1096-0031.1988.tb00476.x
- 190 Clarke, J. A., & Middleton, K. M. (2008). Mosaicism, modules, and the evolution of birds:
- 191 results from a Bayesian approach to the study of morphological evolution using discrete
- 192 character data. *Systematic Biology*, 57(2), 185–201. doi:10.1080/10635150802022231
- 193 Congreve, C. R., & Lamsdell, J. C. (2016). Implied weighting and its utility in palaeontological
- 194 datasets: a study using modelled phylogenetic matrices. *Palaeontology*, 59(3), 447–462.
- 195 doi:10.1111/pala.12236
- 196 Estabrook, G. F., McMorris, F. R., & Meacham, C. A. (1985). Comparison of undirected
- 197 phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, 34(2),
- 198 193–200. Retrieved from <https://www.jstor.org/stable/2413326>
- 199 Giribet, G. (2003). Stability in phylogenetic formulations and its relationship to nodal support.
- 200 *Systematic Biology*, 52(4), 554–564. doi:10.1080/10635150390223730
- 201 Goloboff, P. A. (1993). Estimating character weights during tree search. *Cladistics*, 9(1), 83–91.
- 202 doi:10.1111/j.1096-0031.1993.tb00209.x
- 203 Goloboff, P. A. (1995). Parsimony and weighting: a reply To Turner And Zandee. *Cladistics*,
- 204 11(1), 91–104. doi:10.1111/j.1096-0031.1995.tb00006.x

- 205 Goloboff, P. A. (1999). Analyzing large data sets in reasonable times: solutions for composite
206 optima. *Cladistics*, 15(4), 415–428. doi:10.1006/clad.1999.0122
- 207 Goloboff, P. A., Farris, J. S., Källersjö, M., Oxelman, B., Ramírez, M. J., & Szumik, C. A.
208 (2003). Improvements to resampling measures of group support. *Cladistics*, 19(4), 324–332.
209 doi:10.1016/S0748-3007(03)00060-4
- 210 Goloboff, P. A., Farris, J. S., & Nixon, K. C. (2008). TNT, a free program for phylogenetic
211 analysis. *Cladistics*, 24(5), 774–786. doi:10.1111/j.1096-0031.2008.00217.x
- 212 Guillaume, T., & Cooper, N. (2016). Effects of missing data on topological inference using a
213 Total Evidence approach. *Molecular Phylogenetics and Evolution*, 94, 146–158.
214 doi:10.1016/j.ympev.2015.08.023
- 215 Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees.
216 *Bioinformatics*, 17(8), 754–755. doi:10.1093/bioinformatics/17.8.754
- 217 Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of*
218 *Mathematical Statistics*, 22(1), 79–86. doi:10.1214/aoms/1177729694
- 219 Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological
220 character data. *Systematic Biology*, 50(6), 913–925. doi:10.1080/106351501753462876
- 221 MacKay, D. M. (1953). Quantal aspects of scientific information. *IEEE Transactions on*
222 *Information Theory*, 1(1), 60–80. doi:10.1109/TIT.1953.1188569
- 223 Nixon, K. C. (1999). The Parsimony Ratchet, a new method for rapid parsimony analysis.
224 *Cladistics*, 15(4), 407–414. doi:10.1111/j.1096-0031.1999.tb00277.x
- 225 O'Reilly, J. E., Puttick, M. N., Parry, L., Tanner, A. R., Tarver, J. E., Fleming, J., ... Donoghue, P.
226 C. J. (2016). Bayesian methods outperform parsimony but at the expense of precision in the

- 227 estimation of phylogeny from discrete morphological data. *Biology Letters*, 12(4),
- 228 20160081. doi:10.1098/rsbl.2016.0081
- 229 Owen, M., & Provan, J. S. (2011). A fast algorithm for computing geodesic distances in tree
- 230 space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1), 2–13.
- 231 doi:10.1109/TCBB.2010.3
- 232 Pattinson, D. J., Thompson, R. S., Piotrowski, A. K., & Asher, R. J. (2015). Phylogeny,
- 233 paleontology, and primates: do incomplete fossils bias the tree of life? *Systematic Biology*,
- 234 64(2), 169–186. doi:10.1093/sysbio/syu077
- 235 Penny, D., & Hendy, M. (1985). The use of tree comparison metrics. *Systematic Zoology*, 34(1),
- 236 75–82. doi:10.2307/2413347
- 237 Puttick, M. N., O'Reilly, J. E., Tanner, A. R., Fleming, J. F., Clark, J., Holloway, L., ...
- 238 Donoghue, P. C. J. (2017). Uncertain-tree: discriminating among competing approaches to
- 239 the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society B: Biological*
- 240 *Sciences*, 284(1846), 20162290. doi:10.1098/rspb.2016.2290
- 241 Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical*
- 242 *Biosciences*, 53(1–2), 131–147. doi:10.1016/0025-5564(81)90043-2
- 243 Salisbury, B. A. (1999). Strongest evidence: maximum apparent phylogenetic signal as a new
- 244 cladistic optimality criterion. *Cladistics*, 15(2), 137–149. doi:10.1006/clad.1999.0099
- 245 Sand, A., Holt, M. K., Johansen, J., Brodal, G. S., Mailund, T., & Pedersen, C. N. S. (2014).
- 246 tqDist: a library for computing the quartet and triplet distances between binary or general
- 247 trees. *Bioinformatics*, 30(14), 2079–2080. doi:10.1093/bioinformatics/btu157
- 248 Sansom, R. S., & Wills, M. A. (2013). Fossilization causes organisms to appear erroneously

- 249 primitive by distorting evolutionary trees. *Scientific Reports*, 3, 2545.
- 250 doi:10.1038/srep02545
- 251 Sansom, R. S., Wills, M. A., & Williams, T. (2017). Dental data perform relatively poorly in
- 252 reconstructing mammal phylogenies: Morphological partitions evaluated with molecular
- 253 benchmarks. *Systematic Biology*, 66(5), 813–822. doi:10.1093/sysbio/syw116
- 254 Smith, M. R. (2017). Ternary: an R package to generate ternary plots.
- 255 Smith, M. R., & Caron, J.-B. (2015). *Hallucigenia*'s head and the pharyngeal armature of early
- 256 ecdysozoans. *Nature*, 523(7558), 75–78. doi:10.1038/nature14573
- 257 Steel, M. A., & Penny, D. (1993). Distributions of tree comparison metrics—some new results.
- 258 *Systematic Biology*, 42(2), 126–141. doi:10.1093/sysbio/42.2.126
- 259 Wright, A. M., & Hillis, D. M. (2014). Bayesian analysis using a simple likelihood model
- 260 outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS*
- 261 *ONE*, 9(10), e109210. doi:10.1371/journal.pone.0109210

262 FIGURE LEGENDS

263 **Figure 1. Quartet divergence, accuracy and precision of trees recovered by different**
 264 **analytical methods.** Average number of quartets, across all 100 simulated datasets, that are
 265 unresolved, the same as the generative tree, or resolved differently to the generative tree. Each
 266 marked point corresponds to the strict consensus of all trees that are suboptimal by a given
 267 amount under the specified given method.

268 **Figure 2. Partition resolution of trees recovered by different analytical methods.** Average
 269 number of quartets, across all 100 simulated datasets, that are unresolved, the same as the
 270 generative tree, or resolved differently to the generative tree. Each marked point corresponds to
 271 the strict consensus of all trees that are suboptimal by a given amount under the specified given
 272 method.

273 **Figure 3. Four measures of tree quality.** Congreve and Lamsdell (2016) equate tree quality
 274 with the number of incorrect nodes. Looking only at most resolved trees, equal weights
 275 outperforms implied weights on this measure. But then under any measure of accuracy that
 276 neglects precision, trees can be improved by continuing to collapse their least-supported nodes,
 277 so the optimal tree is entirely unresolved. Favouring precision alone does not distinguish
 278 between methods that can produce perfectly-resolved trees, regardless of whether such trees are
 279 accurate. The most informative (vertically highest) trees, in contrast, strike a balance between
 280 precision and accuracy.