Running head: COMPARING PHYLOGENETIC METHODS

Bayesian and parsimony approaches reconstruct equally informative trees from simulated morphological datasets

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*

Tree topologies are a primary output of phylogenetic analysis. As such, it is important to select a method of tree reconstruction that produces useful and instructive tree topologies. Morphological systematists generally advocate either parsimony methods, using either equal or implied weighting, or Bayesian methods, which employ an explicit probabilistic model of evolution. The performance of these methods has been evaluated by simulating morphological datasets from a known tree topology, and calculating how accurately each method reconstructs the generative tree.

Here I contend that this focus on accuracy risks neglecting another aspect in which trees can be informative: precision. If both of these aspects of information are considered simultaneously, recommendations for phylogenetic practice are quite different. Parsimony methods with implied weighting can rival the performance of Bayesian approaches, whereas trees identified using ‘equal weights’ are less informative and more likely to be misleading. Summary or consensus trees generated by any method can usually be made informative be collapsing poorly supported nodes.

*Keywords*.—phylogenetic inference; parsimony analysis; equal weights; implied weights; Bayesian phylogenetic methods; information content

Morphological phylogeny is the principal means by which observations of fossil taxa are applied to questions of evolutionary history. Parsimony analysis has long been the analytical approach of choice, but recent years have seen a welcome exploration of likelihood-based methods analogous to those used so widely with molecular data. Because true evolutionary trajectories are seldom known, however, it is difficult to evaluate how effectively either approach actually reconstructs evolutionary history.

One pragmatic way to address this question is to ask which method can best reconstruct a known tree from which a dataset has been artificially simulated. The first study to pose this question (Wright and Hillis 2014) favoured a likelihood-based approach, but comes with the notable caveat that the best-performing model, the Mk model (Lewis 2001), is the one used to generate the data. Later studies (O’Reilly *et al.* 2016; Puttick *et al.* 2018) established that a likelihood-based approach was still more accurate when a non-Mk model was used to simulate datasets, though it will always be difficult to demonstrate that any model used to simulate data directly encapsulates all of the processes that contribute to the properties of real morphological datasets (Goloboff *et al.* 2018*a*, *b*). Subsequent work has consistently found that the Mk model is more accurate than equally-weighted parsimony, which in turn is more accurate than parsimony under implied weights (Congreve and Lamsdell 2016; Puttick *et al.* 2017, 2018; O’Reilly *et al.* 2018).

Importantly, a focus on accuracy alone can lead to an erroneous prejudice against methods that, by default, produce highly-resolved trees (Brown *et al.* 2017). It has been suggested that any inherent tendency for precision might be countered after completing an analysis by collapsing the least-supported nodes, which are the most likely to be inaccurate (Goloboff and Szumik 2015; Goloboff *et al.* 2018*a*). This recognizes an inherent trade-off between accuracy and precision, which represent two different but fungible components of the information contained within a tree (Mackay 1953). Despite an implicit acknowledgement that any increase in accuracy comes at the expense of precision, the barter between these two quantities has not been rigorously explored.

Focussing on accuracy alone (as do Congreve and Lamsdell 2016) is clearly unsatisfactory: no tree is more accurate than a single polytomy that resolves no relationships incorrectly (because it resolves none at all). The other extreme is to value precision alone (Carpenter 1988), which represents a preference for error over ignorance, as better-resolved trees will typically contain a higher proportion of erroneous nodes. We should prefer instead trees that contain as much information as possible about the true tree, where both precision and accuracy contribute to the total information content of a tree. By way of analogy, a sundial on a bright day (which is somewhat precise and somewhat accurate) conveys more temporal information than a stopped clock (perfectly precise, but entirely inaccurate) or the imprecise (yet perfectly accurate) statement “it is daytime”.

To evaluate trees based on their information content, we require a measure that allocates higher values to trees that more informatively describe a comparison tree (in our case, the tree that was used to generate a simulated dataset). The value of this measure should not be influenced by any information-neutral barter between precision and accuracy, such as a gain in accuracy obtained at an equivalent loss of precision by collapsing a selection of weakly supported nodes. This requirement necessitates that the measure can be applied to incompletely resolve trees.

Because phylogenetic trees are not vectors, there is no natural scale against which they can be compared (Penny and Hendy 1985). Whilst it is possible to measure distances between edge-weighted trees in multi-dimensional space (Billera *et al.* 2001; Owen and Provan 2011), parsimony methods only evaluate topologies, rather than trees with meaningful edge lengths. As such, there are a variety of measures of comparing topologies (Kuhner and Yamato 2015), each with limitations. A subset of these measures (Penny and Hendy 1985; Steel and Penny 1993; Nye *et al.* 2006) are only defined when both trees are bifurcating, and thus cannot be applied to trees that are not perfectly resolved.

The quartet metric counts the number of four-taxon trees that are present as subtrees of each tree under consideration (Estabrook *et al.* 1985; Steel and Penny 1993). It is in many regards preferable to the more widely used symmetric difference measure (Penny and Hendy 1985), a refinement of the Robinson-Foulds partition metric (Robinson and Foulds 1981), which counts the proportion of edges (i.e. bipartitions or splits) present in each tree that are absent in the other. The quartet metric has a much greater range of possible values, allowing it to distinguish between relatively minor differences, whereas the limited range of integer values available to the symmetric difference measure constrains it ability to distinguish similar trees from more dissimilar trees. Moreover, the symmetric difference measure exhibits some counterintuitive properties: its value can reach its maximum value in a pair of trees that differ only in the position of a single tip (Steel and Penny 1993); and 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 point. This undesirable bias does not occur under the quartet metric.

Both of these methods are nevertheless able to distinguish between precision and accuracy. Given a pair of trees, a quartet or node *E* will satisfy one of five conditions (Estabrook *et al.* 1985): it 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*). (2*d* + *r*1 + *r*2) / 2*E* provides an information-based measure of tree dissimilarity that reflects both accuracy and precision; this quantity corresponds to the quartet dissimilarity metric or the normalised symmetric difference metric for quartets and nodes respectively (Bandelt and Dress 1986). These metrics reflect the number of quartet or node modifications taken to change from one tree into the other: each of the *d* + *r*1 entities that are present in the first tree must be removed, and each of the *d* + *r*2 entities unique to the second tree must be added. This view suggests an analogy with the Kullback-Leibler divergence (Kullback and Leibler 1951), though as neither quartets nor nodes are independent, they do not satisfy the statistical properties of Shannon-Weiner information. A high value of either metric may represent high resolution (but some misinformation) or high accuracy (tempered by low resolution), or some intermediate compromise.

I used the quartet dissimilarity and normalised symmetric difference metrics to revisit the results of two simulation studies that evaluate different methods of phylogenetic reconstruction, in order to establish whether the results obtained remain valid when precision is considered alongside accuracy as a desideratum.

# Methods

## Data analysis

Congreve and Lamsdell (2016; CL hereafter) simulated 55-character matrices from a bifurcating 22-tip tree using a Markov *k­*-state 1 parameter model (Mk1, Lewis 2001) with a gamma parameter (Wright and Hillis 2014). O’Reilly et al. (2016; OR hereafter) simulated character matrices containing 100, 350 and 1000 characters from a bifurcating 75-tip tree using a modified HKY85 (Hasegawa *et al.* 1985) model.

I used TNT v1.1 (Goloboff *et al.* 2008) to conduct parsimony searches on each of these matrices under equal and implied weights, using the parsimony ratchet (Nixon 1999) and sectorial search (Goloboff 1999) heuristics (search options: xmult:hits 20 level 4 chklevel 5 rat10 drift10). I took a strict consensus of all optimal trees obtained under equal weights, and under implied weights (Goloboff 1993) at the concavity constants used in each respective study (CL: *k* = 1, 2, 3, 5 and 10; OR: *k* = 2, 3, 5, 10, 20 and 200). For each dataset I generated a further strict consensus of all trees that were optimal under any of the concavity constants – excluding the unreasonable value of *k* = 1, as this represents the extreme philosophy (Smith and Caron 2015) that each step beyond the first makes a negligible contribution to tree score.

I also generated majority-rule consensus trees in MrBayes 3.2.2 (Huelsenbeck and Ronquist 2001) using an Mk model (Lewis 2001), with rates distributed according to a gamma parameter. I combined results from four runs of four chains, sampling every 10 000 generations for 10 000 000 generations, then discarding the first 40% of samples as burn-in (topology parameter: 0.999 < PSRF < 1.001; ESS > 400). Scripts are provided in the Supplementary Information.

With the results of each dataset, and for each analytical configuration, I generated 20 further trees by progressively lowering the resolution of the most resolved tree. Under the Mk model, I collapsed clades with a posterior probability of < *x*%, with *x* varying uniformly from 50–100. In parsimony analyses, strict consensus trees were produced from forests of trees that were suboptimal by up to *x* steps (TNT command subopt *x*; bbreak;), with *x* corresponding to the integers 1..20 for equal weights, and drawn from a logarithmic distribution (0.730...19, 2.5×10−3→1×100) for implied weights.

For each tree, I calculated the condition of each quartet and partition relative to the generative tree. Quartet methods used the new R package SlowQuartet, which 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, but supports polytomies. A quicker O(*n* log *n*) solution (Sand *et al.* 2014) was applied to pairs of bifurcating trees. Partition measures were calculated using the R package phangorn (Schliep 2011). To generate summary statistics, *s*, *d*, and *r2* 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*.

## Visualization

The overall information content of a tree reflects both its precision and is accuracy. Plotting tree scores against resolution (O’Reilly *et al.* 2016; Puttick *et al.* 2017, 2018) has the drawback that precision affects both of these quantities; such plots will be dominated by autocorrelation, making them difficult to interpret. As a measure of tree score reflects which of three exclusive states (correctly resolved, incorrectly resolved, or unresolved) is represented by each countable unit (a quartet or a node), the natural plotting device is a ternary plot. The three corners of each plot then correspond 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*). Plots were generated in R using the Ternary package (Smith 2017), oriented such that the *y* direction corresponds to tree quality, and the *x* direction to precision.

Each tree recovered maps to a single point in ternary space, which denotes the accuracy and precision of that particular tree. Collapsing the least-supported node in a given tree generates a new tree that is less precise and either more accurate (if the collapsed node happened to be incorrect) or less accurate (if the collapsed node was a node that also appears in the reference tree). This new, less resolved tree will correspond to a point closer in the *x* direction to the fully-unresolved corner of ternary space; collapsing subsequent nodes will yield a path that depicts the gain or loss of accuracy as precision is progressively sacrificed. To visualize the net effect of this barter, I have plotted the mean path for each dataset under each analytical configuration.

# Results

With the Congreve and Lamsdell datasets, 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* ∈ {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. 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 accuracy of implied weights (*k* ∈ {2, 3, 5, 10}) and Bayesian analysis are statistically indistinguishable (at *p* = 0.01) at any given level of precision, whereas equal weights and implied weights with *k* = 1 generate trees that are significantly worse. Relative to the more robust 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.

The same patterns can be observed in the O’Reilly *et al.* datasets (Figure 3): at any given level of precision, the best trees generated by the *Mk* model are similar in accuracy to those generated under implied weights (except with very small values of *k*), but are more accurate than those generated using equal weights.

# Discussion

The strict preference for accuracy (i.e. minimization of the 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 4). 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): the best-resolved tree of equal weights falls closer to this edge than the best-resolved trees under equal weights, but this simply reflects the poor resolving power of equal weights; this approach would prefer a single polytomy that resolves no nodes and therefore obtains the optimal score of zero incorrect nodes.

Divergence metrics offer a more meaningful way to evaluate methods. Using an approach that quantifies all aspects in which a tree might be informative – i.e. both accuracy and precision – it is clear that equal weights and the extreme implementation of implied weights with *k* = 1 are substantially less accurate than other methods at any given level of precision (Figure 4).

In general, the increase in accuracy attainable by intelligently reducing resolution (Goloboff 1995; Salisbury 1999; Goloboff and Szumik 2015) more than compensates for any reduction in precision. Practitioners are likely to improve the overall quality of their results by collapsing the most weakly supported nodes, which are more likely than not to represent misinformation. There is a limit, of course, to how much resolution might be profitably sacrificed; interestingly, the choice of tree comparison metric has a large impact on where this optimal trade-off lies. As a rule of thumb, implied weights parsimony will generate trees of a similar accuracy to Mk trees if they are reduced to an equivalent resolution by collapsing their least-supported nodes – removing a reason to prefer Bayesian analysis to parsimony (cf. O’Reilly *et al.* 2016). Still more accuracy might be yielded from each collapsed node if more sophisticated measures of node support (e.g. Giribet 2003; Goloboff *et al.* 2003) prove more effective at identifying incorrect nodes.

# Implications for practice

I recommend that future simulation studies employ the quartet dissimilarity metric to compare trees with the generative tree (or, if there is a particular reason to count nodes, the normalised symmetric difference metric), so that comparisons reflect information from both precision and accuracy. Ternary diagrams represent a natural method of displaying the relative contributions of accuracy and precision to tree quality.

Analysis on these principle demonstrates that, at least in the simulation studies analysed herein, the results of parsimony analysis are neither more nor less useful than those of probabilistic analyses in reconstructing evolutionary history. The caveat is that parsimony analysis must employ a sensible weighting scheme; implied weighting should be employed with a moderate concavity (*k*) value. Extremely low values (*k* < 3), or extremely high values (i.e. equally weighted parsimony, which is mathematically equivalent to implied weights with an infinite concavity constant) are likely to yield results that are less informative about evolutionary history, with this lack of information representing a combination of imprecision and inaccuracy; results arising under such values do not merit biological interpretation.

As neither Bayesian nor parsimony analyses generate consistently superior results, researchers may wish to explicitly compare the results of both methods; observations common to both methods are particularly likely to be well supported by underlying data.

# Acknowledgements

The TNT software is supported by the Willi Hennig Society.

# Data availability

The SlowQuartet package will be submitted to the CRAN repository when the review process is complete. Its vignettes and provide detailed examples of situations where existing tree distance measures produce undesirable results, and depict analytical results for each individual Congreve & Lamsdell dataset.

Reviewers can access the SlowQuartet package at https://github.com/ms609/SlowQuartet

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

Reviewers can access this repository at: https://figshare.com/s/46744779f750495e527b

# References

BANDELT, H. J. and DRESS, A. 1986. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics*, **7**, 309–343.

BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. 2001. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, **27**, 733–767.

BROWN, J. W., PARINS-FUKUCHI, C., STULL, G. W., VARGAS, O. M. and SMITH, S. A. 2017. Bayesian and likelihood phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on Puttick *et al*. *Proceedings of the Royal Society B: Biological Sciences*, **284**, 20170986.

CARPENTER, J. M. 1988. Choosing among multiple equally parsimonious cladograms. *Cladistics*, **4**, 291–296.

CONGREVE, C. R. and LAMSDELL, J. C. 2016. Implied weighting and its utility in palaeontological datasets: a study using modelled phylogenetic matrices. *Palaeontology*, **59**, 447–462.

ESTABROOK, G. F., MCMORRIS, F. R. and MEACHAM, C. A. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, **34**, 193–200.

GIRIBET, G. 2003. Stability in phylogenetic formulations and its relationship to nodal support. *Systematic biology*, **52**, 554–564.

GOLOBOFF, P. A. 1993. Estimating character weights during tree search. *Cladistics*, **9**, 83–91.

———. 1995. Parsimony and weighting: a reply To Turner And Zandee. *Cladistics*, **11**, 91–104.

———. 1999. Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics*, **15**, 415–428.

——— and SZUMIK, C. A. 2015. Identifying unstable taxa: efficient implementation of triplet-based measures of stability, and comparison with Phyutility and RogueNaRok. *Molecular Phylogenetics and Evolution*, **88**, 93–104.

———, FARRIS, J. S. and NIXON, K. C. 2008. TNT, a free program for phylogenetic analysis. *Cladistics*, **24**, 774–786.

———, TORRES, A. and ARIAS, J. S. 2018*a*. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics*, **34**, 407–437.

———, TORRES GALVIS, A. and ARIAS, J. S. 2018*b*. Parsimony and model-based phylogenetic methods for morphological data: comments on O’Reilly *et al*. *Palaeontology*, **61**, 625–630.

———, FARRIS, J. S., KÄLLERSJÖ, M., OXELMAN, B., RAMÍREZ, M. J. and SZUMIK, C. A. 2003. Improvements to resampling measures of group support. *Cladistics*, **19**, 324–332.

HASEGAWA, M., KISHINO, H. and YANO, T. aki. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

HUELSENBECK, J. P. and RONQUIST, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

KUHNER, M. K. and YAMATO, J. 2015. Practical performance of tree comparison metrics. *Systematic Biology*, **64**, 205–214.

KULLBACK, S. and LEIBLER, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.

LEWIS, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, **50**, 913–925.

MACKAY, D. M. 1953. Quantal aspects of scientific information. *IEEE Transactions on Information Theory*, **1**, 60–80.

NIXON, K. C. 1999. The Parsimony Ratchet, a new method for rapid parsimony analysis. *Cladistics*, **15**, 407–414.

NYE, T. M. W., LIÒ, P. and GILKS, W. R. 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, **22**, 117–119.

O’REILLY, J. E., PUTTICK, M. N., PISANI, D. and DONOGHUE, P. C. J. 2018. Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. *Palaeontology*, **61**, 105–118.

———, ———, PARRY, L., TANNER, A. R., TARVER, J. E., FLEMING, J., PISANI, D. and DONOGHUE, P. C. J. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biology Letters*, **12**, 20160081.

OWEN, M. and PROVAN, J. S. 2011. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **8**, 2–13.

PENNY, D. and HENDY, M. 1985. The use of tree comparison metrics. *Systematic Zoology*, **34**, 75–82.

PUTTICK, M. N., O’REILLY, J. E., TANNER, A. R., FLEMING, J. F., CLARK, J., HOLLOWAY, L., LOZANO-FERNANDEZ, J., PARRY, L. A., TARVER, J. E., PISANI, D. and DONOGHUE, P. C. J. 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society B: Biological Sciences*, **284**, 20162290.

———, ———, PISANI, D. and DONOGHUE, P. C. J. 2018. Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. *Palaeontology*.

ROBINSON, D. F. and FOULDS, L. R. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

SALISBURY, B. A. 1999. Strongest evidence: maximum apparent phylogenetic signal as a new cladistic optimality criterion. *Cladistics*, **15**, 137–149.

SAND, A., HOLT, M. K., JOHANSEN, J., BRODAL, G. S., MAILUND, T. and PEDERSEN, C. N. S. 2014. tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics*, **30**, 2079–2080.

SCHLIEP, K. P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

SMITH, M. R. 2017. Ternary: an R package to generate ternary plots. doi:10.5281/zenodo.1068997.

——— and CARON, J.-B. 2015. *Hallucigenia*’s head and the pharyngeal armature of early ecdysozoans. *Nature*, **523**, 75–78.

STEEL, M. A. and PENNY, D. 1993. Distributions of tree comparison metrics—some new results. *Systematic Biology*, **42**, 126–141.

WRIGHT, A. M. and HILLIS, D. M. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE*, **9**, e109210.

# Figure legends

**Figure 1.** **Status of quartets in trees recovered from Congreve and Lamsdell datasets.** Points denote the average number of quartets that are the same as the generative tree (*Same*), resolved differently to the generative tree (*Different*), or not resolved (*Unresolved*). Each series indicates the effect of progressively collapsing the least-supported nodes in trees generated by analysis of datasets simulated by Congreve and Lamsdell (2016) under the specified analytical parameters. An increase in the *x* direction corresponds to a decrease in quartet dissimilarity (more informative trees); an increase in the *y* direction corresponds to a decrease in precision (less resolved trees).

**Figure 2. Status of partitions in trees recovered from Congreve and Lamsdell datasets.** Points denote the average number of nodes that are the same as the generative tree (*Same*), resolved differently to the generative tree (*Different*), or not resolved (*Unresolved*). Each series indicates the effect of progressively collapsing the least-supported nodes in trees generated by analysis of datasets simulated by Congreve and Lamsdell (2016) under the specified analytical parameters. An increase in the *x* direction corresponds to a decrease in symmetric difference (more informative trees; cf. Robinson-Foulds distance); an increase in the *y* direction corresponds to a decrease in precision (less resolved trees).

**Figure 3. Status of quartets in trees recovered from O’Reilly *et al.* datasets.** Points denote the average number of quartets that are the same as the generative tree (*Same*), resolved differently to the generative tree (*Different*), or not resolved (*Unresolved*). Each series indicates the effect of progressively collapsing the least-supported nodes in trees generated by analysis of datasets with 100, 350 or 1000 characters simulated by O’Reilly *et al.* (2016) under the specified analytical parameters. An increase in the *x* direction corresponds to a decrease in quartet dissimilarity (more informative trees); an increase in the *y* direction corresponds to a decrease in precision (less resolved trees).

**Figure 4. Measuring tree quality.** Congreve and Lamsdell (2016) equated tree quality with the number of incorrect nodes. On this measure, the most resolved trees obtained under equal weights are better than the most resolved trees under implied weights, but worse than a single polytomy. Reducing the precision of trees obtained under implied weights by collapsing the nodes with the least support produces trees that are more accurate (higher proportion of correct nodes; lower number of incorrect nodes). At any given level of precision, equal weights produces the least accurate trees. The most informative (vertically highest) trees strike a balance between precision and accuracy.