Syst. Biol. 64(2):205–214, 2015

DOI:10.1093/sysbio/syu085
Advance Access publication November 4, 2014

Practical Performance of Tree Comparison Metrics

MARY K. KUHNER* AND JON YAMATO

Department of Genome Sciences, Box 355065, University of Washington, Seattle, WA 98195-5065, USA.
**Correspondence to be sent to: Department of Genome Sciences, Box 355065, University of Washington, Seattle, WA 98195-5065, USA; E-mail: mkkuhner@uw.edu*

Received 22 September 2014; reviews returned 26 September 2014; accepted 30 October 2014
Associate Editor: David Posada

Abstract.—The phylogenetic literature contains numerous measures for assessing differences between two phylogenetic trees. Individual measures have been criticized on various grounds, but little is known about their comparative performance in typical applications. We evaluate the performance of nine tree distance measures on two tasks: 1) distinguishing trees separated by lesser versus greater numbers of recombinations, and 2) distinguishing trees inferred with lower versus higher quality data. We find that when the trees being compared are similar, measures that make use of branch lengths are superior, with the branch-length version of the Robinson–Foulds metric performing best. In contrast, for dissimilar trees topology-only measures are superior, with the Alignment metric of Nye et al. performing best. We also apply the measures to a mammalian dataset and observe that the best metric depends on whether branch-length information is of interest. We give practical recommendations for choosing a tree distance metric in different applications. [Phylogenetics; tree comparison; tree distance metrics.]

INTRODUCTION

Phylogeneticists often want to assess how similar or dissimilar two phylogenetic trees are to each other. For example, simulation studies of phylogeny-inference algorithms need to measure how close each algorithm comes to recovering the true tree. A large number of tree comparison measures have been proposed for this purpose, starting with the symmetrical difference measure of Robinson and Foulds (1979). Criticisms of individual measures have been raised (e.g., Penny et al. 1982; Huson et al. 2011), but to our knowledge no systematic survey has been made of their performance. The difficult part of such a survey is deciding what constitutes good or bad performance of a tree comparison measure.

Steel and Penny (2003) consider the usefulness of various measures on theoretical grounds. They conclude that the optimal measure depends on the expected null distribution of trees and the degree of difference among trees being compared, but it is not clear how to relate these conclusions to studies of actual data. In this study, we use simulated and real data to explore the performance of tree comparison measures in tasks relevant to their use in phylogenetic studies.

In many ways, the most natural tree comparison measures are those based on finding the minimum number of rearrangement steps required to transform one tree into the other. Possible rearrangement steps include nearest-neighbor interchange (NNI), subtree pruning and regrafting (SPR), and tree bisection and reconnection (TBR). Unfortunately such measures are seldom used in practice for large studies as they are expensive to calculate if the trees are dissimilar. NP-completeness has been shown for distances based on NNI (Li and Zhang 1999), TBR (Allen and Steel 2001), and SPR (Bordewich and Semple 2004). Most published uses of tree comparison measures, therefore, use simpler measures, very often the topological distance of Robinson and Foulds (1981) ("RF distance"). In this study we explore the usefulness of these simpler, polynomial-time distances.

The importance of choosing appropriate tree distance metrics is shown by an example from our research (MK Kuhner and JR McGill, submitted for publication). We used the topological measure RF (Robinson and Foulds 1981) and the branch-length measure RFL (Robinson and Foulds 1979) to evaluate trees inferred with varying levels of correction for DNA sequencing error. For low levels of sequencing error, RFL showed a clear pattern of superiority of proper correction to over- or undercorrection, whereas correction made no difference to the RF scores. In contrast, for extremely high levels of sequencing error, RFL indicated that proper correction and overcorrection were equivalent, whereas RF showed that overcorrection destroyed topological recovery. Only by using two measures with complementary qualities were we able to form a complete picture.

We consider nine measures of resemblance among trees. Seven are naturally measures of distance, whereas two are measures of similarity; we convert the latter to measures of distance by subtracting from their maxima. One measure, the similarity score of Hein et al. (2005), is not a metric as the distance from tree A to tree B is not necessarily equal to the distance from B to A; we average these two distances to create a symmetrical version. With these modifications, all of the measures are metrics.

We present two experiments based on simulated data, each embodying a criterion for "better" and "worse" metrics, as well as a sample application based on real data.

N-away: Approximating a rearrangement distance.—A natural way to think of distances between trees is in

TABLE 1. Distance measures used

Abbr.	Citation	Basis
Topology only		
RF	Robinson and Foulds (1981)	Number of clades not shared
Trip	Critchlow et al. (1996)	Proportion of triplets not shared
MAST	Gordon (1980)	Size of maximum shared subtree
Align	Nye et al. (2006)	Mismatches in best alignment of branches
Node	Williams and Clifford (1971)	Difference in pathlengths between pairs of tips
Topology and branch lengths		
RFL	Robinson and Foulds (1979), Kuhner and Felsenstein (1994)	Sums of differences in branch lengths
Sim	Hein et al. (2005)	Chance that random points share the same clades
TripL	This study	Sum of discrepant branch lengths among triplets
Branch lengths only		
Int	this study	Summed differences between time-interval lengths

terms of the minimum number of rearrangements of a certain type needed to transform one tree into the other. Such rearrangement-based distances are disfavored in practice as they are expensive to calculate. However, it is straightforward to generate trees separated by a given number of rearrangements, and these can be used to test which of the easily computable metrics best approximates the difficult-to-compute rearrangement distance. We call this experiment "*n*-away" as it asks whether a given metric can distinguish, for example, trees four rearrangements apart from trees five rearrangements apart.

We chose recombinational rearrangement as the basis of our rearrangement distance, to which the simple metrics will be compared. Our strategy is to create successive trees along a simulated chromosome using Hudson's coalescent with recombination (CwR) model (Hudson 1983). The recombinational distance between any two of these trees is simply the number of recombinational breakpoints between them. While it would seem simpler to use familiar phylogenetic rearrangements such as SBR or NNI, we needed to insure that initial and rearranged trees had branch lengths drawn from the same distribution, and that the clock assumption was maintained. CwR rearrangement meets both of these criteria. It is also biologically relevant, since one potential cause of disagreement among inferred trees is the action of recombination or recombination-like processes such as hybridization and horizontal gene transfer.

In the *n*-away experiment, the quality of a metric is defined as its ability to correctly separate smaller from larger recombinational distances. We explore performance both for short distances (very similar trees) and longer distances (increasingly dissimilar trees).

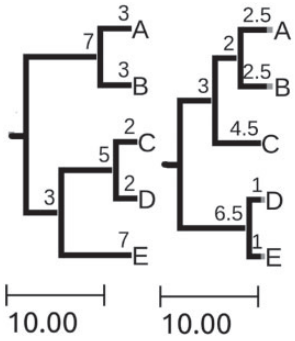
Bullseye: Distinguishing better from worse inferences of a tree.—A common use of distance metrics is to evaluate the performance of different phylogenetic algorithms on simulated data by comparing inferred trees to the true tree (e.g., Kuhner and Felsenstein 1994). A good metric is one that can reliably distinguish a better from a worse inference. We model this by generating data from a known tree and then inferring the tree with successively smaller and smaller subsets of the data. A well-performing metric will be able to determine that trees made with more data are generally closer to the truth, and will be able to do so over as wide a range as possible—that is, it will be informative about both nearly correct inferences and very poor ones. We call this experiment "bullseye" as we envision the inferred trees as rings around the truth, with trees from rich datasets in the innermost ring and trees from meager datasets in the outermost one.

Mammalian dataset.—We also present results for a practical application in which trees inferred from 10 mammalian loci were compared to the tree inferred from their concatenated sequences. Loci producing results most similar to the concatenated result could be preferred for sequencing in additional taxa, whereas loci producing highly discordant results should be examined for evidence of paralogy, incomplete lineage sorting, or other perturbing forces. Though we do not know which metric produces the "correct" answer, we can observe which metrics are able to pick a preferred gene and whether they agree on this choice.

METHODS

General rationale.—We studied clocklike, rooted trees because several of our distance metrics are only defined on such trees, and because inference of clocklike trees is central to areas such as phylogenetic dating and coalescent analysis. We also required the trees to be fully bifurcating. Further study will be needed to see how the performance of tree distance metrics changes when trees are unrooted, nonclocklike, or multifurcating.

Distance metrics.—Metrics used in this study are summarized in Table 1. A pair of example trees with distances between them under each metric are shown in Figure 1.



Metric	Raw Score	Transformed Score
RF	4.0	-
Trip	0.5	-
MAST	4.0	1.0
Align	1.0	-
Node	1.28	-
RFL	39.53	-
Sim	0.39	0.61
TripL	9.98	-
Int	3.80	-

FIGURE 1. Metric scores for comparison of two simple trees. The column headed "raw score" gives the score under the original form of the metric. The column headed "transformed score" gives the score as used in this article: measures of similarity are transformed into measures of distance. When no transformed score is noted, the raw score was used.

The symmetrical difference metric (Robinson and Foulds 1981), also called the Robinson–Foulds metric or RF (but note these names have also been applied to the metric of Robinson and Foulds (1979)), counts the number of branch partitions that appear in one tree but not the other, scoring 1 for each nonmatched partition. Since both trees are scored, the maximum score is twice the number of partitions in each tree. We consider here the rooted-tree version (each branch is identified by the set of tips it leads to), which has a maximum of $2(n-2)$ where n is the number of tips.

Users of published software for distance metrics should be aware that implementations vary in their use of normalizing factors: for example, some implementations of the RF metric divide it by 2. This makes no difference to use of RF within a study, but caution is needed when comparing results across studies.

Two branch-length-aware variants of RF have been published. Robinson and Foulds (1979) proposed that for each branch the score is incremented by the absolute value of the difference in corresponding branch lengths between the two trees, with a branch not found in a particular tree treated as having a length of 0. Kuhner and Felsenstein (1994) used the same approach, but summing the squares of the differences rather than the absolute values; this weights long branches more heavily. Both are instances of a general algorithm in which the absolute difference is raised to a power k before summing, with Robinson and Foulds (1979) having $k=1$ and Kuhner and Felsenstein (1994) having $k=2$. We will call this general algorithm RFL for "RF algorithm with Lengths."

The RF-like metrics are all based on the idea of shared clades or branches defined by possession of exactly the same tips. Moving one tip will, therefore, sometimes lead to a huge jump in score. An alternative family of methods considers triplets (or quartets in unrooted trees) and asks how many triplets are shared. Moving one tip will affect only those triplets containing it, so these methods are expected to be more stable.

The basic Triplet ("Trip") distance was defined by Critchlow et al. (1996) as follows. Enumerate all possible triplets of the tips. Determine the topology of that triplet in A and B (by deleting all other tips) and score 1 point for each triplet whose topology differs. Sum this score over triplets. This is a topology-only distance. (Implementations differ in whether the result should be divided by the number of triplets; we chose to do so.) This method is the rooted-tree equivalent of the Quartet method of Estabrook et al. (1985).

A novel branch-length version of the Trip distance can be defined as follows: In a clocklike, rooted tree, each triplet contains two pieces of length information, namely the length from the tips to the first node and the length from the first node to the second. The first-node lengths from each tree are compared and the absolute values of their differences summed, with a triplet having a different topology in the two trees treated as having branch lengths of zero. The second-node lengths are similarly compared, and the results summed across all triplets. As with the RF family, an infinite number of related methods can be created by raising the differences to a power k before summing. We call this family of distances "TripL" for "Triplet with Lengths."

We consider a novel metric which does not use topology information at all, but focuses on the distribution of interval lengths. The tree is broken into intervals with an interval boundary at the tips and at each node time. Starting at the tips, each interval is examined in turn: for example, we examine the time interval between the tips and the most tipward internal node in tree A and tree B, without consideration of whether these nodes correspond topologically. The absolute differences of these interval lengths are summed. Thus, any two trees with identical node times will compare as identical, regardless of topology. A family of related metrics can be produced by raising the differences to a power k before summing. We call this family Int for "Interval." It is inspired by the use of interval lengths in coalescent theory.

The Maximum Agreement SubTree ("MAST") method (Gordon 1980) measures similarity between trees as the number of tips in the largest subtree identical between the two trees. Its maximum score, indicating a perfect match, is the size of the tree itself; its minimum is deterministic but complex to calculate, and a particular tree may not be able to achieve the minimum no matter what tree it is compared with. MAST, which is natively a similarity score, can therefore be converted to a distance by subtracting from the total size of the tree, but the maximum of the resulting distance will vary from case to case. We implemented MAST according to the algorithm of Goddard et al. (1994), which is more straightforward than faster versions known in the literature, and converted it into a distance by subtracting from the maximum.

The tree alignment metric ("Align") of Nye et al. (2006) considers all the ways in which a one-to-one mapping of branches in the two trees could be made. For each mapping, it calculates a dissimilarity score

for the clades separated by each branch, and takes the dissimilarity score of the optimal mapping as the distance between the trees. Because of the emphasis on clades, this method appears related to RF. However, it is more flexible as it uses information about clades that are almost the same, as well as clades that are precisely the same.

The nodal or path distance ("Node") of Williams and Clifford (1971) considers each possible pair of tips in turn, and asks how many nodes are traversed in drawing the minimal path from one tip to the other in A and in B. The sum of the differences in these minimal path lengths is the distance between A and B. The terms in this difference can be raised to a power k before summing. Williams and Clifford's version has $k=1$; a $k=2$ variant was proposed by Penny et al. (1982) who called it the Path Difference Metric.

Hein et al. (2005) propose a similarity measure based on the probability that a point chosen randomly in A will be on a branch leading to the same set of tips as a point chosen randomly in B. Since the density of points to be chosen depends on the total length of each tree, this score must be normalized. The authors normalize the similarity of A to B by dividing it by the probability that two points chosen at random in A will yield the same set of tips. This implies that the similarity of A to B is not necessarily equal to the (differently normalized) similarity of B to A. To convert Hein's score to a distance metric, we average the A/B and B/A similarities, and subtract from 1; we call the resulting distance "Sim" as it is derived from a measure of similarity. The Sim measure uses branch lengths, but is not sensitive to their magnitudes, only to their ratios in the two trees. Two trees which differ only by a scaling factor in their branch lengths will test as identical. Sim cannot be computed when either tree has a total branch length of zero, as the normalization would divide by zero. Cases where one or both inferred trees were of zero length were, therefore, dropped from consideration for this metric.

For this study, we implemented all metrics in Python. Our programs to compute, score, and graph the metrics are archived on Dryad at http://dx.doi.org/10.5061/dryad.g9089.

Handling of ties.—Many of the metrics yield large numbers of ties, particulary the topology-only metrics, and it was thus essential to handle ties correctly. Ties might be either dropped from consideration or counted as successes, half-successes, or failures. We found that dropping them from consideration (as often recommended for sign tests) or treating them as successes gave a huge apparent advantage to the topology-only metrics by excusing them from giving an opinion on difficult cases. Conversely, treating ties as failures caused the topology-only metrics to perform "worse than chance" on similar trees: their performance would be drastically enhanced by simply adding a random tiebreak. To avoid both extremes, we treated a tie as a half success.

Recombinational distance (n-away experiment).—We are interested in the ability of these metrics to differentiate trees at different rearrangement distances. To produce trees at different recombinational rearrangement distances, we simulated the ancestral recombination graph (ARG) of a long chromosomal region and considered successive local trees along this region. As a proxy for recombinational distance, we used the number of recombination breakpoints separating two local trees.

We generated 5000 random ARGs using the *m*s program (Hudson 2002) with the θ parameter and recombination rate parameter set to 100. Each ARG consisted of 20 tips with 40,000 bp of sequence each. At this level of recombination, the vast majority of recombination locations along the sequence represent only one recombination. Therefore, adjacent local trees along the sequence almost always have a recombinational distance of 0 or 1 (the distances of 0 arise from "invisible" recombinations where the two recombinant lineages coalesce with one another). The number of intervening breakpoints is thus an approximate maximum for the number of intervening recombinations. This number was determined by counting local trees in the *m*s output file.

Inference accuracy (bullseye experiment).—We are also interested in the ability of these metrics to tell a better from a worse reconstruction of the tree. To measure this, we generated random trees using the rantree.c program (Felsenstein J., unpublished data) under either a coalescent (Kingman 1982) or branching process (Yule 1924) model. Coalescent trees have relatively short tipward and long rootwards branches, and are typical of within-population phylogenies; branching process trees have less extreme branch lengths, and are typical of between-species phylogenies. We used a tree scaling factor of 0.1, chosen to produce ample SNPs for successful phylogeny inference. We generated random DNA data on these trees with the rectreedna.c program (Felsenstein J., unpublished data) using a Kimura two-parameter model (Kimura 1980) with a transition/transversion ratio of 2.0. These programs are archived on Dryad at http://dx.doi.org/10.5061/dryad.g9089.

We began with simulated datasets of length 2000 bp for trees of 5, 10, or 20 tips, and then successively cut down the datasets 200 bp at a time to produce datasets of 1800, 1600, … , 200 bp. For each dataset, we inferred the maximum likelihood tree with PAUP* 4.0 (Swofford 2003) under the molecular-clock assumption and using the same mutational model used to simulate the data. When multiple trees tied for best, we arbitrarily chose the first tree listed. In practice, such ties mainly indicate the

presence of zero-length branches; although other kinds of ties are possible, they are rare.

While it is not guaranteed in any given case that the tree inferred from a long sequence will be more accurate than trees inferred from a shorter subsequence of it, maximum likelihood inference with a correct mutational model is known to be statistically consistent (Yang 1994). That is, it converges on the correct answer as the amount of data goes to infinity. Therefore, we expect that on average trees inferred from longer sequences will be more accurate, as long as we use the same mutational model to generate the data and to infer the tree. This is the rationale behind the bullseye experiment.

We found that small differences (200 or 400 bp) in dataset size were very difficult for any metric to detect. We therefore present results based on a span of 600 bp; that is, comparing inference of a particular tree at 2000 bp with its inference at 1400 bp, inference at 1800 with inference at 1200, and so on. Some sets of inferred trees were, therefore, used in two comparisons, leading to bumpiness in the graph.

Speed considerations prevented use of maximum likelihood inference for trees of more than 20 tips. We performed an analysis of trees of 50 tips using the same approach, but with the clocklike distance matrix method UPGMA as the inference algorithm, based on distance matrices calculated using the correct mutational model. PAUP* 4.0 was used to compute both distances and UPGMA trees. We randomized resolution of UPGMA ties, as the alternative of breaking ties by taxon input order spuriously increased the resemblance of inferred trees to each other.

Scaling factors.—Three of our metrics involve summing differences among lengths in the two trees: RFL, TripL, and Int. Statistics of this kind often sum either the absolute values of the differences or their squares, but in principle the differences could be raised to any power k before summing.

For each of our benchmarks, we did a preliminary experiment to determine good values of k for each of these three metrics. For all three metrics, the n-away experiment showed a monotonic relationship between k and success rate: the lowest value of k tested (0.1) was the most successful. These results are presented in Supplementary Materials S3–5 available on Dryad at http://dx.doi.org/10.5061/dryad.g9089. This occurs because in the n-away context the trees are known without error, and thus any discrepancy in the length of a branch is proof that that branch has been affected by a rearrangement. In fact, the optimal approach for n-away would be to score 0 for branches that have identical lengths and 1 for branches that have nonidentical lengths.

However, this "optimal"-k value would be catastrophic if there were any uncertainty in branch length measurement, as it would consider essentially all branches to be nonidentical. As expected, we found very different results for the bullseye experiment (Supplementary Materials S7-9 available on Dryad at http://dx.doi.org/10.5061/dryad.g9089). Intermediate values of k were preferred, though the exact results varied by metric, tree size, and tree type (coalescent or branching process). We chose to base our choice of k-values on the bullseye results to avoid the artificiality of branch lengths known without error. We examined the graphs in Supplementary Materials S7–9 available on Dryad at http://dx.doi.org/10.5061/dryad.g9089 and chose values that appeared optimal: 0.85 for RFL, 1.25 for Int, and 0.65 for TripL.

The Node metric has a scaling factor of k which is applied to differences in counts of intervening nodes rather than differences in branch lengths. Our preferred value was $k=2$ (as proposed by Penny et al. (1982)) but the choice of k made little difference for this metric (Supplementary Materials S6, S10 available on Dryad at http://dx.doi.org/10.5061/dryad.g9089). Interestingly, in contrast to the results for the three branch length metrics, larger values of k were preferred to smaller ones even in the n-away experiment.

Mammalian data.—Data were obtained from J. Felsenstein (personal communication). He selected DNA sequences for protein coding loci found on human chromosome 8 for which the OrthoMaM database (Ranwez et al. 2007) had sequences for all 40 of its included mammals. We took the first 10 of these loci in alphabetical order, namely ADAM7, ASH2L, CHMP4C, DERL1, ERLIN2, AP3M2, CA8, DEPTOR, E2F5, and FZD6. We also created a dataset by concatenating the sequences from these 10 loci. Loci varied in length from 702 bp to 2331 bp with a mean of 1317 bp. Optimal mutational models were chosen using jModelTest2 (Darriba et al. 2012), which finds the best model using maximum likelihood tree inference from PhyML (Guindon and Gascuel 2003). We considered 11 possible mutational models, each allowing or disallowing unequal allele frequencies, per-site rate variation and invariant sites (88 models in total) and chose the model preferred by Akaike's Information Criterion (AIC) for each locus separately and for the concatenated data. The preferred models were GTR+G+I (7 loci and the concatenated dataset) and GTR+G (3 loci). The individual loci and concatenated data were each used to infer clocklike maximum likelihood trees with PAUP* (Swofford 2003) under the model of highest AIC. When multiple trees were inferred the first tree was arbitrarily chosen. Each distance metric was used to compare the inferred tree for each locus to the inferred tree for the concatenated dataset.

RESULTS

It would seem natural to present the scores of each metric as a function of recombinational distance or dataset quality. However, the resulting graphs are dominated by the large differences in scale among the metrics. While some of the distances could be rescaled

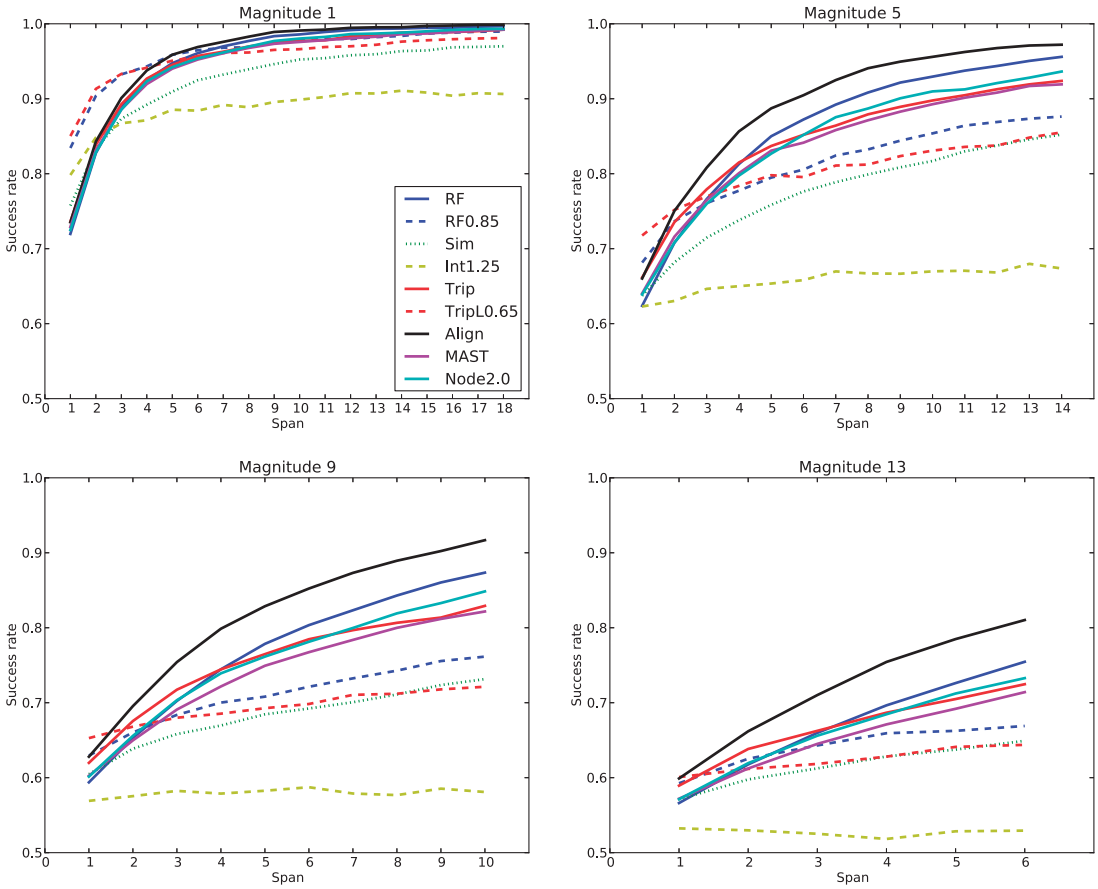


FIGURE 2. *N*-away experiment. Lines show the proportion out of 5000 trials in which the metric correctly ordered distances of magnitude and magnitude + span, with ties counted as half successes. Legend in first panel applies to all panels; abbreviations as given in Table 1. Solid lines show topology-only metrics and dashed lines show branch-length metrics.

for easier comparison, the length-using metrics have no natural maxima, and when we rescaled based on the highest observed value of the metric the results were dominated by noise in that value. We therefore focus on measuring the success of each metric in ordering trees of different distance or quality; this success measure is independent of the scale of the metric and, therefore, readily comparable across different metrics.

N-away benchmark.—The *n*-away experiment assessed how well each distance metric reflected recombinational distance. We test this by scoring how often the metric can correctly order two comparisons. For example, we can ask how often the metric correctly indicates that the distance between tree 0 (the origin) and tree 1 is less than the distance between tree 0 and tree 5.

In comparisons of this kind two numbers can be varied: The magnitude of the first distance in the comparison, and the span between the first and second distances. Magnitude indicates how well the metrics perform on relatively similar versus relatively dissimilar trees, whereas span indicates how good the metrics are at detecting small versus large differences. We organize our graphs by magnitude, and within magnitude by span.

Figure 2 shows, for a range of magnitudes, the ability of each metric to distinguish a distance of that magnitude from a distance a given span further away. For very short spans (comparisons among very similar distances) the metrics which use branch lengths (dashed lines) outperformed the purely topological metrics (solid lines). For comparisons with longer spans this advantage was rapidly reversed. The topology metric Align was consistently better than the others, especially for large magnitudes. The branch-length-only metric Int performed consistently poorly, especially for larger spans and/or magitudes, suggesting extreme vulnerability to saturation.

All subfigures of Figure 2 show sorting accuracy above the random level of 50% for all metrics. Since one concern about distance metrics, particularly RF, is that they may lose effectiveness at very large distances, we additionally tested recombinational distances from 20 to 38. Results, organized by span, are presented in Supplementary Figure S1 available on Dryad at http://dx.doi.org/10.5061/dryad.g9089. While metrics Int and Node have minimal ability to distinguish trees at these distances, all other metrics retain some ability to do so. Surprisingly, the RF metric can distinguish between the distance of 20 recombinations and 33

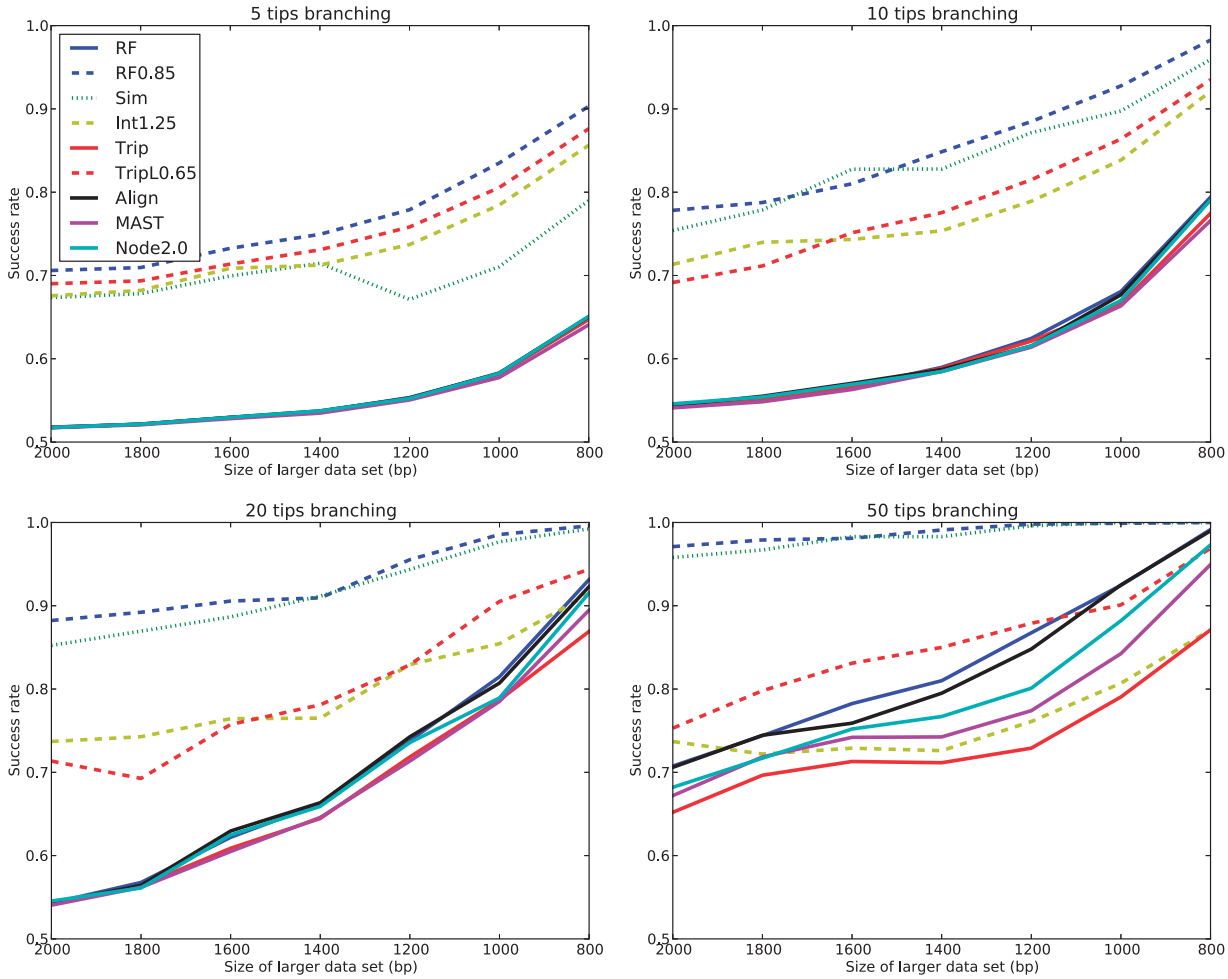


FIGURE 3. Bullseye experiment for branching-process trees. Lines show the proportion of trials in which the metric correctly ordered distances for comparisons of the true tree with trees inferred from the given sequence length and from a sequence 600 bp shorter, with ties counted as half successes. Legends as in Figure 2. Solid lines show topology-only metrics and dashed lines show branch-length metrics. Results for 5–20 tip trees based on maximum likelihood; results for 50-tip trees based on UPGMA. Sample sizes: 5 and 10 tips 10,000 trees; 20 tips 5955 trees; 50 tips 1000 trees.

recombinations with around 83% accuracy, showing that it is not saturated at 20 recombinations, even though it can separate 20 recombinations from 21 with only 55% accuracy (50% accuracy would be obtained from a random guess). The best-performing metric, Align, did even better. Trees can become surprisingly divergent while still retaining enough common features for relative similarity to be meaningfully assessed.

Bullseye benchmark.—The bullseye experiment assessed how well each distance metric tracked inference accuracy, where accuracy was varied by changing the amount of data available for inference. We considered both coalescent and branching process trees under the assumption that the different distribution of branch lengths might change the relative performance of the metrics, but the results were very similar. We present branching-process results in Figure 3 and coalescent results in Supplementary Figure S2 available on Dryad at http://dx.doi.org/10.5061/dryad.g9089.

In contrast to the *n*-away results, bullseye results were dominated by superiority of branch length metrics over topology-only metrics. The Sim metric had poorer performance for trees of five tips as cases with zero-length inferred trees had to be excluded and this apparently excluded cases on which other metrics were successful; otherwise it behaved similarly to the other branch-length metrics. By 20 tips, Int and TripL were lagging well behind the other branch-length metrics, and for the smallest datasets topology-only metrics were approaching their level.

The poor ability of topology-only metrics to distinguish larger from smaller data sets is straightforward to explain: often the same topology was recovered for both datasets, leaving no basis for preferring one to the other. Length-using metrics (including the length-only metric Int) had more information available.

The results shown for trees of 50 tips were based on UPGMA rather than maximum likelihood inference for

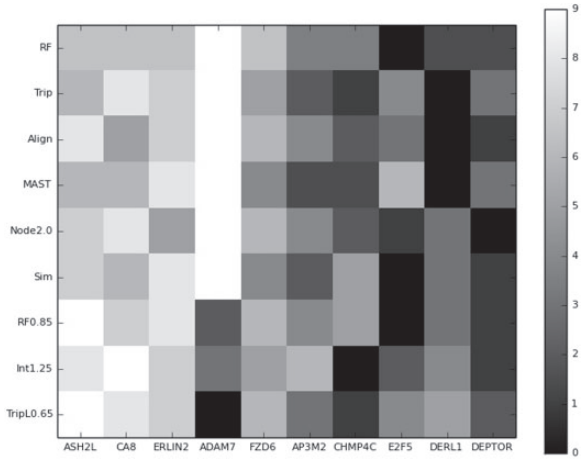


FIGURE 4. Heatmap showing relative ranking of 10 mammalian loci as compared with the overall dataset. Lighter shades indicate higher rank (the tree for that locus was relatively closer to the concatenated tree). Loci are listed in order of their mean rank across all metrics.

speed. (While very fast maximum-likelihood algorithms such as RAxML are available, they do not infer clocklike trees.) With 50 tips the RFL and Sim metrics correctly ordered better and worse trees almost all of the time; TripL and Int did much more poorly. The topology-only metrics did less well, with RF and Align offering the best performance.

Real data application.—We used each metric to compare the maximum likelihood inferred trees for 10 mammalian loci to the tree produced when the 10 sequences were concatenated (the "overall" tree). Figure 4 shows the ranking of the 10 loci by each metric, with lighter shades indicating a higher ranking (closer agreement between the inferred tree for this locus and the overall tree) and darker shades a lower ranking.

Locus ADAM7 shows a striking discrepancy. Every topology-only metric, as well as the Sim metric that is sensitive only to relative branch lengths, ranked the inferred tree for this locus as most similar to the overall tree. Every metric sensitive to absolute branch lengths ranked it among the least similar. The inferred height of the tree for this locus (0.377) was nearly twice the inferred height of the overall tree (0.211). Apparently ADAM7 is very similar topologically to the consensus of the 10 loci, and has branch lengths in the same proportions, but more substitutions per site.

For other loci, the rankings produced by the various metrics were much more similar: loci ASH2L, CA8, and ERLIN2 were broadly preferred by all metrics.

DISCUSSION

Our two simulation experiments gave somewhat discordant results. This can be understood by noting that the n-away experiment looked at trees separated by up to 11 rearrangements. The inferred trees of the bullseye experiment generally differed much less than this, often displaying exactly the same topology with slight differences in branch length. The bullseye results (Fig. 3) are, therefore, analogous to the leftmost entries in the *n*-away results (Fig. 2). In both bullseye and the leftmost *n*-away results, a definite advantage for length-using metrics is seen; length-using metrics can accurately classify trees as closer/further or better/worse when they are too similar to be distinguished by topology-only metrics. However, the *n*-away experiment as well as the rightmost portions of the bullseye experiment suggest that this pattern reverses for highly dissimilar trees.

Another key difference between the two experiments is that in *n*-away the trees were known without error. A branch that was undisturbed by rearrangement would, therefore, have exactly the same length in both trees. It is easy to develop specialized metrics for such perfect-knowledge cases that will fail catastrophically when branch length inference involves error (compare Supplementary Figures S3–5 with Supplementary Figures S7-9 available on Dryad at http://dx.doi.org/10.5061/dryad.g9089). The bullseye experiment, which includes realistic levels of error, favors intermediate values of the scaling parameter *k*. Apparently a too-high value of *k* loses information by overweighting the contribution of the few long branches in the tree, whereas a too-low value overinterprets noise in the short branches.

We were inspired by this observation to attempt to create tree distance metrics involving ratios of branch lengths rather than differences, but were not able to find a satisfactory metric of this type. The difficulty is that very short or, especially, zero-length branches lead to explosively large distances. However, there is a potential niche for a novel branch-length-using metric that downweights longer branches.

Interpretation of the mammalian dataset is impeded because we do not know the true topology or branch lengths. (The topology of the inferred concatenation tree groups marsupials with monotremes and is unlikely to be correct.) Most loci show broadly consistent ranks across metrics (Fig. 4). If ADAM7 were removed from consideration, loci ASH2L, CA8, and ERLIN2 suggest themselves as the best proxies for the overall data, and this view is shared by all metrics. However, the strongly divided rankings for ADAM7, which has a typical topology and proportions but atypical branch lengths, offer a reminder that the choice of metric can strongly influence the results of a study. If a locus were being sought to determine the topological relationships of additional mammals ADAM7 would be an excellent choice, and a topology-only metric would recognize this, as would Sim. If it were being sought as a proxy for genome-wide mutation rates it would be a poor choice, and branch-length metrics other than Sim would indicate as much.

Our practical conclusions from this study are:

(i) To differentiate similar trees, metrics using branch lengths are preferable to those using only topologies. For example, a researcher studying the relative effectiveness

of different phylogeny algorithms may find that a topology-only metric makes all algorithms look the same, whereas a length-using metric will allow them to be ranked. Among branch-length metrics RFL offered the best performance.

(ii) To detect relationships among highly dissimilar trees, metrics using only topologies are preferable to those using branch lengths. Among the topology-only metrics, Align had the best performance. Among branch-length metrics RFL was again the best, but it was inferior to all topology-only metrics. If a researcher was, for example, comparing host and parasite phylogenies that were only slightly correlated, a topology-only measure would be more useful than any branch-length measure.

(iii) The choice between topology-only and branch-length metrics should take into account the purpose for which the measurements are wanted. The sensitivity of branch-length metrics to small differences among trees is a valuable feature when such differences are important, but a mere distraction if only topological accuracy matters. For cases where branch-length proportions are important but absolute branch lengths are not, as in comparing mtDNA trees to nuclear-locus trees or genetic trees to linguistic trees, the Sim metric should be considered.

(iv) For the branch-length metrics, the value of the scaling parameter k does matter. Arbitrary choices of k as seen in Kuhner and Felsenstein (1994) should be avoided. In general, versions close to $k=1$ seem more useful than $k=2$.

(v) Triplet-based distances did not perform well, either with or without branch lengths. We speculate that this is because they are preferentially sensitive to the bottommost branchings in the tree (a large proportion of triplets contain these branchings). Alternatives to equal weighting of triplets should be tried.

The RF metric has been criticized for too rapidly reaching a plateau (Penny et al. 1982), but in this study it performed well, as did its branch-length variant RFL. Apparently neither recombinational rearrangement nor deteriorating tree inference easily drive RF to its maximum, even though the median RF distance between two random trees is in fact at the maximum. In contrast, the Trip and TripL distances, which might be expected to plateau more slowly, did not do particularly well. This, as well as the failure of the RFL $k=2$ measure (proposed by one of the authors on theoretical grounds in 1994) show that theoretical reasoning about distance metrics is not a good substitute for testing their performance.

An alternative to the use of any of these polynomial-time tree comparison metrics would be use of recently developed programs for fast approximate computation of the SPR distance (e.g., Goloboff 2007). Further study will be needed to judge SPR's performance relative to the methods tested here. Like other topology-only methods, SPR will not be useful in resolving extremely similar trees, but it may perform well for trees of intermediate dissimilarity. For highly dissimilar trees current SPR heuristics struggle to provide accurate distances (Goloboff 2007).

Program availability.—Unpublished programs used to simulate trees and DNA, and to calculate and graph distance metrics, are archived on Dryad at http://dx.doi.org/10.5061/dryad.g9089.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.g9089.

Funding

This work was supported by the US National Science Foundation (grant number DEB-1256731 to M.K.K.).

Acknowledgments

We thank Joseph Felsenstein and Kerry Bubb for helpful discussions, and Joseph Felsenstein for providing simulation code, giving permission to archive it, and providing curated mammalian data. David Posada, Hirohiso Kishino, and three anonymous reviewers provided useful suggestions on submitted forms of the manuscript.

References

Allen B.L., Steel M. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. Ann. Combinat. 5:1–15.

Bordewich M., Semple C. 2004. On the computational complexity of the rooted subtree prune and regraft distance. Ann. Combinat. 8: 409–423.

Critchlow D.E., Pearl D.K., Qian C. 1996. The triples distance for rooted bifurcating phylogenetic trees. Syst. Biol. 45:323–334.

Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: More models, new heuristics and parallel computing. Nat. Methods. 9:772.

Estabrook G.F., McMorris F.R., Meacham C.A. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst. Zool. 34:193–200.

Goddard W., Kubicka E., Kubicki G., McMorris F.R. 1994. The agreement metric for labeled binary trees. Math. Biosci. 123:215–226.

Goloboff P.A. 2007. Calculating SPR distances between trees. Cladistics 23:1–7.

Gordon A.D. 1980. On the assessment and comparison of classifications. In: Tomassine R. editor. Analyse de donnes et informatique, Le Chesnay, France: INRIA. P. 149–160.

Guindon S., Gascuel O. 2003. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. Syst. Biol. 52:696–704.

Hein J., Scierup M.H., Wiuf C. 2005. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford: Oxford University Press.

Hudson R.R. 1983. Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. 23:183–201.

Hudson R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Huson D.H., Rupp R., Scornavacca C. 2011. Phylogenetic networks: Concepts, algorithms and applications. New York: Cambridge University Press.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

Kingman J.F.C. 1982. On the genealogy of large populations. J. Applied. Prob. 19A:27–43.

Kuhner M.K. Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.
Li M., Zhang L.X. 1999. Twist-rotation transformations of binary trees and arithmetic expressions. J. Algorith. 32:155–166.
Nye, T.M.W., Lio P., Gilks W.R. 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 22:117–119.
Penny D., Foulds L.R., Hendy M.D. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature 297:197–200.
Ranwez V., Delsuc F., Ranwez S., Belkhir K., Tilak M., Douzery E. 2007. Orthomam: A database of orthologous genomic markers for placental mammal phylogenetics. BmC Evol. Biol. 7:241.
Robinson D.F., Foulds L.R. 1979. Comparison of weighted labeled trees. Lect. Note. Math. 748:119–126.
Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Steel M.A., Penny D. 2003. Distributions of tree comparison metrics–some new results. Syst. Biol. 42:126–141.
Swofford D.L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sunderland (MA): Sinauer Associates.
Williams W.T., Clifford H.T. 1971. On the comparison of two classifications of the same set of elements. Taxon 20: 519–522.
Yang Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. 43:329–342.
Yule G.U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil. Trans. Roy. Soc. Lond. B 213:21–87.