Phylogenetics

Information theoretic Generalized Robinson-Foulds metrics for comparing phylogenetic trees

Martin R. Smith

Department of Earth Sciences, Lower Mountjoy, Durham University, Durham, DH1 3LE, UK

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The Robinson-Foulds (RF) metric is widely used by biologists, linguists and chemists to quantify similarity between pairs of phylogenetic trees. The measure tallies the number of bipartition splits that occur in both trees—but this conservative approach ignores potential similarities between almost-identical splits, with undesirable consequences. 'Generalized' RF metrics address this shortcoming by pairing splits in one tree with similar splits in the other. Each pair is assigned a similarity score, the sum of which enumerates the similarity between two trees. The challenge lies in quantifying split similarity: existing definitions lack a principled statistical underpinning, resulting in misleading tree distances that are difficult to interpret. Here, I propose probabilistic measures of split similarity, which allow tree similarity to be measured in natural units (bits).

Results: My new information theoretic metrics outperform alternative measures of tree similarity when evaluated against a broad suite of criteria, even though they do not account for the non-independence of splits within a single tree. Mutual clustering information exhibits none of the undesirable properties that characterise other tree comparison metrics, and should be preferred to the RF metric.

Availability: The methods discussed in this paper are implemented in the R package 'TreeDist', archived at https://dx.doi.org/10.5281/zenodo.3528123

Contact: martin.smith@durham.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phylogenetic trees represent the history of bifurcating lineages, such as species, languages, or cancerous cells. The comparison of phylogenetic trees underpins a diverse set of scientific enquiries (summarized in Böcker *et al.*, 2013; Bogdanowicz and Giaro, 2013, 2017). Of the many methods that compare labelled phylogenetic trees with unweighted edges (e.g. Farris, 1973; Finden and Gordon, 1985; Meacham and Estabrook, 1985; Penny and Hendy, 1985; Steel and Penny, 1993; Bluis and Shin, 2003), the most widely used is the split-based symmetric distance (RF metric) of Robinson and Foulds (1981).

The RF metric quantifies the dissimilarity between a pair of trees by summing the number of splits (bipartitions of the set of leaves, corresponding to clades in rooted trees) that are unique to either tree. This uncompromising position overlooks potential similarities between non-identical splits, with the undesirable consequence that two trees that differ only in the position of a single leaf may exhibit the maximum RF distance.

'Generalized' RF metrics mitigate this issue by pairing similar splits between two trees. Each pair of splits is allocated a similarity score; the sum of scores when splits are optimally matched defines the overall similarity between two trees. Split similarity can be expressed as the size of the largest split consistent with two input splits (Bogdanowicz and Giaro, 2013, 2012). More sophisticated methods normalize this score against the Jaccard index (Nye *et al.*, 2006), or raise this normalized value to some arbitrary exponent (Böcker *et al.*, 2013). All such values lack a natural unit; depend on chosen exponent and the number of leaves; and fail to reflect similarities in split geometry.

I suggest that the information content (*sensu* Shannon, 1948) of the largest, or, better still, the most informative, split consistent with both input splits represents a more principled statistical definition of split similarity, measurable in non-arbitrary units (bits). Even simpler would be to directly measure the information that two splits hold in common,

using either the phylogenetic (*sensu* Steel and Penny, 2006) or clustering (Meilă, 2007; Vinh *et al.*, 2010) concepts of information. Each of these approaches gives rise to a novel generalized RF metric.

2 System and methods

Let T_1 and T_2 be unrooted phylogenetic trees with labelled leaves X. (Note that an unrooted tree can be converted to an equivalent rooted tree by attaching a dummy leaf to the root node.) Each edge of an unrooted tree divides the leaf set X into two subsets A and B, and can be expressed as a *split* $S = A|B$. A split is *trivial* if A or B contains fewer than two leaves; each non-trivial split in a tree corresponds to an internal edge. Two splits *conflict* if they cannot both occur in a single tree.

A *pairing* (S_1, S_2) contains one non-trivial split from each tree. A *matching*, m, is a set of pairings in which no split occurs more than once. The *score* of a matching is the sum of the scores allocated to each of its pairings. A matching is *optimal* if its score is the best possible for two trees; this score provides the overall similarity score for those trees.

2.1 Phylogenetic information

Define the *phylogenetic probability* of a split $S = A|B$, $P_{Phy}(S)$, as the probability that a uniformly chosen binary tree on X contains the split,

$$P_{Phy}(S) = \frac{(2|A| - 3)!!\,(2|B| - 3)!!}{(2|X| - 5)!!}$$

where the double factorial n!! = n × (n – 2)!!, with 1!! = 0!! = 1. The *phylogenetic information content* of a split is then $h(S) = -\log(P_{Phy}(S))$ (Steel and Penny, 2006). Base two logarithms yield units of bits.

The phylogenetic information content of a split can be used to score a pairing of a split in T_1 with an identical split in T_2. Assigning pairings of non-identical splits a score of zero yields an 'information-corrected' Robinson-Foulds (ICRF) similarity measure.

The Matching Split Distance (Bogdanowicz and Giaro, 2012) can be modified by scoring each pairing (S_1, S_2) with the phylogenetic information content of the most informative split congruent with both S_1 and S_2—namely, the most informative of ($A_1 \cap A_2$, $B_1 \cap B_2$) and ($A_1 \cap B_2$, $B_1 \cap A_2$). I term the score of the optimal matching the *Matching Split Information (MSI) score*.

Let $P_{Phy}(S_1, S_2)$ denote the probability that a uniformly chosen binary tree includes the splits S_1 and S_2. If $S_1 = S_2$, then $P_{Phy}(S_1, S_2) = P_{Phy}(S_1)$. If S_1 and S_2 conflict, $P_{Phy}(S_1, S_2) = 0$. Otherwise, without loss of generality, label S_1 and S_2 such that $A_1 \supseteq A_2$ (and consequently, $B_1 \subseteq B_2$). Any tree including S_1 and S_2 contains one edge that divides A_1 and B_1, and one edge that divides A_2 and B_2. Severing these edges (such that each becomes two edges, each terminating at a new, distinctly labelled, 'dummy' leaf) will result in three unrooted binary trees, whose leaves comprise: (i), B_1 plus a 'dummy' leaf ($|B_1| + 1$ leaves); (ii), A_2 plus a 'dummy' leaf ($|A_2| + 1$ leaves); (iii) the leaves that belong to neither B_1 nor A_2, plus two 'dummy' leaves ($|A_1| - |A_2| + 2$ leaves). Because there are $(2n - 5)!!$ unrooted binary trees on n leaves,

$$P_{Phy}(S_1, S_2) = \frac{(2(|B_1|+1) - 5)!!\,(2(|A_2|+1) - 5)!!\,(2(|A_1| - |A_2| + 2) - 5)!!}{(2|X| - 5)!!}$$

Shared phylogenetic information is the information common to S_1 and S_2, defined as $h_{shared} = 0$ if S_1 and S_2 conflict, $h(S_1) + h(S_2) - h(S_1, S_2)$ otherwise, where $h(S_1, S_2) = -\log(P_{Phy}(S_1, S_2))$. The *different phylogenetic information* $h_{different}$ is $h(S_1) + h(S_2)$ if S_1 and S_2 conflict, $h(S_1, S_2) - h_{shared}$ otherwise. Summing h_{shared} across each pairing in an optimal matching yields the *shared phylogenetic information (SPI) score*.

The SPI score measures how much the information shared between splits in a pair of trees narrows down the set of candidates that could be the historically 'true' tree, corresponding to the philosophy that phylogenetics seeks to reconstruct the single tree that accurately represents historical events.

2.2 Clustering information

A split $A|B$ is a bipartition and, hence, a *clustering* that divides leaves X into exactly two clusters, A and B. Let $P_{Cl}(A)$ denote the probability that a randomly selected leaf belongs to A, $P_{Cl}(A) = |A| \div |X|$, and $P_{Cl}(B)$ the corresponding probability for cluster B. The *entropy associated with S* is given by $-P_{Cl}(A) \log P_{Cl}(A) - P_{Cl}(B) \log P_{Cl}(B)$. The *mutual clustering information* (Meilă, 2007; Vinh *et al.*, 2010) between paired splits S_1 and S_2, $I_{Cl}(S_1;S_2)$, describes the extent to which knowledge of A_1 and B_1 reduces uncertainty regarding the composition of A_2 and B_2; that is to say, how much more likely is an observer to assign a leaf to the correct cluster in S_2 if they know which cluster it belongs to in S_1? This is given mathematically by:

$$I_{Cl}(S_1;S_2) = P_{Cl}(A_1, A_2) \log \frac{P_{Cl}(A_1, A_2)}{P_{Cl}(A_1)P_{Cl}(A_2)} + P_{Cl}(A_1, B_2) \log \frac{P_{Cl}(A_1, B_2)}{P_{Cl}(A_1)P_{Cl}(B_2)} +$$

$$P_{Cl}(B_1, A_2) \log \frac{P_{Cl}(B_1, A_2)}{P_{Cl}(B_1)P_{Cl}(A_2)} + P_{Cl}(B_1, B_2) \log \frac{P_{Cl}(B_1, B_2)}{P_{Cl}(B_1)P_{Cl}(B_2)}$$

where $P_{Cl}(A_1, A_2) = |A_1 \cap A_2| \div |X|$ denotes the probability that a point belongs to cluster A_1 in S_1 and to A_2 in S_2.

The *mutual clustering information (MCI) score* of two trees is the score of the optimal matching when each pairing (S_1, S_2) is assigned the score $I_{Cl}(S_1;S_2)$. The MCI score corresponds to a viewpoint that sees the goal of phylogenetics as reconstructing relationships between leaves; simply put, the metric measures the extent to which two trees agree on how leaves should be grouped.

2.3 Calculating distance from similarity

These information-based measures indicate the extent to which splits in one tree contain information about splits in the other; higher values signify more similar trees. A similarity score can be converted to a distance by subtraction from a maximum value, which may also be used to normalize the scores. There are several approaches to calculating such a maximum, not all of which satisfy the axioms of metric space.

The information content (or entropy, for mutual clustering information) of all splits in the least informative of two trees provides an upper bound on similarity, but subtracting a similarity score from this maximum would result in a distance score of zero when one tree contains only a subset of the splits in the other, despite the trees being non-identical, so does not yield a metric.

A more suitable maximum, by analogy with the variation of information metric (Meilă, 2007), is half the information content (or entropy) of all the splits in both trees; subtracting similarity scores from this value gives a distance that is trivially shown to satisfy $d(x, y) = 0 \Leftrightarrow x = y$ and $d(x, y) = d(y, x)$, and can be shown to satisfy the triangle inequality by lemma 3.2 of Bogdanowicz and Giaro (2012). I term the corresponding distance metrics to the MSI, SPI and MCI the matching split information distance (MSID), the phylogenetic information distance (PID), and the clustering information distance (CID).

In cases where one tree can be deemed 'correct'—for instance, studies that test the efficacy of phylogenetic methods by analysing datasets simulated on a known tree (Kuhner and Felsenstein, 1994; Smith,

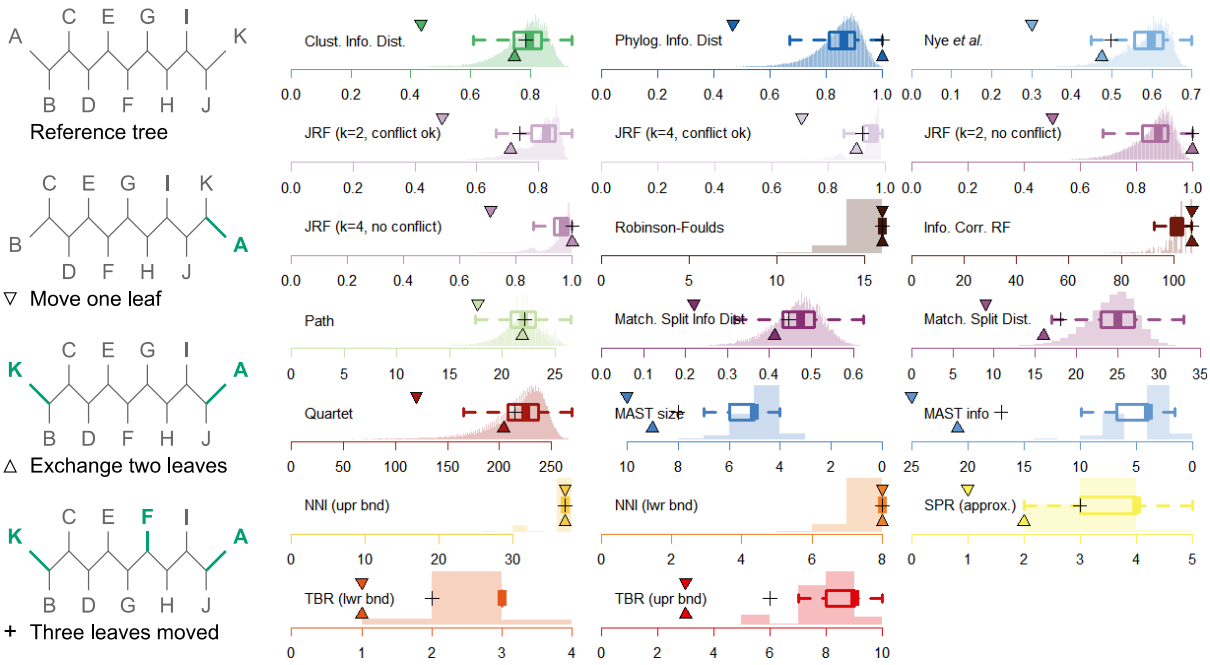

Figure 1. Trees produced by moving one leaf (▽), exchanging two leaves (△), and moving three leaves (+) should exhibit progressively greater distances from the pectinate starting tree, whilst being less distant than trees drawn at random. Histograms and box-and-whisker plots depict distances between the reference tree and 100 000 trees drawn at random from the uniform distribution of binary trees on the same leaves.

2019a), or comparisons between inferred trees and a trusted reference topology (Pompei *et al.*, 2011)—a metric may not be desired. Subtracting the similarity score from the information content of all splits in the 'correct' tree gives the distance from the 'correct' tree to a reconstructed tree, reflecting the amount of information that has been correctly reconstructed.

3 Results

I evaluated the performance of the information-based tree distance metrics against a suite of desiderata. The question of which tree distance method is 'best' depends somewhat on context, but my proposed criteria go some way to evaluate whether a metric is consistent, readily interpretable, versatile, and quick to calculate. I compared the information-based metrics to some popular alternative measures: the generalised RF metric of Nye *et al.* (2006); the Jaccard-Robinson-Foulds (JRF) metric (Böcker *et al.*, 2013), with $k = 4$ and $k = 2$ (for computational efficiency, 'arboreal' matchings were not imposed; pairings of conflicting splits were either permitted, 'conflict-ok', or prohibited, 'no-conflict'); the Matching Splits (MS) distance (Bogdanowicz and Giaro, 2012); the Quartet divergence (QD) (Estabrook *et al.*, 1985; Smith, 2019a); the path distance (Steel and Penny, 1993) (also termed the patristic distance, nodal distance, cladistic distance or tip distance); approximations to the NNI, SPR and TBR rearrangement distances (Li *et al.*, 1996; Hein, 1990); the size (MAST) and phylogenetic information content (MASTI) of the maximum agreement subtree; and the symmetric partition (RF) distance (Robinson and Foulds, 1981). Distances were calculated using the R packages 'phangorn', 'TreeDist', 'TBRDist', and 'Quartet' (Schliep, 2011; Sand *et al.*, 2014; Whidden and Matsen, 2017; Smith, 2019b, 2019c, 2020a); reproducible analyses are archived (Smith, 2020b). All analyses were conducted on unrooted binary trees, but can be generalised to rooted trees by designating one leaf as the 'root'.

3.1 Consistency

In this context, a *consistent* metric assigns higher distances to trees that are more different. Because phylogenetic trees occupy a non-Euclidian geometry, it can be difficult to rank the degree to which certain pairs of trees differ. Nevertheless, uncontroversial examples can be derived by modifying one tree to create a second: larger modifications should correspond to larger differences between an original and modified tree.

3.1.1 Length of move

A tree can be modified by repositioning a single subtree: the further a subtree is moved, the more different the resulting tree. I generated seventeen eleven-tip trees that differ only in the position of their eleventh leaf (the smallest possible subtree). The *move length* between any pair of these trees is defined as the number of nodes that must be traversed to travel from the position of the eleventh leaf in one tree to its position in the second. I define a *mis-ordering* as a case where the distance between a pair of trees is not strictly larger than the distance between every tree pair with a shorter move length, or a case where the distance is not strictly smaller than that between every tree pair with a greater move length.

In this test, the RF, ICRF, JRF, MSID and NNI distances contained no mis-orderings. Mis-orderings were increasingly frequent in the CID (8 of a possible 545), Nye et al. (24), PID (84), quartet (94), path (126) and MS (354) distances. Because the SPR, TBR and MAST distances allocate all non-identical tree pairs in this test an equal distance score, they are insensitive to move length.

Distance	MSID	PID	CID	Nye	Jco2	Jco4	Jnc2	Jnc4	RF	ICRF	MS	QD	MAST	NNI	SPR	TBR	Path
Length of move:																	
Mis-orderings (0–545)	0	84	8	24	0	0	0	0	0	0	354	94	480	480	480	480	126
No. leaves moved:																	
Inconsistent cases (0–289)	23	0	0	24	7	17	7	17	289	146	34	17	17	289	289	289	0
No. moves made:																	
1 < 2 < 3 moves	OK	2=3	OK	OK	OK	OK	2=3	2=3	1=2=3	1=2=3	OK	OK	OK	1=2=3	OK	1=2	OK
Saturation: 11-leaf trees with																	
max score (1–100 000)	1	36	1	2	1	1	36	36	86336	13044	11	1	1	86336	2388	7474	1
Sensitivity:																	
Distinct values (1–100 000)	24167	26478	28939	4381	27789	28781	19221	20488	6	208	28	200	7	16	5	7	302
Shape independence:																	
r^2	0.013	0.010	0.010	0.014	0.004	0.002	0.010	0.004	0.019	0.248	0.066	0.000	0.008	0.016	0.010	0.001	0.485
Leaf addition clusters:																	
Mean rank (1–14)	4.62	6.50	4.46	4.66	5.40	8.08	7.36	9.52	13.84	16.34	7.26	5.14	–	13.38	12.18	11.58	14.14
LLI clusters:																	
Mean rank (1–14)	5.45	13.75	7.10	4.15	5.00	6.75	11.48	10.83	10.98	16.03	3.83	9.55	–	11.15	9.23	11.40	11.90
SPR clusters:																	
Mean rank (1–14)	9.46	7.10	5.96	3.76	2.60	4.70	2.76	5.96	11.24	15.58	14.84	12.38	–	10.82	13.00	10.72	16.14
Cluster recovery:																	
Mean rank	6.51	9.15	5.84	4.19	4.30	6.51	7.20	8.77	12.02	15.98	8.64	9.02	–	11.78	11.47	11.23	14.06
Bullseye subsampling: Success																	
rate (50 leaves, 0–1000)	639	651	650	644	631	635	642	629	410	638	636	626	498	444	406	398	633
Bullseye miscoding: Success																	
rate (50 leaves, 0–1000)	902	937	941	961	955	931	965	929	781	785	777	809	622	802	613	693	704
Bullseye subsampling:																	
Accuracy (ρ)	0.69	0.71	0.71	0.70	0.70	0.67	0.71	0.67	0.60	0.64	0.67	0.62	0.52	0.60	0.33	0.44	0.62
Bullseye miscoding:																	
Accuracy (ρ)	0.91	0.94	0.94	0.94	0.94	0.93	0.94	0.93	0.90	0.75	0.79	0.86	0.79	0.90	0.72	0.84	0.60
Manual rearrangement:																	
Kendall's τ	0.78	0.80	0.85	0.80	0.78	0.75	0.77	0.73	0.67	0.50	0.70	0.82	0.80	0.84	0.73	0.85	0.57
Units	Bits	Bits	Bits	None	None	None	None	None	Arbi-trary	Bits	None	Bits	Arbi-trary	Operations			None
Random distances:																	
IQ range (% of median)	2.10	1.00	1.60	1.90	1.20	0.60	1.20	0.50	0.00	3.50	5.90	1.70	21.10	0.00	3.00	2.60	10.60
Polytomies	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	No	No	No	OK
Calculation speed:																	
20 leaves / µs	35	31	78	45	73	86	60	38	21	37	28	1800	5300	120	130	370	13
Calculation speed:																	
50 leaves / µs	190	110	410	200	430	1200	300	760	26	66	68	2100	23000	140	1000	710	20

Table 1. Evaluation of metrics against the desiderata summarised in the discussion. Darker greens denote better performance against each criterion. The best-performing NNI and TBR approximation was used in each case. Jco*X*, Jnc*X* denote JRF 'conflict ok' and 'no conflict' variants, with k = *X*.

3.1.2. Number of leaves moved

Moving a small subtree a specified distance should represent a smaller change than moving a large subtree the same distance (unless the larger subtree contains a sizeable proportion of the leaves in the tree). I added a pair of leaves adjacent to the eleventh leaf in each of the seventeen previous trees, to produce trees $T_1..T_{17}$. From these trees, I created two further sets of trees by moving either one leaf (the eleventh, for set $U_1..U_{17}$) or two leaves (the newly added pair, for set $V_1..V_{17}$) to each of the nineteen edges. I removed from each set $V_1..V_{17}$ the two trees that could alternatively have been generated by moving a single leaf, and also removed the corresponding trees in $U_1..U_{17}$. I consider a metric to be consistent where the distance from T_i to each tree in U_i (i.e. one leaf moved a certain distance) is strictly less than the distance from T_i to the corresponding tree in U_i (i.e. two leaves moved to the same location).

The PID, CID, 'conflict-ok' JRF and path measures are consistent in all 289 cases. Inconsistency is increasingly frequent in the 'no-conflict' JRF distance (7 cases, k = 2; 17, k = 4); the MAST and quartet distances (17); MSID (23), Nye (24), MS (34) and ICRF (146). The NNI, SPR, TBR and RF measures assign equal distances regardless of the number of leaves moved, so never satisfy this aspect of consistency.

3.1.3. Number of moves made

A single move of a certain length ought to result in a smaller difference than two moves of the same magnitude. As a simple example, using two moves to exchange a pair of leaves represents a larger change than using one move to reposition one leaf adjacent to the second—but the RF, ICRF, NNI and TBR metrics assign these changes the same score (Fig. 1). Moving a third leaf should further increase tree distance, but when that leaf lies between two that were previously exchanged, methods that prohibit pairings of conflicting splits do not recover higher distances (Fig. 1).

It is possible to broaden the scope of this criterion and employ the number of rearrangements as a measure of tree dissimilarity (e.g. the '*n*-away' of Kuhner and Yamato, 2015, and the SPR distance). However, because this approach does not consider the magnitude of individual

moves, it does not fully capture the difference between two trees: for example, one major rearrangement that moves many leaves a large distance may change a tree more profoundly than many minor rearrangements that move few leaves short distances.

3.1.4. Saturation

For most measures, fewer than 0.04% of 100 000 uniformly sampled 11-leaf trees are allocated the maximum distance from a pectinate reference tree (Table 1). The exceptions are the SPR (2.4%), TBR (7.4%), ICRF (13%), RF (86%) and NNI (86%) distances, which thus have a limited capacity to discriminate between very different pairs of trees.

3.1.5. Sensitivity

High resolution allows the discrimination of small differences in tree similarity. In a comparison of 100 000 uniformly sampled 11-leaf trees to a pectinate reference tree, fewer than eight distinct distance values were reported by the MAST, MASTI, NNI (lower bound), SPR, TBR and RF measures. Progressively more sensitivity was expressed in the MS (28 distinct values), Quartet (200), ICRF (210) and path (300) measures, but generalized RF distances displayed much greater sensitivity (Nye et al: 4 400 distinct values; JRF and information-based measures: 19 000–29 000 values; Table 1).

3.1.6. Independence from tree shape

For each of the four tree shapes on eight leaves, I labelled leaves at random until I had generated 100 distinct trees. I measured the distance from each tree to each of the other 399 trees, and fitted a linear model to evaluate what proportion of the variance between distances could be explained by the shape of the trees being compared. Tree shape has essentially no influence on the quartet distance, and makes a negligible (< 2%) contribution to most other distances (RF, JRF, PID, MSID, CID, Nye, MAST, NNI, SPR and TBR). However, tree shape accounts for 6.6% of the variance in the MS distance, 25% of the variance of the ICRF distance, and 48% of the variance in the path distance.

3.1.7. Consistent cluster recovery

Distance measurements allow clusters of similar trees to be identified. I tested each metric in its ability to recover clusters of similar trees generated using three approaches (Lin *et al.*, 2012). For the first test, I generated 500 datasets of 100 trees with $n = 40$ leaves. Each set of trees was created by randomly selecting two k-leaf 'skeleton' trees, where k ranges from $0.3\,n$ to $0.9\,n$. From each skeleton, 50 trees were generated by adding each of the remaining $n - k$ leaves in turn at a uniformly selected point on the tree. For the second and third test, each dataset was constructed by selecting at random two 40-leaf trees. From each starting tree, I generated 50 trees by conducting k leaf-label interchange (LLI) operations (test two) or k subtree prune and regraft (SPR) operations (test three) on the starting tree. An LLI operation swaps the positions of two randomly selected leaves, without affecting tree shape; an SPR operation moves a subtree to a new location within the tree.

For each dataset, I calculated the distance between each pair of trees. Trees were then partitioned into clusters using five methods, using the R packages 'stats' and 'cluster' (R Core Team, 2019; Maechler *et al.*, 2019): spectral clustering (von Luxburg, 2007), partitioning around medioids (Reynolds *et al.*, 2006), and hierarchical clustering using complete, single and average linkage (Stockham *et al.*, 2002; Lin *et al.*, 2012). I define the success rate of each distance measure as the proportion of datasets in which every tree generated from the same skeleton was placed in the same cluster. Figure 2 shows the success rate for each experiment, averaged across all five clustering methods. The Nye, CID and 'conflict-ok' JRF methods perform best across all three experiments (mean rank across all clustering methods and experiments: 3.5; 5.5; 3.7–6.1 / 20); the SPR, TBR, NNI, RF and path distances are consistently worst (mean rank > 10). The 'no-conflict' JRF and PID, which both penalize the pairing of conflicting splits, are in the bottom quartile of methods under this experiment, though they perform better than most methods under the other two experiments.

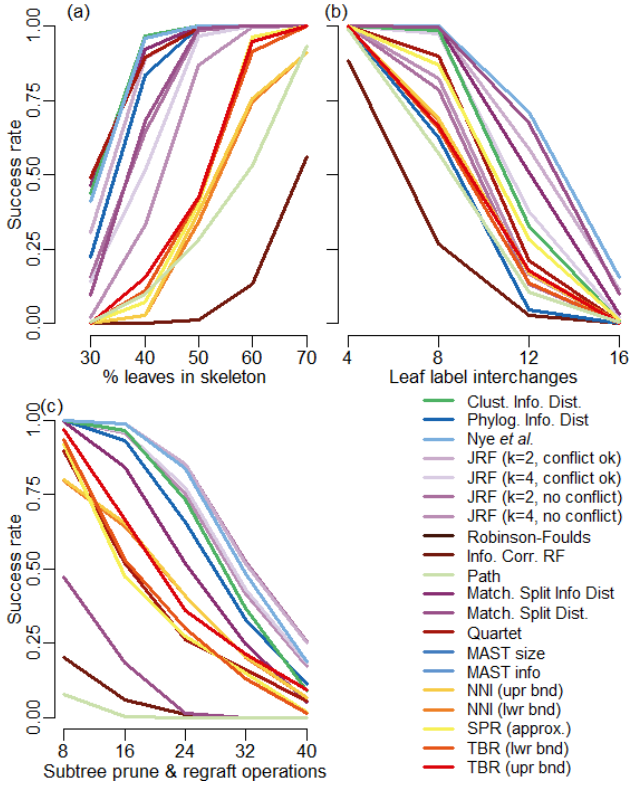

Figure 2. Cluster recovery results. Successful cluster recovery rate for each tree distance metric under (a), test one: add $40 - k$ leaves to two k-leaf skeletons; (b), test two: conduct k LLI operations on two 40-leaf skeletons; (c), test three: conduct k SPR operations on two 40-leaf skeletons.

3.1.8. Consistent with trees inferred from simulated data

I implemented 'bullseye' tests (Kuhner and Yamato, 2015) by drawing 1 000 n-leaf trees at random from the uniform distribution ($n = 5, 10, 20, 50$), and using these trees to simulate phylogenetic characters under the Jukes-Cantor model. I then inferred trees from increasingly degraded versions of these datasets in which either a subset of characters had been removed, or a subset of character state tokens modified. On the basis that a decrease in dataset quality produces a concomitant deterioration in the quality of inference, consistent distance metrics should rank trees inferred from datasets that are more degraded as being further from the reference tree (Kuhner and Yamato, 2015).

I used two approaches to simulating and degrading datasets. For the 'subsampling' approach, I simulated sequences of 2 000 base pairs (using 'phangorn' function simSeq()), and degraded matrices by deleting 200 base pair positions at a time, leaving sequences of 1 800, 1 600, …, 200 base pairs. For the 'miscoding' approach, I simulated 2 000 binary characters, and degraded matrices by switching the state of 2%, 4%, ...,

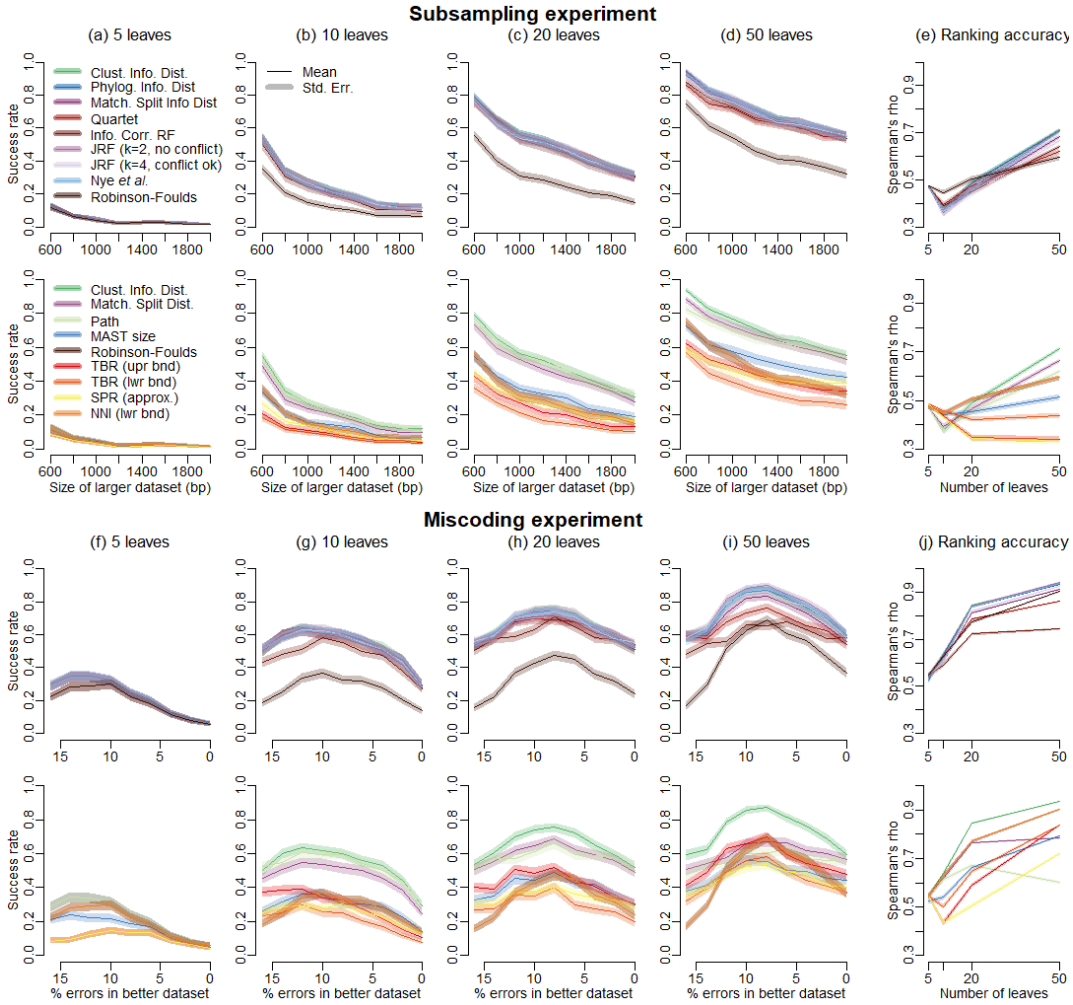

Figure 3. Results of 'bullseye' tests. (a–d), success rate for each tree distance metric when comparing pairs of datasets differing in size by 400 base pairs; (f–i), success rates when comparing pairs of datasets differing by 2% in proportion of erroneous tokens; (e, j), accuracy of each method in ranking trees inferred from progressively degraded datasets according to the degree of dataset degradation. Good methods are accurate, with high success rates.

18% of tokens (0↔1), selected randomly across leaves and characters. Trees were inferred from the resultant datasets using a parsimony optimality criterion with an implied weight concavity constant (Goloboff, 1993) of six. Tree search was conducted in TNT (Goloboff and Catalano, 2016, available with the sponsorship of the Willi Hennig Society) using the tree drift, tree fusing, sectorial search (Goloboff, 1999) and parsimony ratchet (Nixon, 1999) algorithms. Parsimony was chosen over model-based tree reconstruction methods due to its computational efficiency.

I used two methods to evaluate the performance of each distance measure. Firstly (after Kuhner and Yamato, 2015), I paired each dataset with a dataset that had 400 fewer characters ('subsampling' experiment), or 2 percentage points more tokens switched ('miscoding' experiment). For each pair, I recorded a success where a tree inferred from the more degraded dataset was further from the reference tree, and a failure otherwise. On this basis, the CID, PID, MSID, and JRF, quartet and ICRF metrics were equally good (within error) in the subsampling experiments (Fig. 3a–d); they were also the best methods in the miscoding experiment (Fig. 3f–i), though the MSID, quartet and ICRF distances displayed progressively worse performance in larger trees. The path and MS methods performed relatively well; the MAST, RF and rearrangement distances markedly worse.

Secondly, I used Spearman's rank correlation to compare the ranking of dataset quality with the ranking of distances of inferred trees. With larger trees (20+ leaves), the CID, PID, MSID, and JRF performed best in the subsampling experiment (Fig. 3e) and miscoding experiment (Fig. 3j).

3.1.9. Artificial tree rearrangement

I generated a chain of 100 50-leaf trees, starting from a pectinate tree and deriving each tree in turn by performing an SPR operation on the previous tree. A consistent measure of tree similarity should correlate with the number of SPR operations separating a pair of trees in this chain. This said, because one SPR operation may counteract some of the difference introduced by a previous one, perfect correlation is unlikely. A Kendall's rank correlation coefficient of $\tau_B > 0.8$ was obtained between the number of SPR operations and the CID, TBR, NNI (tight bound), quartet and PID measures (Table 1). τ_B exceeded 0.7 for the MAST, Nye, MSID, JRF, SPR and MS measures, and was lower for the RF and NNI (loose bounds) ($\tau_B = 0.67$), path (0.57) and ICRF (0.50) measures.

3.2 Readily interpretable

Although distances do not need meaningful units in order to rank pairs of trees in order of similarity, it is nevertheless desirable for the value of a tree distance metric to have a straightforward meaning (Bogdanowicz and Giaro, 2013) whose value is easily contextualized.

3.2.1. Units

The RF, JRF, MS, MAST, Nye and path distances employ arbitrary units, making their values difficult to interpret. The RF metric is particularly problematic, as its seemingly straightforward units are in fact biased to afford increased significance to splits that are less even, and thus more likely to be contradicted by chance (Bogdanowicz and Giaro, 2013; Smith, 2019a). Equivalently, the NNI, SPR and TBR distances can only be considered to have meaningful units ('number of rearrangement operations') if all rearrangements are treated as equal in magnitude.

The value of the MS distance is also difficult to interpret, because compatible splits (e.g. AB|CDEFGHI=ABCDE|FGHI, difference = 3) can be marked as more different than perfectly incompatible splits (AB|CDEFGHI=AE|BCDFGHI, difference = 2).

The quartet distance is more promising: each quartet statement specifies one of three equiprobable relationships between four leaves and thus represents $\log_2(3)$ bits of information—notwithstanding double-counting of information arising when non-independent quartets are treated as independent. Even if the absolute value of the quartet distance has been considered confusing, its normalized value is simple to comprehend—it reflects the probability that the relationships between four randomly selected leaves are the same on both trees, and thus has a clearly defined range ([0, 1]) and expected value (⅓).

Explicitly information-based measures are also measured in bits; this natural unit corresponds to an equivalent degree of difference for any pair of trees, no matter their number of leaves. This said, the non-independence of splits within a given tree leads again to some double-counting of information. The mutual information between entire trees, rather than one split at a time, would be more meaningful still—if only it were readily calculated.

3.2.2. Context

In order to evaluate whether a given distance is large or small, it is necessary to understand the range of values that a metric can take. In most circumstances, the expected value of a pair of random trees provides better context than the maximum distance that is theoretically possible: it is more common to ask how likely a certain tree difference is to have arisen by chance than how far it lies from a maximum value.

The expected similarity of a pair of trees sampled from the uniform distribution can only be calculated exactly for the quartet metric. I approximated the expected similarity for other distances by taking the median distance between 1 000 pairs of uniformly sampled *n*-leaf trees. For most measures, random tree pairs have a score very close to the median (interquartile range < 4% of median value). The exceptions are the MS, MAST, path and MASTI distances, whose wider interquartile range (respectively 5.9%, 8.5%, 11%, and 14% of median value) makes it difficult to evaluate the meaning of these metrics with reference to the expected value for a random tree pair. Moreover, as the diameters of the MS and path metrics cannot be easily calculated, their absolute values are also difficult to interpret.

3.3 Versatile

Rearrangement distances are not defined on non-binary trees; other metrics can compare trees that contain polytomies, and are thus more versatile.

3.4 Quick to calculate

Where large numbers of tree comparisons are required, computational efficiency is important. I recorded the time taken to calculate the distance between 990 pairs of 20-leaf and 50-leaf trees in R on a desktop computer with 8.47 GB of RAM and an Intel Core™ i7-3770 3.40GHz CPU. Starting from a pectinate tree, I conducted 44 SPR operations, recording the tree after each operation, and compared each pair of non-identical trees; the resultant 990 comparisons thus include tree pairs with both small and large distances. The fastest methods were the path (mean from ten iterations: 20 leaves, 13 µs; 50 leaves, 20 µs), RF (21 & 26 µs) and MS (28 & 68 µs); all other methods had a run time of ≤ 80 µs (20 leaves) and ≤ 500 µs (50 leaves) except the JRF with $k = 4$ ('no-conflict': 38 & 760 µs; 'conflict-ok': 86 & 1 200 µs), rearrangement (NNI: 120 & 120 µs; SPR: 130 & 1 000 µs; TBR: 370 & 710 µs), quartet (1 800 & 2 100 µs) and MAST (5 300 & 23 000 µs) distances.

4 Evaluation

Most tree distance metrics exhibit minimal correlation with one another (adjusted $r^2 < 0.3$; Supplementary Figs 1–2), indicating that different measures capture different aspects of tree similarity and encounter different biases and errors. In particular, the widely used Robinson-Foulds and SPR metrics exhibit counterintuitive behaviour in a wide range of situations, whereas the path distance performs particularly poorly in practical applications (Table 1). The use of these unreliable and potentially misleading measures of tree similarity should be discouraged.

Correlation is somewhat higher between generalized Robinson-Foulds distances, reflecting their shared approach of matching similar splits. Jaccard-Robinson-Foulds distances converge on the RF distance as k increases, and, where pairings of conflicting splits are permitted and $k \to 1$, on the Nye *et al.* distance. In turn, the Nye *et al.* method correlates closely (adjusted $r^2 > 0.7$) with the information-based MSID, PID and CID distances, which each avoid the issues that distort the RF metric. Because the PID does not recognize any similarity between conflicting splits, it still results in counterintuitive tree distances in certain cases. As such, the CID and MSID are the only natural measures of splitwise tree similarity to display all the expectations of behaviour considered herein (Table 1). They are continuous, and thus capable of unbridled precision; they are difficult to saturate, with actively contradictory trees receiving higher distance scores than random trees; they are readily normalized; and partition size effects are explicitly accounted for by the information-theoretic underpinning. Pragmatically, they are among the most consistent methods for ranking the similarity of trees inferred from simulated datasets. The CID marginally outperforms the MSID on each criterion considered herein, justifying a slightly longer calculation time. As such, I recommend mutual clustering information as the most appropriate measure of tree similarity based on matching splits between trees, and its complement, the clustering information distance, as an intuitive and meaningful metric for tree distance.

Acknowledgements

The manuscript was much improved through the comments of T.M.W. Nye and three anonymous referees.

Conflict of Interest: none declared.

References

Bluis,J. and Shin,D.G. (2003) Nodal distance algorithm: Calculating a phylogenetic tree comparison metric. *Proceedings - 3rd IEEE Symposium on BioInformatics and BioEngineering, BIBE 2003*, 87–94.

Böcker,S. *et al.* (2013) The generalized Robinson-Foulds metric. In, Darling,A. and Stoye,J. (eds), *Algorithms in Bioinformatics. WABI 2013. Lecture Notes in Computer Science, vol 8126*. Springer, Berlin, Heidelberg, pp. 156–169.

Bogdanowicz,D. and Giaro,K. (2017) Comparing phylogenetic trees by matching nodes using the transfer distance between partitions. *Journal of Computational Biology*, **24**, 422–435.

Bogdanowicz,D. and Giaro,K. (2012) Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9**, 150–160.

Bogdanowicz,D. and Giaro,K. (2013) On a matching distance between rooted phylogenetic trees. *International Journal of Applied Mathematics and Computer Science*, **23**, 669–684.

Estabrook,G.F. *et al.* (1985) Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, **34**, 193–200.

Farris,J.S. (1973) On comparing the shapes of taxonomic trees. *Systematic Zoology*, **22**, 50.

Finden,C.R. and Gordon,A.D. (1985) Obtaining common pruned trees. *Journal of Classification*, **2**, 255–276.

Goloboff,P.A. (1999) Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics*, **15**, 415–428.

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

Goloboff,P.A. and Catalano,S.A. (2016) TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics*, **32**, 221–238.

Hein,J. (1990) Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*, **98**, 185–200.

Kuhner,M.K. and Felsenstein,J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, **11**, 459–468.

Kuhner,M.K. and Yamato,J. (2015) Practical performance of tree comparison metrics. *Systematic Biology*, **64**, 205–214.

Li,M. *et al.* (1996) Some notes on the nearest neighbour interchange distance. In, Cai,J.-Y. and Wong,C.K. (eds), *Computing and Combinatorics*. Springer, Berlin, Heidelberg, pp. 343–351.

Lin,Y. *et al.* (2012) A metric for phylogenetic trees based on matching. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **4**, 1014–1022.

von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat Comput*, **17**, 395–416.

Maechler,M. *et al.* (2019) cluster: Cluster Analysis Basics and Extensions. *Comprehensive R Archive Network*, **2.1.0**.

Meacham,C.A. and Estabrook,G.F. (1985) Compatibility methods in systematics. *Annual Review of Ecology and Systematics*, **16**, 431–446.

Meilă,M. (2007) Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, **98**, 873–895.

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

Nye,T.M.W. *et al.* (2006) A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, **22**, 117–119.

Penny,D. and Hendy,M. (1985) The use of tree comparison metrics. *Systematic Zoology*, **34**, 75–82.

Pompei,S. *et al.* (2011) On the Accuracy of Language Trees. *PLoS ONE*, **6**, e20109.

R Core Team (2019) R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria.

Reynolds,A.P. *et al.* (2006) Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *J Math Model Algor*, **5**, 475–504.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

Sand,A. *et al.* (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.

Shannon,C.E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423, 623–656.

Smith,M.R. (2019a) Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. *Biology Letters*, **15**, 20180632.

Smith,M.R. (2019b) Quartet: comparison of phylogenetic trees using quartet and bipartition measures. *Comprehensive R Archive Network*, doi:10.5281/zenodo.2536318.

Smith,M.R. (2019c) TBRDist: Calculate Rearrangement Distances Between Unrooted Phylogenetic Trees. *Comprehensive R Archive Network*, doi:10.5281/zenodo.3548333.

Smith,M.R. (2020a) TreeDist: Distances between Phylogenetic Trees. *Comprehensive R Archive Network*, doi:10.5281/zenodo.3528123.

Smith,M.R. (2020b) TreeDistData: Analysis of Phylogenetic Tree Distance Measures. *Zenodo*, doi:10.5281/zenodo.3697901.

Steel,M.A. and Penny,D. (1993) Distributions of tree comparison metrics—some new results. *Systematic Biology*, **42**, 126–141.

Steel,M.A. and Penny,D. (2006) Maximum parsimony and the phylogenetic information in multistate characters. In, Albert,V.A. (ed), *Parsimony, Phylogeny, and Genomics*. Oxford University Press, Oxford, pp. 163–178.

Stockham,C. *et al.* (2002) Statistically based postprocessing of phylogenetic analysis by clustering. *Bioinformatics*, **18**, S285–S293.

Vinh,N.X. *et al.* (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, **11**, 2837–2854.

Whidden,C. and Matsen,F.A. (2017) Calculating the Unrooted Subtree-Prune-and-Regraft Distance. *preprint at arXiv*, 1511.07529v3.