Information Content of Trees: Three-taxon Statements Inference Rules and Dependency

Valentin Rineau, René Zaragüeta, Jérémie Bardin

[1]Valentin Rineau*, [2] René Zaragüeta and [3]Jérémie Bardin

[1] *Center for Theoretical Study, Charles University & Czech Academy of Sciences, Jilská 1, 110 00 Praha 1, Czech Republic.*

[2] *Sorbonne Université, CNRS, MNHN, EPHE, Université des Antilles, Institut de Systématique, Evolution, Biodiversité, ISYEB, CP 48, 57 rue Cuvier, 75005 Paris, France.*

[3] *Sorbonne Université, Centre de Recherche en Paléontologie – Paris (CR2P), 4 place Jussieu, barre 46-56 5ème étage, case 104, 75005 Paris, France.*

*Corresponding author. Email: valentin.rineau@gmail.com.

ABSTRACT

The three-taxon statement (also called triplet) is the fundamental unit of rooted trees in phylogenetic systematics. Various supertree and phylogenetic methods use three-taxon statements that are minimal rooted statements of degree of kinship relationships. Because of their fundamental role in phylogenetics, three-taxon statements are present in methodological research of various disciplines in evolutionary biology, as in consensus methods, supertree methods, species-tree methods, distance metrics, phylogenetics, and cladistic biogeography. Three-taxon statements are thus widely used. However, their theoretical properties have been poorly investigated. As a result, three-taxon statements methods are subject to important flaws related to information redundancy. Correcting these biases is essential to improve the efficiency of methods using three-taxon statements. Our aim is to study the behavior of three-taxon statements and the interactions among them in order to enhance their performance in phylogenetic studies. We have identified new types of very specific interactions between three-taxon statements responsible of the emergence of redundancy and dependency in trees. We propose for the first time a classification of three-taxon statements interactions and trace the link between those and the emergence of dependency and redundancy. A new fractional weighting procedure for suppressing redundancy of three-taxon statements is proposed. Our method is subsequently empirically tested in the supertree framework using simulations. We show that three-taxon statements using fractional weights perform drastically better than classical supertree methods such as MRP or methods using unweighted three-taxon statements. Our study shows that appropriate fractional weighting of three taxon statements is an efficient measure of phylogenetic information content for rooted trees. Fractional weighting is of critical importance for removing redundancy in any method using three-taxon statements, as in consensus, supertrees, distance metrics, and phylogenetic or biogeographic analyses.

Keywords: fractional weighting, information content, three-taxon statements, triplets, cladistics, phylogenetic analysis, supertree methods, consensus trees.

Phylogenetic trees are at the heart of most evolutionary biology studies and their reliability is of critical importance. It has long been recognized in the phylogenetic setting the need to understand and manipulate the information contained in phylogenies. Any tree (taxon tree, gene tree, biogeographic tree, etc.) can be fruitfully decomposed into subunits of several kinds (Adams 1986; Vach 1994), as components (rooted subtrees with all taxa and a single informative node; Nelson & Platnick 1981, Wilkinson et al. 2004), three-taxon statements (3ts thereafter; rooted subtrees with three taxa and a single informative node; Adams 1986), splits (bipartitions of a set of taxa; Farris 1970), or quartets (bipartitions of 4 taxa; Strimmer and von Haeseler 1996). These subunits are traditionally considered in the phylogenetic framework as information bricks (e.g. Mickevich & Platnick 1989; Nelson 1979; Nelson & Ladiges 1992; Wilkinson 1994a), and can be used to measure the information content of a phylogenetic tree.

Here we focus on rooted trees. Within a rooted tree, the 3ts (see Nelson & Ladiges 1991a, b; Nelson & Platnick 1991) take a special place as they represent the minimal (rooted) phylogenetic information. They have also been called “triad” in Adams (1986) and Vach (1994), “triplet” in Wilkinson (1994a), or still “rooted triple” in (Bryant 2003) in the consensus and supertree literature. A 3ts is a statement in which given three distinct taxa, two are related compared to a third one. For example, given three taxa *a*, *b* and *c*, *a* and *b* are more closely related to each other than either is to *c*: (*a*,*b*),*c*.

Because of their atomic nature, 3ts are at the heart of many methods. For example, the method of phylogenetic reconstruction commonly referred to as three-taxon analysis (Nelson and Ladiges 1991b; Nelson and Platnick 1991) decomposes hierarchical characters into 3ts matrices and then selects the cladograms that are congruent with the maximal amount of those 3ts (Zaragüeta et al. 2012). The same principle is applied in cladistic biogeography and the construction of areagrams (i.e. cladograms of biogeographic areas; Nelson and Ladiges 1991a). The use of 3ts in the supertree framework was first proposed by Williams (2004). Currently, several supertree methods (Ranwez et al. 2010; Sevillya et al. 2016; Dannenberg et al. 2019; Mavrodiev et al. 2019) explicitly handle 3ts. Another use of 3ts is also for consensus tree representations (Aho et al. 1981; Adams 1986; Nelson & Ladiges 1994; Wilkinson 1994a; Bryant 2003; Cao et al. 2009). The 3ts permit to precisely quantify the information content of a phylogenetic structure (Mickevich and Platnick 1989; Nelson and Ladiges 1992; Williams and Humphries 2003). The retention index based on 3ts (Kitching et al. 1998) allows measuring the proportion of hierarchical relationships that are found in the optimal cladograms. 3ts-based distance metrics have been used to compare topologies among sets of phylogenetic trees (Grand et al. 2013; Kuhner and Yamato 2015). Finally, 3ts are also explicitly used for coalescent-based species tree estimation from gene tree (Liu et al. 2010; Islam et al. 2020), and for phylogenetic network reconstruction (Poormohammadi et al. 2020).

The use of 3ts requires their independence. Nelson and Ladiges (1992) showed that the independence among a 3ts set could be achieved by using a fractional weighting to remove redundancy. Wilkinson et al. (2004) criticized this procedure by showing that certain relationships among 3ts generating dependency were not taken into account in the computation of fractional weighting. However, these relationships were neither clearly defined nor formalized in their work. A thorough work on relationships among 3ts in order to propose new operational solutions to remove dependency becomes essential today. The goal of this paper is dual: i) to develop the relationship between dependency and relationships among 3ts and ii) to propose a weighting method to improve the quality and accuracy of phylogeny analyses using 3ts.

In the first part of this paper, we summarize a historical review of the procedures proposed in the literature to take into account dependency in trees. We highlight main challenges and discuss the solutions that have been proposed in the past. In the second section, we demonstrate the link between tree-shape and dependency within phylogenetic trees. We recognize different types of relationships between internal nodes that lead to dependency. In the third part, we exhaustively describe, formalize, and classify the different relationships among couples of 3ts, and how to accordingly combine 3ts in trees or decompose trees in 3ts. Our work on tree-shape and on 3ts is then used in a fourth part to propose a new procedure called *corrected fractional weighting* that removes all types of dependency between 3ts. This procedure is needed to fulfil the requirement of independence on which are based most of the analyses using phylogenetic information. Finally, we used simulated datasets to compare different supertree methods that use different representation of phylogenetic information and show that methods using fractional weighting performs significantly better.

Dependency and Weighting 3ts: an historical review

The use of 3ts as a unit of measurement for the information content of phylogenies requires that they be independent of each other (i.e. absolute independence *sensu* Wilkinson et al. 2004). Performing an analysis based on non-independent data leads to biases that can have important consequences on the quality of the result. Independence between two relationships implies that the truth of one does not involve the truth or falsity of the other. The 3ts *a*(*bc*) and *a*(*bd*) are independent because the truth of one tells nothing about the truth of the other. On the other hand, the 3ts $a(bc)$, $a(bd)$, and $a(cd)$ are not independent relative to each other, because the truth of two of these 3ts necessarily implies the truth of the third (Nelson and Ladiges 1992):

$$a(bc) + a(bd) + a(cd) \rightarrow a(bcd)$$

$$a(bc) + a(bd) \rightarrow a(bcd)$$

$$a(bc) + a(cd) \rightarrow a(bcd)$$

$$a(bd) + a(cd) \rightarrow a(bcd)$$

Relationships among 3ts may lead to dependency and therefore to redundancy. In the given example, the information content of these three 3ts cannot be 3; since the information given by the three 3ts is identical to that given by two of them, there is a repetition of information that leads to the overvaluation of 3ts.

Fractional weighting (FW; Nelson and Ladiges 1992) was the first attempt to correct the overweighting of 3ts. Since any two 3ts are needed among the three, the amount of information carried by each 3ts is not one but 2/3. The corrected value of the set is 2. A component containing t taxa of which n are included in the informative node contains $n(t-n)(n-1)/2$ 3ts, but only requires $(n-1)(t-n)$ 3ts to be reconstructed. This quantity $(n-1)(t-n)$ corresponds to the phylogenetic value of a component sensu Nelson and Ladiges (1992). In this case, the value of a 3ts is $2/n$, i.e. the number of independent 3ts divided by the total number of 3ts.

The presence of several nested internal nodes implies another type of dependency. The first attempt of fractional weighting by Nelson and Ladiges (1992) was applied component by component, herein *fractional weighting per component* ($\text{FW}_{\text{comp.}}$) did not take into account the dependence. Indeed, in FW$_{\text{comp.}}$, the weight of a 3ts present in several components equals the sum of its weights for each component in which it appears. The tree *a*(*b*(*cd*)) can be decomposed into two components and then into 3ts:

a(*b*(*cd*) → *a*(*bcd*) + *ab*(*cd*)

a(*bcd*) → *a*(*bc*) + *a*(*bd*) + *a*(*cd*) (weight for each 3ts: 2/3)

ab(*cd*) → *a*(*cd*) + *b*(*cd*) (weight for each 3ts: 1)

The information content of the tree is 4. 3ts weights are subject to redundancy: 3ts *a*(*cd*) is overweighted (5/3) because it is present in both components *a*(*bcd*) and *ab*(*cd*) while it is present only once in *a*(*b*(*cd*)). For this reason, Nelson and Ladiges (1992: 491) describe FW$_{\text{comp.}}$ and *Uniform Weighting* (sum of 3ts per component, see table I) as "*both misleading*". A 3ts cannot be present more than once in a tree, and therefore must not have a weight greater than 1. The table I shows another problematic example with the tree *a*(*bc*(*de*)). To overcome this difficulty, Nelson and Ladiges (1992) propose a different way to calculate the absolute value of a 3ts called herein *total fractional weighting* (FW$_{\text{total}}$). It is a reappraisal of the idea that the weight of a 3ts is the number of independent 3ts divided by the total number of 3ts, but at the scale of the whole tree, by summing the weights of 3ts for each component: "*For each of their cladograms, we list the number of independent three-taxon statements and the absolute value of each of all possible statements, stated as the ratio of independent statements to all possible statements*" (Nelson and Ladiges 1992: 491). The FW$_{\text{total}}$ becomes for a 3ts the sum of all independent 3ts for each component / the sum of all 3ts for each component. For C components deducible from a phylogenetic tree:

$$3is_{total\ FW} = \frac{\sum_{i=1}^{C}(n_i - 1)(t - n_i)}{\sum_{i=1}^{C}\frac{(t - n_i)(n_i - 1)n_i}{2}} = \frac{2}{\sum_{i=1}^{C} n_i}$$

As a result, all the 3ts of a tree have the same weight. The four 3ts of the tree *a*(*b*(*cd*)) have a weight of 1. Nelson and Ladiges conclude by considering this calculation of the absolute value of a 3ts as the most relevant (Nelson and Ladiges 1992: 494).

Wilkinson et al. (2004) gave a new attempt to describe the dependencies between 3ts using Dekker's (1986) dyadic inference rules. According to Dekker, only two inferences are possible from pairs of quartets to deduce additional quartets. By transposing these inference rules to 3ts, Wilkinson et al. (2004: 994) propose three rules for inferring 3ts exemplified as follows:

b(*cd*) + *b*(*ce*) → *b*(*de*)

b(*cd*) + *c*(*de*) → *b*(*ce*)

d(*ab*) + *a*(*de*) → *e*(*ab*)

The original Dekker rules imply that any tree containing two quartets involved by the dyadic inference rules also contains an additional quartet. Wilkinson et al. (2004) point out that for 3ts, several rules can intervene from a single couple. For example:

b(*cd*) + *c*(*de*) → *b*(*ce*)*

b(*cd*) + *b*(*ce*)* → *b*(*de*)

Which results in:

b(*cd*) + *c*(*de*) → *b*(*c*(*de*)) → *b*(*cd*) + *c*(*de*) + *b*(*ce*) + *b*(*de*) (fig. 4b in Wilkinson et al. 2004)

This implies that the presence of some couples of 3ts can imply the truth of two secondary 3ts. For Wilkinson et al. (2004), this type of dependence is not taken into account by Nelson and Ladiges' fractional weighting. Since the tree *b*(*c*(*de*)) needs only two 3ts to be rebuilt, its weight should be 2. Consequently, the FW_{total} can only take into account one type of dependence corresponding to the first rule presented above. These rules lead to the problems of dependency inside a set of 3ts; the purpose of weighting being to suppress them.

However, Wilkinson et al. do not give all the possible combinations that lead to additional 3ts. This is justified by the fact that any fully resolved tree with *t* taxa can be fully entailed by specific sets of *t*-2 3ts (Steel, 1992). Consequently, any dichotomous tree includes only *t*-2 independent 3ts, each triplet having a weight of 6 / (N^2 - N). Wilkinson et al. named this procedure *Minimal Weighting* (MW), which aimed to take into account all possible inference rules without having to define them. The arguments put forward in favor of the MW are: (i) an identical weight for all 3ts because there is no argument to weight differentially the 3ts from a tree, (ii) a linear evolution of the total weight of a tree with the number of taxa (polynomial with $FW_{comp.}$ and FW_{total}), and (iii) an evolution of the total weight of a tree that is not influenced by tree shape.

Dependency and Tree Shape

Our aim being to find a way of deleting dependency issues in phylogenetic trees, we must define in a first place where the dependency can arise. Here we show for the first time how the tree shape can modify the dependency between nodes, and by corollary altering the information content of phylogenies. To fulfil this aim, we need to precisely define the various types of nodes of phylogenetic trees. There are three types of internal nodes : apical nodes, orthologous nodes, and paralogous nodes.

Definition 1. *Apical nodes* are nodes that only include leaves. The tree *b*(*c*(*de*)) contains only one apical node, (*de*).

Definition 2. *Orthologous nodes* (sensu Zaragüeta et al. 2004, also called asymmetric or unbalanced nodes) include only one internal node as direct descendant. In the tree *b*(*c*(*de*)), the node (*cde*) is orthologous because it leads only to the apical node (*de*), and the root node (*bcde*) because it immediately leads only to (*cde*).

Definition 3. *Paralogous nodes* (sensu Zaragüeta et al. 2004, also called symmetric or balanced nodes) are nodes that include two or more internal nodes as direct descendants. The tree (*ab*)(*cd*) contains one paralogous node, the root (*abcd*), which contains two internal nodes (*ab*) and (*cd*).

The relationships between internal nodes in the tree can be linked to that of dependence. Orthologous nodes show dependency where a node is differentiated in another node, while paralogous nodes show occurrences of independence, where a node is divided into several independent and separate lineages. Dependency, i.e. inclusion, is a kind of information that may be taken into account, in contrast to independency, i.e. disjunction. Therefore, a pectinate tree contains more information that a symmetric tree containing independent nodes. As we believe as Williams and Ebach (2008: 213) that in a phylogenetic analysis, a multi-state character is more informative than the pair of corresponding binary characters (because the relationships between states are explicitly showed only in the multi-state character), we can say more broadly that a tree carries more information than the sum of all its components. The dependence or independence between nodes should have an impact on the information content, and thus on 3ts weights. Consequences for 3ts are explained in the next section.

Classification of 3ts statements relationships

Relationships among 3ts can be classified relatively to dependency. We propose here a first complete classification of the types of relationships between pairs of 3ts – dyadic relationships. We show that the specifics of relationships among 3ts depends on their common leaves and positions. Two 3ts can be identical or not, compatible or not, combinable or not. Combinability can itself be decomposed in four possible relationships. All definitions are given below.

Definition 1. Two distinct 3ts are *compatible* if they can be true at the same time, i.e. if it exists *at least* one tree containing both.

Corollary. Two 3ts are incompatible if they cannot be true at the same time, i.e. the truth of one necessarily implies the falsity of the other. Two 3ts are incompatible if they have exactly the same taxa but state different relationships (e.g. $a(bc)$ and $b(ac)$).

Definition 2. Two compatible 3ts are *combinable* if they can be unambiguously gathered in a single tree. In other words, two 3ts are combinable if and only if they are present in the strict consensus of all possible trees containing them. Formally, two 3ts with their respective taxa sets D = $\{u, v, w\}$ and E = $\{x, y, z\}$ are combinable if it exists an informative strict consensus which taxa set F satisfies $|F| = |D \cup E| = 4$. We note that 3ts are combinable if and only if they have two taxa in common.

Corollary. If two 3ts are not combinable, the above-mentioned strict consensus is a bush (a tree without any informative node). Combinability logically implies compatibility, but the reverse is not true: compatible couples of 3ts sharing less than two taxa are non-combinable. The couple $a(bc)$ and $a(de)$ is non-combinable because a is the only common taxon, and entails several possibilities, as (*de*)(*a*(*bc*)) or *a*(*b*(*d*(*ce*))) and which strict consensus is a bush (no unambiguous common statement between all possible trees).

Four relationships among combinable 3ts exist, with specific properties, depending on the location of their common taxa (Figure 1).

Definition 3. Two combinable 3ts are in *in-in nts relationship* if both common taxa are located inside the apical node of both 3ts, and the combination of the 3ts results in a single n-taxon statement (i.e. a tree with a single informative node; nts thereafter; Wilkinson 1994).

The example of *a*(*cd*) + *b*(*cd*) allows us to build the tree *ab*(*cd*). The tree can be decomposed into the two 3ts that allowed its construction:

$$a(cd) + b(cd) \rightarrow ab(cd) \rightarrow a(cd) + b(cd)$$

Information conveyed by each 3ts allows an additional taxon to be added as a sister group to the apical node (*cd*). The fact that no additional 3ts is generated implies that no dependency is generated: that is the reason two-leaf apical nodes never cause fractional weight, as Nelson and Ladiges (1992) noticed.

Definition 4. Two combinable 3ts are in *in-out nts relationship* when one of the common taxa is connected directly to the root and the other is connected to the apical node in both 3ts. The combination of two 3ts in an in-out nts relationship leads to build a nts with a single informative node of three leaves.

The couple *a*(*bc*) and *a*(*cd*) illustrates the in-out nts relationship, whose analysis highlights the implication of a third 3ts:

$$a(bc) + a(cd) \rightarrow a(bcd) \rightarrow a(bc) + a(cd) + a(bd)$$

The in-out nts relationship between two 3ts generates a third 3ts with two taxa in common with the other two 3ts, at the same positions. The truth of these two 3ts implies the truth of a third. Among the three 3ts $a(bc)$, $a(cd)$, and $a(bd)$, it can be seen that whatever the two 3ts chosen, they will be in the in-out nts relationship and will involve the third one. It is precisely this kind of relationships which is taken into account by Nelson and Ladiges' FW to correct the redundancy. This type of dependency is the only one present within the components. If a new 3ts is added and forms additional in-out nts relationships, then all 3ts can be combined, as in-out relationship is transitive:

$$a(bc) + a(cd) + a(de) \rightarrow a(bcde) \rightarrow a(bc) + a(cd) + a(de) + a(bd) + a(be) + a(ce)$$

Three 3ts are linked by two in-out relationships. Each in-out relationship generates an additional 3ts:

$$a(bc) + a(cd) \rightarrow a(bd)$$

$$a(cd) + a(de) \rightarrow a(ce)$$

The generation of the secondary 3ts leads to the appearance of two new in-out dependency relationships, leading to the same 3ts:

$$a(bd) + a(de) \rightarrow a(be)$$

$$a(bc) + a(ce) \rightarrow a(be)$$

We call the initial 3ts as primary 3ts, $a(bc)$, $a(cd)$, and $a(de)$ in the example. The 3ts generated from the combination of primary 3ts are called secondary 3ts, here $a(bd)$ and $a(ce)$. The 3ts generated by the combination of secondary 3ts are called tertiary 3ts, here $a(be)$ and so on (to n-ary 3ts for the nth level of combination). For the combination of the set of three primary 3ts, a total of six n-ary 3ts are entailed. This is why coding a set of six 3ts $a(bc)$, *a*(*cd*), *a*(*de*), *a*(*bd*), *a*(*be*), *a*(*ce*), requires a weight of ½ for each 3ts since only three 3ts among the six are needed to build the node (given that all terminals are represented among the set). From the three primary 3ts we can deduce all the secondary and tertiary 3ts. The corollary is that 3ts may not satisfy the pairwise compatibility theorem (Estabrook et al. 1976; Wilkinson 1994b). Indeed, the fact that all primary 3ts are pairwise compatible does not imply that all n-ary 3ts are compatible.

Definition 5. Two combinable 3ts are in *orthologous relationship* when both have a common taxon connected to the apical node, and when the other common taxon is connected to the apical node in one of the 3ts and to the root in the other. Unlike with the in-in and in-out nts relationships, the trees constructed from two 3is linked by an orthologous relationship have two informative nodes, one of which is orthologous (see fig. 4b in Wilkinson et al. 2004). If the tree *a*(*b*(*cd*)) is decomposed into its two components *ab*(*cd*) and *a*(*bcd*), the first generates only one in-in nts relationship and the other three in-out nts relationships (Appendix 1a). The combination of these two components implies the existence of two new couples of 3ts that were not present in the independent components:

b(*cd*) (from *ab*(*cd*)) + *a*(*bd*) (from *a*(*bcd*)) → *a*(*b*(*cd*)) → *b*(*cd*) + *a*(*bd*) + *a*(*cd*) + *a*(*bc*)

b(*cd*) (from *ab*(*cd*)) + *a*(*bc*) (from *a*(*bcd*)) → *a*(*b*(*cd*)) → *b*(*cd*) + *a*(*bd*) + *a*(*cd*) + *a*(*bc*)

These two couples linked by an orthologous relationship produce the complete tree *a*(*b*(*cd*)). If combined, 3is linked by orthologous relationships, therefore, lead to inter-node dependency. This shows that the quantification of the information contained in a tree is not equal to the sum of the information of each component. Methods that decompose trees in their components without taking into account the relationships between these components, such as the supertree method MRP (Baum 1992; Ragan 1992) or the biogeographical method BPA (Wiley 1986; Brooks 1990), are thus flawed. The information conveyed by the inclusion (dependency) of one node in another must be taken into account.

Definition 6. Two combinable 3ts are in *paralog relationship* when among the two taxa they have in common, one is connected to the root in the first 3ts and the apical node in the second 3ts, and the other is connected to the apical node in the first 3ts and to the root in the second. The paralogous relationship appears when at least two informative nodes are disjoint (i.e. if there is a paralogous node in the tree where the 3ts come from).

The following example illustrates a paralogous relationship between two 3ts that generates a symmetric tree (Appendix 1b):

$$c(ab) + a(cd) \rightarrow (ab)(cd) \rightarrow c(ab) + a(cd) + d(ab) + b(cd)$$

This combination allows two 3ts, $d(ab)$ and $b(cd)$, to be deduced.

Some general corollaries may be listed from the set of above described relationships:

- Any hierarchical tree with at least one informative node can be analyzed into a set of 3ts with no incompatible relationships.
- The four dyadic relationships that are in-in, in-out, orthologous, and paralogous are necessary and sufficient to reconstruct any phylogenetic tree that can be deduced from a set of 3ts, including Bryant and Steel’s “irreducible high order rules” (1995: 445–452) that are not formalized but only exemplified, their finding being the result of an incomplete axiomatization.

— Any tree containing an informative node and more than one leaf outside this informative node contains in-in relationships (Fig. 1a).

— Any tree containing an internal informative node and more than two leaves connected to that node contains in-out relationships (Fig. 1b).

— Compatible but not combinable 3ts (zero to one taxon in common) are present in a tree as long as there is an informative node comprising more than three leaves. As soon as there are at least two informative nodes in a tree, paralogous (fig. 1c) and/or orthologous relationships (fig. 1d) appear. It is, therefore, possible to find all these types of relationships within the same tree (Appendix 1c).

Weighting procedures

Formalized types of relationships among 3ts that lead to dependency are the only way to remove redundancy in 3ts decomposition. Firstly, we will show that MW fails to remove all instances of dependency. Secondly, we will propose a new procedure, grounded on the FW rationale, that succeeds at removing every dependency from any 3ts set.

Minimal weighting

Wilkinson et al. (2004) designed the MW in order to remove all dependencies issues using a simple formula (*t*-2). However, there are two arguments for rejecting the MW.

The first concerns the identical weight given to each triplet, as in Nelson and Ladiges' $\mathrm{FW}_{\mathrm{total}}$ computation. This weight implies that among the four 3ts of (*b*(*c*(*de*))), only two are necessary to reconstruct the tree. The total weight of the tree (its information content) is 2, and each 3ts has an information content of ½. Wilkinson et al. (2004) rationale implies that each of the 3ts has the same information content. If this was true, any set of two 3ts among *b*(*cd*), *c*(*de*), *b*(*ce*), and *b*(*de*), would allow to reconstitute the tree, which is not correct. Here is the result of the different combinations of 3ts:

(1) $b(cd) + c(de) \longrightarrow b(c(de))$

(2) $c(de) + b(ce) \longrightarrow b(c(de))$

(3) $b(cd) + b(ce) \longrightarrow b(cde)$

(4) $b(cd) + b(de) \longrightarrow b(cde)$

(5) $b(ce) + b(de) \longrightarrow b(cde)$

(6) $c(de) + b(de) \longrightarrow bc(de)$

(1) and (2) are the only couples that allow the tree to be fully reconstituted. The other couples allow only one of the two components to be rebuilt: the couples (3) to (5) allow rebuilding *b*(*cde*), and the couple (6) allows rebuilding *bc*(*de*). In order to keep only the minimum number of 3ts, the statement "any 2 among 4" is incorrect. Therefore, MW is inconsistent with Wilkinson et al.'s (2004) rationale.

The second argument against MW is a strong limitation in its conception: it can only handle fully dichotomous trees. The total weight of a tree of *t* taxa is *t*-2 for a totally bifurcating tree and, to our knowledge, no analytical solution has so far been proposed to include multifurcations. Using the total weight formula of *t*-2 for polytomous trees implies that the information content of a given tree is identical to that of any of its subtrees as long as they have the same leaves. The trees (*ab*)(*cd*), *ab*(*cd*) and (*ab*)*cd* all have an information content of 2 according to this formula. This conclusion is clearly illogical.

Corrected fractional weighting

To successfully eliminate dependency among a 3ts set, the different types of 3ts relationships and the scale at which they operate (intra-node or inter-node) must be taken into account. MW fails to recognize in detail the different types of relationships. Both $\mathrm{FW}_{\mathrm{total}}$ and $\mathrm{FW}_{\mathrm{comp.}}$ takes into account a single relationship (in-out nts), which is insufficient. Our strategy to suppress the dependency is set up in sequential steps, each step corresponding to a type of 3ts relationship (table II). We will explain here a corrected version of the FW ($\mathrm{FW}_{\mathrm{cor.}}$) that take into account all types of dependency defined above.

The first step corresponds to the procedure set out by Nelson and Ladiges in 1992. It consists in removing the dependency linked to in-out relationships within each component itself. The formula for the number of independent 3ts for a component is (t-n)(n-1), t being the total number of taxa and n being the number of taxa connected to the informative node (Nelson and Ladiges, 1992). The factor (t-n) corresponds to the number of leaves connected directly to the root. Increasing their number only generates in-in relationships. The factor (n-1) is linked to in-out relationships. For a($bcde$), the factor (n-1) is 3. Only three 3ts are needed to rebuild the tree. The factor (t-n) is 1 and acts as a multiplier according to the number of leaves at the root. For example, by adding a leaf f to the root, (n-1) does not change. (t-n) on the other hand, equals 2:

$$a(bcde)\ [3] + f(bcde)\ [3] \rightarrow af(bcde)\ [6]$$

The numbers in square brackets are the weights of the corresponding trees. The minimum number of pairs of leaves pairs (e.g. a,b and c,d) necessary for each leaf to be in a pair and each pair to have a leaf in common with at least one another pair is (n-1). (n-1) is then multiplied by the number of leaves connected to the root (t-n) to give the formula (t-n)(n-1). (t-n) does not change between the formula of the number of independent 3ts and the formula of the total number of 3ts. On the other hand, the value (n-1) changes: to have the total number of 3ts, we need to search for all the combinations of two leaves among n, i.e.:

$$\binom{n}{2} = \frac{n!}{2(n-2)!} = \frac{n(n-1)}{2}$$

By replacing (n-1) by $\binom{n}{2}$ in (t-n)(n-1), we obtain the total number of 3ts:

$$\text{Number of 3ts for a component} = \frac{n(n-1)(t-n)}{2}$$

Nelson and Ladiges' FW$_{\text{comp.}}$ is, therefore, FW$_{\text{cor.}}$'s first step in eliminating redundancy. This step of removing intra-node redundancy requires the analysis of each component independently. For each component, all its 3ts are generated, and the weight of 2/n (number of independent 3ts / total number of 3ts) is assigned to each 3ts. This weight eliminates overweight due to in-out relationships.

For example, component *a*(*bcd*) can be decomposed into three 3ts: *a*(*bc*), *a*(*cd*), *a*(*bd*). Each of the three pairs of 3ts forms an in-out nts relationship which logically implies the third 3ts. Each 3ts is therefore overweighed because it exists both as a primary and secondary 3ts. More generally, a 3ts present in a component is overweighed if it can be generated from other 3ts. This first step makes it possible to define the phylogenetic weight of each node.

FW$_{\text{cor.}}$ is obtained by removing the dependency linked to in-out nts relationships, as in FW$_{\text{comp}}$, and then to orthologous relationships among 3ts of different nodes (inter-node redundancy). Since 3ts have already been weighted in the previous step, the correction of the orthologous dependency (i) concerns inclusions between informative nodes are present in the tree, and (ii) revises the weight of the 3ts assigned in the first step.

Nelson and Platnick (1991) showed that every 3ts generated from a fully pectinate tree has a weight of 1, applying FW or not because there is no redundancy between the 3ts of a fully pectinate tree. The correction of inter-node redundancy requires checking the 3ts of each pair of nodes involving direct inclusion. Two 3ts in an orthologous relationship will generate secondary 3ts that already exist as primary ones. Using the example of the tree *a*(*b*(*cd*)) analyzed in Figure 2, Nelson and Ladiges (1992: p. 491) showed that the statement *a*(*cd*) occurs in both series. This 3ts is the only one having a weight greater than 1 5/3 (2/3 + 1): it is the overweighed 3ts. The component *a*(*bcd*) bears the redundant 3ts because it is the only component of the tree that contains this particular 3ts corrected by $\mathrm{FW}_{\mathrm{comp}}$. The other component consists only of in-in relationships: its 3ts are therefore all essential to the reconstruction of the tree and have a weight of 1. The analysis of the two orthologous relationships of the tree highlights the overweight of *a*(*cd*):

$$a(bc) + b(cd) \rightarrow a(bd) + a(cd)$$

$$a(bd) + b(cd) \rightarrow a(bc) + a(cd)$$

The secondary 3ts generated are *a*(*bd*), *a*(*bc*) and *a*(*cd*) (twice). All the 3ts resulting from orthologous relationships of both components allow generating only one component: *a*(*bcd*). However, among all the 3ts, *a*(*cd*) appears twice while the others appear only once because the orthologous relationship that overweighed this particular 3ts produces also a supplementary secondary 3ts, the only 3ts present in both components. The uncertainty of ‘two 3ts out of three’ from the component *a(bcd)* can then be resolved: the two necessary 3ts are *a*(*bd*) and *a*(*bc*). *a*(*cd*) is deleted from the component *a*(*bcd*), and its weight (2/3) is equally distributed among the others — if its weight was not distributed between the remaining 3ts, the result would be that knowledge about orthologous relationship between nodes would lead to a loss of phylogenetic information. In (*a*(*b*(*cd*))), the weight of each of the four 3ts becomes 1. The ambiguity is resolved. The analysis of inter-node relationships allows us to correct the weights of 3ts from intra-node relationships because the inclusion of one node in another carries specific phylogenetic information.

To exemplify how polytomies are handled by this procedure, we apply these two steps to the tree *a*(*bc*(*de*)) (Figure 3 and Table I). The six 3ts generated by the first component are weighted 1/2 (in-out relationships) and the three 3ts generated by the second component are weighed 1 (in-in relationships). The total information content of the tree is 6. It is noted that *a*(*de*) is present in both components, and can, therefore, be removed from the component *a*(*bcde*). After correction, the total weight of the tree is still 6, but the weights of the 3ts have been corrected. This tree shows a polytomy. Unlike the weights of the 3ts derived from the tree (*a*(*b*(*cd*))), which all end up balanced at 1, the weights of some of the 3ts derived from (*a*(*bc*(*de*))) are still fractional, even if the correction makes them tend towards 1. The dichotomous pectinate structure is the most informative: it contains the maximum number of components, and these components are all dependent on each other. Uncertainty increases with the number of polytomies, and with the number of branches involved in polytomies.

The last type of dependency is related to paralogous relationships. Since paralogous relationships only concern independent, i.e. disjoint, groups, there is no dependency to manage at the level of 3ts for correcting 3ts weights. Indeed, paralogous relationships never provide redundancy of any particular 3ts.

An example with the analysis of the tree ((*ab*)(*cd*))) will illustrate this point (Appendix 1b):

((*ab*)(*cd*))) → (*ab*)*cd* + *ab*(*cd*)

(*ab*)*cd* → *c*(*ab*) + *d*(*ab*)

ab(*cd*) → *a*(*cd*) + *b*(*cd*)

The weighting schemes proposed in the literature relay on the number of nodes. (i) The tree has a weight of 2 (corresponding to MW). Taken *independently*, each component of the tree also has a weight of 2; the choice of this weighting leads to accepting that the tree (*ab*)(*cd*)) contains the same amount of phylogenetic information as (*ab*)*cd* or *ab*(*cd*). (ii) The tree has a weight of 4 and can be divided into two components without any loss of information (corresponding to the FW). However, since these two components are disjoint, they are independent. Here, the tree information is the sum of the information of its two independent components. Both components have a weight of 2. Paralogous relationships do not create differential weights between the 3ts because no 3ts is overrepresented by n-ary statements. Therefore, there is no justification for correcting in any way the weight of the 3ts intra-node.

Following Nelson and Ladiges (1992: 492): '*For cladogram 11, there is one internal node, ((AB)(CD)), with four statements, each with an absolute value of ¾*'. The authors provided no justification of their result (their result cannot be reached nor with the $\mathrm{FW}_{\mathrm{comp.}}$, nor with the $\mathrm{FW}_{\mathrm{total}}$). However, as already pointed out by Nelson and Ladiges, if two totally dichotomous trees with the same number of leaves (more than four) are compared, the tree with more paralogous nodes will have less information content. The information content does need to reflect that. The information is maximal in pectinate trees because they are devoid of independent couples of nodes. For example, the weight of *a*(*b*(*c*(*de*))) is 10 whereas the weight of *a*((*bc*)(*de*)) is 9. In conclusion, our amended version $\mathrm{FW}_{\mathrm{cor.}}$ is the unique weighting scheme that takes into account dependency emerging both from intra-node (in-out relationships) and inter-node redundancy (ortholog relationships). Our proposal of $\mathrm{FW}_{\mathrm{cor.}}$ has been implemented in the software LisBeth (Zaragüeta et al. 2012).

SIMULATIONS

MATERIAL AND METHODS

Following theoretical foundations of the relationship between dependency and weighting developed in the previous parts, our aim here is to simulate datasets and analyses to empirically test the efficiency of $FW_{cor.}$ versus current alternative methods. Each simulation consists in 1) generate a tree (named hereafter ‘the original tree’), 2) produce two subtrees, 3) generate noise in the subtrees, 4) use the subtrees with different weighting schemes to 5) reconstruct an optimal tree (and produce a strict consensus if several optimal trees exist) and 6) compare this optimal tree (or the strict consensus of all optimal trees) with the original one to assess the relative efficiency of the different weighing schemes (Fig. 4). Each original tree is randomly generated using the python package ete3 (Huerta-Cepas et al. 2016) and contains 11 terminal taxa. Two subtrees are generated. Each is a copy of the original tree from which a defined number *m* of terminal taxa has been removed (*m* goes from 0 to 8; the choice of taxa is random). For each generation, 1000 simulations are run, leading to a total of 9000 runs. All the runs were repeated by introducing a varying amount of random noise by taxa permutations to simulate homoplasy and/or distortion. The different amount of distortion is included by pruning and regrafting a number *n* of taxa (*n* goes from 0 to 5). The choice of the pruned taxa and the regrafting locations are both random. The 9,000 runs were thus repeated 6 times leading to 54,000 runs. To combine both subtrees in the optimal tree, we used two supertree methods to reconstruct the optimal tree, one based on a parsimony approach (i.e. Matrix Representation with Parsimony) using PAUP 4.0a (Swofford 2003), and the other based on a 3ta approach (Three-taxon analysis; Nelson and Platnick 1991) with three different weighing schemes: $FW_{cor.}$, $FW_{comp.}$, and MW (in our simulations MW and $FW_{total.}$

give always the exact same results because the trees of each analysis always have the same number of terminal taxa), using Lisbeth (Zaragüeta et al. 2012). Note that when all trees of an analysis have the same number of taxa, the total fractional weight would lead to the same results as the minimal weight because the weights of all 3ts are the same for a given tree. The total number of runs is increased to 216,000. The ability of the methods to reconstruct the original tree is assessed by three complementary metrics. Triplet distance metrics (Tavares 2018) were preferred to Robinson-Foulds, because the latter is strongly influenced by wildcard taxa. This is a common problem in supertree simulation analyses (Penny et al. 1982). Finally, the triplet distance metrics can be decomposed for our purpose in true resolutions (*TR*) and false resolutions (FR; Grand et al. 2013; Rineau et al. 2018). The first one, named *TR*, is the number of 3ts founded in both the original and optimal trees divided by the number of 3ts of the original tree. This can be viewed as the proportion of the 3ts from the original tree that is present in the optimal one. The second one, named *FR*, is the number of 3ts founded in the optimal tree but not in the original tree, divided by the number of 3ts of the optimal tree. This can be viewed as the proportion of the 3ts from the optimal tree that were not present in the original one. At last, the efficiency (*TR* – *FR*; range: [-100; 100]), allow to synthesize the results. -100 is the case where all phylogenetic relationships are false and the tree is completely resolved (i.e. dichotomous); 100 where all phylogenetic relationships are true and the tree is completely resolved. A null result means that there are as many true and false phylogenetic relationships, regardless of how well resolved the optimal tree is. In order to characterize the quantity of information brought by optimal trees, we counted internal nodes and 3ts from optimal tree for maximizing precision, hereafter named *constrict resolution* and *constrict information content* respectively.

The results were then analyzed using R software (R Development Core Team 2008). We first explored the simulations by performing a PCA to provide a global overview of the covariations between the variables involved in the analyses. These variables consist in the starting parameters (i.e. number of permutations and number of taxa in subtrees), the number of optimal trees as well as their number of taxa, both the resolution and the information content of the strict consensus and finally, the 3 measures of the results quality (i.e. true resolutions, false resolutions and efficiency). Then, we describe in more details the relationships between the variables of interest. As those relationships show an important degree of complexity, we were not able to model them in a satisfactory manner. As a consequence, we used linear models to describe and test the relative efficiency of the methods and weighting schemes as well as other variables that can modify their relative efficiencies.

Results and Discussion

The two first axes of the PCA represent 73 % of the dataset variance. The first axis represents mostly the contribution of TR and four positively correlated variables that are: the constrict resolution, the constrict information content, the number *n* of taxa in subtrees, and the number of taxa in the optimal tree (related to the more or less important overlapping of the taxa sets of the subtrees; Fig. 5). For a given number of permutations, the amount of TR compared to the number of taxa in subtrees has an S-shaped distribution (see SM 1a), consequently TR tends to a horizontal asymptote (e.g. 100% of TR for 0 permutations when the number of taxa in subtrees tends to the one in the original tree). As the number of permutations increases, the asymptote decreases. The second axis of the PCA is mostly structured by FR (Fig. 5). As expected, the number of permutations (i.e. the quantity of noise) correlates mainly with this axis. In details, for a given number of permutations, the number of FR firstly increases with the number of taxa in subtrees and then decreases (SM 1b). Obviously, the efficiency (TR – FR) positively correlates to the first axis (thus to TR) and negatively to the second axis (thus to FR) (Fig. 5, SM 1c). Finally, the number of optimal trees does not correlate much with the other variables.

The relationships between all variables were thoroughly explored. The relative amount of TR and FR depends also greatly on the number of taxa in the optimal tree (SM 2). For a given number of taxa in the optimal tree, most of the runs fall on a line going from 0% TR and 100% FR to 0% FR and a percentage *p* of TR. This percentage *p* reaches 100% when the number of taxa in the optimal tree equals the one in the original tree. These lines represent the maximal resolution reachable with fully bifurcating optimal trees.

As concluded from the PCA, the variance of the dataset can be mostly summed up by TR, FR, the number of taxa in subtrees, and the number of permutations. The relationships between those variables are mostly monotonous. Thus, we computed two linear models with respectively TR and FR as dependent variables and as independent variables for both models: the number of taxa in subtrees, the number of permutations, and the different types of analyses (MRP, 3ta with $FW_{cor.}$, $FW_{comp., or}$ MW/ FW_{total}). These models thus have three variables and we included all the interactions terms. The homoscedasticity is not verified. For example the variance of TR drastically increases in relation to the number of taxa in subtrees. As a consequence, we do not expect from these models to test for the effects but rather to investigate main tendencies and to observe specific interactions between the different types of analyses and the other variables (summaries of linear models are provided in SM 3). We use the function *interact_plot* from the R package *jtools* to visualize to three interactions as well as main effects from isolated variables at the same time (Fig. 6). Linearity checks are provided in the supplementary information (SM 4). The relationships are consistent with what inferred from the PCA. The increase of TR with the number of taxa in subtrees depends on the number of permutations. The higher the number of permutations is, the lower the increase of TR in relation to the number of taxa in subtrees. The main results with the substantial effect are that FW$_{cor.}$ and FW$_{comp.}$ are indistinct from each other and better compared to MRP and MW as the number of taxa in subtrees increases. The increase of FR with the number of permutations also depends on the number of taxa in subtrees and the type of analysis. For a few taxa in subtrees, the analyses are indistinct no matter the number of permutations. With both the increase of the number of taxa and permutations, differences between analyses increase with the order of FR amount: MW, MRP, and indistinctively FW$_{cor.}$ and FW$_{comp.}$ (Fig. 6). The efficiency decreases with the number of permutations. It also increases with the number of taxa in subtrees, especially at a low number of taxa. The numerous taxa in subtrees are, the higher the slope of this relationship is. In other terms, the efficiency does not really increase with the number of taxa in subtrees for too noisy datasets. This being said, FW$_{cor.}$ and FW$_{comp.}$ have better efficiencies compared to MRP and MW as the number of taxa in subtrees increases (Fig. 6).

In conclusion, FW$_{cor.}$ and FW$_{comp.}$ provide more resolved trees in comparison to MRP and MW. As the number of true resolutions (TR) is generally higher than the number of false resolutions (FR), efficiency is a bit better for FW$_{cor.}$ and FW$_{comp}$. This leads us to recommend these weighting methods for the purpose of supertree reconstruction methods, cladistics methods, or metrics using triplets. These are very preliminary results and they need to be confirmed in other conditions such as the type of noise, the tree shape. Future works will have to study more precisely the differential effect of FW$_{cor.}$ versus FW$_{comp.}$, which was indistinguishable in our simulations. The absence of significant differences between FW$_{cor.}$ and FW$_{comp.}$ may be related to the very low number of source trees. The higher the number of source trees is, the more incongruent relationships there is. It is the incongruence that will lead to the choice of this or that 3ts, and the choice should be biased with the FW$_{comp.}$ which will weight some 3ts to more than 1, which is theoretically flawed.

Conclusion

Our paper focuses on the different types of relationships within 3ts sets, some of them being directly responsible for the emergence of dependency. We also proposed a method for evolutionary biologists whose aim is to completely eliminate these different types of dependency from which all methods based on three-taxon statements currently suffer.

In-out nts, orthologous and paralogous relationships lead to the redundancy of some 3ts. We propose for the first time a method of 3ts weighting based on these different types of redundancy. Ideally, any phylogenetic or supertree analysis should only include datasets whose elements are independent of each other, to eliminate any possibility of bias. We have shown that the simple decomposition of a tree into its components leads to a loss of information and that our method of corrected fractional weighting is the only procedure able to fully delete all the redundancies. We have also shown using computer simulations that the absence of weighting can significantly reduce the effectiveness of any method based on the 3ts. We also proposed the first empirical comparison between supertree component methods (MRP) and 3ts (three-taxon analysis), confirming the efficiency of methods which decompose trees into 3ts. More precisely, the most efficient methods are those which weight 3ts through a decomposition step into components, as Fractional Weighting. Finally, understanding the dependency relationships between the 3ts will permit to improve the accuracy and efficiency of supertree methods, consensus methods and phylogenetic analysis methods based on 3ts. Triplet distance metrics should expect the same handling. To conclude, we hope to open new debates and to fuel methodological discussions on how phylogenetic information should be weighted in order to improve the accuracy of analyses.

ACKNOWLEDGEMENTS

Valentin Rineau and René Zaragüeta conceived the ideas, Valentin Rineau acquired and analysed the datasets, and led the writing of the manuscript, Jérémie Bardin conducted the statistical analyses. All authors contributed critically to the drafts and gave final approval for publication. We deeply thank the Editor in Chief Bryan Carstens, the Associate Editor Claudia Solis-Lemus, David Williams, Evgeny Mavrodiev and an anonymous reviewer for their fruitful comments on previous drafts, and to Stephane Prin and Paul Zaharias for discussions on 3ts combination procedures. We also thank Stephane Prin, Paul Zaharias, and Rachel Vautrin for comments on the draft and corrections. This work has been founded by the UMR 7207, Centre de Recherche en Paléontologie – Paris, France.

LITERATURE CITED

Adams E.N. 1986. N-trees as nestings: Complexity, similarity, and consensus. Journal of Classification. 3:299–317.

Aho A.V., Sagiv Y., Szymanski T.G., Ullman J.D. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM Journal on Computing. 10:405–421.

Baum B. 1992. Combining Trees as a Way of Combining Data Sets for Phylogenetic Inference, and the Desirability of Combining Gene Trees. Taxon. 41:3–10.

Brooks D.R. 1990. Parsimony Analysis in Historical Biogeography and Coevolution: Methodological and Theoretical Update. Systematic Zoology. 39:14–30.

Bryant D. 2003. A classification of consensus methods for phylogenetics. DIMACS series in discrete mathematics and theoretical computer science. 61:163–184.

Bryant D., Steel, M. 1995. Extension Operations on Sets of Leaf-Labelled Trees. Advances in applied mathematics. 16:425–453.

Cao N., Bourdon E., El Azawi M., Zaragüeta R. 2009. Three-item analysis and parsimony, intersection tree and strict consensus: a biogeographical example. Bulletins de la société géologique de France. 180:13–15.

Dannenberg K., Jansson J., Lingas A., Lundell E.M. 2019. The approximability of maximum rooted triplets consistency with fan triplets and forbidden triplets. Discrete Applied Mathematics. 257:101–114.

Dekker M.C.H. 1986. Reconstruction methods for derivation trees. Unpublished Master Thesis. Department of Mathematics and Computer Science, Vrije Universiteit, Amsterdam.

Estabrook G.F., Johnson Jr.C.S., McMorris F.R. 1976. A mathematical foundation for the analysis of cladistic character compatibility. Mathematical Biosciences. 29:181–187.

Farris J.S. 1970. Methods for Computing Wagner Trees. Systematic Zoology. 19:83–92.

Fitch W.M. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology. 19:99–113.

Grand A., Corvez A., Duque Velez L.M., Laurin M. 2013. Phylogenetic inference using discrete characters: performance of ordered and unordered parsimony and of three-item statements. Biological Journal of the Linnean Society. 110:914–930.

Huerta-Cepas J., Serra F., Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution. 33:1635–1638.

Islam M., Sarker K., Das T., Reaz R., Bayzid M. S. 2020. STELAR: A statistically consistent coalescent-based species tree estimation method by maximizing triplet consistency. BMC genomics, 21: 1–13.

Kitching I.J., Forey P.L., Humphries C.J., Williams D.M. 1998. Cladistics: The Theory and Practice of Parsimony Analysis. Second edition. Oxford: Oxford University Press.

Kuhner M.K., Yamato J. 2015. Practical Performance of Tree Comparison Metrics. Systematic Biology. 64:205–214.

Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutionary Biology. 302:1–18.

Mavrodiev E.V., Williams D.M., Ebach M.C. 2019. On the Typology of Relations. Evolutionary Biology. 46:71–89.

Mickevich M.F., Platnick N.I. 1989. On the information content of classifications. Cladistics. 5:33–47.

Nelson G. 1979. Cladistic analysis and synthesis: principles and definitions, with a historical note on Adanson's Familles des Plantes (1763–1764). Systematic Biology, 28:1–21.

Nelson G., Ladiges P.Y. 1991a. Standard assumptions for biogeographic analysis. Australian Systematic Botany. 4:41–58.

Nelson G., Ladiges P.Y. 1991b. Three-area statements: standard assumptions for biogeographic analysis. Systematic Biology. 40:470–485.

Nelson G., Ladiges P.Y. 1992. Information content and fractional weight of three-item statements. Systematic biology. 41:490–494.

Nelson G., Ladiges P.Y. 1994. Three-item consensus empirical test of fractional weighting. Systematics Association Special Volume. 52: 193–209.

Nelson G., Platnick N.I. 1981. Systematics and Biogeography: Cladistics and Vicariance. Columbia University Press. New York.

Nelson G., Platnick N.I. 1991. Three-Taxon Statements: A More Precise Use of Parsimony? Cladistics. 7:351–366.

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.

Poormohammadi H., Zarchi M. S., Ghaneai H. 2020. NCHB: A Method for Constructing Rooted Phylogenetic Networks from Rooted Triplets based on Height Function and Binarization. Journal of Theoretical Biology, 2–35.

Prin S. 2012. Structure mathématique des hypothèses cladistiques et conséquences pour la phylogénie et l'évolution. Avec une perspective sur l'analyse cladistique. Unpublished PhD Thesis.

R Development Core Team. 2008. R: A language and environment for statistical computing. Vienna, Austria. R Foundation for Statistical Computing.

Ragan M. 1992. Phylogenetic Inference Based on Matrix Representation of Trees. Molecular Phylogenetics and Evolution. 1:53–58.

Ranwez V., Criscuolo A., Douzery E.J.P. 2010. SUPERTRIPLETS: a triplet-based supertree approach to phylogenomics. Bioinformatics. 26:i115–i123.

Rineau V., Zaragüeta R., Laurin M. 2018. Impact of errors on cladistic inference: simulation-based comparison between parsimony and three-taxon analysis. Contributions to Zoology. 87:25–40.

Sevillya G., Frenkel Z., Snir S. 2016. Triplet MaxCut: a new toolkit for rooted supertree. Methods in Ecology and Evolution. 7:1359–1365.

Strimmer K., von Haeseler A. 1996. Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies. Molecular Biology and Evolution. 13:964–969.

Swofford D.L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts: Sinauer Associates.

Tavares B.L. 2018. A synopsis of comparative metrics for classifications. arXiv preprint arXiv:1804.03929.:37.

Vach W. 1994. Preserving consensus hierarchies. Journal of Classification. 11:59–77.

Wiley E.O. 1986. Methods in vicariance biogeography. In: Hovenkamp P., editor. Systematics and Evolution: a matter of Diversity. Utrecht. p. 283–306.

Wilkinson M. 1994a. Common Cladistic Information and its Consensus Representation: Reduced Adams and Reduced Cladistic Consensus Trees and Profiles. Systematic Biology. 43:343–368.

Wilkinson M. 1994b. Three-taxon statements: when is a parsimony analysis also a clique analysis? Cladistics. 10:221–223.

Wilkinson M., Cotton J., Thorley J. 2004. The Information Content of Trees and Their Matrix Representations. Systematic Biology. 53:989–1001.

Williams D.M. 2004. Supertrees, components and three-item data. In: Bininda-Emonds O., editor. Phylogenetic supertrees: combining information to reveal the tree of life. Springer Science & Business Media. p. 389–408.

Williams D.M., Ebach M.C. 2008. Foundations of systematics and biogeography. Springer Science & Business Media.

Williams D.M., Humphries C.J. 2003. Component Coding, Three-item Coding, and Consensus Methods. Sys. Biol. 52:255–259.

Zaragüeta R., Lelièvre H., Tassy P. 2004. Temporal paralogy, cladograms, and the quality of the fossil record. Geodiversitas. 26:381–389.

Zaragüeta R., Ung V., Grand A., Vignes-Lebbe R., Cao N., Ducasse J. 2012. LisBeth: New cladistics for phylogenetics and biogeography. Comptes Rendus Palevol. 11:563–566.

CAPTIONS

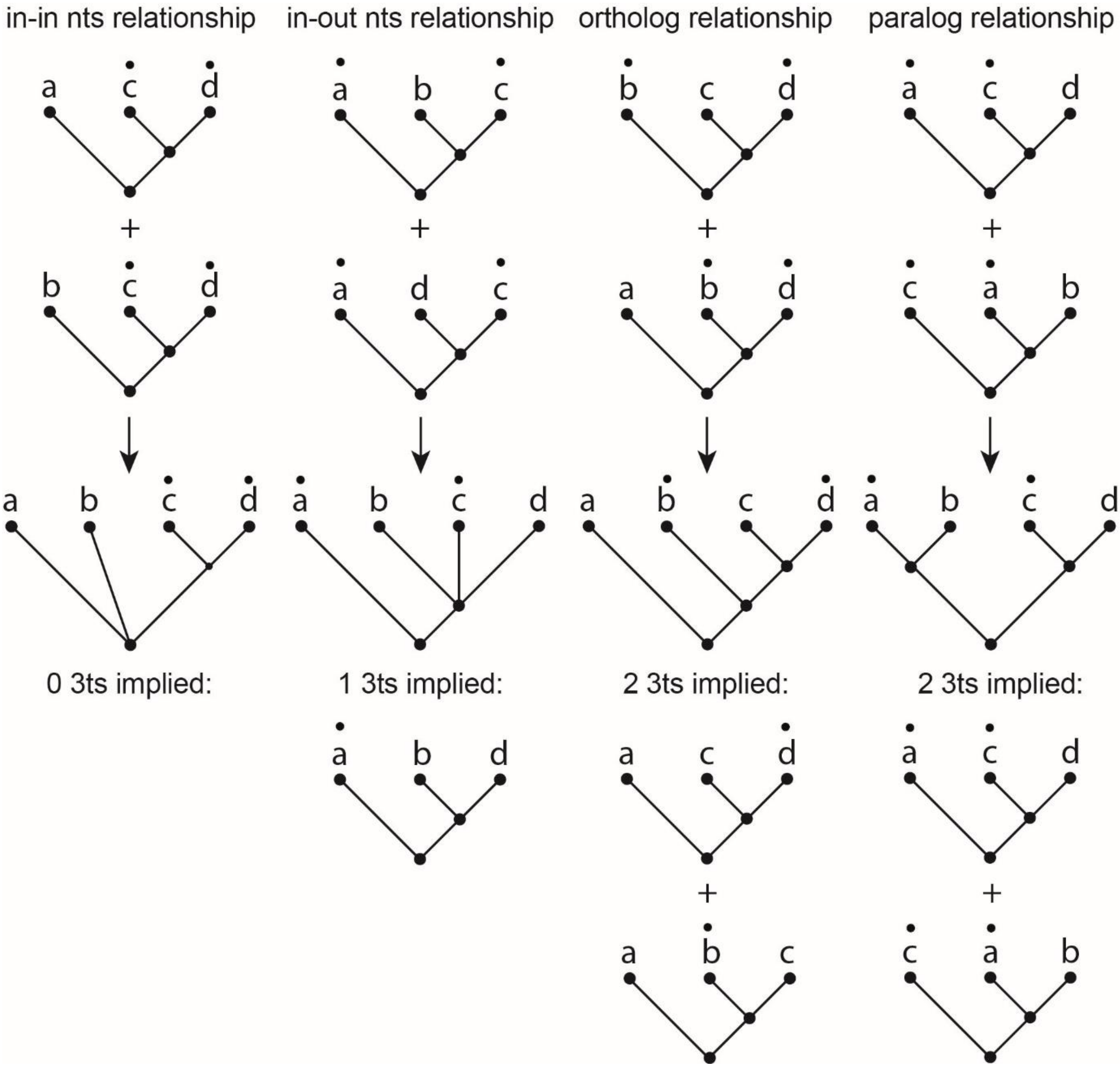
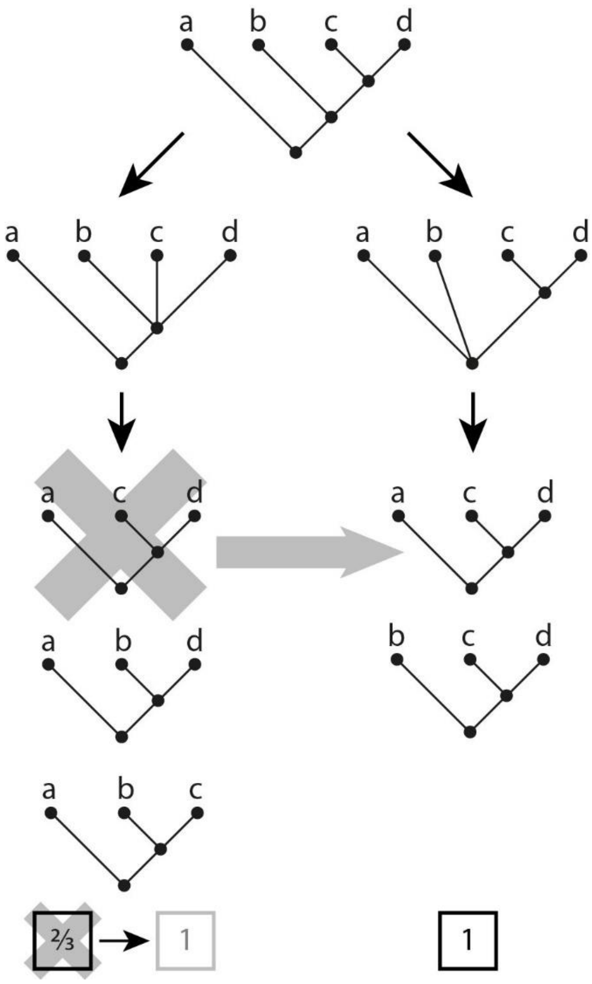


Figure 1.

Four possible types of relationships between pairs of 3ts can be combined. Two 3ts sharing two taxa in common can be combined into a 4-taxon tree from which it is possible to deduct 0, 1 or 2 3ts. The black dots correspond to the leaves common to both 3ts. a. in-in nts relationship. b. in-out nts relationship. c. ortholog relationship. d. paralog relationship.

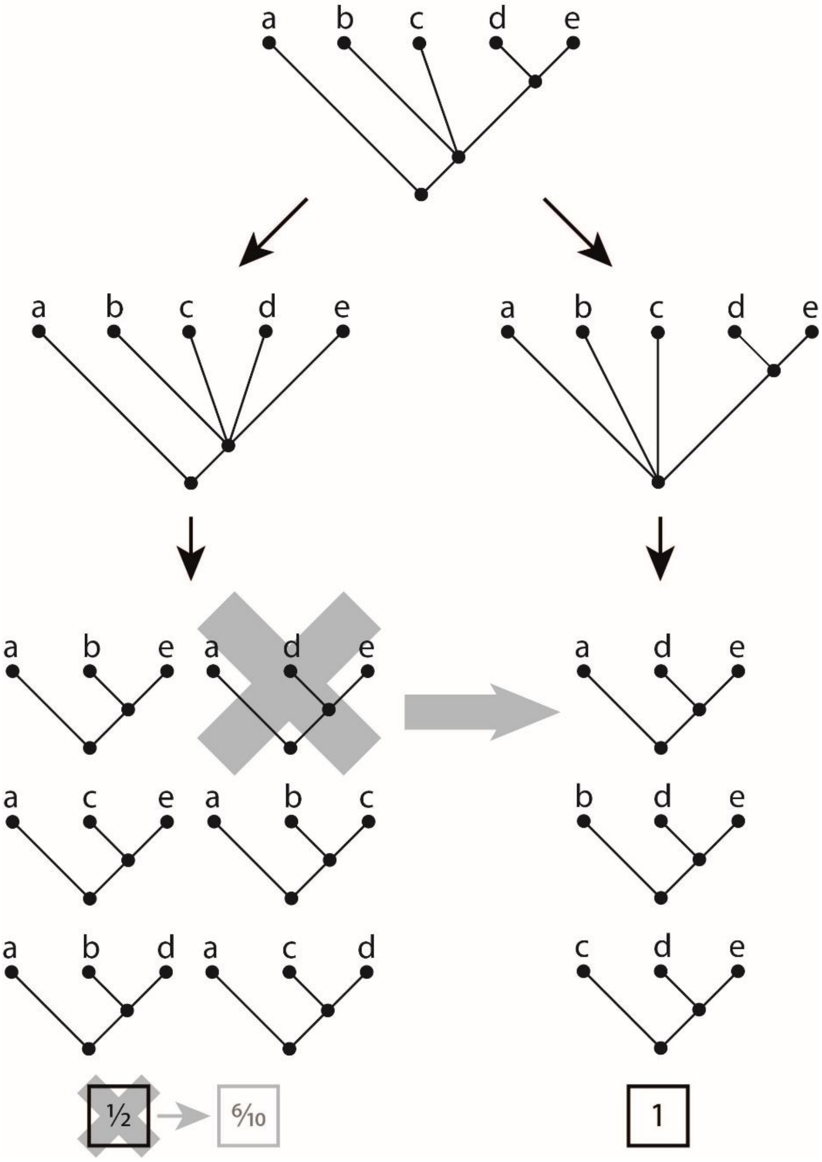


750

751 Figure 2.

752 Fractional weighting procedure with correction of inter-node redundancy on (*a*(*b*(*cd*))).

753 The dichotomous pectinate tree only produces 3ts of weight 1.



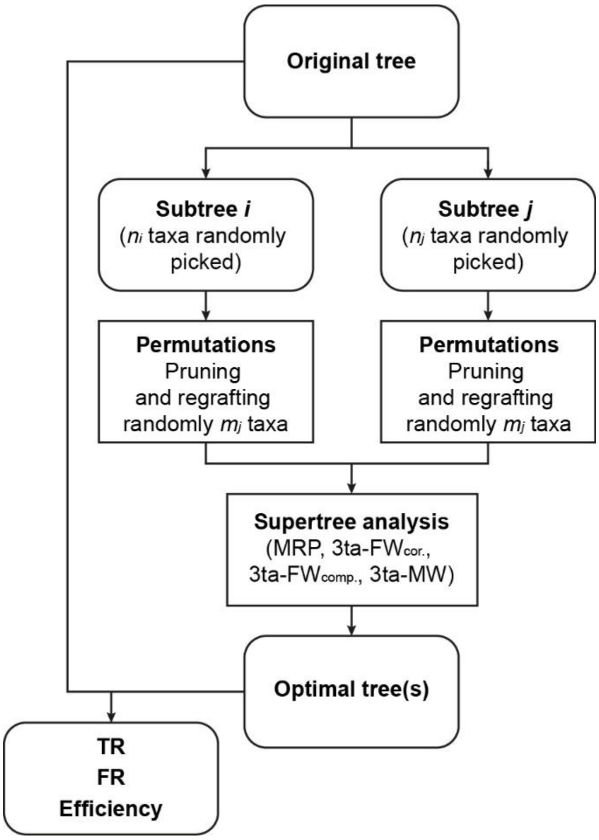
754

755 Figure 3.

756 Fractional weighting procedure on (*a*(*bc*(*de*)). The tree is not completely resolved, which

757 leads to uncertainty amongst 3ts and the persistence of fractional weightings for five of them.

758


Original tree
Subtree *i* (n_i taxa randomly picked)
Subtree *j* (n_j taxa randomly picked)
Permutations Pruning and regrafting randomly m_i taxa
Permutations Pruning and regrafting randomly m_j taxa
Supertree analysis (MRP, 3ta-FW$_{cor.}$, 3ta-FW$_{comp.}$, 3ta-MW)
Optimal tree(s)
TR
FR
Efficiency

759

760 Figure 4.

761 Flow-chart describing the steps for each run of the simulations: generation of an original tree, sampling of two subtrees with *n* taxa, pruning and regrafting of taxa, supertree analysis, comparison of the optimal tree(s) with the original tree in order to calculate *TR*, *FR*, and efficiency.

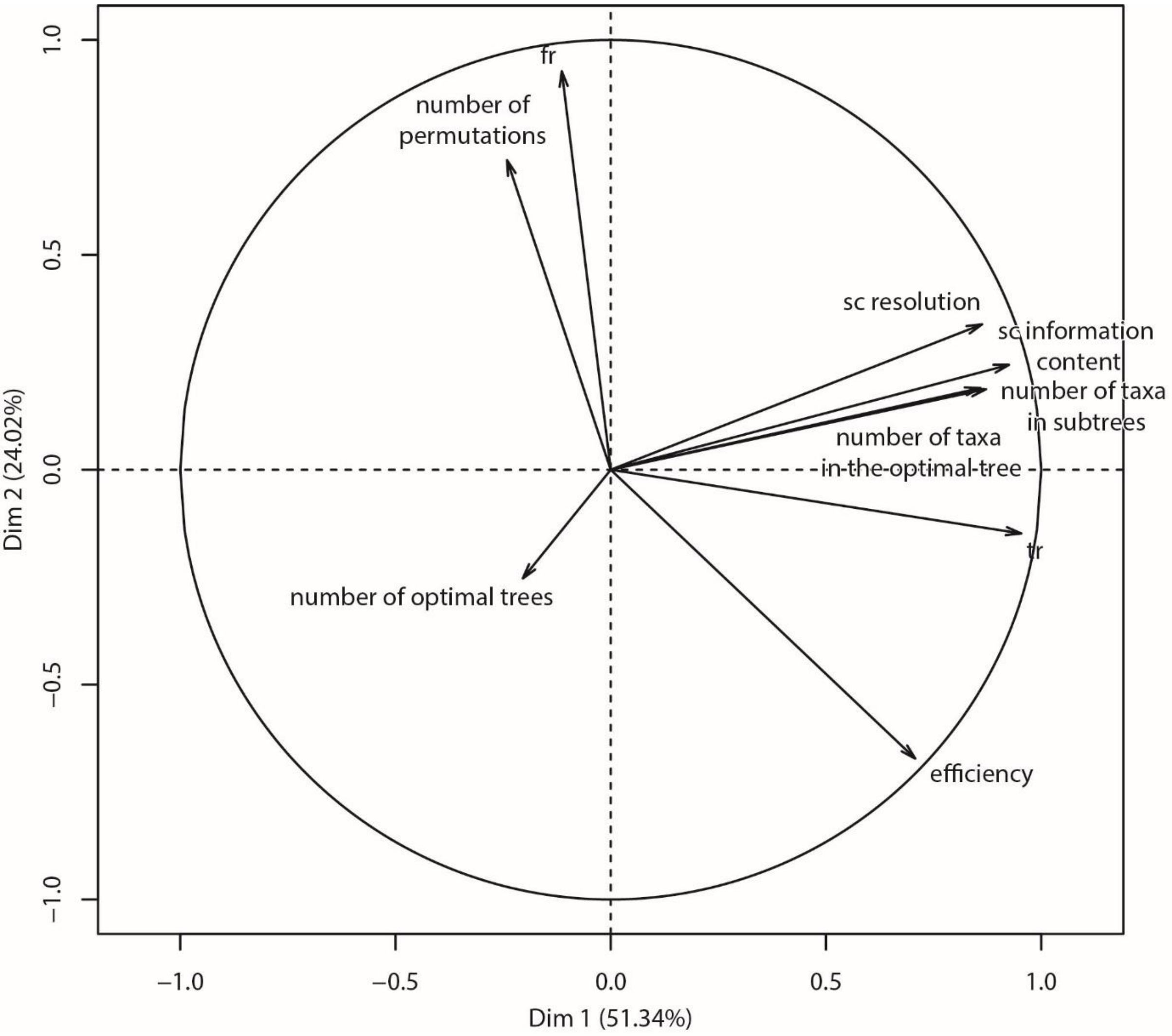


Figure 5.

Variables factor map of the PCA for the two first principal components (*PCA* function from *FactoMineR* R package). The first axis represents 51.34% of the variance and correlates positively with *TR* (true resolutions), numbers of taxa in subtrees and optimal tree, strict consensus (sc) information content and resolution. The second axis represents 21.02% of the variance and correlates positively with FR (false resolutions) and the number of permutations. The variance of the efficiency is split between the first and second axis.

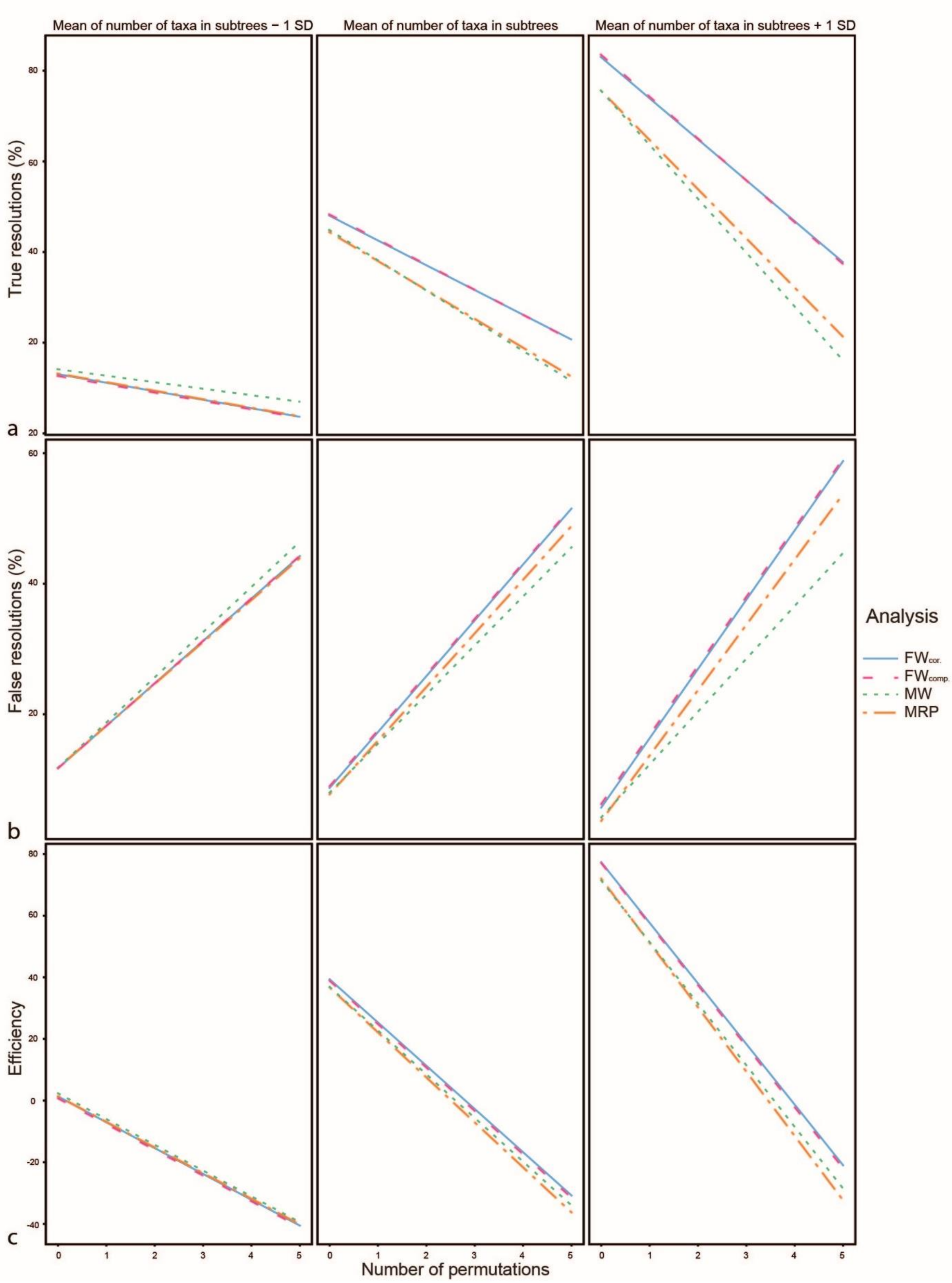

Mean of number of taxa in subtrees − 1 SD
Mean of number of taxa in subtrees
Mean of number of taxa in subtrees + 1 SD
True resolutions (%)
False resolutions (%)
Efficiency
a
b
c
80
60
40
20
60
40
20
80
60
40
20
0
-20
-40
0
1
2
3
4
5
Number of permutations
Analysis
FWcor.
FWcomp.
MW
MRP

Figure 6.

Interaction plots illustrating three linear models all including as independent variables the number of taxa in subtrees, the number of permutations and the types of analysis (function *interact_plot* from *jtools* R package. a. TR (number of true resolutions) as dependent variable. b. FR (number of false resolutions) as dependent variable. c. Efficiency as dependent variable. Types of lines represent the types of analysis, solid line: $FW_{cor.}$, dashed line: $FW_{comp.}$, pointed line: MW, dashed and pointed line: MRP.

Table I.

Decomposition of the tree (*a*(*bc*(*de*)) into 3ts and calculation of their weights using different methods.

3ts	Fractionnal weighting proposed herein	Nelson and Ladiges (1992) total fractionnal weighting	Nelson and Ladiges (1992) fractionnal weighting per component	Uniform weighting	Minimal weighting (Wilkinson et al. 2004)
(d,e),c	1	6/8	1	1	3/8
(d,e),b	1	6/8	1	1	3/8
(d,e),a	1	6/8	3/2 (1+1/2)	2	3/8
(b,c),a	6/10 (=1/2+1/10)	6/8	1/2	1	3/8
(b,d),a	6/10 (=1/2+1/10)	6/8	1/2	1	3/8
(b,e),a	6/10 (=1/2+1/10)	6/8	1/2	1	3/8
(c,d),a	6/10 (=1/2+1/10)	6/8	1/2	1	3/8
(c,e),a	6/10 (=1/2+1/10)	6/8	1/2	1	3/8
Total	6	6	6	9	3

Table II.

Pseudocode detailing the steps of removing the dependency of a set of 3ts from a tree by fractional weighting.

Pseudocode. Corrected_Fractional_weighting(tree)

```
if tree as informative nodes then:
    for each informative node:
        generate the component associated to the informative node
        calculate the number of leaves (n) branched to the informative node of the component
        decompose the component in all its implied 3ts
        attach a weight of 2/n to each 3ts
    if number of informative nodes > 1:
        for each couple of informative nodes directly included into each other from least inclusive to most inclusive nodes:
            calculate the sum of weights (t) of the 3ts from the least inclusive node that are common with the most inclusive node
            suppress the 3ts from the least inclusive node that are common with the most inclusive node
            calculate the number of remaining 3ts (i)
            for each remaining 3ts:
                Add t/i to the weight of the 3ts
    return the computed 3is with their associated fractionnal weighting
else
    return
```

Nelson and Ladiges (1992) fractionnal weighting per component

Weighting correction proposed herein

Appendix 1

a. Analysis of the different types of relationships between 3ts from a pectinate tree (*a*(*b*(*cd*))) and its components (*ab*(*cd*)) and (*a*(*b*(*cd*)) taken independently. b. Analysis of the different types of relationships between 3ts from a symmetric tree ((*ab*)(*cd*)) and its components (*ab*(*cd*)) and ((*ab*)*cd*) taken independently. c. Analysis of the different types of relationships between 3ts from a mixed tree (((*ab*)(*cd*))*e*) and its components ((*ab*)*cde*), (*ab*(*cd*)*e*), and ((*abcd*)*e*) taken independently. The different types of relationships are in-in, in-out, orthologous (o), paralogous (p). An empty cell means 3ts are compatible but not combinable.

a.

Component 1:
(a,b,(c,d))

	b(c,d)	a(c,d)
b(c,d)	.	in-in
a(c,d)		.

Component 2: (a,(b,c,d))

	a(b,d)	a(c,d)	a(b,c)
a(b,d)	.	in-out	in-out
a(c,d)		.	in-out
a(b,c)			.

Tree: (a,(b,(c,d))) (component 1 + 2)

	b(c,d)	a(b,d)	a(c,d)	a(b,c)
b(c,d)	.	o	in-in	o
a(b,d)		.	in-out	in-out
a(c,d)			.	in-out
a(b,c)				.

b.

Component 1:
(a,b,(c,d))

	a(c,d)	b(c,d)
a(c,d)	.	in-in
b(c,d)		.

Component 2:
((a,b),c,d)

	c(a,b)	d(a,b)
c(a,b)	.	in-in
d(a,b)		.

Tree: ((a,b),(c,d)) (component 1 + 2)

	c(a,b)	a(c,d)	d(a,b)	b(c,d)
c(a,b)	.	p	in-in	p
a(c,d)		.	p	in-in
d(a,b)			.	p
b(c,d)				.

c.

Component 1: ((a,b),c,d,e)

	c(a,b)	d(a,b)	e(a,b)
c(a,b)	.	in-in	in-in
d(a,b)		.	in-in
e(a,b)			.

Component 2: (a,b,(c,d),e)

	a(c,d)	b(c,d)	e(c,d)
a(c,d)	.	in-in	in-in
b(c,d)		.	in-in
e(c,d)			.

Component 3: ((a,b,c,d),e)

	a(b,c)	a(c,d)	a(d,e)	a,(b,d)	a,(b,e)	a,(c,e)
a(b,c)	.	in-out		in-out	in-out	in-out
a(c,d)		.	in-out	in-out		in-out
a(d,e)			.	in-out	in-out	in-out
a,(b,d)				.	in-out	
a,(b,e)					.	in-out
a,(c,e)						.

Tree: (((a,b),(c,d)),e) (component 1 + 2 + 3)

	c(a,b)	d(a,b)	e(a,b)	a(c,d)	b(c,d)	e(c,d)	e(a,c)	e(a,d)	e(b,c)	e(b,d)
c(a,b)	.	in-in	in-in	p	p		o		o	
d(a,b)		.	in-in	p	p			o		o
e(a,b)			.				in-out	in-out	in-out	in-out
a(c,d)				.	in-in	in-in	o	o		
b(c,d)					.	in-in			o	o
e(c,d)						.	in-out	in-out	in-out	in-out
e(a,c)							.	in-out	in-out	
e(a,d)								.		in-out
e(b,c)									.	in-out
e(b,d)										.

Supplementary Material

SM 1.

Boxplots of true resolutions (TR), false resolutions (FR) and efficiency in regard to the number of taxa in subtrees, split by the type of analysis (MRP: matrix representation with parsimony, $\text{FW}_{\text{cor.}}$: 3ta with corrected fractional weighting, MW: 3ta with minimal weighting, $\text{FW}_{\text{comp.}}$: 3ta with fractional weighting per component) and the number of permutations.

SM 2.

Movie of a three dimensions block diagram with the percentages of true and false resolutions (TR and FR), the number of taxa in the optimal tree (also highlighted by colors). Darkening of each color depends on the resolutions of the optimal tree, dark corresponds to poorly resolved trees and bright corresponds to well-resolved tree.

SM 3.

Summaries of the three linear models containing formulas of the models (bold), the estimates of each coefficient as well as standard error, t-values, and p-values.

SM 4.

Loess smoothed lines indicating a deviation from linearity effect from the three linear models.