Syst. Biol. 59(5):584–593, 2010

DOI:10.1093/sysbio/syq044
Advance Access publication on September 3, 2010

An Intuitive, Informative, and Most Balanced Representation of Phylogenetic Topologies

WATARU IWASAKI[1,*] AND TOSHIHISA TAKAGI[1,2,3]

[1]*Department of Computational Biology, The University of Tokyo, Kashiwa, Chiba 277-8568, Japan;* [2]*Database Center for Life Science, Research Organization of Information and Systems, Bunkyo-ku, Tokyo 113-0032, Japan; and* [3]*National Institute of Genetics, Research Organization of Information and Systems, Mishima, Shizuoka 411-8540, Japan;*
**Correspondence to be sent to: Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan; E-mail: iwasaki@k.u-tokyo.ac.jp.*

Received 24 June 2009; reviews returned 29 September 2009; accepted 29 April 2010
Associate Editor: Mark Holder

Abstract.—The recent explosion in the availability of genetic sequence data has made large-scale phylogenetic inference routine in many life sciences laboratories. The outcomes of such analyses are, typically, a variety of candidate phylogenetic relationships or tree topologies, even when the power of genome-scale data is exploited. Because much phylogenetic information must be buried in such topology distributions, it is important to reveal that information as effectively as possible; however, existing methods need to adopt complex structures to represent such information. Hence, researchers, in particular those not experts in evolutionary studies, sometimes hesitate to adopt these methods and much phylogenetic information could be overlooked and wasted. In this paper, we propose the centroid wheel tree representation, which is an informative representation of phylogenetic topology distributions, and which can be readily interpreted even by nonexperts. Furthermore, we mathematically prove this to be the most balanced representation of phylogenetic topologies and efficiently solvable in the framework of the traveling salesman problem, for which very sophisticated program packages are available. This theoretically and practically superior representation should aid biologists faced with abundant data. The centroid representation introduced here is fairly general, so it can be applied to other fields that are characterized by high-dimensional solution spaces and large quantities of noisy data. The software is implemented in Java and available via http://cwt.cb.k.u-tokyo.ac.jp/. [Centroid wheel tree; centroid representation; phylogenetic tree; probability distribution; traveling salesman problem.]

In this era of 1000 sequenced genomes, it is routine in most life sciences laboratories to search sequence databases for evolutionarily related sequences and build phylogenetic trees, which often become very large. The most common outcomes of such analyses are a variety of candidate tree topologies, even when the power of genome-scale data is exploited (Ciccarelli et al. 2006). More than one topology is usually produced by a number of methods, for example, the traditional bootstrap analysis (Felsenstein 1985) and the more modern Bayesian Markov chain Monte Carlo method (Huelsenbeck et al. 2001). Biologists today, therefore, tend to be faced with more and more candidate phylogenetic topologies.

Figure 1 schematically illustrates two widely used representations for summarizing phylogenetic topology distributions. Consider a phylogenetic relationship of four clades is investigated with three candidate topologies inferred as shown (Fig. 1a). The most commonly adopted representation is to simply put bootstrap value or posterior probability on each internal branch of the "best tree", which is usually the tree with the highest likelihood or that created using the original sequence alignment (Fig. 1b). These values are called "supports", defined as the proportion that the candidate topologies contain the corresponding "splits", that is, the bipartition systems of all clades. An apparent drawback of this representation is that it cannot convey information on candidate splits that are absent from the best tree. For example, although the second topology in Figure 1a is actually supported almost as strongly as is the first, Figure 1b does not reveal anything about it and focuses only on the first topology. To avoid such potentially misleading bias, researchers sometimes adopt another common representation, the consensus tree (Fig. 1c). This tree consists of splits with support greater than an arbitrary value. However, this representation also fails to reveal valuable information about the distribution. Although neither the first nor the second topology in Figure 1a is supported with enough confidence individually, they are still more strongly supported than that shown on the third, but this information is not reflected at all in Figure 1c.

To better represent topology distributions, three major approaches have been intensively investigated. The first is to simply provide reliability information on every possible split, for example, in the form of bar graphs, by abandoning use of geometrical structures. One of the most established methods in this category is related to spectral analysis (Hendy and Penny 1993; Charleston 1998). Given sequence or distance data, this method estimates the support for every possible split by correcting for the effects of parallel and multiple substitutions, independently of the choice of a phylogenetic topology. Then, the estimated support values are presented as a bar graph along with a tree that best describes the data. Hence, for splits that do not appear in this tree, the phylogenetic relationships between them need to be inferred by researchers. This becomes particularly painstaking when clade numbers grow because, in this case, the numbers of possible splits grow exponentially and the relationships between them become much more complicated. The second approach presents multiple topologies instead of just one. Some methods try to choose as few topologies as possible to represent the distribution (Wilkinson 1996; Stockham et al.

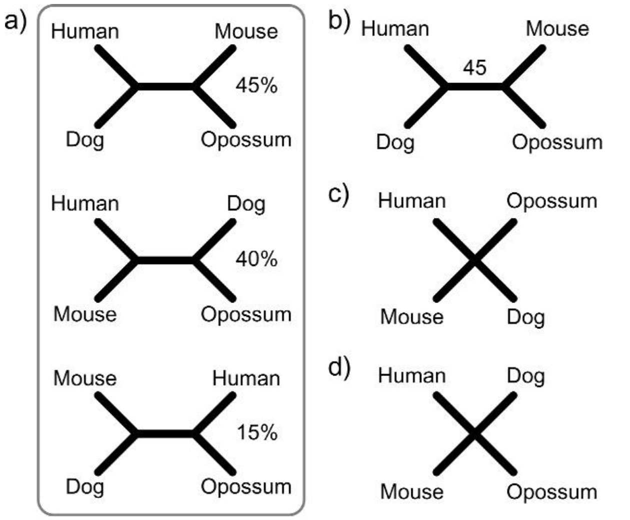

FIGURE 1. Commonly used representations of phylogenetic topology distributions. a) Three candidate topologies for the four clades and their example occurrence probabilities. b) The best tree representation. c, d) Two possible consensus tree representations. They are formally the same, though it is possible to develop a visualization method that treats them as different. More specifically, (a) can be better represented by (d) than (c) (see text).

2002; Bonnard et al. 2006), whereas others regard each topology as a data point and project these points onto some space to visualize the distribution (Hillis et al. 2005; Nye 2008). A similar drawback of these methods is that they ultimately require researchers to investigate the multiple topologies and combine their information, by themselves, to interpret the distribution. The third approach is the application of phylogenetic networks, which try to visualize information within networks instead of trees (Huson and Bryant 2006). This approach comprises an array of sophisticated methods, including so-called reticulate networks, which are particularly effective in illustrating some specific evolutionary events such as hybridizations, horizontal gene transfers, and recombination (Huson and Kloepper 2005). Nonetheless, phylogenetic networks also have several drawbacks. First, from the viewpoint of biology, trees would be still more appropriate than networks for representing some basic aspects of evolution, such as the evolution of species and most eukaryotic genes (Galtier and Daubin 2008). Second, from the viewpoints of mathematics and statistics, it is sometimes difficult to interpret phylogenetic networks (Galtier and Daubin 2008). Complicating the situation is the fact that they are based on evolutionary models that vary considerably between individual methods (Huson and Bryant 2006) and techniques whose statistical properties are not clear (e.g., greedy algorithms; Holland and Moulton 2003; Bryant and Moulton 2004; Huson et al. 2004). Finally, from the viewpoint of intuitiveness, they can become just complex, especially for non-experts of evolutionary studies (see, e.g., Huson et al. 2004). In light of this, researchers have sometimes hesitated to adopt the use of phylogenetic networks.

In this paper, we propose the "centroid wheel tree" (CWT) representation, which best reflects the entire distribution of candidate phylogenetic topologies and which can be interpreted intuitively even by nonexperts. The CWT is based on the sound mathematics of "centroid representation", which we introduce as a theoretically balanced representation of probability distributions. This is an extension of the "centroid estimation", by which estimation is carried out with the "centroid" of candidate solutions instead of the best solution (Carvalho and Lawrence 2008). This technique has recently been reported to be effective in bioinformatics when applied to problems characterized by high-dimensional solution space and considerable noise (Hamada et al. 2009; Joshi et al. 2009). We show that the problem of finding the CWT can be solved within the framework of the traveling salesman problem (TSP), a rigorously studied branch of optimization problems for which very sophisticated program packages are available (Applegate et al.). The name wheel tree is taken from the English common name of *Trochodendron aralioides* (an eudicot native to East Asia) for two of its defining characteristics: leaves growing in wheel-like shapes and a mixture of advanced and specialized characters in the eudicot family.

FORMULATION AND PROPERTIES

Key Idea

The key idea behind the CWT is to assign special meaning to circular orderings of branches around multifurcating nodes (i.e., nodes adjacent to ≥4 branches) in consensus trees. These nodes are generated by collapsing weakly supported edges to zero lengths while the consensus trees are being built. Traditionally, in the context of phylogenetic analysis, and in the wider field of graph theory, only connectivity is considered and such circular orderings in layout are usually ignored. For example, the consensus trees in Figure 1c,d have been treated identically. However, humans can interpret the two trees differently. In Figure 1d, clade human can be interpreted as more distant from opossum than from mouse and dog. What is most important is that, in this miniature data set from Figure 1a, human appears next to opossum less frequently and would be represented by Figure 1d more adequately than by Figure 1c. Hereafter, we call such multifurcations that consider circular orderings of the branches "wheel nodes", and consensus trees that contain wheel nodes "wheel trees."

Formulation

The concrete procedure for building CWTs is given in Figure 2a. The procedure accepts phylogenetic trees containing the same set of taxonomic units and first builds a consensus tree with a given threshold. It can also accept weight values that reflect observed frequencies or probabilities of the trees. Then, for each multifurcating node (v) on the consensus tree, the best circular ordering

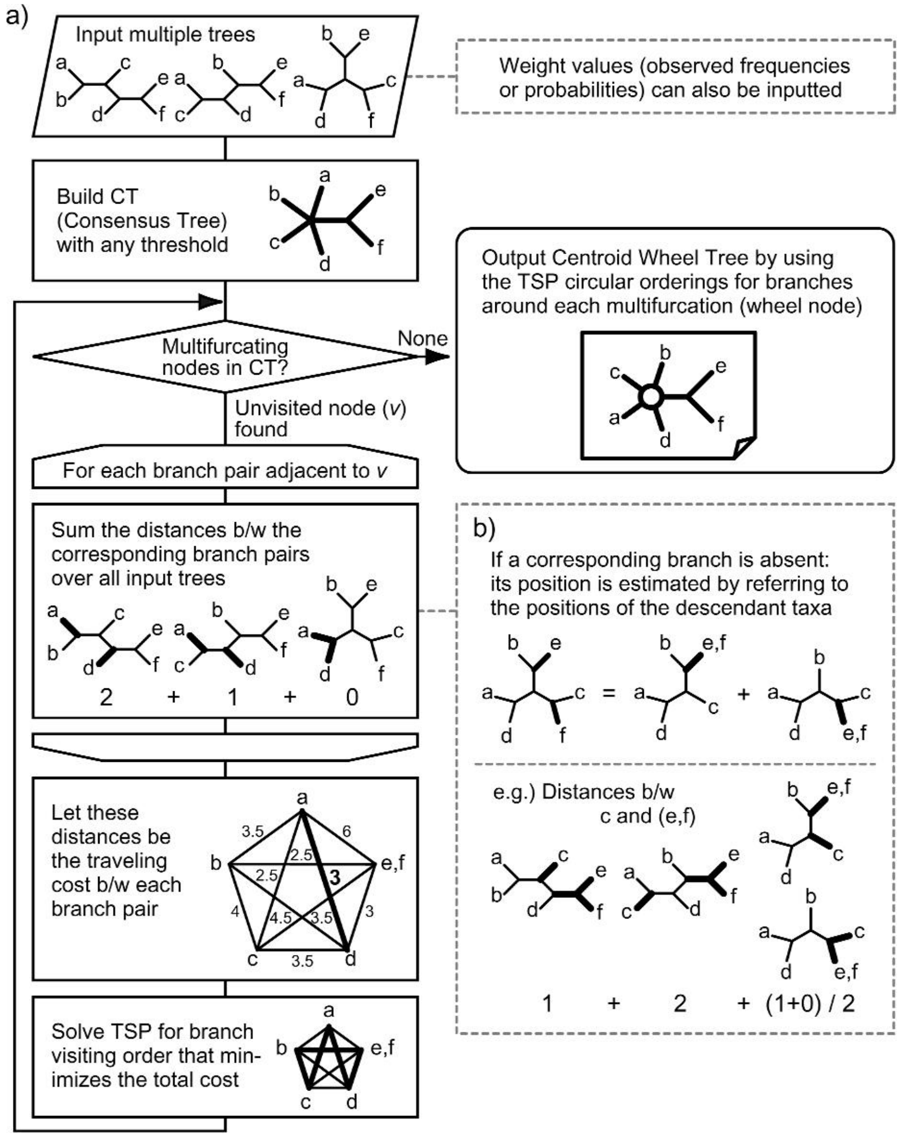
a)
Input multiple trees
Weight values (observed frequencies or probabilities) can also be inputted
Build CT (Consensus Tree) with any threshold
Multifurcating nodes in CT?
None
Output Centroid Wheel Tree by using the TSP circular orderings for branches around each multifurcation (wheel node)
Unvisited node (v) found
For each branch pair adjacent to v
Sum the distances b/w the corresponding branch pairs over all input trees
2 + 1 + 0
Let these distances be the traveling cost b/w each branch pair
3.5
6
2.5
2.5
3
4
4.5
3.5
3
3.5
Solve TSP for branch visiting order that minimizes the total cost
b)
If a corresponding branch is absent: its position is estimated by referring to the positions of the descendant taxa
=
+
e.g.) Distances b/w c and (e,f)
1 + 2 + (1+0) / 2
a
b
c
d
e
f
e,f

FIGURE 2. Procedure for building CWTs. a) The procedure first builds a consensus tree with a given threshold. Then, all branches adjacent to its multifurcating nodes are circularly ordered to best represent the topology distributions, as close branches in the input trees become successive. This is done by calculating distance between each branch pair (in this example, one between a and d is 3; thick lines). The best orderings are calculated as the solutions of the traveling salesman problem. b) For trees that do not contain all splits adjacent to each multifurcating node, their positions are estimated by referring to the positions of the descendant taxa.

of the branches adjacent to v (*v-branches*) is calculated through the following steps. First, for each branch pair of *v-branches*, sum the distances (the minimum number of edges) between the corresponding branch pairs over all input trees. See the next paragraph for trees that do not contain all splits of *v-branches*. Second, let these values be the "traveling cost" between each branch pair. Third, solve the TSP tour that minimizes the total traveling cost to obtain the best circular ordering of *v-branches*. Intuitively, this traveling cost was designed to become large if distant splits in the input trees are successive around the wheel nodes. In other words, the TSP solution minimizing the total cost makes close splits in the input trees successive as much as possible, embodying the key idea described in the previous paragraph. After repeating the above for all multifurcating nodes, finally visualize the consensus tree by using the TSP circular orderings for branches around each multifurcation (wheel node).

It is possible that the input trees contain a tree T that does not contain all splits of *v-branches*. In this case, the distances on T are defined by converting T into a set of trees containing those splits (Fig. 2b). The basic idea is that, for example, in Figure 2b, the branch leading to (e,f) would be observed at the positions of taxa e and f with the same probabilities. More precisely, for each branch b of *v-branches*, collect all its descendant taxa on the consensus tree by regarding v as the ancestor and choose one of them randomly. Then, mark the external branch adjacent to this taxon on the tree T as corresponding to b. After repeating these steps for all branches of *v-branches*, remove all unnecessary parts from T (external branches that are not marked and internal nodes that are attached to only two branches). The distances between each pair of *v-branches* are then calculated by using the processed tree. Finally, by repeating the above steps for all possible combinations of the taxa choice and averaging the distances, the expected distance between each split pair is obtained. Note that, if T contains all splits of *v-branches*, the procedure above always results in the same tree, in which topological relationships among the branches corresponding to *v-branches* are the same as in the original T. Therefore, this is a natural extension of the original definition of distance, letting the circular orderings reflect the information in the whole topology distribution as much as possible.

It should also be noted that although trees are treated as unrooted throughout this paper, it is not difficult to root a CWT because it retains its tree shape and our current implementation can actually accept rooted trees (see Software Availability). The more precise pseudocode and formulation are given in the Appendix.

CWT is the Most Balanced Representation

In this section, we show that the CWT produced by the procedure outlined above not only embodies the key idea we sketched out but, mathematically, is the centroid representation, which is the most balanced representation of the topology distributions regarding the loss functions we define below. We also provide an intuitive explanation in the last paragraph.

Let us first introduce the concept of the centroid representation. Let θ, D, and $P(\theta|D)$ be unobserved data, observed data, and the posterior probability of θ given D, respectively. Traditionally, a maximum a posteriori estimator $\hat{\theta}_{\mathrm{MAP}} = \mathrm{argmax}_{\theta} P(\theta|D)$ is often adopted to represent the whole distribution of candidate θ. However, $\hat{\theta}_{\mathrm{MAP}}$ may be an ineffective representation of collections of small-probability solutions in high-dimensional spaces with considerable noise (Carvalho and Lawrence 2008), features that are shared by phylogenetic inference. Instead, the centroid representation tries to capture the distribution of the posterior probability mass in the ensemble. Given a "loss function" $L(\varphi, \theta)$ that quantifies how each θ deviates from the representation φ, the centroid representation is defined by $\varphi_{\mathrm{Centroid}} = \arg\min_{\varphi} \sum_{\theta} L(\varphi, \theta) P(\theta|D)$, which is theoretically the best for minimizing the expected value of the loss function. If $\varphi_{\mathrm{Centroid}}$ itself is a candidate solution, it is called the "centroid estimator," which is introduced in detail in (Carvalho and Lawrence 2008), along with important extensions to the Hamming loss function.

A CWT is the centroid representation regarding the two loss functions below. Assume that a wheel tree (Φ) and one of the input trees ($T_i \in \{T_1, \ldots, T_N\}$) are given, and let $l_{\mathrm{Layout}}(\Phi, v, T_i)$ be the sum of the expected distances between any T_i branch pair whose corresponding branch pair is successive around a wheel node v in Φ. Then the "layout incongruity loss function" $L_{\mathrm{Layout}}(\Phi, T_i)$ is defined as the total sum of $l_{\mathrm{Layout}}(\Phi, v, T_i)$ over every wheel node, to quantify how all circular orderings in Φ agree with T_i. Among all possible Φ, a CWT minimizes the expected loss function against all input trees $T_1, \ldots, T_N$. In addition, the traditional consensus tree that CWT relies on is also a centroid representation regarding the split incongruity loss function, which quantifies how the split sets of two trees disagree with each other. See the Appendix for details and proofs.

It is useful to provide some intuitive explanation of the fact that a CWT is the centroid representation here. Imagine a heap of stones, each of which has a drawing of a phylogenetic topology and whose mass is proportional to the probability of the tree. If we place the stones so that similar topologies are close to each other, a map of a topology distribution is created, as in Hillis et al. (2005) and Nye (2008). If this distribution is unimodal and unbiased, the heaviest stone or the best tree is anticipated to be at the "center" of the distribution and represent it well, although sometimes this assumption does not hold true, as discussed in the references above. Instead, the CWT representation tries to find the centroid (or the center of gravity) of the topologies, by placing them on a light plate on which each position corresponds to a wheel tree. In classical mechanics, the centroid is the point that balances the moment of gravity and minimizes the sum of the products of mass and squared distances. This clearly resembles the definition of a CWT, as we use the cost functions defined above instead of the squared distances, to quantify how close a wheel tree and a phylogenetic topology are. Therefore, the CWT representation can be regarded as the centroid that balances the background topology distributions best.

APPLICATIONS

Visualization of CWTs

In addition to making the circular orderings around multifurcating wheel nodes best represent the topology distributions, it is useful if some statistical information on them is available. Our current implementation of the CWT program provides visualization as in Figure 3.

First, as in ordinary phylogenetic representations, numbers at internal branches are the proportions of the input trees containing the corresponding splits. Second, numbers around the wheel nodes indicate the proportions that the flanking splits constitute in a monophyletic group. For example, in Figure 3, Z_1% of the trees have the 3-furcation {a | d | b,c,e,f} (i.e., the three splits {a | b,c,d,e,f}, {d | a,b,c,e,f}, and {a,d | b,c,e,f}). Third, numbers within wheel nodes indicate the extent to which each circular ordering of the branches naturally represents the candidate topologies. In other words, the numbers indicate the proportion of the input topologies that can be restored by just "pulling out" branches without changing the orderings. (See the right-hand

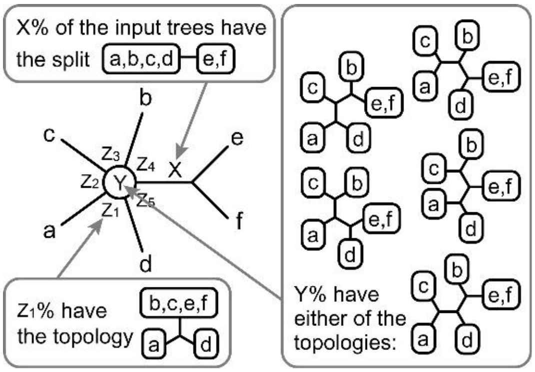

FIGURE 3. Information that the CWT representation conveys. Branches around multifurcating nodes (wheel nodes) are circularly ordered by the procedure in Figure 2. As in the ordinary phylogenetic representations, numbers at internal branches (X) are the support values of the corresponding splits. Numbers around wheel nodes ($Z_1, \ldots, Z_5$) indicate the proportions of the flanking splits that constitute a 3-furcation. Numbers within wheel nodes (Y) indicate the proportions of the input trees for which their circular ordering is the best (i.e., does not require traversing the same branches three times to visit all splits). Note that the input trees that do not have all the split sets adjacent to the wheel node are also used to estimate the values within and around the wheel nodes. Options for displaying the strict supports excluding them and the average distances are also available.

box in Figure 3 for an example. As another example, the value for the wheel node in Figure 1d is 85%, which is the total support of the first and second topologies in Figure 1a). Mathematically, this equals the proportion of the ordering that constitutes a shortest TSP tour, given the distance matrices derived using each topology only.

Note that, at the default settings, the values within and around the wheel nodes are calculated "on the assumption of the consensus tree topology." For the input trees that do not have all the split sets adjacent to the wheel node, the proportion of topologies within all possible combinations of the corresponding branch positions (see "Formulation" and Fig. 2b) that contain the corresponding topologies are added to those values. Therefore, they are not strict support values (like the ones at internal branches) but expected values based on the assumption of the consensus tree topology. For the users' sake, the current implementation also offers options for showing the strict values that input trees contain the corresponding topologies as well as the average distances used for the TSP calculation.

CWT Applied to a Real Data Set

Figure 4a is a CWT representation derived from a real data set. The input trees were 246 phylogenetic trees obtained by applying the maximum-likelihood method to reliable single-copy orthologs conserved among 21 fungal species (downloaded from FUNYBASE; Marthey et al. 2008); a 60% threshold was used in building the consensus tree.

The wheel node indicated by the thick arrow connects the four splits Ago, Kla, X, and Y. This node indicates that, given the topology of the consensus tree, 73% of the input trees are expected to contain either of the splits {Ago,Kla | X,Y} or {Kla,X | Ago,Y}, and 39% and 34% contain the 3-furcations {Ago | Kla | X,Y} and {Kla | X | Ago,Y}, respectively. Neither the best tree nor the consensus tree representation provides such detailed information, unless multiple trees are used or the tree shapes are abandoned as in the phylogenetic network representation (Fig. 4b). Note that the values on the opposite sides of the wheel nodes are the same. This is because in cases of 4-furcating nodes, for example, if the splits Ago and Kla are adjacent (i.e., {Ago | Kla | X,Y}) then X and Y are adjacent (i.e., {X | Y | Ago, Kla}). This property does not hold true without the consensus tree topology assumption, because the input trees can contain trees that do not contain the splits X and Y and in such cases trees with {Ago | Kla | X,Y} can be without {X | Y | Ago,Kla} (Fig. 4a, left sides of the dotted rectangles). It is also worthwhile to examine the average distance around each wheel node used in the TSP calculation (Fig. 4a, right sides of the dotted rectangles); in cases of 4-furcating nodes, the distances between splits are equal to the proportions that they do not constitute in a monophyletic group.

In addition, because a CWT uses a tree shape, taxonomic groups at multiple levels can be recognized fairly intuitively. In particular, a CWT can suggest the existence of taxonomic groups with support below the consensus-tree threshold, in addition to those making splits on the consensus trees, as successive branches around wheel nodes. Figure 5a is a CWT representation of the same data set as in Figure 4, based on an 80% consensus tree. To avoid a cluttered appearance because of the many multifurcations, a color visualization option that uses colors instead of numbers is used to show the support values around the wheel nodes. For example, the thick gray arc lines α, β, γ, and δ indicate the class Eurotiomycetes, class Sordariomycetes, order Hypocreales, and subphylum Agaricomycotina, respectively. An interesting application is the star-like CWT, obtained by specifying high threshold values where no split is supported at a level above the threshold (Fig. 5b; branch lengths are ignored and only the topology is shown). The star-like CWT shows "the optimal sequential ordering" of all taxa based on the distribution of the input topologies and, as a result, many biological groups appear as successive branches around the wheel node (thick gray arc lines). By virtue of the sophistication of the present TSP solvers (Applegate et al.), it takes only a few seconds to obtain such star-like CWTs of this size.

Characteristics of CWTs Clarified by an Artificial Data Set

To help understand how the CWT representation works, we show an artificially created example that sheds light on the characteristics of CWTs. Figure 6a is a data set containing 15 trees, and Figure 6b,c are the produced CWT representations based on the 50%

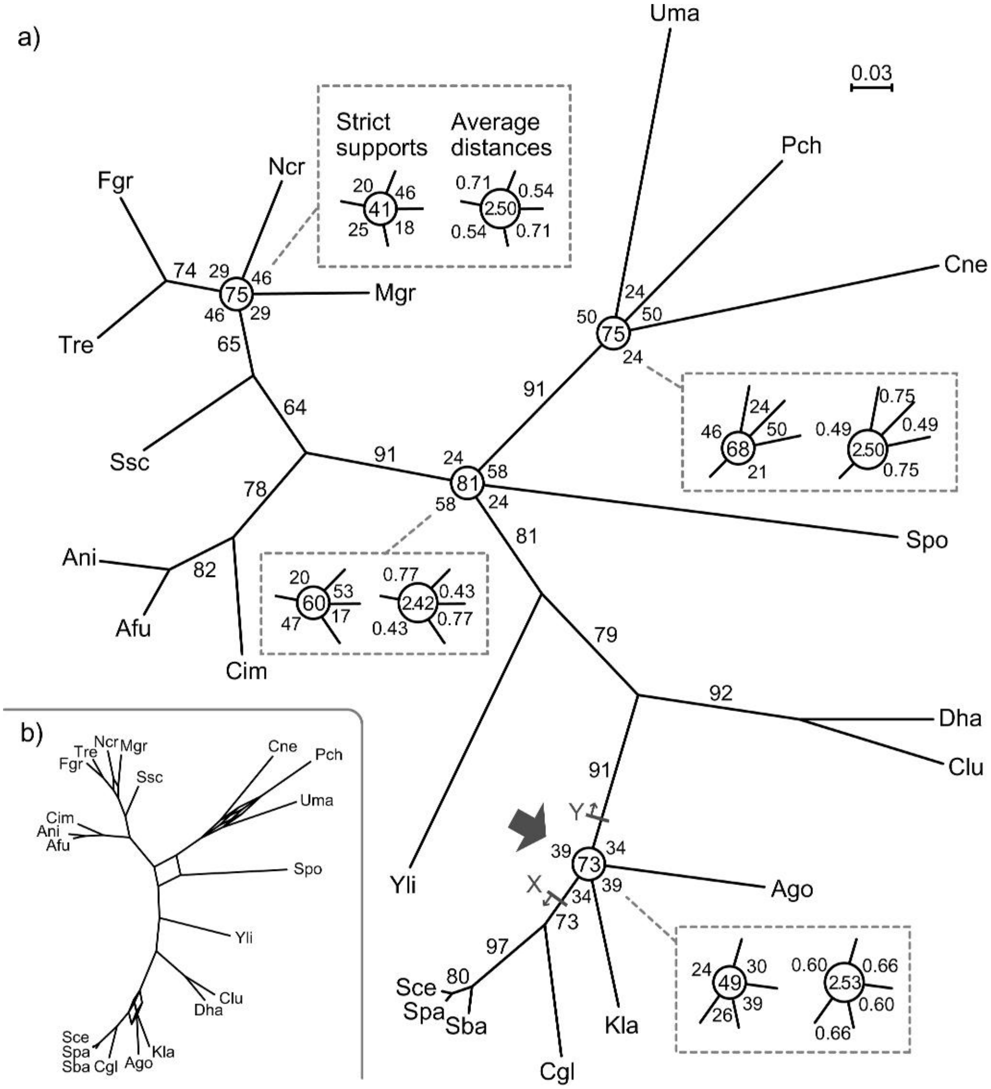

FIGURE 4. CWT for 264 trees based on single orthologs conserved over 21 fungal species. a) CWT based on the 60% consensus tree. The values at the wheel nodes include the estimated values for trees that do not contain all splits adjacent to them (default option). For each wheel node, the strict values excluding them and the average distances are shown in the dotted rectangles on the left and right sides, respectively. For the branch lengths, those stored in the original database (Marthey et al. 2008) are directly incorporated in their original units. b) Consensus phylogenetic network with a threshold of 0.2 (Holland and Moulton 2003). The scientific names of the species are as follows: Ago, *Ashbya gossypii*; Afu, *Aspergillus fumigatus*; Ani, *A. nidulans*; Cgl, *Candida glabrata*; Clu, *C. lusitaniae*; Cim, *Coccidioides immitis*; Cne, *Cryptococcus neoformans*; Dha, *Debaryomyces hansenii*; Fgr, *Fusarium graminearum*; Kla, *Kluyveromyces lactis*; Mgr, *Magnaporthe grisea*; Ncr, *Neurospora crassa*; Pch, *Phanerochaete chrysosporium*; Sba, *Saccharomyces bayanus*; Sce, *S. cerevisiae*; Spa, *S. paradoxus*; Spo, *Schizosaccharomyces pombe*; Ssc, *Sclerotinia sclerotiorum*; Tre, *Trichoderma reesei*; Uma, *Ustilago maydis*; Yli, *Yarrowia lipolytica*.

majority-rule consensus tree showing support values and average distances, respectively. The strict option does not change the support values because there is no nontrivial split exceeding the 50% threshold (Fig. 6d) and the consensus tree contains trivial splits only. Figure 6e shows the sums of the distances between each split pair, which are used for the TSP calculation for the circular ordering determination.

Two notable characteristics of CWT are demonstrated with this example. First, though the most frequent split is {c,e | a,b,d} (Fig. 6d), the distance between c and e is not the smallest (Fig. 6e). In other words, they are given the highest priority if we construct a consensus tree by adopting highly supported splits in a greedy manner (*Greedy consensus tree*, Fig. 6f) but not if we build the CWT. This is because the CWT considers more information than the existences of splits, that is, it also considers the distances between them. In this example, though the split {a,b | c,d,e} appears only once, the splits a and b are always within a distance of 1 in all trees and thus the cost for making them successively becomes small. Such differences can be made clear by visualizing and comparing the support values and the average distances (Fig. 6b,c; we recommend trying these options,

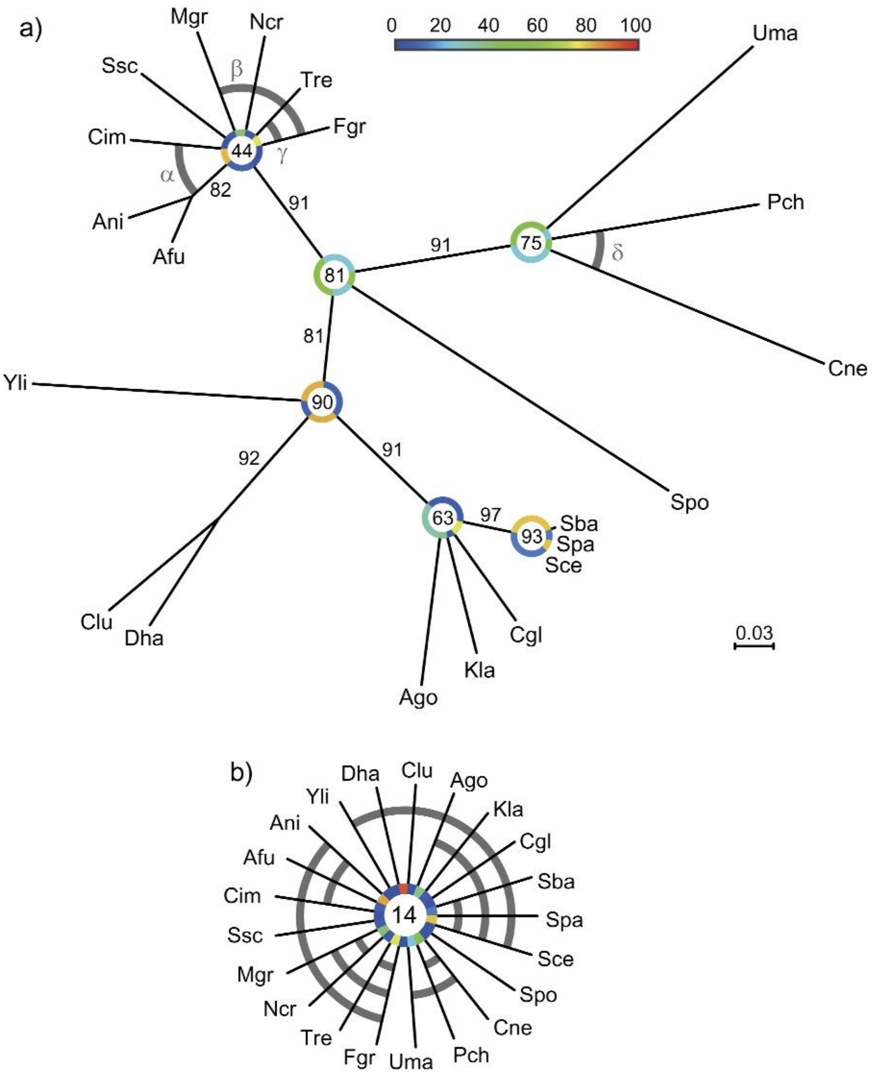

FIGURE 5. Taxonomic groups as successive branches around wheel nodes. a) CWT based on the 80% consensus tree for the same data set as in Figure 4. This is the color visualization, which uses colors instead of numbers to show values around the wheel nodes, to avoid a cluttered appearance. Thick gray arc lines indicate taxonomic groups that appear as successive splits around the wheel nodes. The support values include the estimated values for trees that do not contain all splits adjacent to the wheel nodes (default option). b) A star-like CWT based on the 100% consensus tree. The 21 species line up in the optimal order based on the cost function and the TSP solution. Many taxonomic groups constitute successive branches around the wheel node (gray arcs).

especially if weak support values are displayed). Though the support for the 3-furcation {c | e | a,b,d} is over twice as ones of {a | d | b,c,e} and {b | d | a,c,e}, the distances between c and e, a and d, and b and d are the same. Such information cannot be easily obtained by glancing at the input trees (Fig. 6a), the split sets (Fig. 6d), or the greedy consensus tree (Fig. 6f). The other characteristic of CWTs that can be seen in this example is that, although the smallest distance is between the splits a and b (Fig. 6e), they are not successive around the resultant wheel node (Fig. 6b,c). This occurs because of the intrinsic nature of TSP: choosing locally optimal paths does not always derive the globally best tour, and in this example the total distance of any tour containing the path a–b exceeds that of the tour a–c–e–b–d (Fig. 6e). This is intentional and reflects the very objective of a CWT: the most balanced representation of the topology distributions.

Software Availability

The software for building a CWT was implemented in Java, and is available at http://cwt.cb.k.u-tokyo.ac.jp/. Users can run the program on the Web or download it under the GNU General Public License. To run the program locally, it is necessary to install Java SE 5.0 (or higher), Concorde (one of the best exact TSP solvers currently available; Applegate et al.), Phyutility Java archive; Smith and Dunn 2008), Apache XML Graphics Commons Java archives, and Args4j Java archive.

The software accepts any set of rooted or unrooted phylogenetic trees in the Newick format covering the

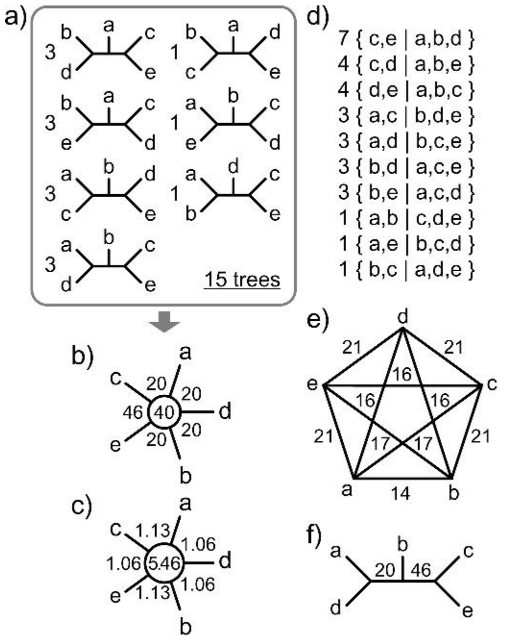

FIGURE 6. Characteristics of CWTs demonstrated with an artificial data set. a) The 15 input tree topologies. b) The CWT based on the 50% majority rule consensus tree. The support values do not change by the strict value option because this CWT contains trivial splits only. c) The CWT displaying the average distances. Because they use more information than the existence of splits, the distances are not proportional to the values in (b). For example, the distances can differ at positions with the same support values. d) The observed frequencies of every non-trivial split. Note that none exceeds the 50% threshold. e) The sums of the distances between each split pair for the TSP calculation. This figure illustrates the following characteristics of CWTs: The strongest supported split does not always correspond to the closest pair, and the closest pair is not always adopted in the circular ordering, which is based on the best TSP tour. f) Greedy consensus tree for the same data set.

same taxa set, with one tree per line. A real value separated by space at the beginning of each line can be provided to indicate the weight of the tree. A threshold value for building the consensus tree should be designated. If a threshold >100 is given, no split is supported stronger than the threshold and thus a star-like CWT is produced. Options to use rooted trees, show strict support values, show average distances, use the color visualization, and output only topologies are available.

The software produces the CWT representation as .nhx, .ps, and .png files. The .nhx files are in the New Hampshire eXtended format (http://www.phylosoft.org/NHX/), which is an extension of the conventional Newick format. The extra information described in the main text is given by using ":XN=" tags wrapped by "[&&NHX" and "]", which mean "custom data associated with nodes" in the NHX format. For example, if (i) a 4-furcating wheel node is surrounded by Clade1 to Clade4, (ii) the values flanked by Clade1 and Clade2, Clade2 and Clade3, Clade3 and Clade4, and Clade4 and Clade1 are Val1, Val2, Val3, and Val4, respectively, and (iii) the value within the wheel node is ValC, the notation is "(Clade1, Clade2, Clade3, Clade4)[&&NHX: XN=ValC|Val1,Val2,Val3,Val4];" (in the case that the node is the root) or "(Clade2, Clade3, Clade4)[&&NHX: XN=ValC|Val1,Val2,Val3,Val4]Clade1" (otherwise). That is, circular orderings of branches are designated by appearance order and statistical information is given in the square brackets. ":B=" tags can also be inserted before the ":XN=" tags, to represent confidence values for parent branches. .ps (PostScript) and .png (Portable Network Graphics) files provide 2-dimensional visualizations of the CWT. The visualization is based on a modified radial layout that follows the TSP-based circular orderings of branches.

CONCLUSION AND PERSPECTIVES

In this paper, we introduced the CWT representation, which provides rich information on phylogenetic topology distributions in a highly intuitive and most balanced manner. Examples based on real and artificial data sets were also provided to show the advantages and characteristics of CWTs. Better phylogenetic representations are of increasing importance because DNA sequencing technologies are advancing at an exponential rate, leading to ubiquitous demands for interpreting large-scale phylogenetic analyses. In addition, because a CWT conceptually resembles an intermediate data structure in phylogenetic network construction (Dress and Huson 2004), CWT could possibly be used to extend them. Furthermore, because the basic concept of the centroid representation is fairly general, it may also be applied to fields beyond the life sciences that are tending toward high-dimensional solution space, considerable noise, and data explosions.

FUNDING

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" (17017002) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and Grant-in-Aid for Young Scientists (Start-up) (21810005) from the Japan Society for the Promotion of Science.

ACKNOWLEDGMENTS

The authors thank M. Hamada, K. Kin, Y. Watanabe, J. Sullivan, M. Holder, and the anonymous reviewers for valuable suggestions and comments on the manuscript.

REFERENCES

Applegate D., Bixby R.E., Chvátal V., Cook W. http://www.tsp.gatech.edu/concorde.html.

Berry V., Gascuel O. 1996. On the interpretation of bootstrap trees: appropriate threshold of clade selection and Induced gain. Mol. Biol. Evol. 13:999–1011.

Bonnard C., Berry V., Lartillot N. 2006. Multipolar consensus for phylogenetic trees. Syst. Biol. 55:837–843.

Bryant D., Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Carvalho L.E., Lawrence C.E. 2008. Centroid estimation in discrete high-dimensional spaces with applications in biology. Proc. Natl. Acad. Sci. U S A. 105:3209–3214.

Charleston M.A. 1998. Spectrum: spectral analysis of phylogenetic data. Bioinformatics. 14:98–99.

Ciccarelli F.D, Doerks T., von Mering C., Creevey C.J., Snel B., Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311:1283–1287.

Dress A.W., Huson D.H. 2004. Constructing splits graphs. IEEE/ACM Trans Comput Biol. Bioinform. 1:109–115.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 39:783–791.

Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Phil. Trans. R. Soc. Lond. Ser. B. 363:4023–4029.

Hamada M., Kiryu H., Sato K., Mituyama T., Asai K. 2009. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics 25:465–473.

Hendy M.D., Penny D. 1993. Spectral Analysis of Phylogenetic Data. J Classif. 10:5–24.

Hillis D.M., Heath T.A., St John K. 2005. Analysis and visualization of tree space. Syst Biol. 54:471–482.

Holder M.T., Sukumaran J., Lewis P.O. 2008. A justification for reporting the majority-rule consensus tree in Bayesian phylogenetics. Syst Biol. 57:814–821.

Holland B., Moulton V. 2003. Consensus networks: a method for visualising incompatibilities in collections of trees. In: Benson G, Page R., editors. WABI 2003, LNBI 2812 . Berlin (Germany): Springer. p. 165–176.

Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 294:2310–2314.

Huson, D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Huson D.H., Dezulian T, Klopper T., Steel M.A. 2004. Phylogenetic super-networks from partial trees. IEEE/ACM Trans. Comput. Biol. Bioinform. 1:151–8.

Huson D.H., Kloepper T.H. 2005. Computing recombination networks from binary sequences. Bioinformatics 21 (Suppl 2):ii159–ii165.

Joshi A., De Smet R., Marchal K., Van de Peer Y., Michoel T. 2009. Module networks revisited: computational assessment and prioritization of model predictions. Bioinformatics. 25:490–6.

Margush T., McMorris F.R. 1981. Consensus n-Trees. Bull. Math. Biol. 43:239–244.

Marthey S., Aguileta G., Rodolphe F., Gendrault A., Giraud T., Fournier E., Lopez-Villavicencio M., Gautier A., Lebrun M.H., Chiapello H. 2008. FUNYBASE: a FUNgal phYlogenomic dataBASE. BMC Bioinform. 9:456.

Nye, T.M. 2008. Trees of trees: an approach to comparing multiple alternative phylogenies. Syst. Biol. 57:785–794.

Smith S.A., Dunn C.W. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–6.

Stockham C., Wang L.S., Warnow T. 2002. Statistically based postprocessing of phylogenetic analysis by clustering. Bioinformatics 18 (Suppl 1):S285–S293.

Wilkinson M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Mol. Biol. Evol. 13:437–44.

APPENDIX

Definitions

Let $T_1, \ldots, T_N$ be N input phylogenetic trees and $w_1, \ldots, w_N$ be their weights. Each phylogenetic tree is defined by $T_i = (V_i, E_i)$, where V_i is the set of all external and internal nodes and E_i is the set of all branches (edges). Hereafter, all trees are assumed to have the same set of taxa $X = \{x_1, \ldots, x_M\}$ on the external nodes.

Any branch e on a tree $T = (V, E)$ splits the taxa set X into two groups. We call such a bipartition system $Split(e, T)$. Let $SplitSet(T) = \{Split(e, T) | e \in E\}$ and $Freq(s)$ be the weighted frequency of the trees that contain the split s in $T_1, \ldots, T_N$, i.e., $Freq(s) = \sum_{i=1}^{N} \delta_i w_i$ where $\delta_i = 1$ *if* $s \in SplitSet(T_i)$ and $\delta_i = 0$ *otherwise*. $AdjBranches(v, T)$ is the set of branches directly attached to node v on T, and $MultiFurcatingNodes(T) = \{v \in V || AdjBranches(v, T)| \geqslant 4\}$. If $e \in AdjBranches(v, T)$, $Descendants(e, v, T)$ is the subset of X that can be traversed from the node adjacent to e that is not v, without crossing e. Let $ExtBranch(x, T)$ be an external branch on T attached to the taxon x.

For $X_{sub} \subseteq X$ and $\{x_a, x_b\} \subseteq X_{sub}$, we define $Dist(x_a, x_b, X_{sub}, T)$ as follows. First, recursively remove any external branch e from T if $e \notin \{ExtBranch(x, T) | x \in X_{sub}\}$. Second, recursively remove any internal node v if $|AdjBranches(v, T)| = 2$ by replacing $AdjBranches(v, T)$ with one internal branch to keep the tree T connected. Then $Dist(x_a, x_b, X_{sub}, T)$ is the minimum number of edges on the converted tree T' that separate $ExtBranch(x_a, T')$ and $ExtBranch(x_b, T')$. Note that $ExtBranch(x_a, T')$ and $ExtBranch(x_b, T')$ themselves are not counted (i.e., if they are adjacent each other, then $Dist(x_a, x_b, X_{sub}, T) = 0$).

Pseudocode for Obtaining CWTs

Given $T_1, \ldots, T_N$, $w_1, \ldots, w_N$, and a threshold value γ for building the consensus tree, the following pseudocode produces a CWT. Note that the function c is symmetric. i.e., $c(e_p, e_q) \equiv c(e_q, e_p)$.

```
Construct the consensus tree $T_C = (V_C, E_C)$, where
$SplitSet(T_C) = \left\{ s | Freq(s) \geqslant \gamma \sum_{i=1}^{N} w_i \right\}$.
For each $v \in MultiFurcatingNodes(T_C)$ {
    $\{e_1, e_2, \ldots, e_k\} = AdjBranches(v, T_C)$
    For each $\{p, q\} \subset \{1, 2, \ldots, k\}$ {
        Set $c(e_p, e_q) = 0$
        For each $T_i \in T_1, \ldots, T_N$ {
            For each taxa set $X_{sub} = \{x_1, x_2, \ldots, x_k\}$ that
             $\forall_{1 \leqslant t \leqslant k}\ x_t \in Descendants(e_t, v, T_c)$ {
            $c(e_p, e_q) = c(e_p, e_q) + Dist(x_p, x_q, X_{sub}, T_i) \cdot w_i / \prod_{t=1}^{k} |$
             $Descendants(e_t, v, T_c)|$
            }
        }
    }
    Find the TSP cycle $(e_{f(1)} e_{f(2)} \cdots e_{f(k)})$ for $AdjBranches$
$(v, T_C)$, which minimizes $\sum_{t=1}^{k-1} c(e_{f(t)}, e_{f(t+1)}) + c(e_{f(k)}, e_{f(1)})$
}
Lay out $T_C$ by following the TSP-cycle orderings for the
branches around $MultiFurcatingNodes(T_C)$.
```

Consensus Tree is the Centroid Representation

It can be shown that the consensus tree is in fact the centroid representation of all candidate trees regarding the loss function of "split incongruity", which quantifies the degree of disagreement between split sets of two trees. More precisely, given a tree representation

$T' = (V', E')$ and a candidate tree $T_i = (V_i, E_i)$, the "split incongruity loss function" is defined as

$$L_{split}\left(T', T_i\right) = \xi \cdot FP\left(T', T_i\right) + FN\left(T', T_i\right),$$

where

$$\begin{aligned} FP\left(T', T_i\right) &= \left|\left\{s \left| s \in SplitSet\left(T'\right) \wedge s \notin SplitSet\left(T_i\right)\right.\right\}\right| \\ FN\left(T', T_i\right) &= \left|\left\{s \left| s \notin SplitSet\left(T'\right) \wedge s \in SplitSet\left(T_i\right)\right.\right\}\right| \end{aligned}.$$

This function becomes large if T' contains splits that are absent from T_i ("false positives") and vice versa ("false negatives"). ξ designates the relative penalties for the two types of errors and usually $\xi \geq 1$ because by convention false positives are more undesirable, given that they would exaggerate weak phylogenetic signals. To obtain the centroid representation, it is necessary to know the posterior probability $P(T_i|D)$ in addition to the loss function and, for example, both bootstrap and Bayesian methods have been developed to give approximate values for $P(T_i|D)$ (Felsenstein 1985; Huelsenbeck et al. 2001). If we define $w_1, \ldots, w_N$ as $w_i \propto P(T_i|D)$, the "split incongruity centroid tree" $T_{SplitCentroid}$ is the tree that fulfills

$$SplitSet\left(T_{SplitCentroid}\right) = \left\{ s \left| \frac{\xi}{1+\xi} \leqslant \frac{Freq\left(s\right)}{\sum_{i=1}^{N} w_i} \right. \right\}$$

(Berry and Gascuel 1996; Holder et al. 2008; Margush and McMorris 1981). Therefore, $T_{SplitCentroid}$ with the penalty parameter ξ is the consensus tree with the threshold $\xi/(1+\xi)$. In other words, the consensus tree with the threshold γ is $T_{SplitCentroid}$ with the penalty parameter $\gamma/(1-\gamma)$. Note that if $1 \leq \xi$ then $1/2 \leq \gamma < 1$.

CWT is the Centroid Representation in Double Metrics

As was already described, the consensus tree $T_C = (V_C, E_C)$ still possesses ambiguity with regard to branch orderings around the multifurcating nodes. Each layout can be specified by a function $f(v,t) : V_C \times \mathrm{N} \rightarrow \mathrm{N}$ that is defined for $v \in MultiFurcatingNodes(T_C)$ and $t = 1, \ldots, k(v, T_C)$, where $k(v, T_C) = |AdjBranches(v, T_C)|$ and $\{e_1, e_2, \ldots, e_{k(v,T_C)}\} = AdjBranches(v, T_C)$, and the cycle $(e_{f(v,1)} e_{f(v,2)} \cdots e_{f(v,k(v,T_C))})$ specifies the circular ordering of $AdjBranches(v, T_C)$. In the following paragraph, we show that, among every possible layout of consensus trees, the CWT is the centroid representation regarding the loss function of "layout incongruity".

Let $\Phi = (T_C, f)$ be a layout-specified tree representation of T_C. Then, the "layout incongruity loss function" is defined as

$$L_{Layout}\left(\Phi, T_i\right) = \sum_{v \in MultiFurcatingNodes(T_C)} l_{Layout}\left(\Phi, v, T_i\right),$$

where

$$\begin{aligned} l_{Layout}(\Phi, v, T_i) = &\sum_{t=1}^{k(v,T_C)-1} \\ &\left(\sum_{\{X_{sub}=\{x_1, x_2, \ldots, x_{k(v,T_C)}\} | \forall_{1 \leqslant t \leqslant k(v,T_C)}\ x_t \in Descendants(e_t, v, T_c)\}} \right. \\ &\left. \frac{Dist(x_{f(v,t)}, x_{f(v,t+1)}, X_{sub}, T_i)}{\prod_{t=1}^{k} |Descendants(e_t, v, T_c)|} \right) \\ &+ \sum_{\{X_{sub}=\{x_1, x_2, \ldots, x_{k(v,T_C)}\} | \forall_{1 \leqslant t \leqslant k(v,T_C)}\ x_t \in Descendants(e_t, v, T_c)\}} \\ &\frac{Dist(x_{f(v,k(v,T_C))}, x_{f(v,1)}, X_{sub}, T_i)}{\prod_{t=1}^{k} |Descendants(e_t, v, T_c)|}. \end{aligned}$$

Intuitively, for each split pair s_a and s_a that correspond to successive branches around each multifurcating node v in Φ, l_{Layout} sums the distances on the tree T_i between the expected positions of the corresponding branches on the assumption of the topology T_C. Then, L_{Layout} sums it for all wheel nodes to quantify how well Φ represents T_i. Then the "layout incongruity centroid tree" $\Phi_{LayoutCentroid}$ is the tree representation that minimizes the expected L_{Layout} for all $T_1, \ldots, T_N$:

$$\begin{aligned} &\sum_{i=1}^{N} L_{Layout}\left(\Phi, T_i\right) P\left(T_i \left| D\right.\right) \\ &\propto \sum_{i=1}^{N} \left(\left(\sum_{v \in MultiFurcatingNodes(T_C)} l_{Layout}\left(\Phi, T_i, v\right) \right) w_i \right) \\ &= \sum_{v \in MultiFurcatingNodes(T_C)} \left(\sum_{i=1}^{N} l_{Layout}\left(\Phi, T_i, v\right) \cdot w_i \right) \\ &= \sum_{v \in MultiFurcatingNodes(T_C)} \left(\sum_{t=1}^{k(v,T_C)-1} c(e_{f(v,t)}, e_{f(v,t+1)}) \right. \\ &\qquad \left. + c(e_{f(v,(v,T_C))}, e_{f(v,1)}) \right) \end{aligned}$$

where c is the traveling cost calculated in the pseudocode. The final term is a minimum if we choose f to minimize the term in the bracket for each multifurcating node v, because they are independent of each other. Such an f is exactly the TSP tour obtained in the pseudocode; therefore, CWT is $\Phi_{LayoutCentroid}$ and the most balanced according to the two measures of split incongruity and layout incongruity.