CONSENSUS TECHNIQUES AND THE COMPARISON OF TAXONOMIC TREES

EDWARD N. ADAMS III

Abstract

Adams, E. N. III (Dept. of Computer Science, Univ. of Illinois, Urbana, Ill. 61801) 1972. Consensus techniques and the comparison of taxonomic trees. Syst. Zool. 21: 390–397.—A new problem in the science of classification is presented, along with its solution. The problem is to combine the information in several taxonomic trees into a single tree. The solution is a computational method for computing a tree which represents only that information shared by all the rival trees. Such a method, called the *consensus* method, can be used to "compare" several rival tree representations or to compute a more stable tree from slightly perturbed variants of the original data. A method is defined and demonstrated for each of two different types of trees: rooted, fully labelled trees, and rooted trees with unlabelled internal nodes. [Classification; trees.]

Every classificatory field of science has the problem of comparing rival classifications of a set of objects. If the classification takes the form of a tree (hierarchical classification, cluster analysis, dendrogram) the problem is particularly difficult. The approach that has been followed is to compute a numerical index of agreement between trees, such as correlation of the trees with each other (cophenetic correlation), or calculation of a "distance" between two rival trees (Robinson, 1971). But an additional approach should also be followed, asking the question, can we combine the information from rival classifications into a new, hopefully more accurate classification? Such a *consensus* of the rivals is useful both in tree comparison and tree discovery.

The consensus of two or more trees is a tree representing only that information that is shared by all of the trees. The consensus is a conservative estimate of a compromise classification, because any information not represented in *all* the rivals is not represented in the consensus. If the rivals are quite dissimilar (e.g., if one rival bears very little relation to "reality") then the consensus contains very little information. But if the rivals are all very similar, then the consensus contains a great deal of information.

When a classification becomes generally accepted as an accurate representation of "reality" in some field, and then new methods producing different results are introduced, it is necessary to have an objective procedure to determine how well the new ones approximate the standard. Such a comparison is needed to evaluate the methods producing the new classification. Where there is no standard ("best") classification in a field, however, an objective comparison procedure permits corroboration of the results of different methods (or people).

In both cases, consensus can be an informative technique. In the first case, if the consensus of the "best" tree and the "new" tree is identical to the "best" tree, then the "new" tree is at least as precise as the standard and agrees with it in all regards. If the consensus is identical to the "new" tree, the "new" tree is correct to the level of detail it embodies. Otherwise the "new" tree is incorrect in some regards, and the consensus tree shows exactly which parts of the tree are correct. In the second case, if the consensus is identical to one of the rivals, that rival is less precise than the other, but they agree as much as they can. If the consensus is not identical to any of the rivals, it represents only their agreements. Investigators evaluating methods or people's work can see how well the standard is approximated, and investigators studying the classified objects benefit from a con-

sensus tree which combines the information of several different methods.

Consensus can also be useful in the discovery process. A classification is dependable only to the extent that it is stable with respect to small perturbations in the similarity data. Taxonomists ought to, as a matter of course, compare the result of a classification procedure with the results of the procedure on slightly perturbed data. This policy is stated principally by proponents of minimal spanning tree (MST) methods, such as Jardine and Sibson (1971). Taxonomists recognize the need for stability, but current stability measures do not yield completely adequate descriptions of stability.

The stability of the classification is made more visible by consensus techniques. If two rival trees on perturbed data are compared, the scientist can get a feeling of "this classification is stable" or "this classification is not stable." But unless the rivals are identical, he is still faced with the question, "What *is* the classification of these objects?" If "it" is stable, then what exactly is "it"; what tree will he propose as the true classification? If "it" is not stable, has he a starting point for making further guesses at the true classification? An answer to both questions is "the consensus tree." After all, if the scientist is willing to believe that these rivals adequately represent the true classification, he believes that their areas of disagreement are small and can be ignored. Since the consensus of these rivals ignores exactly the areas of disagreement, the consensus is his best approximation to the real classification.

If, on the other hand, the rivals come out dissimilar, the consensus at least shows him whatever tree structure they have in common. This is good for two reasons: the scientist gets results from a stability check without having to choose among dissimilar classifications, each having spurious precision; and he suppresses artifacts due to imposition of tree structure on data that quite possibly has little tree structure.

The notion of consensus of trees depends on the ability to find all "information represented in all the rivals," but this latter presupposes the definition of "information represented by a tree." Clearly this depends on the type of tree. For example, a tree whose leaves and internal nodes are all named can be interpreted as giving *parenthood*, *dominance*, or *inclusion* information about the named nodes. However, a tree with unlabelled internal nodes has no names for parents, so parenthood is not represented as much as *brotherhood*. Similarly, a tree with links of specified lengths represents a more absolute measure of brotherhood than a tree with links of unspecific length. Further, MST techniques produce trees without roots, so dominance is hard to define, whereas other trees have roots. "Information represented by a tree" and "information shared by two trees" must be defined separately for each type of tree.

In drawings and discussions of trees in this paper, the "root" of a tree (the ultimate ancestor) is at the top, and the leaves (the final descendants) are at the bottom. If a node has more than one immediate ancestor, the structure is no longer a tree and we call it an "ingrown tree." If the top node of a tree is removed, the structure is no longer a tree, but is a collection of trees, termed a "forest."

Trees with labelled internal nodes have generally been produced by experts in their fields, based on extensive study of an encyclopedic data base. Data exists for all the present organisms (or languages or cultural objects, etc.) as well as for extinct or ancestral forms (through fossil records, written accounts, etc.), so there is good reason to consider the ancestor nodes of the tree to be just as real objects as the terminal nodes. Therefore the information shown in a fully labelled tree is *ancestry*. Since the time separations are rarely precise, we define the information contained in a fully labelled tree as the set of ancestor-descendant relationships without regard to time separation. Since this is merely a set,

the intersection of several such sets constitutes the information shared by several rival trees.

A paper-and-pencil algorithm for computing the consensus of several fully labelled trees follows:

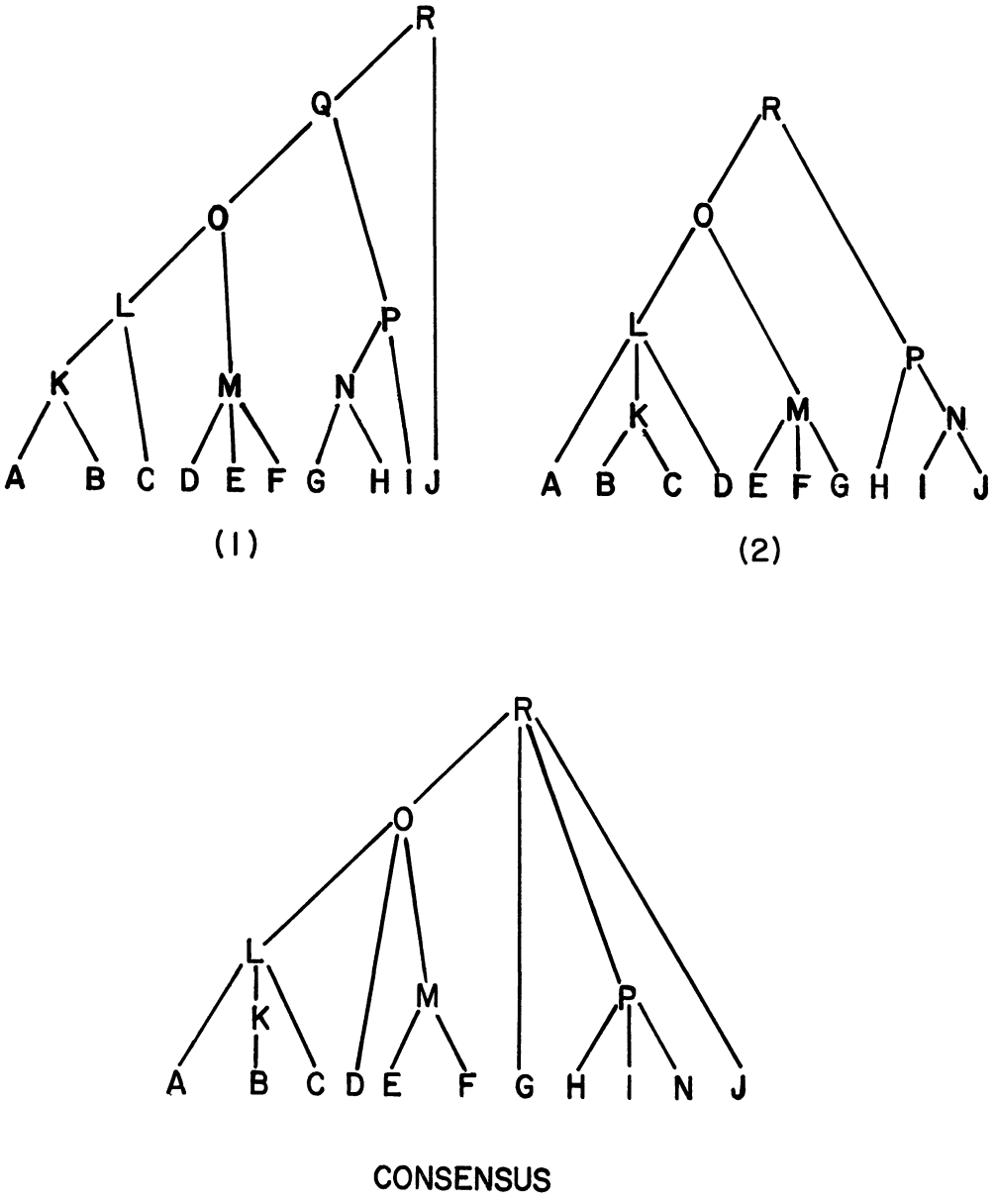
Call one rival Tl (it doesn't matter which one).
Starting at the top of Tl, pick a node (call it N) whose Tl ancestors have already been picked. N's nearest ancestor on the censensus tree C is defined as follows:
Discard N if it is not present on all rival trees.
Find which of N's ancestors on Tl are also its ancestors on all the other trees. They will be N's ancestors on C, with the furthest from the root being its most recent ancestor.
Draw a line from the most recent ancestor to N. If N has no ancestors common to all rival trees, then it is a root of C. (It is quite permissible for any of the rivals as well as C to be a forest rather than a tree.) If N's ancestors do not all lie in one line from the root of C, then C does not have a perfect tree structure. Draw a line from the deepest one, and if there are more than one deepest one, choose one of them arbitrarily (this is the only possible case of nonuniqueness). See the discussion below about lack of pure tree structure.

If we consider a forest R to be a set of ordered ancestor-descendant pairs, R corresponds to a transitive, antisymmetric, irreflexive relation with the tree property: (for all nodes S, P, Q) if (S, Q) and (P, Q) are in R than either (S, P) or (P, S) is in R. This requirement states that if two objects are ancestors of a third, one of them is an ancestor of the other. This is what makes such a relation into a tree structure, rather than an ingrown tree. If several rival sets R correspond to such relations on the same domain of objects, the consensus C is merely the intersection of all the sets R.

It is easy to prove that C will be a tree (namely, it will be a transitive, antisymmetric, irreflexive relation satisfying the tree property above) unless there are nodes Q, S, T such that (Q, T) and (S, T) are in all the rivals, (Q, S) is in some, and (S, Q) is in all the rest. This causes Q and S to be the ancestors of T without either being the ancestor of the other. Two kinds of situations can give rise to this situation. In one, one taxonomist will have to give the name Q to a genus and S to a species, while another would name the genus S and the species Q. The other situation arises in typical application to MST's. The ingrown tree defined in either case still represents all the ancestor information of the rival trees, but until most taxonomists feel comfortable with ingrown trees, the consensus should be defined as the tree derived from the ingrown tree by pruning the fewest number of links. The paper-and-pencil method above does exactly this.

This consensus definition has many good points. It can be computed for a group of rival trees all at once; large trees can be conveniently handled by hand; a computer program to do this would execute in time proportional to $n \log n$, where n is the number of nodes in a tree; the consensus tree is unique (except for an ingrown tree where the rejoining branches are the same length): forests and trees can be used without modification; the method is completely insensitive to the order in which the trees are selected, and the order in which nodes are chosen. An example calculation demonstrates the method and its results. Figure one shows two rival trees labelled (1) and (2). The names of the nodes are the letters A through R.

The top node of (1) is R. It exists on both rivals but it has no ancestors on both trees, so R is a root of (3), the consensus tree. Next, since we are using tree (1) as Tl, pick either Q or J. Q has the ancestor R on (1) but does not exist on (2), so Q



LV

FIG. 1.—Two rival, fully labelled trees and the resulting consensus tree. For details, see text p. 392–394.

does not appear on C. J has the ancestor R on (1) and N, P, R on (2). The only ancestor shared by both trees is R, so draw a line from R to J in (3). Now we can choose either O or P. O has ancestors Q and R on (1), but only R on (2), so its ancestor on (3) is R. Draw a line from R to O. P's ancestors in (1) are Q and R, but only R in (2), so link P with R. L has ancestors O, Q, R on (1) and O, R on (2), so its ancestors are O and R, with the link going to the furthest from the root. Connect L to O. The computation continues, linking M to O, N to P, K to L, B to K, C to L, D to O, E to M, F to M, G to R, H to P, and I to P. Notice the node K with only one son B. This result indicates that the rivals agree that K is ancestor to B, although they could not agree on any other descendants of K. Also note that although the rivals disagree completely about the descendants of N, they both agree that N is descended from P. So N is not discarded, as it should be if it were merely a hypothetical construct.

Although some taxonomic trees are of this form, many trees do not represent information about ancestorhood. These are frequently generated by numerical taxonomy methods, whether manual or performed by computer. In these cases, the internal nodes are manufactured by the method, and are merely representations of the fact of brotherhood of the subtrees joined at the node.

The information about a set of terminals is shown by the branching under the nearest common ancestor (called the *least upper bound* or LUB) of that set. All the terminals of the set under one branch from the node are related. So the information agreed on by the LUBS of a set in rival trees is that certain terminals are related in both trees. Suppose the LUB of the set ABCDEFGHI in one tree branches two ways, grouping the terminals ABCDE and FGHI, and the LUB in the other tree branches three ways grouping them AIE, BCD, FGH. They agree on the closeness of AE, of BCD and of FGH. But they disagree about the inclusion of I in either group 1 or group 3, and about the combination of groups 1 and 2. On the consensus tree, the set splits into AE, BCD, FGH, and I. To build further branching on the consensus, treat each of these subsets in turn in exactly the same way as the original set. That is, find the LUB of BCD in one tree and LUB of BCD in the other tree, and determine the branching of the consensus from that. Then do the same for FGH, and AE, and I. (Actually, since a two member set splits in two, there is no need to formally construct the LUB and observe the splitting; similarly, a one-member set will not split any further.) Mathematically, consider the LUB of a set as a partition of that set expressed by the branching at that node. The partition of the LUB in the consensus tree is expressed by the cross product of the partition from one rival LUB with the partition of the other rival LUB. So, if there are n groups in the partition of one LUB and m groups in the partition of the other LUB, then there are mn groups in the product partition, one for each intersection of a group from one tree and a group from the other. If there is a third rival with p groups in the partition of the LUB of the set, then there are mnp groups in the product partition. Null intersections are then ignored.

Thus the recursive definition of consensus for a set of terminals of rooted trees with unlabelled internal nodes is:

1) if the set contains exactly one terminal, the consensus is that terminal itself.

2) otherwise find the LUB of the set on each rival tree. Construct the product partition as above. The consensus of the set is a node with a branch leading to each consensus of a non-null group of the product partition.

Figure 2 shows two rival trees on the set (1, 2, Z5, Y3, 38, Z10, 20, 49). The labels on the internal nodes come about through operation of the algorithm, and are merely a convenience for bookkeeping. We are interested in the entire set, so the roots of both trees are the LUB's. If there were an extra outlying object on one of the rivals,

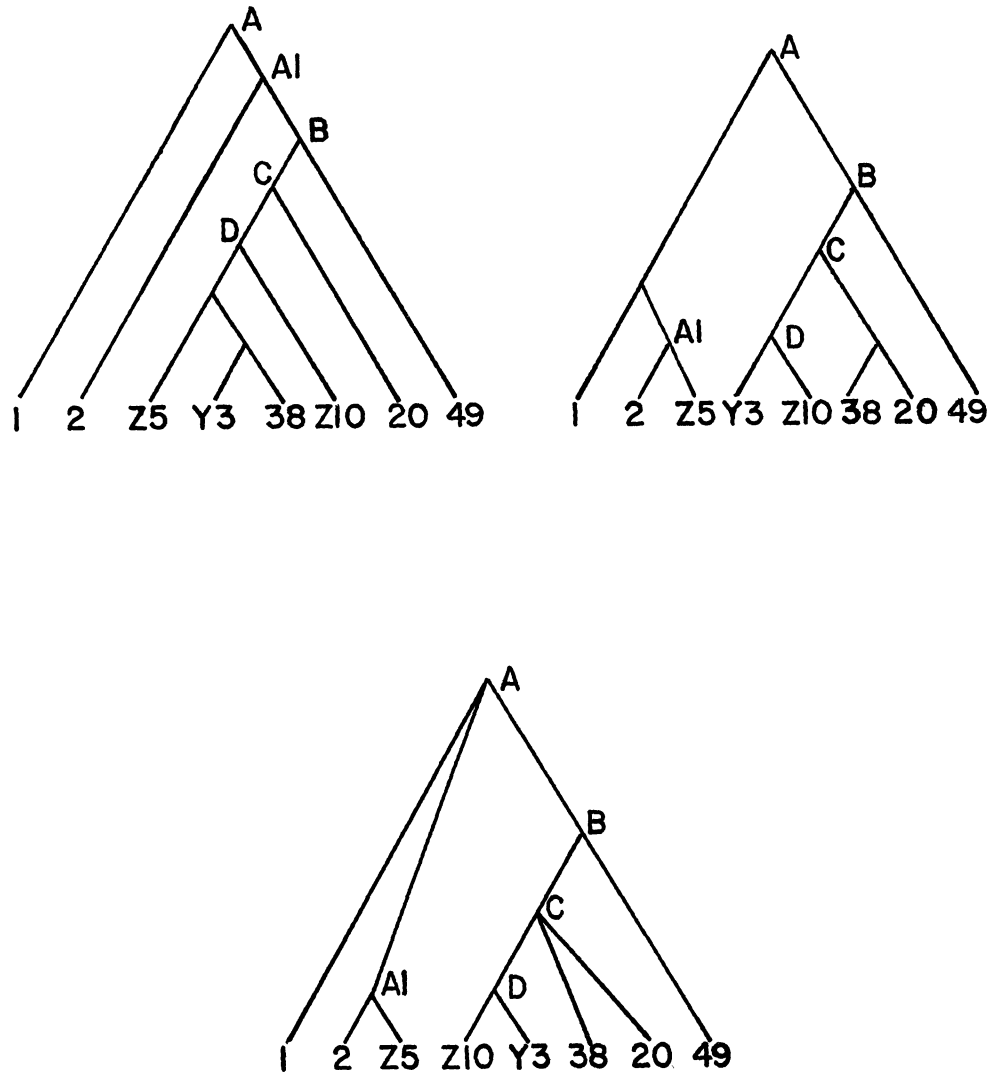


FIG. 2.—A consensus tree from two rival trees with unlabelled internal nodes. For details, see text p. 394–396.

split off before anything else, then the root would not be the LUB of the set we are interested in, and we would have to search down until we found it. We mark the LUB's with the letter A for ease of reading. The partition of A on the left rival is (1) and (2, Z5, Y3, 38, Z10, 20, 49). On the right rival, the partition is (1, 2, 5) and (Y3, 38, Z10, 20, 49). The product partition is computed as follows:

(1) versus (1, 2, Z5) = (1)
(1) versus (Y3, 38, Z10, 20, 49 = the null set
(2, Z5, Y3, 38, Z10, 20, 49) versus (1, 2, Z5) = (2, Z5)

(2, Z5, Y3, 38, Z10, 20, 49) versus (Y3, 38, Z10, 20, 49) = (Y3, 38, Z10, 20, 49)

Thus there are three subsets of the original set. We draw three branches from the consensus node labelled A. One corresponds to the leaf (1), one to the set (2, Z5), and one to the set (Y3, 38, Z10, 20, 49). Now treat these subsets just as the original set above. Consider the set (2, Z5). Its LUB's are marked A1 on the rivals. Notice A1 is construed to dominate *only* 2 and Z5, even though it looks more important on the left rival tree. The partition of A1 is (2) and (Z5) on both trees, so the node corresponding to this set splits into 2 nodes, one pointing to (2) and one to (Z5). Actually, this process is unnecessary for 2-element sets, since the LUB will automatically divide it into 2 one-element sets. We did this just for illustration. Now consider the remaining set (Y3, 38, Z10, 20, 49). Its LUB's are marked B on the trees. The partition of the B's is (49) and (20, Z10, 38, Y3) on both trees. Notice again that although Z5 is under this node on the left tree, we ignore it now, because that set is covered in some other branch of the consensus. Since the partitions of the B's are identical, their product partition will remain the same. Thus we draw two lines from B, one corresponding to the leaf (49) and one to the set (20, Z10, 38, Y3). The LUB of this set is labelled C on both trees. The partitions are (20) and (Z10, 38, Y3) on the left tree and (38, 20) and (Z10, Y3) on the right tree. Taking all combinations of intersections, we get 3 subsets: (20), (38), and (Z10, Y3). Thus three lines are drawn from C, one to (20), one to (38), and one to (Z10, Y3). After finishing the automatic treatment of this two-member set, we get the finished consensus tree. The labels which we generated during the process give a subjective feel for the correspondences between nodes in the rival trees and the consensus trees. Notice that not all the nodes in the rival trees were labelled. This happened because some of the terminals were outliers in only one tree, and were filtered out early in the process. Notice also, parallel branches in the consensus need not correspond to parallel branches in the rival trees (see A1 and B). In fact, although it is not shown here, it is quite possible for a rival tree node to be the LUB of different sets, and thus be labelled several different times..

The interpretation of the consensus tree is straightforward. At the uppermost level of the set, both trees agree on the coherence of the set (49, 20, Z10, 38, Y3) and of the set (2, Z5). They do not agree about the way in which these sets go together with each other or with terminal 1. About the larger of these sets, both trees agree that the (20, Z10, 38, Y3) is coherent, but 49 is not agreed on as a participant. In this case, in fact, they agree that 49 is an outlier, but this is bonus information.

This example is a highly abbreviated version of an example from Ivimey-Cook (1969), with Z5 replacing a small subtree which was identical in the rival trees, Z10 replacing a large subtree which was identical in the rival trees, Z10 replacing a large subtree which was identical in the rivals, and Y3 replacing a very large subset, described by two different trees in the two rivals. As a result of comparing these two rivals with an earlier description by a regular taxonomist, Ivimey-Cook decided that a separate group should be made for 1, 2, and Z5. The consensus tree in figure 2 suggests that although 2 and Z5 do form a separate group, 1 is possibly an outlier. Since his comparison covered one more rival tree, it is not surprising that the conclusions are not exactly the same.

There are good features to this consensus definition. The definition is a constructive algorithm for computing consensus; it is a fast paper-and-pencil algorithm even for large trees; it yields a unique result; the result is always a tree; the method is completely insensitive to the order in which the subsets of a set are looked at; and although a forest cannot be used as is, any forest can be converted to a tree by creating a fictitious father node dominating all of the trees of

the forest, the consensus computed, and the father node of the consensus tree discarded.

Although comparison of taxonomic trees has been a well-researched problem, the approaches have all been towards computing from rival trees a *number* denoting their similarity. Consensus is a complementary approach, where we compute from rival trees a *revised tree* denoting the areas of their similarity. Such an approach to the recognized problems of tree comparison, classification stability, and classification discovery is desirable, for the same reason that the concept of *mean* is a desirable approach to the problem of comparison of several multidimensional observations of some object. The advantage in dimensional over taxonomic analysis is the ease of constructing a revised estimate from a set of observations. I hope that these fast algorithms for computing consensus can provide taxonomic analysis with the same convenience.

REFERENCES

IVIMEY-COOK, R. B. 1969. The phenetic relationships between species of *Ononis*, p. 69 to 90. *In* A. J. Cole (ed.) Numerical Taxonomy. Academic Press, N.Y.

JARDINE, N. AND R. SIBSON. 1971. Mathematical Taxonomy. Wiley, New York.

ROBINSON, D. F. 1971. Comparison of Labeled Trees with Valency of Three. J. Combinatorial Theory, Ser. B, 11:105.

(*Received March, 1972*
Revised July, 1972)