Reprinted with an erratum from:

Alfredo Rizzi, Maurizio Vichi,
Hans-Hermann Bock (Eds.)

Advances in Data Science and Classification

Proceedings of the 6th Conference of the
International Federation of Classification Societies (IFCS-98)
Università "La Sapienza", Rome, 21-24 July, 1998

The Information Content of Consensus Trees

Joseph L. Thorley[1], Mark Wilkinson[1,2] and Mike Charleston[3]
1. School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK
2. Department of Zoology, Natural History Museum, London SW7 5BD, UK
3. Department of Zoology, University of Oxford, Oxford 0X1 3PS, UK

Abstract: Phylogenetic Information Content, a class of measures of the information provided by consensus trees based on the number of permitted resolutions of the consensus, is introduced. A formula for the number of permitted resolutions of Adams consensus trees is derived and a proof given. We argue that maximising PIC measures provides a sensible criterion for choosing among alternative consensus trees and we illustrate this for consensus trees of cladograms.

Keywords: Consensus, Phylogeny, Adams resolutions, Information Content.

1. Introduction

Consensus trees (CTs) are used by systematic biologists, in many different contexts, to summarise graphically the agreement between multiple fundamental phylogenetic trees. The development of many alternative consensus methods raises the issue of choice between methods and/or tree(s). In this paper we derive a class of measures of the information provided by phylogenetic trees, argue that in certain contexts the optimal CTs are those which convey the most information, and illustrate the approach.

2. Phylogenetic Information Content

In order to convey information a CT must prohibit a subset of the possible phylogenetic relationships (Mickevich and Platnick, 1981; Wilkinson, 1994). Given that the information (I) conveyed by an event is

$$I = -\log \frac{\text{probability of the event}}{\sum \text{probabilities of all possible events}} \qquad (1)$$

where the base of the logarithm determines the unit of information (e.g. $\log_2$-*bits*; ln - *nats*), then, under the assumption that all possible phylogenies are bifurcating and equiprobable, the Phylogenetic Information Content (PIC) of a tree is

$$\mathrm{PIC} = -\log \frac{\text{number of permitted bifurcating trees } (n_\mathrm{R})}{\text{number of possible bifurcating trees } (n_\mathrm{T})}. \qquad (2)$$

Because the concept of a phylogenetic tree is a general one, PIC defines a class of tree information measures with specific measures existing for each type of phylogenetic tree. In this paper, we focus exclusively on cladograms, which are the *n*-trees of Bobisud and Bobisud (1972), and the corresponding PIC measure which we refer to as Cladistic Information Content (CIC).

For a CT to provide phylogenetic information, n_R must be less than n_T, i.e., it must be possible to deduce from the CT alone which of the possible trees could not have been represented among the fundamentals. We refer to CTs which, unless totally unresolved, always fulfil this condition as *prohibitive*. *Permissive* CTs are those which permit all possible trees (i.e. $n_\mathrm{R} = n_\mathrm{T}$) irrespective of their resolution. In order to be prohibitive, CTs must be strict *sensu* Wilkinson (1994), i.e., their groupings must represent a particular type of phylogenetic relationship (components, *n*-taxon statements or nestings) that occurs in all the fundamentals.

3. Calculating CIC

The number of possible bifurcating rooted cladograms (n_T) is a function of the number of leaves *n* and is given by RB(*n*)

$$RB(n) = (2n-3)!! \,. \qquad (3)$$

Rohlf (1982) provided equation (4) for calculating the number of permitted bifurcating trees (n_R) for strict CTs of rooted cladograms whose groupings represent components or other *n*-taxon statements, i.e., the CTs produced by the strict component (Sokal and Rohlf, 1981), strict Reduced Cladistic (RCC), Reduced Adams (RAC), Disqualifier-Faithful (DF) (Wilkinson, 1994), and Largest Common Pruned (LCP) (Gordon, 1980) consensus methods.

$$n_R = \prod_{i \in V(T)} RB(d_i) \qquad (4)$$

where d_i is the number of vertices immediately descendant from vertex *i* in the set of vertices *V(T)* of tree *T*.

The number of permitted resolutions (n_{R}) of an Adams CT (Adams, 1972), the only strict CT to capture nestings (Adams, 1986), is

$$n_{\mathrm{R}} = \mu_{root} \qquad (5)$$

where

$$\mu_i = \sum_{(A,B)\,\text{a split of}\,D_i} \frac{\left(\prod_{j\in D_i} \mu_j\right) RB(g_A) RB(g_B)}{\prod_{j\in D_i} RB(h_j)} \qquad (6)$$

where D_i is the set of vertices immediately descendant from vertex *i*, a *split* of a set *D*, is in this case, a partition of *D* into two non-empty, mutually exclusive subsets whose union is *D*, g_A is the number of leaves descendant from subset *A* of *V(T)*, and h_j is the number of leaves in the *j*th subtree descendant from vertex *i*. The proof of this is given in the appendix.

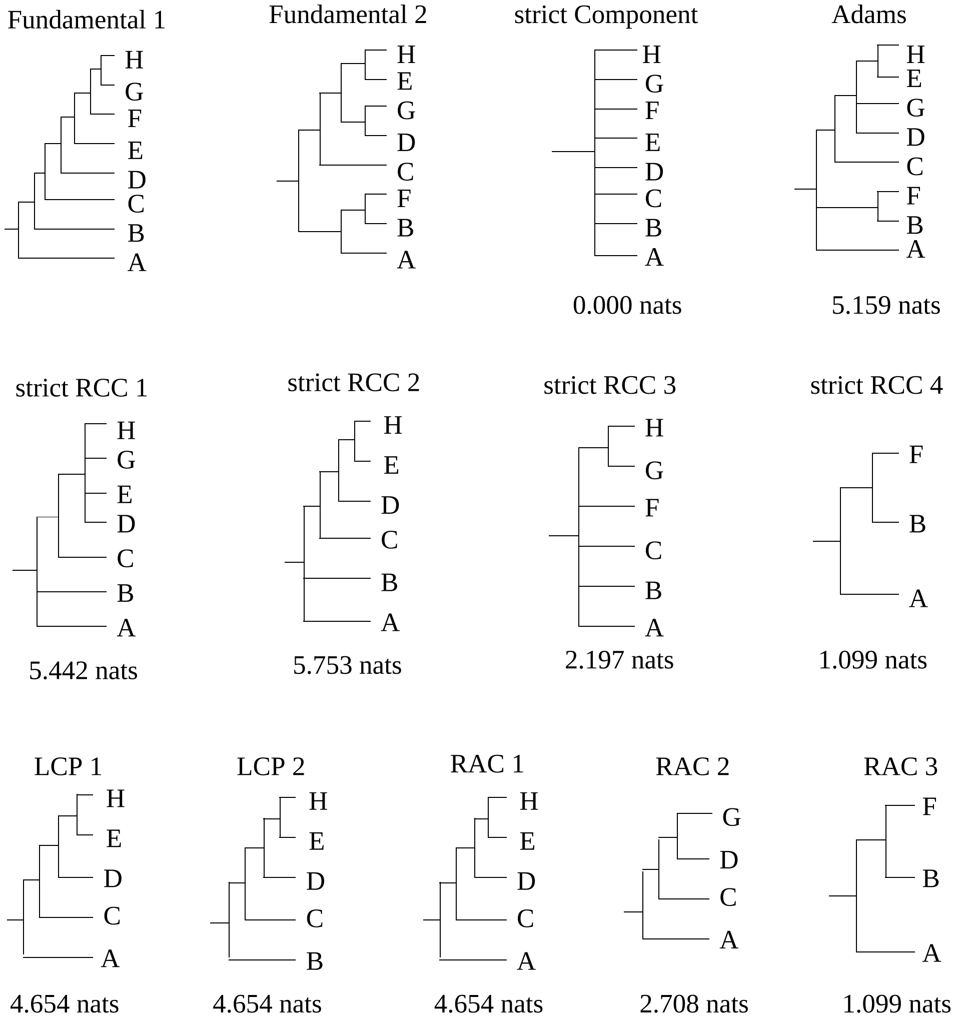
4. Choosing a CT

The production of a consensus profile, a set of CTs, by several consensus methods means that each method is not synonymous with a single CT. Furthermore, the information content of each CT depends upon both the properties of the method and the characteristics of the particular fundamentals under consideration. Hence our emphasis is on the choice of CT as opposed to method, although the question of which method(s) will produce the most informative CT and sets of CT(s) is also discussed.

In Figure 1, the strict component CT is completely unresolved, and therefore provides no information, despite the occurrence of considerable agreement among the fundamentals. Insensitivity of the strict component CT has been well documented (e.g. Adams, 1986; Swofford, 1991). The Adams CT, in contrast, is well resolved and provides CIC = 5.159 nats of information. However, the groupings in the Adams CT represent nestings and thus the polytomies of this tree are more permissive than those in CTs whose groupings represent *n*-taxon statements. The strict RCC 1 and strict RCC 2 CTs provide more information than the Adams despite being less resolved and/or including fewer taxa.

The LCP, RAC and strict RCC consensus methods overcome the problems of insensitivity and ambiguity that affect the strict component and the Adams respectively, by the exclusion from CTs of taxa whose variable position among the fundamentals prevents unambiguous summary of the relationships common to the remaining taxa. Multiple CTs are produced when, as is usually the case, there is more than one way to exclude taxa to achieve unambiguity.

Figure 1: Two fundamental cladograms (a,b), and their strict CTs [Modified and extended from Adams, 1972; Wilkinson, 1994]



LCP trees are produced by pruning single taxa from the fundamental trees until they are rendered identical. RAC trees are produced by pruning branches from the polytomies of Adams CTs until the polytomies can be correctly interpreted as *n*-taxon statements. The strict RCC method first identifies the complete set of non-redundant *n*-taxon statements common to all the fundamentals. These *n*-taxon statements are then represented graphically through the production of a consensus profile.

Because of the way the RCC profile is constructed, no LCP or RAC tree can ever provide more information than the most informative RCC tree (e.g. Figure 1). Given that the strict component CT is included in the strict RCC profile unless it is completely unresolved, then choice of the most informative CT must be between the Adams CT or a tree from the RCC profile. In this example, strict RCC 2 maximises CIC. An example in which the Adams CT is the most informative CT is provided by Adams (1986; Figure 2).

5. Choosing a set of CTs

CIC can also be used to select the set of CTs which provide the most cladistic information. However, the occurrence of the same information in more than one tree means that the combined CIC of a set of CTs can be less than the sum of their individual CICs. If two strict CTs, both representing *n*-taxon statements, have the relationship that all taxa present in one of the trees also occur in the other, then the information common to the two CTs is the cladistic information content of the single CT in their strict RCC profile. However, if this relationship does not hold, then the only currently available way to determine combined CIC is with a brute force approach (i.e. compare every single possible tree with the set of CTs to determine whether it is permitted). The problem of selecting a set of CTs from a consensus profile so as to maximise combined information content appears to be computationally complex.

Applying the brute force approach to the strict RCC profile in figure 1, we find that the pair of CTs which maximise combined cladistic information is {RCC 2, RCC 3} (CIC = 6.954 nats). RCC 1 provides no information that is not already conveyed by RCC 2 and RCC 3, whilst conversely, all the information provided by RCC 4 is non-redundant with respect to the other three trees in the profile. Thus the combined cladistic information content of the entire strict RCC profile is 8.053 (6.954 + 1.099) nats. Because the strict RCC trees graphically represent all *n*-taxon statements common to all the fundamentals, the LCP (combined CIC = 5.753 nats) and RAC (combined CIC = 5.589 nats) profiles will never provide more information than the strict RCC profile. Furthermore, all the information provided by these two alternative profiles will be represented in the RCC profile. In fact, only the Adams CT can convey cladistic information that is non-redundant with respect to the strict RCC profile.

6. Discussion

Measures of information content provide a basis for choosing amongst alternative strict CTs and methods. Due to the probabilistic nature of information, it may also be possible to calculate the phylogenetic information content of majority-rule and other CTs with groupings that represent relationships occurring in less than 100% of the fundamentals.

Although this paper focuses solely on cladistic information, dendrograms, which are equivalent to cladograms with internally ranked nodes, are another very important type of phylogenetic tree. Whilst consensus methods which take into account the rank of internal nodes have been developed (Neumann, 1983; Stinebrickner, 1984) they suffer from problems of ambiguity or are permissive. Prohibitive dendritic consensus methods are currently under development by the authors.

The utility of our measure of phylogenetic tree information content extends far beyond the selection of CT(s). For example, a natural measure of tree similarity is the information common to the set of trees (Mickevich, 1978). Conversely, a measure of tree dissimilarity could be based on the symmetric difference of the information provided by each tree. Thus, PIC defines a class of tree similarity (and dissimilarity) measures for studies of congruence that can also be normalised to produce Consensus Indices (Mickevich, 1978).

Appendix

Proof of equation 6.

We need the following simple lemma:

Suppose we have a forest of k rooted binary trees $T_1,...,T_k$ with leaf sets $L_1,...,L_k$ respectively, such that $\bigcup_{i=1}^{k} L_i = \{1,\ldots,n\}$, and $L_i \cap L_j = \emptyset$ for all $1 \leq i < j \leq k$. Let the number of leaves in each tree T_i be a_i. The number of rooted binary trees T^* with leaf set $\{1,...,n\}$ which contain as subtrees each of the T_i is given by

$$N\left(n;\{a_i\}_{i=1}^{k}\right) = \frac{RB(n)}{\prod_{i=1}^{k} RB(a_i)}.$$

Proof: It is well known that the number of rooted binary trees with n leaves is $(2n-3)!! = (2n-3)(2n-5)...(3)(1) = (2n-3)!2^{2-n}/(n-2)!$. One way in which we may generate all the possible rooted binary trees is by successively adding leaves $1,...,n$ to each of the edges of the tree. By adding a_2 leaves in a similar way to tree T_1 we can create all the possible trees with (a_1+a_2) leaves, containing T_1 as a subtree, in $(2a_1 - 1)(2a_1 + 1)\ldots(2(a_1 + a_2) - 3) = RB(a_1 + a_2)/RB(a_1)$ ways. Similarly, the number of trees with (a_1+a_2) leaves containing T_2 as a subtree is $RB(a_1 + a_2)/RB(a_2)$. The above argument is independent of the underlying trees T_1 and T_2: ergo the number of trees on $n > (a_1+a_2)$ leaves containing T_1 and T_2 is $RB(n)/(RB(a_1)RB(a_2))$. The argument is easily extended to $T_1,...,T_k$.

Let an *Adams resolution* of a vertex *v* in a tree T be a resolution of *v* which is consistent with treating *T* as an Adams consensus tree. Now let μ_i be the number of resolutions of the subtree rooted at vertex *i* of a given Adams consensus tree *T*. Then μ_{root} is the number of Adams resolutions of the complete tree *T*. We proceed from the tips of *T* to the root, finding the number of Adams resolutions of each vertex *j* before we continue to the ancestor of *j*. Let D_j be the set of vertices which are immediately descendant from vertex *j* and let $|D_j| = d_j$. Let L_j be the set of leaf vertices in the subtree T_j rooted at *j*, and $|L_j| = l_j$. Thus $n = l_{root}$ and the degree of vertex *j* is d_j+1. Lastly let *G(A)* be the set of leaves descendant from the vertices in the *set* $A \subseteq V(T)$, with cardinality g_A.

Consider vertex $i \in V(T)$, the immediately descendant vertices $v_j \in D_i$ of which have been resolved into trees $T_1,\ldots,T_{d_i}$. According to Nelson and Platnick's Interpretation 2b (Nelson & Platnick, 1980; Wilkinson, 1994), any Adams resolution of vertex *i* must maintain the branching order of the vertices in $T_1,\ldots,T_{d_i}$, but this resolution has no other constraints. Thus we may take any split (*A*,*B*) of D_i and resolve the vertices descendant from *A* to one side of *i* and the vertices descendant from $B = D_i \setminus A$ to the other side, while maintaining the branching order within each of the $T_1,\ldots,T_{d_i}$. The number of Adams resolutions at vertex *i* is thus the sum over all possible splits (*A*,*B*) of D_i, of the number of rooted binary trees on leaf set $\bigcup_{j \in D_i} L_j$, containing each of the resolved subtrees immediately descendant from *i*. The result follows.

References

Adams, E.N. (1972). Consensus techniques and the comparison of taxonomic trees, *Syst. Zool.*, 21, 390-397.

Adams, E.N. (1986). *N*-trees as nestings: complexity, similarity and consensus, *J.Classif.*, 3, 299-317.

Bobisud, H.M., and L.E. Bobisud. (1972). A metric for classification, *Taxon*, 21, 607-613.

Gordon, A.D. (1980). On the assessment and comparison of classifications, in: *Analyse de Donnees et Informatique*, Tomassone, R. (Ed.), Le Chesnay: INRIA, 149-160.

Mickevich, M.F. (1978). Taxonomic congruence, *Syst. Zool.*, 27, 143-158.

Mickevich, M.F., and N.I. Platnick. (1981). On the information content of classifications, *Cladistics*, 5, 33-47.

Nelson, G., and N.I. Platnick. (1980). Multiple branching in cladograms: two interpretations, *Syst. Zool.*, 29, 86-91.

Neumann, D.A. (1983). Faithful consensus methods for *n*-trees, *Mathematical Biosciences*, 63, 271-287.

Rohlf, F.J. (1982). Consensus indices for comparing classifications, *Mathematical Biosciences*, 59, 131-144.

Sokal, R.R., and F.J. Rohlf. (1981). Taxonomic congruence in the Leptopodomorpha re-examined, *Syst. Zool.*, 30, 309-325.

Stinebrickner, R. (1984). An extension of intersection methods from trees to dendrograms, *Syst. Zool.*, 33, 381-386.

Swofford, D.L. (1991). When are phylogeny estimates from molecular and morphological data incongruent?, in: *Phylogenetic analysis of DNA sequences*, Miyamoto, M.M. and Cracraft, J. (Eds.), Oxford Univ. Press, 295-333.

Wilkinson, M. (1994). Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles, *Syst. Biol.*, 43, 343-368.

Erratum (not included in original article)

The equation for calculating the number of permitted resolutions of an Adams consensus tree is incorrect (Mike Steel, pers. comm.). The error in the proof is the statement that 'The above argument is independent of the underlying trees T_1 and T_2.' In general the number of ways of combining two trees with non-overlapping leaf sets is not independent of their topologies (Constantinescu and Sankoff, 1986). The CIC of the Adams CT in Figure 1 is correct.

Constantinescu, M. and D. Sankoff (1986). Tree enumeration modulo a consensus. *J. Classif.*, 3, 349-356.