the traits that are used as phylogenetic markers truly discrete, tightly integrated, and functionally autonomous modules? (2) At what point is the error associated with relaxing the severity of the criterion for a character significant enough to cause systematic error in phylogenetic reconstruction? Answers to these questions will likely vary from study to study and across systems. The debate is not new and is not one that will likely be settled with a few exchanges in the pages of *Systematic Biology*. Nevertheless, we think it is one that has enduring importance.

We are grateful for the opportunity provided by O'Leary et al. (2003) to voice our perspective. We do not expect the readership to come to a final resolution favoring one view over another. However, for those interested, we simply recommend a rereading of O'Leary and Geisler's (1999) original paper, Naylor and Adams's (2001) reanalysis, and O'Leary et al.'s (2003) rejoinder herein. We leave it to the readership to come to their own conclusions.

REFERENCES

NAYLOR, G. J. P., AND D. C. ADAMS. 2001. Are the fossil data really at odds with the molecular data? Morphological evidence for Cetartiodactyla phylogeny reexamined. Syst. Biol. 50:444–453.

O'LEARY, M. A., J. GATESY, AND M. J. NOVACEK. 2003. Are the fossil data really at odds with the molecular data? Morphological evidence for whale phylogeny (re)reexamined. Syst. Biol. 52:853–864.

O'LEARY, M. A., AND J. H. GEISLER. 1999. The position of Cetacea within Mammalia: Phylogenetic analysis of morphological data from extinct and extant taxa. Syst. Biol. 48:455–490.

First Submitted 10 June 2003; reviews returned 12 June 2003; final acceptance 8 August 2003
Associate Editor: Chris Simon

Syst. Biol. 52(6):865–868, 2003

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150390252297

Matrix Representations with Parsimony or with Distances: Two Sides of the Same Coin?

FRANÇOIS-JOSEPH LAPOINTE,[1] MARK WILKINSON,[2] AND DAVID BRYANT[3]

[1]*Département de Sciences Biologiques, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, Québec H3C 3J7, Canada; E-mail: lapoinf@biol.umontreal.ca*
[2]*Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, U.K.*
[3]*Centre for Bioinformatics, McGill University, Lyman Duff Building, 3775 University, Montréal, Québec H3A 2B4, Canada*

Matrix representation with parsimony (MRP) is a method that takes as input a collection of source trees, recodes them as binary matrices, and returns a tree that is closest to the source trees using a parsimony criterion (Baum, 1992; Ragan, 1992). The average consensus method, on the other hand, takes as input a collection of weighted trees (i.e., with branch lengths) and uses a matrix representation with distances (MRD) analysis to seek the weighted tree that is closest to the source trees using a least-squares criterion (Lapointe and Cucumel, 1997). MRP has been mostly used as an alternative to data combination for assembling supertrees from source trees bearing nonidentical but overlapping sets of leaves (for reviews, see Sanderson et al., 1998; Bininda-Emonds et al., 2002). The average procedure has also been employed to produce supertrees while taking into account branch lengths (Lapointe and Kirsch, 2001). However, both of these approaches can be used to combine source trees with identical leaf sets, in the so-called consensus setting (sensu Bininda-Emonds, 2003). In that particular context, MRP and MRD represent two sides of the same coin, and these methods are closely related consensus techniques. Here, we briefly describe the coding and optimization steps of both approaches to identify their resemblances and differences, and we have used an example to illustrate the equivalence among those seemingly different methods, in the consensus setting. Finally, we note that the close relationship between MRP and MRD may not hold in the supertree context.

CODING TREES FOR MRP

Given a rooted tree t representing the relationships among a set of leaves (taxa) $S = \{1, \ldots, n\}$, there exist a variety of possible binary matrix representations m (e.g., Purvis, 1995; Ronquist, 1996; Wilkinson et al., 2001) corresponding to t. Here, we focus on the representation originally introduced by Ragan (1992). We define a binary matrix m, with n rows representing the leaf set S of t and p columns (or matrix elements, sensu Baum and Ragan, 1993) representing the internal nodes of t. For each such element of m, all terminal taxa (leaves) descending from the corresponding node are scored 1, and all others are scored 0. To polarize the elements, an additional line is added to the matrix to represent an outgroup (or root) with all-zero values. A parsimony analysis of this binary matrix representation will recover the corresponding tree it is encoding (Baum, 1992; Ragan, 1992) when zero-length branches, if there are any, are collapsed. As a special case, a fixed number n of additional columns could be added to the matrix to represent the terminal nodes of t, each one scored 1 for the corresponding taxon and 0 otherwise. These elements are noninformative in a cladistic sense (i.e., they represent

autapomorphic characters), and a parsimony analysis of such an extended matrix will produce the same tree as a binary matrix representation coding only for internal nodes.

Coding Trees for MRD

Given a weighted tree t representing the relationships among a set of leaves (taxa) $S = \{1, \ldots, n\}$, there exist a variety of possible distance matrix representations d encoding t. For example, d can be coded as a square $n \times n$ path-length distance matrix containing for every pair of taxa a and b the sum of the branch lengths along the path between a and b (Buneman, 1971). With no loss of generality, that matrix d can code only for topological relationships by setting all branch lengths to 1. These path-length or branch distance (sensu Zaretskii, 1965) matrices define unrooted trees, but rooted trees also could be considered by adding an internal node (the root) to the tree (and correspondingly in the matrix d) to provide ancestor–descendent relationships. Any distance algorithm applied to this MRD will recover the corresponding tree (whether rooted or not and with or without branch lengths).

Optimization Criterion for MRP

Let $T = \{t_1, \ldots, t_k\}$ be a collection of rooted trees defined on the same set of leaves $S = \{1, \ldots, n\}$. If $\mathrm{L}(t_1, t_2)$ denotes the parsimony fit of the binary matrix representation m_2 of t_2 to the tree t_1, the MRP consensus tree t_c is defined as the solution to the following minimization criterion (Bryant, 2003; Thorley and Wilkinson, 2003):

$$\mathrm{L}(t_c, T) = \sum_{i=1}^{k} \mathrm{L}(t_c, t_i). \quad (1)$$

It is computed by applying a parsimony algorithm to a binary matrix combining the k matrix representations m_i of the corresponding trees t_i of T (Baum, 1992; Ragan, 1992). When more than one parsimonious tree is obtained from the combined matrix, a strict consensus of these trees is computed to synthesize the results.

Optimization Criterion for MRD

Let $T = \{t_1, \ldots, t_k\}$ be a collection of weighted trees (rooted or unrooted) defined on the same set of leaves $S = \{1, \ldots, n\}$, and let $\Delta(t_1, t_2)$ be the sum of squared differences between the path-length distance matrices d_1 and d_2 associated with the trees t_1 and t_2. The MRD consensus tree t_c is the tree (with branch lengths) that minimizes the following criterion (Lapointe and Cucumel, 1997):

$$\Delta(t_c, T) = \sum_{i=1}^{k} \Delta(t_c, t_i). \quad (2)$$

Practically, t_c is obtained by applying a least-squares algorithm (e.g., Cavalli-Sforza and Edwards, 1967) to a matrix $\bar{d}$ of average pairwise distances computed over the k matrix representations with distances encoding the trees of T or to the sum of pairwise distances (see Levasseur and Lapointe, 2002). When more than one solution is obtained, a strict consensus is computed to represent topological agreement among the consensus trees.

MRP Meets MRD

There exists a nice correspondence between binary matrix representations of trees and branch distance matrices. Suppose that all branch lengths of a tree t_i (including terminal ones) are set to 1 and that terminal nodes (leaves) are coded as additional elements in the encoded matrix m_i, the path-length distance $d_i(a, b)$ between two taxa a and b is then equal to the number of matrix elements of m_i for which a and b disagree. In other words, the Hamming (or mismatch) distance between any two rows in the binary matrix representation of a tree is exactly the branch distance between the corresponding taxa.

To illustrate this relationship, consider the trees t_1 and t_2 in Figure 1, with their corresponding binary matrices m_1 and m_2 and branch distance matrices d_1 and d_2. These branch distance matrices are exactly the Hamming distance matrices computed from the corresponding binary matrices. The MRP consensus tree is obtained by combining the matrices m_1 and m_2 representing t_1 and t_2 in a single data set m_{1+2}, and applying a parsimony algorithm. Similarly, the MRD consensus tree is computed from a matrix d_{1+2} summing the branch distance matrices d_1 and d_2 and applying a least-squares algorithm. That matrix (d_{1+2}) can also be derived by computing Hamming distances from the combined binary matrix representation (m_{1+2}), just like for individual matrices. The difference between MRP and MRD then lies in the optimization criterion selected to construct consensus trees from distinct matrix representations (Eqs. 1, 2). In this particular example, although not in general, parsimony and distance algorithms produced the same pair of consensus trees, which were identical to the source trees t_1 and t_2. Consequently, the strict consensus of both MRP and MRD trees is also identical (see Fig. 1).

Discussion

We have shown that MRP and MRD are related methods when used in a consensus setting. Whereas the coding steps of both approaches are equivalent, the optimization steps based on different criteria are not identical. The example presented in Figure 1 illustrates that these methods can lead to the same results in specific situations when used in a coherent fashion. However, this may not always be the case, especially when the trees combined differ in branch lengths and when path-length distances are used instead of branch lengths to compute the MRD consensus trees. Still, a weighted version of MRP, using actual branch length as weights of the

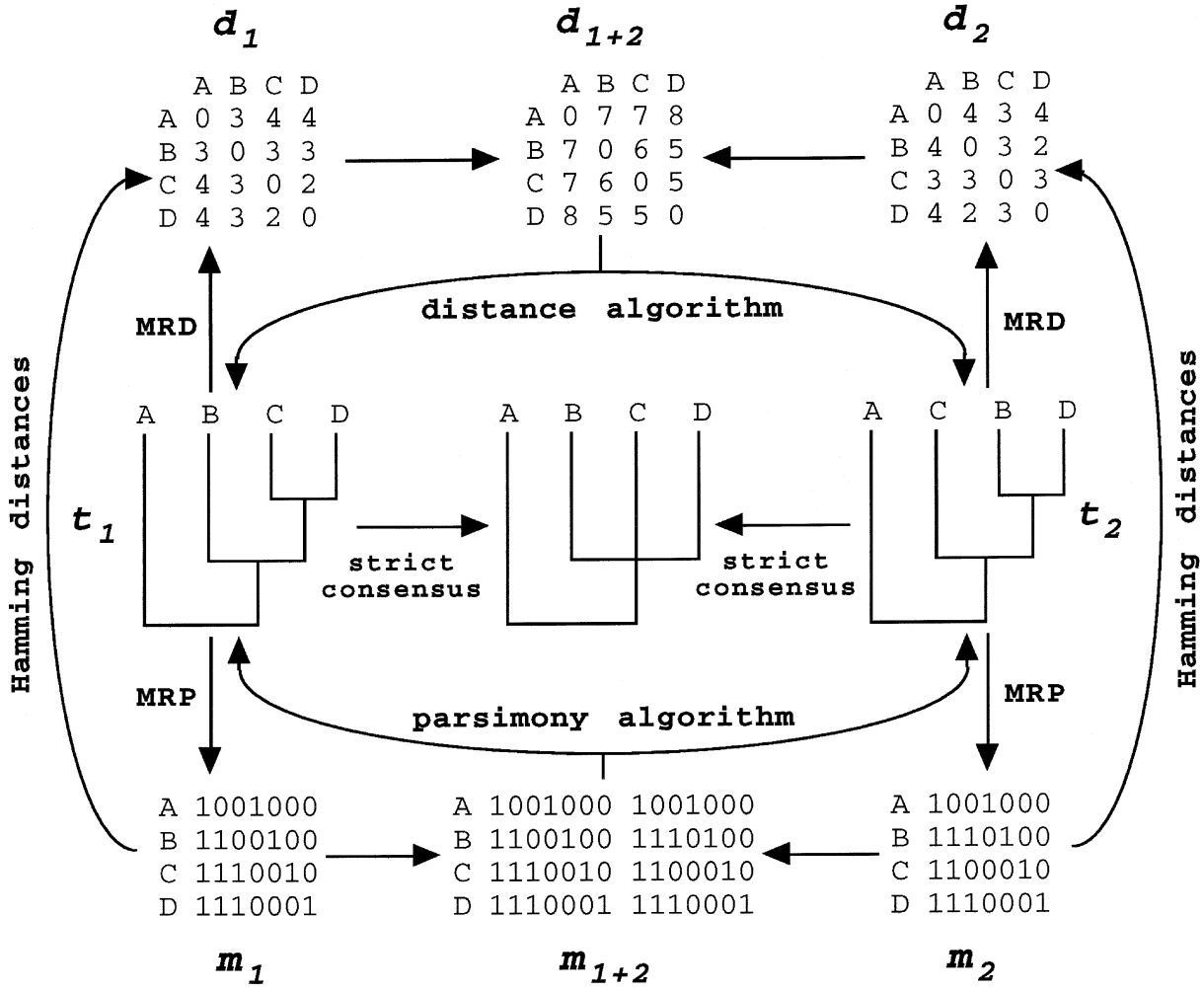

FIGURE 1. The relationship between MRP and MRD. The trees t_1 and t_2 can be coded either with binary matrix representations (m_1 and m_2) or with path-length distance matrices (d_1 and d_2) by setting all branch lengths to 1. Similarly, the branch distance between two taxa in d_1 and d_2 can be directly computed as the Hamming distance between the corresponding rows in m_1 and m_2. The combination of the different matrix representations is used to compute MRP or MRD consensus trees by applying a parsimony or a distance algorithm to m_{1+2} or d_{1+2}, respectively. In the present case, both approaches produce the same two consensus trees, identical to t_1 and t_2. The strict consensus of those trees represents the topological agreement between them.

corresponding matrix elements (and including terminal nodes as extra columns), is equivalent to using path-length distances for MRD. The correspondence in that case is satisfied by accounting for weights when computing Hamming distances among pairs of taxa. MRP could then be used as an alternative consensus method for weighted trees (see also Lapointe, 1998).

Generalization of our demonstration to the supertree setting (sensu Gordon, 1986) is not possible because the relationship between binary matrix representations and branch distance matrices does not hold for trees bearing nonidentical leaf sets. Therefore, MRP and MRD codings are equivalent only when (1) all the trees combined are defined on the same set of leaves (i.e., in the consensus setting), (2) all branch lengths are set to 1, and (3) Hamming distances are used to compute distance matrices from binary matrix representations of source trees. Under these very specific conditions, MRP and MRD differ only in terms of optimization criteria (Eqs. 1, 2). As such, the discrepancies obtained with these methods, if any, can be attributable to the well-known differences between parsimony and distance algorithms in phylogenetic analysis (see Swofford et al., 1996).

ACKNOWLEDGMENTS

We thank Mike Steel and Olaf Bininda-Emonds for helpful comments on the manuscript. This work was supported by NSERC grants 0155251 to FJL and 238975-01 to DB, and by BBSRC grant 40/G18385 to MW.

REFERENCES

BAUM, B. R. 1992. Combining trees as a way of combining data for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10.

BAUM, B. R., AND M. A. RAGAN. 1993. Reply to A. G. Rodrigo's "A comment on Baum's method for combining phylogenetic trees." Taxon 42:637–640.

BININDA-EMONDS, O. R. P. 2003. MRP supertree construction in the consensus setting. Pages 231–242 *in* Bioconsensus (M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds.). American Mathematical Society, Providence, Rhode Island.

BININDA-EMONDS, O. R. P., J. L. GITTLEMAN, AND M. A. STEEL. 2002. The (super)tree of life: Procedures, problems and prospects. Annu Rev. Ecol. Syst. 33:265–289.

BRYANT, D. 2003. A classification of consensus methods for phylogenetics. Pages 163–183 *in* Bioconsensus (M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds.). American Mathematical Society, Providence, Rhode Island.

BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity. Pages 387–395 *in* Mathematics in archeological and historical sciences (F. R. Hodson, D. G. Kendall, and P. Tautu, eds.). Edinburgh Univ. Press, Edinburgh, U.K.

CAVALLI-SFORZA, L. L., AND A. W. F. EDWARDS. 1967. Phylogenetic analysis: Models and estimation procedures. Evolution 32:550–570.

GORDON, A. D. 1986. Consensus supertrees: The synthesis of rooted trees containing overlapping sets of labeled leaves. J. Classif. 3:335–348.

LAPOINTE, F.-J. 1998. For consensus (with branch lengths). Pages 73–80 *in* Advances in data science and classification (A. Rizzi, M. Vichi, and H.-H. Bock, eds.). Springer-Verlag, Berlin.

LAPOINTE, F.-J., AND G. CUCUMEL. 1997. The average consensus procedure: Combination of weighted trees containing identical or overlapping sets of taxa. Syst. Biol. 46:306–312.

LAPOINTE, F.-J., AND J. A. W. KIRSCH. 2001. Construction and verification of a large phylogeny of marsupials. Aust. Mammal. 23: 9–22

LEVASSEUR, C., AND F.-J. LAPOINTE. 2002. A family of average consensus methods for weighted trees. Pages 355–369 *in* Classification, clustering and data analysis: Recent advances and applications (H.-H. Bock, K. Jajuga, and A. Sokolowski, eds.). Springer-Verlag, Berlin.

PURVIS, A. 1995. A modification to Baum and Ragan's method for combining phylogenetic trees. Syst. Biol. 44:251–255.

RAGAN, M. A. 1992. Phylogenetic inference based on matrix representations of trees. Mol. Phylogenet. Evol. 1:53–58.

RONQUIST, F. 1996. Matrix representation of trees, redundancy, and weighting. Syst. Biol. 45:247–253.

SANDERSON, M. J., A. PURVIS, AND C. HENZE. 1998. Phylogenetic supertrees: Assembling the trees of life. Trends Ecol. Evol. 13:105–109.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Mass achusetts.

THORLEY, J. L., AND M. WILKINSON. 2003. A view of supertree methods. Pages 185–193 *in* Bioconsensus (M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds.). American Mathematical Society, Providence, Rhode Island.

WILKINSON, M., J. L. THORLEY, D. T. J. LITTLEWOOD, AND R. A. BRAY. 2001. Towards a phylogenetic supertree of Platyhelminthes? Pages 292–301 *in* Interrelationships of the Platyhelminthes (D. T. J. Littlewood and R. A. Bray, eds.). Taylor and Francis, London.

ZARETSKII, K. 1965. Constructing a tree on the basis of a set of distances between the hanging vertices. Usp. Math. Nauk 20:90–92 (in Russian).

First submitted 5 February 2003; reviews returned 23 April 2003; final acceptance 21 August 2003
Associate Editor: Mike Steel