

Consensus, Networks, and Information Content

Olivier Gauthier¹ and François-Joseph Lapointe²

Keywords: consensus, information, matrix representation, split decomposition, median networks

1 Introduction.

Phylogenetic inference often results in the production of multiple trees on a given set of leaves. Consensus methods are commonly used to identify areas of conflict and agreement among trees or only retain the relationships that are supported either unanimously or by a majority of trees while discarding other, less supported relationships. The choice of a given consensus method can be based on different criteria. From an axiomatic perspective, methods that are pareto and co-pareto should be selected. On the other hand, one may prefer methods that produce consensus containing more information, i.e. consensus trees that are well resolved. In this paper we discuss different ways to produce consensus containing more phylogenetic information in the form of Consensus Networks (CN). The measure of Phylogenetic Information Content (PIC) [1] is extended here to measure the information content of consensus networks.

2 Consensus networks.

Recent years have seen the description of several new methods and algorithms to reconstruct phylogenetic networks (e.g. [2]). These methods can be used to compute CN by first obtaining a Matrix Representation (MR) of the input trees [3]. For example, one can combine the input trees with their average distance matrix and then compute the consensus using split decomposition (SD) [4]. Alternatively, median networks (MN) [6] can be used to compute the consensus of trees represented by their bipartitions [5]. Since MN include all the most parsimonious trees from a given data set unless it is constrained to two dimensions, this CN method produces solutions that include all the trees in the input profile. On the other hand, SD eliminates the relationships that are the least supported (Fig. 1). This difference has important implications when we consider the PIC of the CN.

3 Consensus and information content.

In order to extend the PIC of [1] to CN, we define it as being proportional to the ratio of the number of binary trees that refine the CN to the number of possible binary trees for the same number of leaves. Since networks contain multiple trees that do not contain redundant information, the PIC of a CN is obtained by summing the PIC of the maximally resolved trees contained in the network. In the examples described in Fig. 1, the strict consensus and the CN using MN of the tree input trees contain no phylogenetic information, since both Fig. 1 (a) and (d) can be refined by the three possible trees with four leaves. When one of the topologies is less supported than the other, the CN using SD will have one of the topologies in Fig. 1 (b), which contains 2/3 of the maximum PIC possible with four leaves; i.e. these topologies can be refined by only two of the possible trees with

¹ Département de sciences biologiques, Université de Montréal, C.P. 6128, Succursale A, Montréal, Québec, Canada H3C 3J7, E-mail: olivier.gauthier@umontreal.ca

² Département de sciences biologiques, Université de Montréal, C.P. 6128, Succursale A, Montréal, Québec, Canada H3C 3J7, E-mail: francois-joseph.lapointe@umontreal.ca

four leaves. Hence, although less efficient [7] than the CN using MN, a CN computed with SD will always convey more information than other consensus methods that do not resolve conflict.

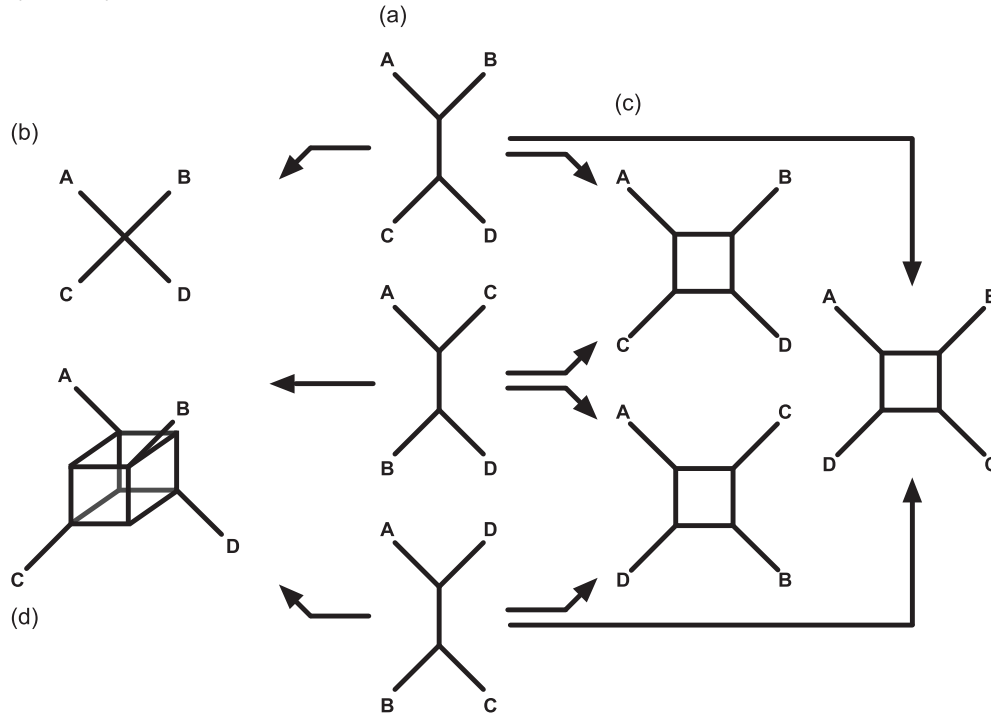


Figure 1: (a) Three input trees, (b) their strict consensus, and consensus networks (CN) using either (c) split decomposition (SD) or (d) median network (MN). The CN using SD will return one of the topologies in (c) depending on which tree in (a) is the least supported, whereas MN will return the topology presented in (d). When all trees are equally supported, SD returns (b).

References

- [4] Bandelt, H.J. and Dress, A.W.M. 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular phylogenetics and Evolution* 1:242-252.
- [6] Bandelt, H.-J., Forster, P., Sykes, B.C. and Richards, M.B. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* 141:743-753.
- [5] Holland, B. and Moulton, V. 2003. Consensus networks: a method for visualising incompatibilities in a collection of trees. In *Proceedings of the Third International Workshop in Algorithms in Bioinformatics (WABI)*, Berlin. Springer. pp. 165-176.
- [3] Lapointe, F.-J., Wilkinson, M. and Bryant, D. 2003. Matrix Representations with Parsimony or with Distances: Two Sides of the Same Coin? *Systematic Biology* 52:965-8668.
- [2] Posada, D. and Crandall, K. A. 2001. Intraspecific gene genealogies: Trees grafting into networks. *Trends in Ecology and Evolution* 16:37-45.
- [1] Thorley, J. L., Wilkinson, M. and Charleston, M.A. 1998. The information content of consensus trees. In Rizzi, A., Vichi, M. and Bock, H.-H., editors, *Advances in data science and classification*, Berlin: Springer. pp. 91-98.
- [7] Wilkinson, M. and Thorley, J.L. 2001. Efficiency of strict consensus trees. *Systematic Biology* 50:610-613.