New algorithm for improving reticulated networks

Alix Boc[1], Abdoulaye Banire Diallo[2], Vladimir Makarenkov[3]

Keywords: phylogenetic tree, reticulated network, distance matrix, least-squares criterion

1 Introduction.

Species evolution has long been assumed to be a branching process uniquely representable by a tree topology. In such a topology, each species can only be linked to its closest ancestor; interspecies relationships are not allowed. A reticulate model can adequately describe such complicated mechanisms as species hybridization or lateral gene transfer in bacteria (see Legendre and Makarenkov 2002). A method and a corresponding software for inferring reticulated networks representing evolutionary relationships among a group of species have been described in Makarenkov (2001) and Makarenkov and Legendre (2002).

In this paper, we present a least-square algorithm for improving the topology of a given reticulated network. The new algorithm includes two main steps. First, it proceeds by addition of new branches to the topology of a classical phylogenetic tree to provide a reticulated network (for more details see, Legendre and Makarenkov 2002). Second, it allows to rearrange the network topology by removing or substituting some of its branches according to the OLS or WLS models.

2 Algorithm.

Assume that we have already a reticulated network constructed from a distance matrix between taxa (e.g. species) using the method from Legendre and Makarenkov (2002). The algorithm we describe here introduces several new operations designed to improve the least-squares (LS) approximation of a given distance matrix by a reticulated network. The new algorithm proceeds by local rearrangements of a network topology improving at each step the value of the LS or WLS coefficients.

First, the LS coefficient can be *improved by removing* an existing branch and adding a new one. All branches in the reticulated network, including those of the original phylogenetic tree, are considered as candidates for removal. The only restriction for this operation is that the resulting network have to remain connected. For a particular branch *ab* of length l_{ab} considered for removal from a reticulated network *R*, we have to find all pairs of taxa that will be affected by this deletion. This means that for any pair of taxa *ij* such that either $dist(ia) + l_{ab} + dist(bj) = dist(ij)$ or $dist(ja) + l_{ab} + dist(bi) = dist(ij)$, we have to recompute the value $dist(ij)$ assuming that the branch *ab* is no longer in *R*.

Second, a branch-removing operation can be followed by a branch-addition operation, what will consist in a *branch-substitution operation*. The pair of branches (removed and added) corresponding to the lowest value of the LS or WLS coefficient can be selected for substitution. These operations may significantly redesign the topology of the initial reticulated network. The time complexity of the a branch-substitution operation is $O(mn^4)$, where *m* is the number of

[1] Department of Informatics, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8. e-mail: boc.alix@courrier.uqam.ca

[2] Department of Informatics, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8. e-mail: diallo.abdoulaye_banire@courrier.uqam.ca

[3] Department of Informatics, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8. e-mail: makarenkov.vladimir@uqam.ca

branches in the reticulated network and *n* is the number of taxa. If only the branch removal operation is considered, we simply have to recompute the value of the LS or WLS criterion and make the decision about the potential branch deletion.

In addition, the new algorithm also proceeds by reassessment of branch length estimates of a reticulated network with a fixed topology. The reassessment loop may be repeated several times to reach the minimum value of the LS or WLS criteria. As this is usually the case, improvement in fit causes increase in time complexity. Thus, if the reassessment procedure is incorporated into the algorithm, the time complexity of each iteration will increase up to $O(pmn^4)$, where *p* is the number of reassessment loops performed over all *m* branches.

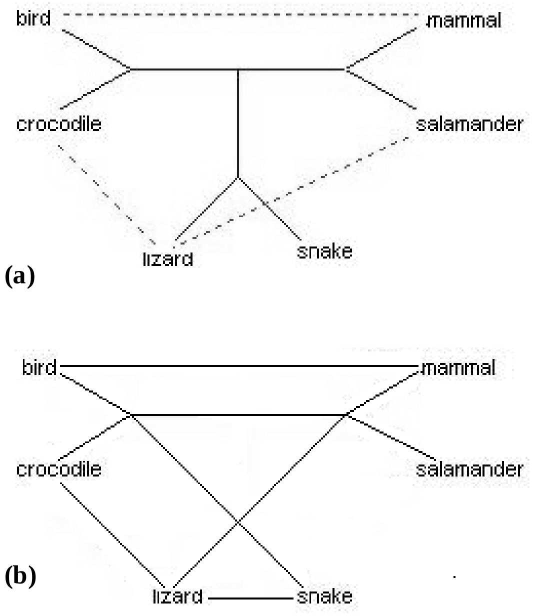

Figure 1: Reticulated network representing phylogenetic relationships among six vertebrate organisms (for details on the vertebrate morphological dataset, see *MacClade* 4.0 manual by D. R. Maddison and W. P. Maddison).
(**a**) First, a phylogenetic tree (full lines) were inferred from a distance matrix among vertebrates using the NJ method by Saitou and Nei (1987); then, three reticulation branches (dashed lines) were added to the phylogenetic tree to create a reticulated network. The reticulation branches linking *lizard* and *crocodile*, *lizard* and *salamander*, and *mammal* and *bird* show that these species are more closely related to one another than it is illustrated by the NJ phylogenetic tree. (**b**) The branch-substitution procedure ran on these data totally redesigned the network topology leading to the following reticulated network.

3 References.

[1] Legendre, P. and Makarenkov, V. 2002. Reconstruction of biogeographic and evolutionary networks using reticulograms, *Systematic Biology* 51:199-216.

[2] Makarenkov, V. 2001. T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks, *Bioinformatics* 17:664-668.

[3] Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. and Evol.* 4:406-425.