ABSTRACT

Orthology and paralogy distinguish whether a pair of genes originated by a speciation or a gene duplication event, whereas xenology refers to horizontal gene transfer. These concepts play a key role in phylogenomics and species tree inference is one of its prevalent tasks. Commonly, species tree inference is performed using sequence-based phylogenetic methods which heavily rely on the initial data sets to be solely composed of 1:1 orthologs. Such approaches are strongly restricted to a small set of genes that provide information about the species tree. In this work, it is shown that the restriction to 1:1 orthologs is not necessary to reconstruct a reliable hypothesis on the evolutionary history of species.

Besides orthology, knowledge on all three major driving forces of gene evolution can be considered: speciation, gene duplication, and horizontal gene transfer. The corresponding concepts of orthology, paralogy, and xenology imply binary relations on pairs of genes. These relations, in turn, convey meaningful phylogenetic information and allow the inference of plausible phylogenetic species trees.

To this end, it is shown that orthology, paralogy, and xenology have to fulfill certain mathematical properties. In particular, they have to be representable as a tree – the so-called gene tree. This work investigates the theoretical concepts of tree representable sets of binary relations to unfold the underlying mathematical structure. Various novel characterizations for those relations are given and the close connection between tree representable sets of binary relations and cographs, symbolic ultrametrics, and so-called *unp* 2-structures is revealed. Based on the novel characterizations, polynomial-time recognition algorithms for tree representable sets of relations are presented. In the case, a set of relations is tree representable, the corresponding tree representation can be found in polynomial time as well.

Moreover, for the NP-complete problems of editing a given set of relations to its closest tree representable set, exact algorithms are developed by means of formulations as integer linear program. Finally, all algorithms have been implemented in the software `ParaPhylo`, a species tree inference method based on orthology and paralogy data. It is demonstrated on simulated data sets, as well as real-life data sets, that non-trivial phylogenies can indeed be reconstructed from tree-free orthology estimates alone.