Reconstructing phylogenetic trees from clustering trees

Costa E (1,*), Vens C (1,2), Blockeel H (1,3)

1 - Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium

2 - Institut National de Recherche Agronomique, UMR 1301, 400 Route des Chappes, 06903 Sophia-Antipolis, France

3 - Leiden Institute of Advanced Computer Science, Universiteit Leiden, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

{eduardo.costa,celine.vens,hendrik.blockeel}@cs.kuleuven.be

TEASER

Some top-down methods for phylogenetic tree construction can be viewed not just as constructing trees, but as identifying constraints that the phylogenetic tree must satisfy. We show that this viewpoint can lead to improved phylogenetic trees.

MOTIVATION

In the context of phylogenetic tree reconstruction, divisive clustering methods can be used to infer phylogenetic trees in a top-down way. These methods have the important advantage of providing an explanation for the resulting topology, since the splits are described by polymorphic locations in the sequences. However, the quality of the resulting trees is rather variable. In this work, we argue that trees induced by top-down methods can be viewed not just as phylogenetic trees, but also as identifying constraints that the real phylogenetic tree must satisfy.

MATERIALS AND METHODS

We analyzed trees inferred by Clus-φ, a distance based method for phylogenetic tree reconstruction based on a conceptual clustering method that extends the well-known decision tree learning approach. Each split defines two subclusters, such that the total branch length of the tree is minimized. However, the split does not define how the subclusters have to be connected. We propose a post-processing method that processes the clustering tree bottom-up, at each split finding the internal branch that connects the two subclusters with a minimal number of mutations.

RESULTS

To evaluate this method we used a number of synthetic datasets generated by an evolutionary process simulator. In general, the post-processed Clus-φ trees are more similar to the underlying target trees of the synthetic datasets than the original Clus-φ trees, which shows that the post-processing step yields a better approximation of the target tree. When we consider Neighbor Joining and Parsimony results in this comparative analysis, we observe that the post-processed Clus-φ trees tend to be better than the NJ trees and are comparable to the parsimony trees.

DISCUSSION

The results show that trees resulting from top-down phylogenetic tree construction can be improved by post-processing them. This post-processing is based on the viewpoint that the methods do not necessarily return the correct tree, but return constraints that the correct tree must satisfy. These constraints allow to guide the search for the tree with a minimal number of mutations in a more exhaustive way than the greedy search performed by parsimony methods. In general, the quality of post-processed Clus-φ trees are comparable to that of parsimony trees.

ACKNOWLEDGMENTS

Eduardo Costa is supported by the Research Foundation – Flanders (FWO) and the GOA Probabilistic Logic Learning. Celine Vens is a Postdoctoral Fellow of the Research Foundation – Flanders (FWO).