

Celine Vens

Abstract

We propose a novel distance based method for reconstruction of phylogenetic trees. Our method is based on a conceptual clustering method that extends the decision tree learning approach. Our method starts from a single cluster and repeatedly splits it into subclusters until all sequences form a different cluster. We assume that a split can be described by referring to particular polymorphic locations, which makes such a divisive method computationally feasible.

To define the best split, our method uses a criterion that is close to Neighbor Joining's (NJ) optimization criterion, namely, constructing a phylogenetic tree with minimal total branch length. Our approach has a number of important advantages.

First, by listing the polymorphic locations at the internal nodes, it provides an explanation for the resulting tree topology.

Second, the resulting trees can be used as classification trees to add new sequences to the phylogeny.

Third, the top-down tree growing process can be stopped before a complete tree is generated, yielding an efficient gene or protein subfamily identification approach.

Our algorithm, called Clus-IT, can be found at <http://www.cs.kuleuven.be/~dtai/clus>.