Using Clustering Trees for Learning Phylogenetic Trees

Celine Vens and Hendrik Blockeel

Department of Computer Science, Katholieke Universiteit Leuven
Celestijnenlaan 200A, 3001 Leuven, Belgium
{celine.vens,hendrik.blockeel}@cs.kuleuven.be

1 Introduction

This paper presents ongoing work on an application of machine learning in phylogenetic analysis, which is the study of evolutionary relatedness among various groups of organisms. Insights in evolutionary relationships are important because they can help to determine the function of uncharacterized genes and they can be used to predict future variants of fast-growing viruses.

More precisely, we focus on the following task: given a set of DNA sequences, and given that they all originate from a single sequence via successive mutations, find the phylogenetic tree that describes the evolutionary process. In Section 2, we explain what phylogenetic trees are and how they are usually built. Section 3 presents clustering trees, which will be used in Section 4 to construct phylogenetic trees. Section 5 summarizes the advantages of our approach.

2 Phylogenetic trees

A phylogenetic tree is a tree that graphically illustrates the evolutionary relationships among various species or organisms. Each leaf node in the tree represents an organism and nodes share a common ancestor if they are believed to originate from the same organism. Fig. 1 shows a phylogenetic tree for the HIV-1 dataset[1].

One of the most popular algorithms for constructing phylogenetic trees is the neighbor-joining method [3]. It takes into account the similarity of the molecular information (e.g., DNA sequence) of organisms. Closely related organisms generally have a high degree of agreement in their molecular structure, while the molecules of organisms distantly related usually show a pattern of dissimilarity. The algorithm starts by calculating the dissimilarity between each pair of sequences (based on edit distance) to produce a pairwise distance matrix. Afterwards, a bottom-up hierarchical clustering algorithm is applied that initially assigns each individual to its own cluster and iteratively joins the two most similar clusters (by constructing a common parent node) until only one cluster remains. The algorithm has a time complexity quadratic in the number of sequences [2].

Other well-known methods to construct phylogenetic trees are the maximum parsimony and maximum likelihood methods, which both use an exhaustive search over all possible phylogenetic trees.

The existing methods have difficulties with modelling convergent evolution, which means that several organisms are similar, not as a result of a single mutation in the past, but because of evolutionary pressure.

[1] *hivALN.phy* file at `http://www.kuleuven.be/aidslab/phylogenybook/dataset.htm`.

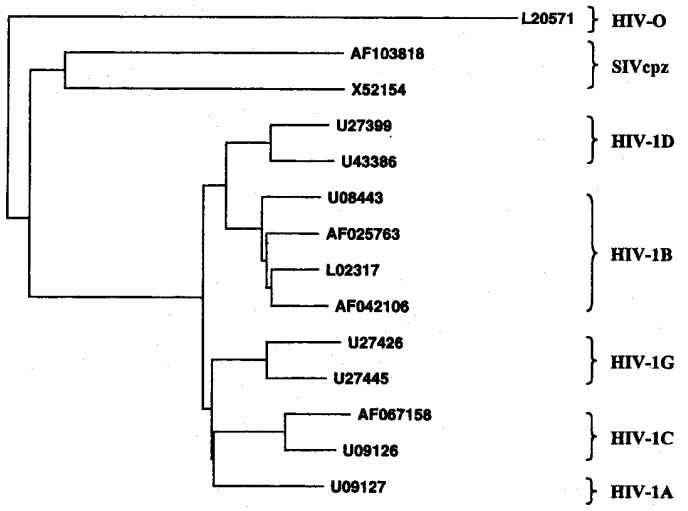


Figure 1: Phylogenetic tree for the HIV-1 dataset output by the neighbor-joining method. The figure is taken from [4]. The HIV-1 subtype each sequence belongs to is shown in the right.

3 Clustering trees

As Blockeel et al. [1] note, a decision tree can be seen as a clustering tree. More precisely, a decision tree is viewed as a hierarchy of clusters: the root node corresponds to one cluster containing all data, which is recursively partitioned into smaller clusters while moving down the tree. The standard "top-down induction of decision trees" (TDIDT) algorithm is easily adopted to grow such clustering trees. In fact, induction of clustering trees generalizes induction of decision trees by ensuring homogeneity according to any set of attributes, instead of to one target attribute, in each subset of the partition on the training instances induced by a split. This involves using a heuristic function that selects in each node the test that minimizes the distance within the resulting clusters in its child nodes. The exact definition of this distance can be instantiated for a given learning task.

The clustering tree framework is implemented in the CLUS system. More information about clustering trees and CLUS can be found at `http://www.cs.kuleuven.be/~dtai/clus`.

4 Using clustering trees for learning phylogenetic trees

A node in a phylogenetic tree means that at some point in the past, a certain mutation gave rise to two separate lines of evolution. In order to find this mutation, we can inspect all positions of the sequences in the set corresponding to the node. Every mutation at each position divides the set into a subset with and without the mutation at that position. The further away these two subsets are, the more likely it is that this mutation happened long ago.

The above reasoning motivates the top-down construction of phylogenetic trees and is implemented by learning a clustering tree, using as a distance function the edit distance between sequences. As such, the proposed method forms clusters in a similar way as the neighbor-joining method, but in a top-down fashion. An important difference is that in the new approach a cluster is defined by a conjunction of simple properties (the conjunction of consecutive mutations that have led to the cluster), instead of by enumerating all elements.

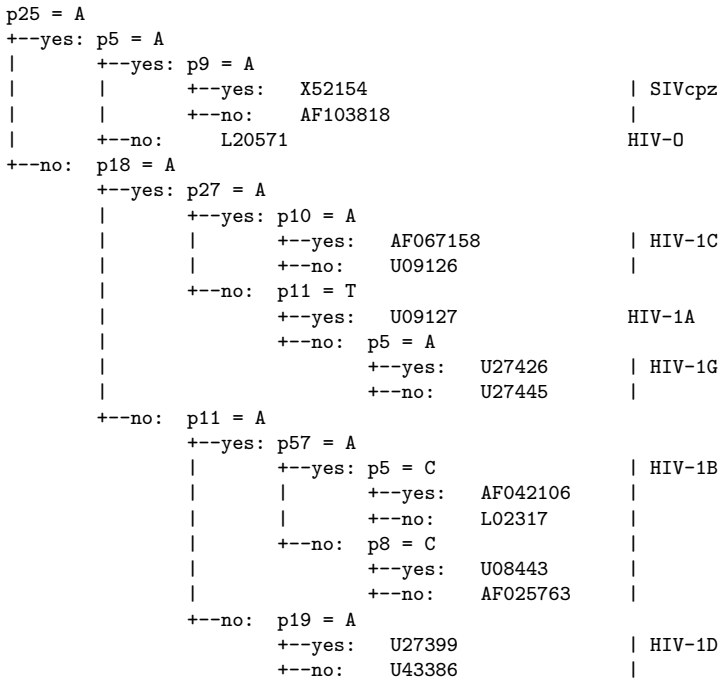
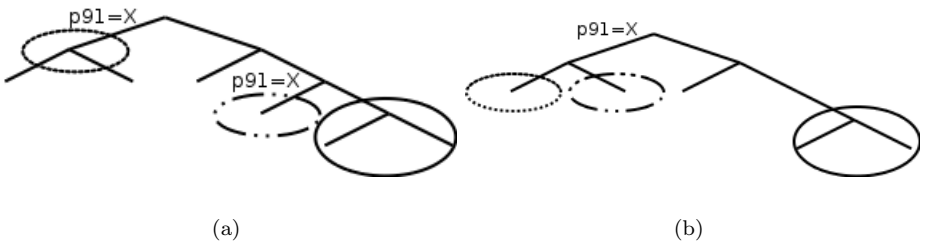


Figure 2: Phylogenetic tree for the HIV-1 dataset output by CLUS. A test $p25 = A$ means that at position 25 of the DNA sequence, there is a mutation into A. The HIV-1 subtype each sequence belongs to is shown in the right.

An initial experiment on the HIV-1 dataset with our method yields the tree shown in Fig. 2. A comparison of this figure to Fig. 1 shows similar structures in the tree, although splits may be slightly different. This is probably due to the distance measure used by the neighbor-joining method, which applies a correction for the occurrence of multiple mutations at the same position and has not yet been incorporated in our distance measure. A second observation is that the neighbor-joining method employs the distance between clusters to visualize a time dimension, which can also be included in our approach.

The time complexity of the proposed method is $O(ndl)$, with n the number of sequences, d the depth of the tree, and l the length of the sequences. As such, the method scales much better than other phylogenetic tree constructing methods in the number of sequences.

There are still some issues to be investigated. If a position is polymorphic (i.e., mutated versions are frequent at that position of the sequences), this may mean two things: (1) there was a mutation long ago, or (2) the mutation occurred multiple times more recently in evolution (e.g., by evolutionary pressure). Obviously, we are interested in selecting mutations of the first kind. Whereas existing methods have problems with this, the proposed heuristic has a tendency to select the right mutations, because if the mutation occurs multiple times at different points in time, it will occur in various dissimilar sequences, hence the intra-cluster distances will be relatively large. Now suppose the true tree contains a mutation that occurred long ago and that was repeated more recently (see Fig. 3(a), denoted by the branches $p91 = X$). If the algorithm chooses the right mutation, then every sequence with the mutation goes into the same cluster, even though some will be very different. Because of this large intra-cluster distance, groups of sequences belonging to different occurrences of the mutation will be very soon split off as a separate cluster (see Fig. 3(b)). Post-processing techniques are needed for detecting this kind of mutation and for reattaching some nodes to other nodes. Another issue that relates to this is to



(a) (b)

Figure 3: (a) True tree. (b) Clustering tree.

investigate the suitability of other heuristic functions, e.g., maximizing the minimum or mean distance between clusters, instead of minimizing the distance within clusters.

5 Summary

Building phylogenetic trees by using clustering trees has several important advantages. First, since a decision tree clusters conceptually, the subclusters are defined by description of the followed mutation paths instead of by enumeration. Second, our approach favours old mutations that truly gave rise to separate lines in evolution, while the existing methods have problems with this. Finally, our approach scales better towards the analysis of many sequences, which can be useful, for instance, in research to the fast mutating HIV where one is confronted with thousands of variants of the virus.

While the method is still being developed, initial experiments have demonstrated that the use of clustering trees, after certain non-trivial adaptations, is applicable to building phylogenetic trees.

Acknowledgements

Celine Vens is supported by the EU FET IST project "Inductive Querying", contract number FP6-516169. Hendrik Blockeel is a post-doctoral fellow of the Fund for Scientific Research of Flanders (FWO-Vlaanderen).

References

[1] H. Blockeel, L. De Raedt, and J. Ramon. Top-down induction of clustering trees. In *Proceedings of the 15th International Conference on Machine Learning*, pages 55–63, 1998.

[2] Isaac Elias and Jens Lagergren. Fast neighbor joining. In *Proc. of the 32nd International Colloquium on Automata, Languages and Programming*, pages 1263–1274. Springer-Verlag, July 2005.

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

[4] Y. Van de Peer and M. Salemi. Phylogeny inference based on distance methods. In A. Vandamme and M. Salemi, editors, *The phylogenetic handbook*, pages 101–136. Cambridge University Press, 2003.