

# Mining Mutation Pathways of HIV considering Phylogenetic Information

Snezhana Dubrovskaya, Jan Ramon, Leander Schietgat,  
Hendrik Blockeel

Dept. of Computer Science, Katholieke Universiteit Leuven,  
Celestijnenlaan 200A, 3001 Leuven, Belgium

`name.surname@cs.kuleuven.be`

The genetic sequence of HIV is only about 10K pairs long, which makes genome sequencing (determining the primary sequence) relatively easy and fast. The accumulation of HIV data during the last decade provides data mining with some interesting challenges. One example is the task of discovering mutation pathways that make HIV resistant against therapy.

It is widely known that the reproduction process of HIV is prone to errors, which causes mutations (genetic variations) in the genome of the virus. The existence of viruses having certain mutations can be explained by two reasons. First, natural evolution is responsible for the genetic variability. A second process is selective pressure. Only viruses which are able to adapt to their environment survive.

Predictive data mining models can give us insight in how the virus develops resistance against some drug. Still, it remains difficult to make a distinction between mutations related to selective pressure (in which we are interested) and mutations related to evolution. Phylogenetic trees can help us to solve this problem. They show the evolutionary relationships among the sequences, which are assumed to have a common ancestor. However, mutations that are related to selective pressure should be ignored while building a phylogenetic tree.

In our project we will develop an algorithm that builds predictive models, taking into account the phylogenetic structure of the sequences. This is a two-way process: first, a regular phylogenetic tree learner tries to make a taxonomy of the HIV sequences according to all mutations. Then, we apply predictive modelling to define some mutations that are related to selective pressure. Then, we implement this evolution model into a new phylogenetic learner, that will ignore (a subset of) the mutations related to selective pressure. Again, some predictive modelling is used to define new mutations. We hope that this process will eventually converge into a reasonable solution.

Our data include 8396 sequences from the Stanford HIV RT and Protease Sequence Database, which is available at: <http://hivdb.stanford.edu/>.