Efficient Estimation of Variable Order Markov Chains
Marcel H. Schulz, Hugues Richard and Martin Vingron

Max Planck Institute for Molecular Genetics, Berlin

Variable order markov chains have been applied to a variety of problems in computational biology, from gene annotation to protein domain detection. Particularly, it showed good performance for classifying protein families represented by domains of Pfam.

The order $L$ of a markov chain determines the length of the context, that is, the $L$ preceding letters which are taken into account to predict the next symbol. The general idea of variable order markov chains is to remove states from the fixed-order state space, which either cannot be reliably estimated or are redundant according to some estimation process. This estimation process acts recursively on the context comparing the gain in descriptive power for a context and its suffixes, like the context tree of Rissanen or the probabilistic suffix tree (PST)[3, 4].

We used enhanced suffix arrays to implement the linear-time and space algorithm of Apostolico and Bejerano for constructing a PST [1] and show that it is extremely fast in practice and up to three orders of magnitude faster than the previous implementation. The implementation is very general and can be used for model estimation with a variety of estimators, e.g. [3, 4].

Additionally, we consider the class of Parsimonious Markov Models [2], which is now described by ambiguous contexts, allowing for the grouping of characters at each position. Estimation of such models is computationally prohibitive, as it scales exponentially with the size of the alphabet. To tackle this problem we reduce the space of considered contexts according to the biochemical properties of amino acids.

Finally, we conduct classification experiments assessing the performance of the various estimation methods.

References


