Towards Discovering Structural Signatures of Protein Folds based on Logical Hidden Markov Models

(Extended Abstract)

Kristian Kersting [a] Tapani Raiko [b] Stefan Kramer [a] Luc De Raedt [a]

[a] *Machine Learning Lab, Computer Science Department,*
University of Freiburg
Georges-Koehler-Alle 079, 79112 Freiburg, Germany
[b] *Laboratory of Computer and Information Science,*
Helsinki University of Technology
P.O. Box 540, Helsinki, FIN-02015 HUT, Finland

Keywords: Hidden Markov Models, Probabilistic Modeling, First-Order Logic, Structural Signatures, SCOP

With the growing number of determined protein structures and the availability of classification schemes, it becomes increasingly important to develop computer methods that automatically extract structural signatures for classes of proteins. We present a new Machine Learning technique, *Logical Hidden Markov Models (LOHMMs)*, and apply it to the task of finding structural signatures of folds according to the classification scheme SCOP.

LOHMMs extend the unstructured model representation of HMMs by incorporating complex, internal structure into the specification of transitions (and therefore of emissions). In LOHMMs, we summarize sets of states by *abstract states*, which are represented by logical atoms. A logical atom then represents all states that can be obtained by instantiating the atoms. Abstract states are connected by *abstract transitions*, which summarize sets of transitions between states. As usual in predicate calculus, we allow for the use of variables as well as unification. They allow us to make abstraction of specific symbols and states. Together, the abstract states and abstract transitions of a logical (hidden) Markov model provide a compact representation of an equivalent HMM. We devised adaptations of the *forward* and *backward procedures*, which enabled us to adapt the Viterbi algorithm for determining the most likely state sequence, and the Baum-Welch algorithm for maximum likelihood estimation of the abstract transition probabilities. The resulting algorithm for learning the abstract transition probabilities is an instance of EM.

We applied LOHMMs to a new dataset extracted from the Protein Data Bank (PDB) and the SCOP (Structural Classification of Proteins) database. The resulting dataset contained the secondary structure of domains (of proteins) for five large fold classes as logical sequences, that is, sequences of logical atoms. The goal was to discover structural characteristics of these selected folds. Our preliminary results indicate that LOHHMs are in fact applicable to the task: The number of parameters of our LOHMM is by an order of magnitude smaller than the number of an equivalent HMM, and the generalization performance is comparable to the one of [TMS01], although the datasets are slightly different. Furthermore, we show that it is easy to extract characteristic patterns from the trained LOHMMs.

Summing up, we believe that LOHMMs show a great promise to become a useful tool for the analysis of structured sequence data in general and for the discovery of structural signatures of protein folds in particular. More details can be found in the long paper [KRKR02].

References

[KRKR02] K. Kersting, T. Raiko, S. Kramer, and L. De Raedt. Towards discovering structural signatures of protein folds based on logical hidden markov models. Technical Report 175, University of Freiburg, Germany, June 2002.

[TMS01] M. Turcotte, S. Muggleton, and M. J. E. Sternberg. The effect of relational background knowledge on learning of protein three-dimensional fold signatures. *Machine Learning*, 43(1/2):81–95, 2001.