Semi-supervised learning of Hidden Markov Models for biological sequence analysis

Hidden Markov Models (HMMs) are probabilistic models widely used in applications in computational sequence analysis. HMMs are basically unsupervised models. However, in the most important applications, they are trained in a supervised manner. Training examples accompanied by labels corresponding to different classes are given as input and the set of parameters that maximize the joint probability of sequences and labels is estimated. A main problem with this approach is that, in the majority of the cases, labels are hard to find and thus the amount of training data is limited. On the other hand, there are plenty of unclassified (unlabeled) sequences deposited in the public databases that could potentially contribute to the training procedure. This approach is called semi-supervised learning and could be very helpful in many applications. We propose here, a method for semi-supervised learning of HMMs that can incorporate labeled, unlabeled and partially-labeled data in a straightforward manner. The algorithm is based on a variant of the Expectation-Maximization (EM) algorithm, where the missing labels of the unlabeled or partially-labeled data are considered as the missing data. We apply the algorithm to several biological problems, namely, for the prediction of transmembrane protein topology for alpha-helical and beta-barrel membrane proteins and for the prediction of archaeal signal peptides. The results are very promising, since the algorithms presented here can significantly improve the prediction performance of even the top-scoring classifiers. Supplementary data are available at Bioinformatics online.