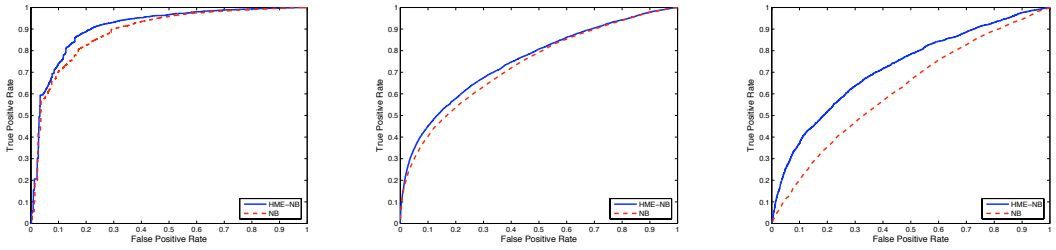
# Using Global Sequence Similarity Improves Biological Site-Specific Classifiers

Jivko Sinapov, Cornelia Caragea, Drena Dobbs and Vasant Honavar

Many prediction problems in Bioinformatics involve the prediction of class labels for each residue in a protein sequence (e.g., prediction of RNA binding residues, post-translational modification sites, etc.). Typically, the classifiers are trained based on local features of each site in the training set of protein sequences. In this work, we present a Hierarchical Mixture of Experts (HME) [1] model that improves such classifiers by taking into account the global sequence similarity between the test and training sequences.

First, we compute the pairwise similarity matrix for each pair of protein sequences based on a global sequence alignment score using the Blosum62 substitution matrix. Using the similarity matrix, we recursively partition the training set of proteins using the Spectral Clustering algorithm [2], resulting in a hierarchical partitioning of the sequences. A hierarchical mixture of experts classifier is trained such that each leaf node in the partitioning contains an expert model trained on data points from the leaf's sequences, while each midpoint node combines the predictions from its children based on the test sequence's cluster memberships. Each expert in the HME classifier is a Naïve Bayes model [3] trained on data points from the sequences that fall within the corresponding leaf node. The HME model with Naïve Bayes (HME-NB) is evaluated against a Naïve Bayes model trained on all data points in the training set of sequences, i.e. no global sequence similarity information is used.

Experiments are conducted on three representative problems: prediction of O-Linked glycosylation sites, prediction of RNA-binding protein sites and prediction of Protein-Protein interface sites in a protein sequence. The features for each residue are extracted by using a sliding window centered on the target residue. Figure 1 shows the ROC curves of the HME-NB and NB models obtained after performing 10-fold cross-validation on the three datasets. The figure shows that HME-NB outperforms NB on all three problems. In particular, considering the pairwise global similarity of the protein sequences significantly improves the performance measures on the Protein-Protein interface dataset.




a) O-Glycosylation b) Protein-RNA c) Protein-Protein

Fig. 1. Receiver Operating Characteristic curves HME-NB (solid, in blue) and NB (dotted, in red) models obtained after performing 10-fold cross-validation on the three problems: a) prediction of O-Linked glycosylation sites, b) prediction of RNA-binding sites, and c) prediction of Protein-Protein interface sites

References

[1] Waterhouse, S. R. and Robinson, A. J., Classification Using Hierarchical Mixture of Experts, *Proceeding of the 1994 IEEE Workshop on Neural Networks for Signal Processing*, 177-186, 1994.
[2] Paccanaro, A., Casbon, J. A., and Saqi, M. A. S. (2006), Spectral Clustering of Protein Sequences, *Nucleic Acids Research*, **34**.
[3] Mitchell, T. M., Machine Learning, McGraw Hill, 1997.