

Kernel-Based Machine Learning Methods in Genome Scale Protein Fold Assignment**Langlois, Robert, Dai, Yang, Lu, Hui****Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, USA**

The three-dimensional structure of a protein can be very useful in probing the function of a gene. Although various structural genomics projects have been developed around the world, currently there is still a four-order of magnitude gap between proteins with known sequence and proteins with known structure. Computer modeling methods have been used in predicting protein structures with high speed. We are developing a kernel-based machine learning package for genome scale protein fold assignment using Support Vector Machines (SVM). The SVM method is a powerful tool in discriminating two classes of objects with fast speed and only few training cases. With various voting systems, this technique has been extended to multi-class classification. SVM also provides a useful alternate to threading and other sequence-based methods in fold assignment. To achieve the high quality prediction, two key components have been developed: biological descriptors and the SVM classifier. Indeed, good descriptors are integral to SVM's ability to separate folds accurately. Further, parameter tuning, multi-class classifier building, and feature selection in the SVM classifier are crucial in dealing with the large number of protein classes. First, in descriptor development, we will adopt the strategy of developing biologically meaningful descriptors. Specifically, many insights can be taken from protein structure analysis and from previous protein structure prediction experiences. Additionally, descriptors have been built from primary sequences, predicted secondary structures, predicted folds, and data mining of Protein Data Bank (PDB). And finally, the descriptors have been tested with several benchmark test sets. With a published benchmark consisting of 27 folds, our SVM protocol can achieve 57% of prediction accuracy using only the amino acid composition, compared with the 49% accuracy in published results. Using a new test set with 53 folds, where each fold consists of at least 20 proteins from SCOP database (at least 10 for training and testing), our SVM protocol can achieve 22% of accuracy with composition alone. Although this is not practical, it is, however, much better than the baseline (random) prediction accuracy of 5%. Nevertheless, when carefully designed descriptors based on predicted secondary structure are added, the accuracy improves to 48%. Lastly, feature selection has also proved to reduce the number of descriptors needed for multi-fold classification and has increased the accuracy.