

5th Annual Joint Bioinformatics Workshop

July 19, 2005
2229 Seamans Center
University of Iowa

by

Peter Zaback
Iowa State University

Improved support vector machine prediction of protein structural features with a substitution matrix based kernel

ABSTRACT

Improved support vector machine prediction of protein structural features with a substitution matrix based kernel

Peter Zaback, Josh Williams, Feihong Wu, Vasant Honavar, and Drena Dobbs

Bioinformatics and Computational Biology Graduate Program, Laurence H Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, 50011

Technological advances continue to widen the chasm between available protein sequence data and associated structural information. Because tertiary structure ultimately determines a protein's function, it is important to develop computational methods capable of accurately predicting structural features from sequence. Recently, support vector machines (SVMs) have been applied to a variety of such classification tasks with success comparable to other state-of-the-art methods. At the heart of the SVM training method is the kernel function, which calculates a 'similarity score' between two instances. SVM training is most successful when the kernel returns a high score for a pair of instances when they are members of the same class, and a low score when they are not. Past approaches have frequently used what we call a sequence identity kernel (SIK), which scores a pair of instances based only on the number of positions at which they share the same residue. This scoring method therefore ignores the varying degrees of physicochemical similarity between amino acids - information that is captured in a wide variety of substitution matrices. We demonstrate that use of a substitution matrix based kernel (SMK) significantly improves accuracy and correlation coefficient in prediction of residue solvent accessibility, when compared with the SIK. Current work is directed at testing whether this approach will similarly improve predictions for other protein structural or functional features (e.g. secondary structural elements, catalytic sites), through the use of appropriate substitution matrices for specific tasks.