Identification of SH2-peptide interactions using support vector machine.

Abstract

Src homology 2 (SH2) domains are structurally conserved protein domains, found in many intracellular signal-transducing proteins. Phosphorylation of tyrosine residues by tyrosine kinases is an important part of signal transduction. SH2 domains are the largest family of the peptide-recognition modules (PRMs) that recognize phosphotyrosine containing peptides. Hence, these domains have a vital role in cellular signaling. Around 120 SH2 domains have been identified in 110 human proteins and each SH2 domain binds with a specific subset of peptides. Therefore, peptide motif recognition by specific SH2 domains is important for understanding its biological function. Currently only a few programs have been published for the prediction of SH2-peptide interactions but most of them are based on position specific weight matrices (PWMs) which ignore modeling the dependencies between the amino acids. Furthermore, these tools either don’t model for all human SH2 domains or/and are not publically available. In the current study we are developing a machine learning approach for prediction of SH2-peptide interactions, which shall be made publically available. An ideal way to study protein-protein interaction using machine learning approach is to use high throughput data to build positive and negative datasets. We used microarray and peptide array library data for making positive and negative datasets. Our program selected important novel features and trained the data with those features. We built up separate models for each SH2 domains (based on the available data). We measured performance in terms of the AUC (area under the ROC curve), with 10 fold cross-validation, which is comparable to existing methods.