Evaluation of Different Biological Data and Computational Classification Methods for Use in Protein Interaction Prediction in Signaling Pathways in Arabidopsis Thaliana

Klein-Seetharaman, Judith[1]; Qi, Yanjun[2]; Gabor, Lisa[3]
[1]Department of Structural Biology; [2]School of Computer Science; [3]Department of Electrical and Computer Engineering
[1]University of Pittsburgh; [2]Carnegie Mellon University; [3]The George Washington University

ABSTRACT: A recent study conducted by this research group concluded that using the correct combination of classifiers and features, supervised machine learning could be used to make predictions regarding protein interactions based on direct and indirect biological datasets for yeast cells. We sought to repeat these results for *Arabidopsis thaliana*, a model organism for flowering plants, and defined “protein interactions” in three overlapping subdivided categories: (1) physical interaction, (2) co-complex relationship, and (3) pathway co-membership. To investigate systematically the utility of different data sources and the way the data is encoded as features for predicting each of these types of protein interactions, we assembled a large set of biological features and varied their encoding for use in each of the three prediction tasks. It is predicted that the importance of different features depends on the specific prediction task and the way they are encoded.