Designing a computational system to predict protein-protein interactions in *Arabidopsis thaliana*

Kamaldeep Singh[1,2], Judith Klein-Seetharaman[3], Yanjun Qi[4]

[1]Bioengineering & Bioinformatics Summer Institute, Dept. of Computational Biology, University of Pittsburgh, 15260
[2]School of Engineering and Computer Science, University of The Pacific, Stockton, CA 95211
[3]Department of Structural Biology, University of Pittsburgh School of Medicine, PA 15260
[4]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

Overview

Purpose:

To build a computational protein-protein interaction prediction system for *Arabidopsis thaliana*

Approaches:

- Integrated direct experimental methods with indirect biological datasets.
- Generated a list of known interactions and a list of non-interacting proteins.
- Designed feature sets used for prediction
- Utilized various classifiers (machine learning methods) to process the data

Results:

Results are based on the score values assigned to the feature sets by machine learning methods. The feature set values were further converted into to graphs and charts using the Weka machine learning software.[1] It was observed that the different computational methods were able to predict protein-protein interactions with a 75% accuracy.

Introduction

A protein's function and its activity are usually modulated by the proteins with which it interacts. Therefore, it is important to understand the mechanism by which proteins communicate and collaborate. However, predicting protein-protein interactions is one of the most challenging problem of the post-genomic era. High-throughput experimental approaches provide some data about protein interactions, but the data is fairly noisy and the results are incomplete and often yield high false-positive and false-negative rates[2]. Therefore, computational methods are often employed in addition to experimental methods in order to improve the success of protein interaction prediction. Here we use computational techniques to predict protein interactions in the model plant *Arabidopsis Thaliana*. Several genes in *A. Thaliana* are similar to those found in crop and medicinal plants therefore a complete proteome of Arabidopsis will be of great industrial use.

Method

There are a multitude of methods that can predict protein-protein interactions but each technique has its strengths and weaknesses. Experimental methods include the Two-Hybrid (Y2H)[2] screens and mass spectrometry techniques such as tandem affinity purification (TAP)[2], and high-throughput mass-spectrometric protein complex identification (HMS-PCI)[2]. These methods have limited screening and specificity capabilities allowing many protein interactions to go undetected. Hence, it was suggested that combining direct experimental data with indirect biological datasets (e.g. sequence data) can optimize the task of predicting protein - protein interactions.[3]

Arabidopsis Thaliana

Dataset Used:

- The Arabidopsis Information Resource (TAIR) to generate:
  - Positive set – pairs of interacting proteins
  - Negative set – noninteracting proteins

Features Used:

- Ortholog data (InParanoid program)
- Gene expression data (Microarray)

Protein1	Protein2	Label	Ortholog	MicroArray
AT1G12220	AT5G13160	1	0	88.518
AT2G01570	AT4G24210	1	0	379.701
AT4G35000	AT4G35450	1	1	101.24
AT3G20780	AT2G26990	1	1	573.883
AT1G26830	AT5G02820	1	0	141.174

Table 1: Features Used

The positive and negative sets were assigned values of 1 and 0, respectively. A pair from either set if present in the ortholog data received a value of 1, and 0 otherwise. For the microarray data the pairs were assigned unique values based on the absolute difference in average intensity.

Results

Weka, a machine learning software was used to analyze and visualize the feature set data. The figure below is of a histogram which contains a total of 3000 protein pars of which approximately 800 are identified as positive.

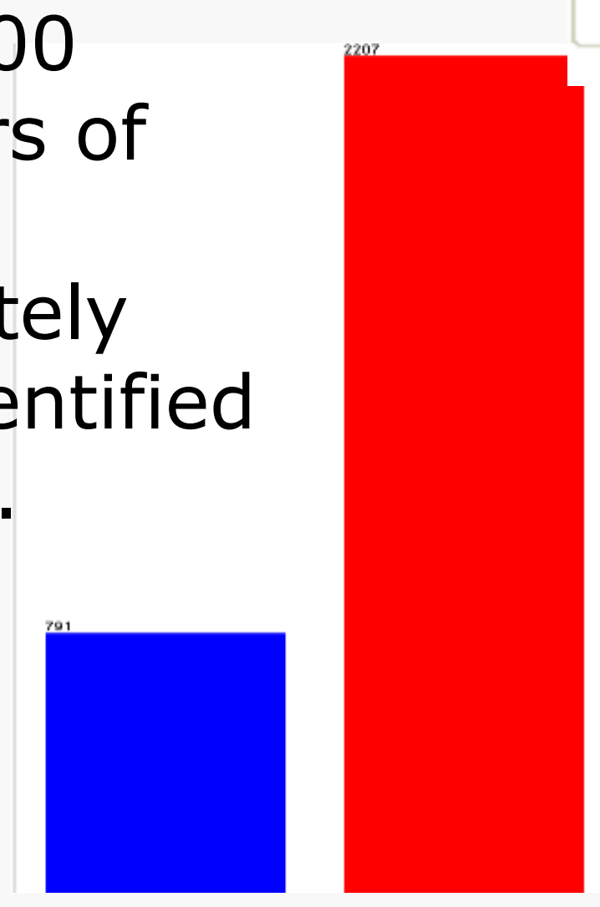

Figure1: Positive vs. Negative interactions

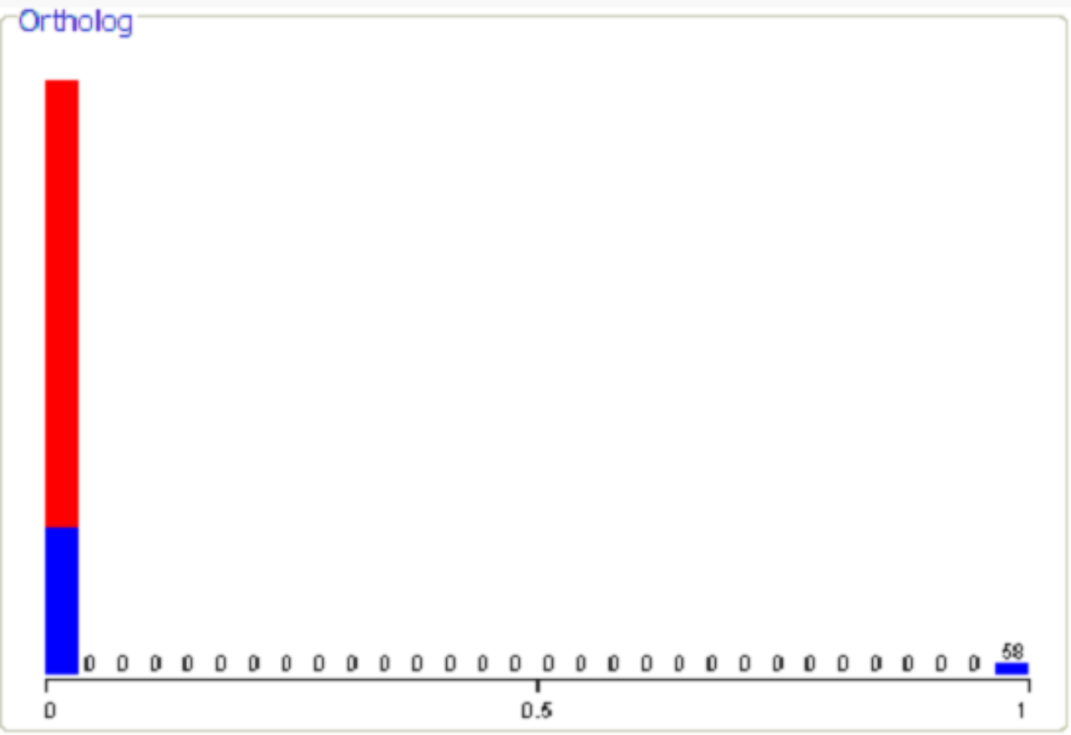

Figure2: Ortholog feature set

In figure above it can be seen that of the 800 positive pairs 58 of them are also present in the ortholog feature set. However, note that none of the pairs from the negative set are identified in the ortholog data.

Figure 3 contains the absolute difference of intensity values between each pair. The x-axis contains the average intensity values and from the y-axis it can be seen that the values for the positive set is approximately half the values of the negative set.

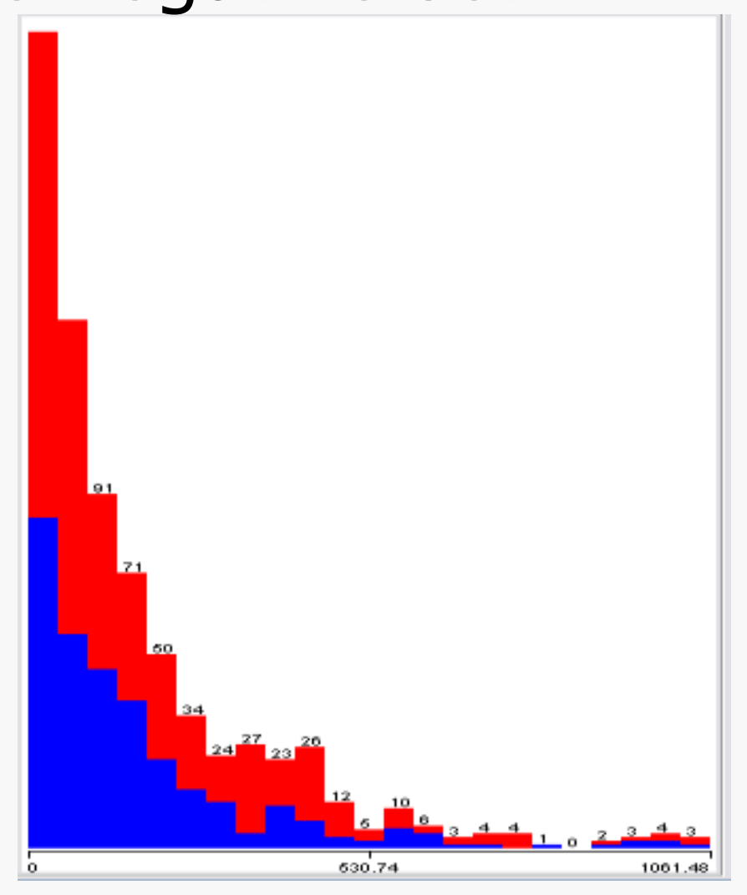

Figure3: MicroArray feature set

Conclusions

All of the various classifiers utilized generated the result value, 75 percent correctly predicted pairs, which indicates that even with just two features the classifiers have some ability to predict the protein interaction pairs in *A. thaliana*.

Classifier	# Correct Instances	Percent Correct
J48	2265	75.5504%
Random Forest	2265	75.5504%
RandomTree	2265	75.5504%
Logistic	2265	75.5504%
SMO	2265	75.5504%

Table 2: Correctly Predicted Interactions

However, predicting interactions using only two feature sets does not allow one to evaluate the different classification methods used for the prediction task. Therefore, in order to differentiate between the various classifiers additional feature sets are required.

Acknowledgements

The national BBSI program (http://bbsi.eeicom.com) is a joint initiative of the NIH-NIBIB and NSF-EEC, and the BBSI @ Pitt is supported by the National Science Foundation under Grant EEC-0234002.

Judith Klein-Seetharaman, Department of Structural Biology at the University of Pittsburgh

Yanjun Qi, Language Technologies Institute of the School of Computer Science at Carnegie Mellon University

Ankur Agarwal, BBSI, 2007

Carnegie Mellon

References

1. Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
2. Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins 2006;63:490-500.
3. Von Mering C, Krause r, SSnel B, Cornell M, Oliver S, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. Nature 2002;417:399-403.