Evaluation of Computational Methods in the Prediction of Protein Interactions in *Arabidopsis thaliana*

Lisa Gabor[1,2], Judith Klein-Seetharaman[3], Yanjun Qi[4]
[1]Bioengineering & Bioinformatics Summer Institute, Dept. of Computational Biology, University of Pittsburgh, 15260
[2]Department of Electrical and Computer Engineering, School of Engineering and Applied Science, The George Washington University, 20052
[3]Department of Structural Biology, University of Pittsburgh, 15260
[4]Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 15213

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.

To investigate systematically the utility of different data sources and the way the data is encoded as features for predicting these interactions, we assembled a large set of biological features and varied their encoding.

Introduction

The interactions that occur between two proteins are essential parts of biological systems. Through a combination of modern robotics, data processing and control software, liquid handling devices, and sensitive detectors, high-throughput methods allow a researcher to effectively conduct millions of biochemical, genetic, or pharmacological tests in a short period of time. Through this process one can rapidly identify active compounds, antibodies or genes which modulate a particular biomolecular pathway. The results of these experiments provide starting points for drug design and for understanding the interaction or role of a particular biochemical process in biology.

High-throughput methods can directly detect the set of proteins that interact in a variety of organisms, however the outcomes often render incomplete results and show a high propensity for false-positive and false-negative rates.

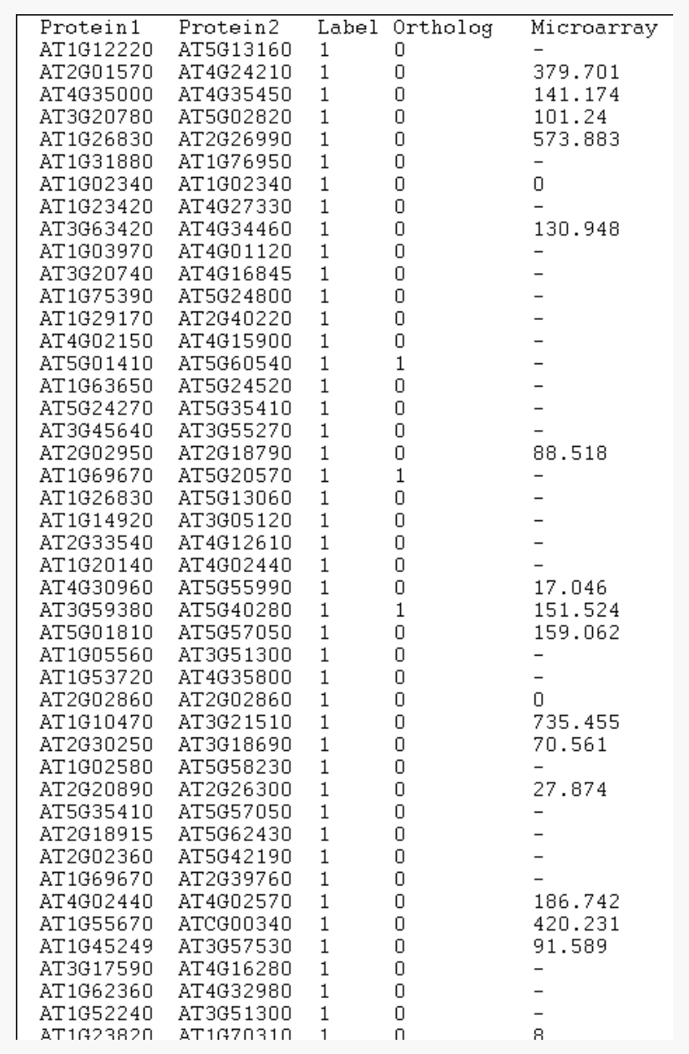
The findings of this preliminary research suggest that these methods and framework for distinguishing protein-protein direct, co-complex, and co-pathway interactions can be extended for organisms where little direct high-throughput information is available, for example, in humans or plants, such as *Arabidopsis thaliana*.

Method

The methods for this project are based upon a series of trials explained in detail in the paper by Qi, Bar-Joseph, and Klein-Seetharaman, which concluded that the Random Forest method of classification yielded results with the lowest propensity for false-positives and negatives. In order to make predictions about the interactions of protein pairs, publicly available information was used to "learn" the software.

The learning of the software was based on two main datasets. The positive set defined the set of proteins already known, through experimental methods, to interact. The negative set was generated by a random sample of all possible combinations of the master list of genes.

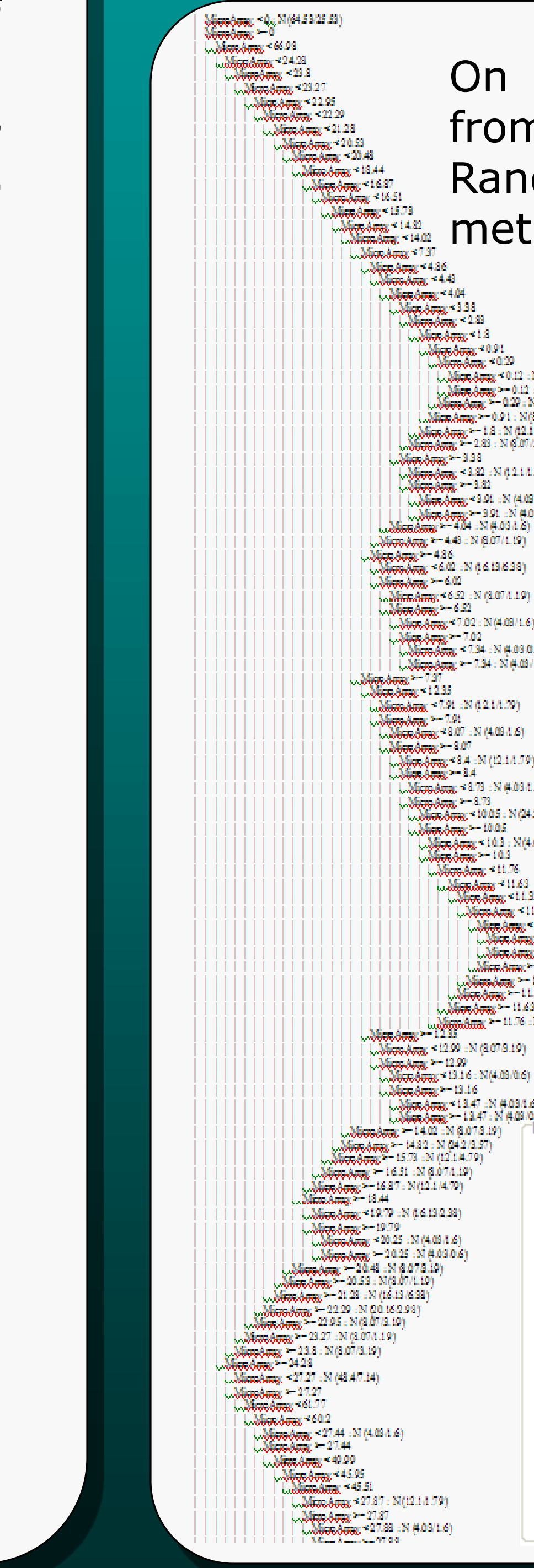
Arabidopsis thaliana


```
Protein1   Protein2   Label Ortholog  Microarray
AT1G12220  AT5G13160  1     0         -
AT2G01570  AT4G24210  1     0         379.701
AT4G35000  AT4G35450  1     0         141.174
AT3G20780  AT5G02820  1     0         101.24
AT1G26830  AT2G26990  1     0         573.883
AT1G31880  AT1G76950  1     0         -
AT1G02340  AT1G02340  1     0         0
AT1G23420  AT4G27330  1     0         -
AT3G63420  AT4G34460  1     0         130.948
AT1G03970  AT4G01120  1     0         -
AT3G20740  AT4G16845  1     0         -
AT1G75390  AT5G24800  1     0         -
AT1G29170  AT2G40220  1     0         -
AT4G02150  AT4G15900  1     0         -
AT5G01410  AT5G60540  1     1         -
AT1G63650  AT5G24520  1     0         -
AT5G24270  AT5G35410  1     0         -
AT3G45640  AT3G55270  1     0         -
AT2G02950  AT2G18790  1     0         88.518
AT1G69670  AT5G20570  1     1         -
AT1G26830  AT5G13060  1     0         -
AT1G14920  AT3G05120  1     0         -
AT2G33540  AT4G12610  1     0         -
AT1G20140  AT4G02440  1     0         -
AT4G30960  AT5G55990  1     0         17.046
AT3G59380  AT5G40280  1     1         151.524
AT5G01810  AT5G57050  1     0         159.062
AT1G05560  AT3G51300  1     0         -
AT1G53720  AT4G35800  1     0         -
AT2G02860  AT2G02860  1     0         0
AT1G10470  AT3G21510  1     0         735.455
AT2G30250  AT3G18690  1     0         70.561
AT1G02580  AT5G58230  1     0         -
AT2G20890  AT2G26300  1     0         27.874
AT5G35410  AT5G57050  1     0         -
AT2G18915  AT5G62430  1     0         -
AT2G02360  AT5G42190  1     0         -
AT1G69670  AT2G39760  1     0         -
AT4G02440  AT4G02570  1     0         186.742
AT1G55670  ATCG00340  1     0         420.231
AT1G45249  AT3G57530  1     0         91.589
AT3G17590  AT4G16280  1     0         -
AT1G62360  AT4G32980  1     0         -
AT1G52240  AT3G51300  1     0         -
AT1G23820  AT1G70310  1     0         8
```
Generated input file for Weka software

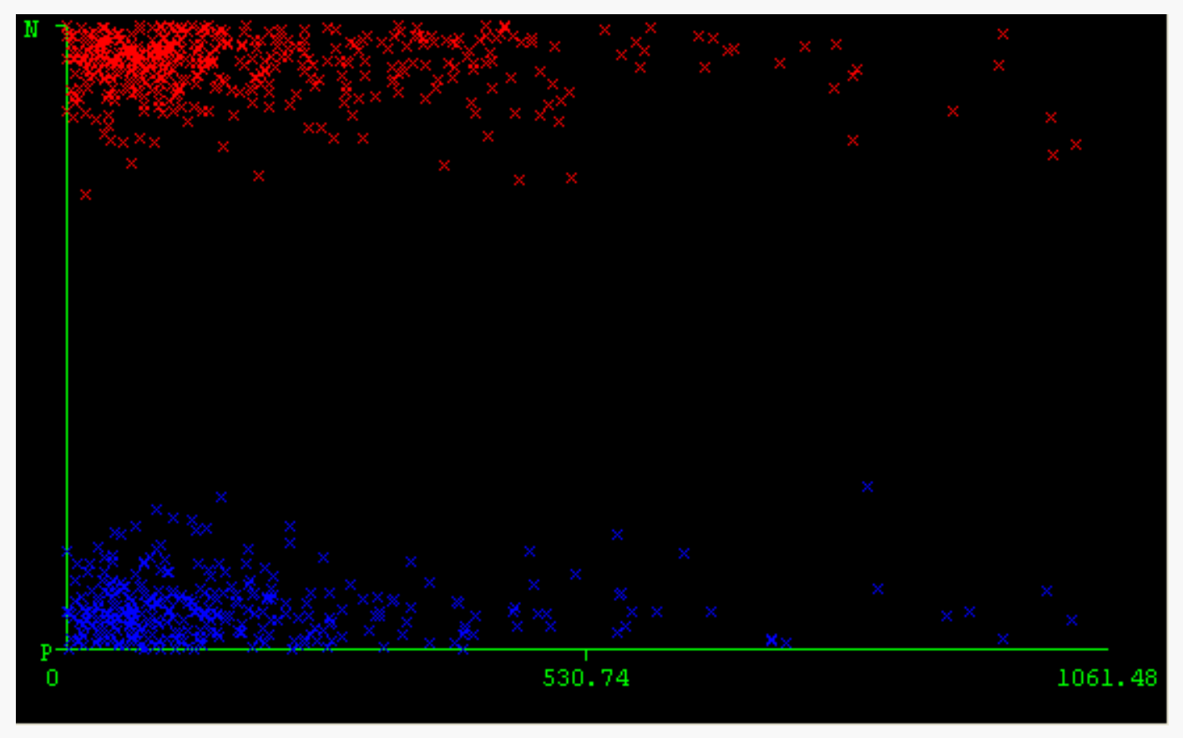
The feature sets were created using other information about the genes of the organism, including ortholog and gene expression data. Using these, comparisons between datasets were , in general, made to assign scores of one (1) or zero (0) to the pairs for interacting or non-interacting predictions, respectively. This is not always the case, however, as in the microarray data, real values of average intensity were assigned to the pairs.

The set of scores and a label column (one for positive set, zero for negative) were combined into an array that could then be used as input into the Weka software, essentially making a prediction about the regarding the probability of the interaction of the two proteins in the specified pair.

Results

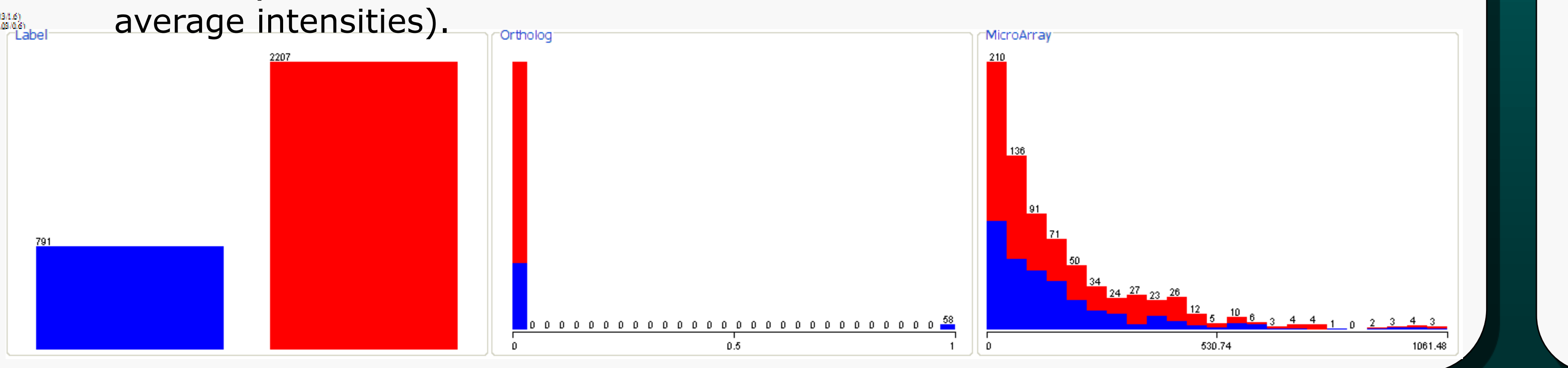


On the left is a sample from the output of the RandomTree classification method.



Shown above is a visualization of the microarray data where the blue "x"s represent the positive dataset and the red represent the negative. The x-axis is the absolute difference in average intensities (where gene expression data was available) of each protein in the given pair.

The three histograms below represent the three categories of data (from left to right): the label (positive or negative) and the two feature sets—the ortholog and microarray data. The label graph shows that the sample contained about 3000 protein pairs, approximately 800 of which were known interactions (positive). In the ortholog representation, only about 60 positive pairs received a score of one. The height of the bars on the Y axis of the microarray data on the right mean the number of pairs, the X axis shows the value range of the microarray feature (the absolute difference of the average intensities).



Preliminary Conclusions

The results at this stage are insufficient to make generalizations about the different classification methods available. As the project progresses, additional feature sets will be added and the advantages and disadvantages of each will become more apparent and distinct. For example:

- Because there is not a distinctive difference in the patterns of the distributions of the microarray visualization, this feature alone could not be used to make an interaction prediction.
- Each of the classifiers gave the same results regarding accuracy in predicting interacting pairs (74%). This is due to the fact that there are not enough feature sets to distinguish the training (see table below).

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%

The result value (75 percent corrected predicted) indicates that even with just two features, it is possible to predict the protein interaction pairs in AT. This partially proves the applicability of extending the method on this species. Adding additional features would definitely achieve better performance.

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

References

von Mering C, Krause R, Snel 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.

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.

Qi Y, Klein-Seetharaman J, Bar-Joseph Z. Random Forest similarity for protein-protein interaction prediction from multiple sources. Pacific Symp Biocomput 2005;10:531-542.

Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, Warfsmann J, Ruepp A. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res 2004;32(Database issue):D41-44.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003;302:449-453.

Breiman L. Random forests. Machine Learn 2001;45:5-32.