Designing a Computational System to Predict Protein-Protein Interactions in Arabidopsis Thaliana

Lisa Gabor, Kamaldeep Singh

Mentor: Judith Klein-Seetharaman

Yanjun Qi

Department of Structural Biology

University of Pittsburgh

Overview

- Introduction and Background
- Purpose
- Methods
- Results
- Conclusions
- Acknowledgements

Introduction

- Predicting protein-protein interactions is one of the most challenging problems of the post-genomic era
- High-throughput methods can be used but are noisy and often yield false-positive/negative results
- Computational techniques can be employed to identify interactions between proteins

Purpose

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

Methods

- High-throughput methods
  - Mass spectrometry and Yeast 2-Hybrid (Y2H), for example
  - Advantages and disadvantages
- Computational methods
  - Machine learning
  - Example

Methods

- Computational projects are based on experimental data available to the public
- Organism-specific databases provide downloadable files
  - InParanoid, NCBI, Gene Ontology (GO)
  - The Arabidopsis Information Resource (TAIR)

Methods

- TAIR is the database of choice for all *A. thaliana* information
  - Leader of *A. thaliana* research and funding
  - “Gold Standard” dataset
- ftp provides downloadable files
  - Files collected from sources like GO, NCBI, private research, etc.
  - Our project…

Methods

- These datasets could be used to make predictions about protein interactions
  - Machine learning
- Positive set—pairs of interacting proteins determined using experimental methods
- Negative set—randomly generated from the master list of all *A. thaliana* genes

Methods

- Feature sets
  - Used to generate arrays of “scores” that will eventually be combined to make a prediction based on some threshold value
  - For example: orthologs, microarray data

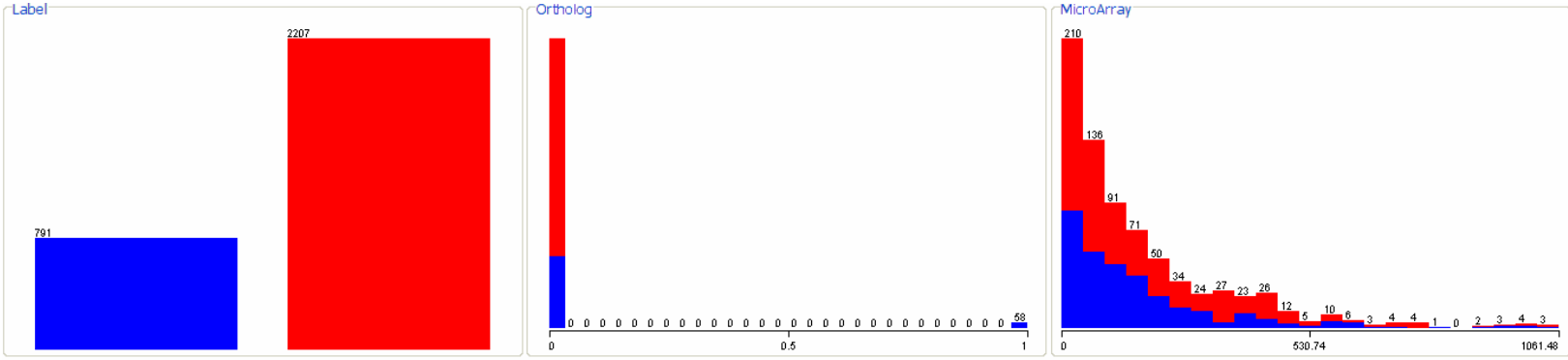
Results

- Results are determined from the score values assigned to each feature set
- Results are not facts!

Results

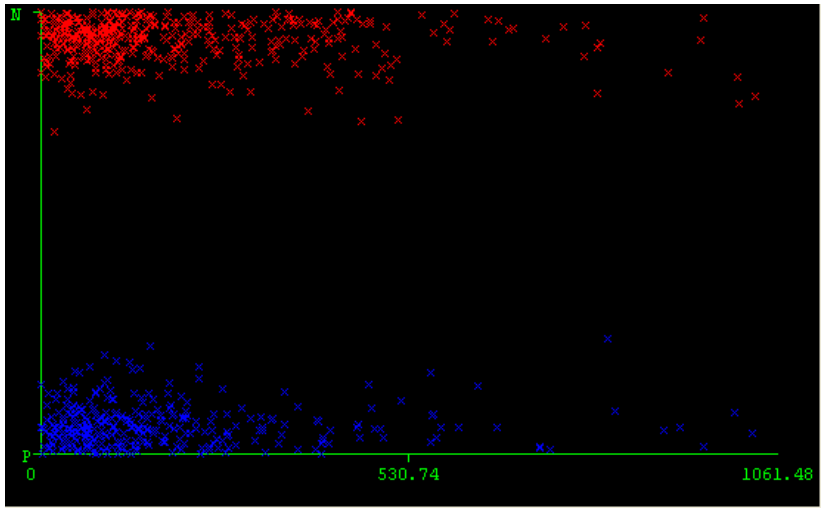
The three categories of data (from left to right):

- Label (positive or negative)
  - shows that the sample contained about 3000 protein pairs, approximately 800 of which were known interactions (positive)
- Two feature sets—the ortholog and microarray data



Results

- Visualization of the microarray data
  - Blue “x”s represent the positive dataset
  - 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.



Conclusions

- The results at this stage are insufficient to make generalizations about classification methods
  - For example:

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%

  - Distinctions will be possible when there are more feature sets (ie: microarray data)
- With the addition of feature sets, conclusions will be possible regarding the classification methods as well as regarding protein interaction predictions

Acknowledgements

Ankur Agarwal

Acknowledgements

- Judith Klein-Seetharaman
  - Department of Structural Biology, University of Pittsburgh, PA
- Yanjun Qi
  - Language Technologies Institute, School of Computer Science, Carnegie Mellon University, PA