


Evaluation of Different Biological Data and Computational Classification Methods for Use in Protein Interaction Prediction

Yanjun Qi, Ziv Bar-Joseph,
Judith Klein Seetharaman
Carnegie Mellon University &
UPMC, Pittsburgh, PA



Outline

- Background
- Goals of high-throughput computing
 - Goals of this particular project
- Materials and Methods
- Results
- Conclusions
- Implications, Applications, and the Future

Background: Terms and Definitions

- High-throughput computing
- Feature
- Protein-Protein Interaction
 - Physical interaction
 - Co-complex relationship
 - Co-pathway relationship
- In silico

Background

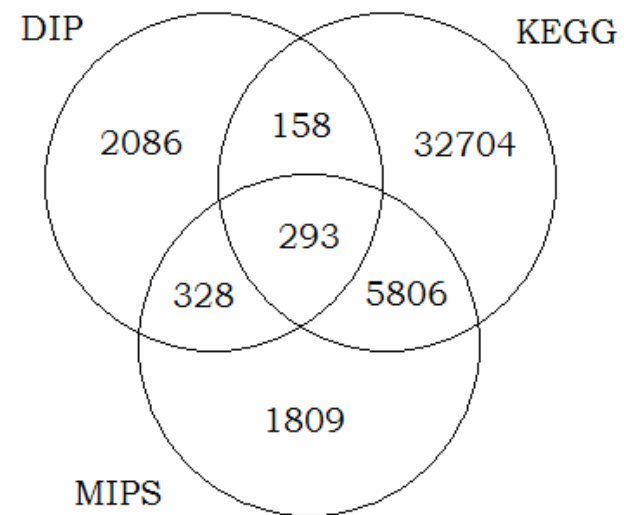
- Protein-protein interactions (PPI) within a cell are highly useful
- High-throughput methods can directly detect the set of proteins that interact in yeast
 - Two-hybrid screens (Y2H)
 - Mass spectrometry methods
- These methods have can have high rates of false- positives/negatives

Background

- It has been shown that some indirect biological datasets contain information on PPI
- Based on these observations, it has been suggested that direct data on PPI can be combined with indirect data in a supervised learning framework
 - Machine learning
- Goal of the project

Materials

- All methods applied require a training and a test set for machine learning
 - “Gold standard set”
- Positive examples
 - DIP
 - MIPS
 - KEGG



Methods

- Preliminary performance and parameter optimization evaluations were performed
- Six classifiers:
 - Logistic regression
 - Random forest (RF)
 - Naïve Bayes
 - Decision tree
 - Support vector machine
 - RF similarity-based k-Nearest-Neighbor
- Each method differs mainly in terms of classifiers, feature sets, and their encodings and gold-standard datasets used

Results and Conclusions

- 30,000 yeast protein-protein pairs were selected to learn the decision model and another test set of the same size was used to evaluate the performance of the trained classifier in the context of the data set and feature encoding used
- The three prediction tasks yielded different success rates for all classifiers, and co-complex prediction appeared to be an easier task than the other two
- The RF classifier consistently ranked as one of the two best methods for all combinations of the features

Applications and the Future

- These methods and framework for distinguishing protein-protein direct, co-complex, and co-pathway interactions can be extended
 - Organisms where little direct high-throughput information is available (ie: humans)