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

Yanjun Qi[1] and Judith Klein-Seetharaman[2]
[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
[2]Department of Structural Biology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15260
Lisa Gabor[3]
[3]School of Engineering and Applied Science, George Washington University, Washington, District of Columbia 20052

Introduction and Background

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 yeast, however the outcomes often render incomplete results and show a high propensity for false-positive and false-negative rates[1].

The task of prediction protein-protein interactions can be segmented into three overlapping categories: (1) whether or not the proteins physically interact, (2) whether or not they are parts of the same complex, and (3) whether or not the two proteins are members of the same pathway[2]. In order to determine the usefulness of the different methods available for predicting protein interactions, a large set of biological features were assembled and their coding was varied for use in each of the three prediction tasks. Six different classifiers were used to evaluate the effectiveness and accuracy in predicting interactions: Random Forest (RF), RF

similarity-based k-Nearest-Neighbor, Naïve Bayes, Decision Tree, Logistic Regression, and Support Vector Machine[2].

Briefly, the logistic regression (LR) has been used to estimate wehther or not a pair of proteins have direct physical interactions using high-throughput features. A recently proposed method (kRF) combines the Random Forest and kNN with a summary feature set using the Database of Interacting Proteins (DIP). The Naïve Bayes (NB) classifier was used to predict co-complex relationships using a summary feature set from the Gene Ontology database, essentiality data, and direct high-throughput interaction experiments. Each method differs mainly in terms of classifiers, feature sets, and their encodings and gold-standart datasets used[2].

The three prediction tasts mentioned previously yielded different success rates for all classifiers, and co-complex prediction appeared to be an easier task than the other two. The study concluded that the RF classifier consistently ranked as one of the two best methods for all combinations of the features. The RF classifier was therefore used to study the three different datasets, including the Database of Interacting Proteins (DIP) for predicting direct physical interactions between protein pairs[3], the Munich Information Center for Protein Sequences (MIPS) for co-complex relationships[4], and the Kyoto Encyclopedia of Genes and Genomes (KEGG) for inferring pathway networks[5].

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[2].

Methods

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 and as such are summarized as follows. Three gold standard datasets were used (as discussed previously): (1) DIPS for predicting direct physical protein-protein interactions, (2) MIPS for predicting co-complex protein pairs, and (3) the KEGG pathway database for predicting co-pathway relationships (it is intuitive that these datasets should overlap). The gold standart set was used to obtain the postive (scores indicating that the proteins *do* interact) examples. An approach detailed by Zhang et al[6]. was followed to identify negative examples. The final gold standard set for these contained one positive interaction for every 600 negative interaction pairs[1].

Performace comparisons on each of the six different classifiers were based on several training and testing procedures. Paramater optimization was conducted for each combination using separate training and test datasets. 30,000 yeast protein-protein pairs were randomly 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. This was repeated with a random sample 25 times for each case and average values were reported in the paper by Qi, Bar-Joseph, and Klein-Seetharaman.

Goals of Research

As explained in the introduction, the Random Forest (RF) classification method was used to determine the variety of ways in which proteins interact in yeast. Based on the evaluation of the different combinations of features and datasets, several conclusions about the prediction tasks and different classifiers were made. It was found that (1) the co-complex relationship was

easiest to predict, which is most likely attributed to the fact that co-complex prediction is an intermediate task between being in the same general pathway and direct physical interactions. (2) The RF classifier performs most favorably overall amongst the six different classifiers across the three prediction tasks. This is related to the fact that it can easily combine different types of data, that RF does not assume each dataset to be completely independent (which is advantageous because many of the datasets are expected to be correlated), and RH is particularly robust against seemingly unexplained and/or missing values and could be used to estimate those missing values[7].

This will be repeated in an attempt to improve the scores by using a more updated dataset. This method will then be applied to the set of human proteins involved in signaling pathways and processes. Interaction prediction scores for all protein-protein pairs for both of these will be made. When the protein-protein interaction scores are collected for the human signaling pathways, it is hoped that these scores will be used to find the potential signaling pathways in humans.

References

[1]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.

[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]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.

[4]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.

[5]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.

[6]Zhang L, Wong S, King OD, Roth FP. Predicting co-complexed protein pairs using genomic and proteiomic data integration. BMC Bioinformatics 2004;5:38.

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