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

Kamaldeep Singh, School of Engineering and Computer Science, University of the Pacific, Stockton, CA 95211
Yanjun, Qi, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
Mentor: Dr. Judith Klein-Seetharaman, Department of Structural Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15260

Introduction

Proteins are unique amongst organic compounds in supporting every reaction occurring in biological systems. They can be thought of as molecular machines within cells performing functions specified by the information encoded in genes. Proteins participate in diverse biological functions ranging from DNA replication, cell division, structural support, and signal transduction, among others. Over the last decade enormous amount of sequencing information has been generated via the Human Genome Project. However, the genome is merely a parts list which nature uses to assemble proteins. It is the proteins that are responsible for the tangible features representing life. In order to recognize and understand the diverse roles of proteins it is first important to understand the mechanism by which the linear sequences provided by the genome project are converted into three-dimensional proteins. Determining protein structure is beneficial as it yields information about its dynamics and ultimately function. Further, it is even more important to understand the mechanism by which a set of proteins communicate and collaborate toward a common cellular function. Identifying interactions between proteins will allow a broader outlook and a more comprehensive understanding of the biological processes occurring in cells. Also, accurate predictions of protein interaction can be useful in making implications about the functions of unknown proteins according to their interacting partners.[1]

Protein-protein interaction can be categorized into three subtasks: (1) physical interaction, (2) interaction amongst proteins in the same complex, and (3) proteins involved in pathway networks.[1] Direct high-throughput experimental methods such as two-hybrid screens and mass spectrometry can predict the set of interactions amongst proteins. [1] However, these methods often yield false-positive and false-negative results. [1] Although, it was shown that indirect datasets can also be used in protein interaction prediction and thus integrating direct high-throughput methods with indirect biological data sources can lead to more accurate results. [2,3]

Predicting protein-protein interactions is one of the most challenging yet motivating problems of the post-genomic era. Hence, researchers are shifting their focus from the study of a single protein to that of the entire proteome. Therefore it becomes necessary to systematically evaluate the different methods that are utilized in predicting protein-protein interactions. Ultimately the method(s) that perform the best in terms of accurately predicting protein-protein interaction can be used to model the proteome of a given organism.

Methods

Recent studies conducted by Yanjun Qi, Ziv Bar-Josepth, and Judith Klein-Seetharaman employed a combination of three different data sets, two feature encoding styles, and six different methods.[1] Each data set was used to predict a particular protein-protein interaction subtask. Database of Interacting Proteins (DIP) was used to predict the physical interaction amongst proteins.[1] For predicting proteins involved in the same complex the Munich Information Center for Protein Sequences (MIPS) complex catalog was used.[1] Lastly, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used to infer protein interaction in pathway networks.[1]

The way the data is encoded was also varied. Two different types of feature encoding; “Summary” and “Detailed” were considered.[1] “Summary” encoding involved similar experiments that were combined to produce a single value.[1] On the contrary, in the “detailed” encoding each experiment was represented separately to provide multiple values.[1]

Additionally, six different classifiers, Random Forest (RF), RF similarity-based k-Nearest-Neighbor (kRF), Naïve Bayes (NB), Decision Tree (DT), Logistic regression (LR), and Support Vector Maching (SVM) were compared in order to highlight the key differences in their performance.[1] These classifiers are all models used to estimate interactions occurring amongst proteins but they differ from one another in that they each take a unique approach towards estimating specific types of interactions (physical, co-complex, co-pathway network). Therefore, different classifiers were considered for each protein prediction subtask and independent of the prediction subtask it was observed that the RF classifier performed the best in terms of accurately predicting interactions. [1]

Goals

It is the goal of this project to build a computational protein-protein interaction prediction system for *Arabidopsis Thaliana*. As mentioned above the Random Forest (RF) classifier has been shown to be the method of choice and therefore this method will be utilized in predicting potential protein-protein interactions for this plant. Several genes in *Arabidopsis Thaliana* are common with crop and medicinal plants therefore a complete proteome of Arabidopsis will be of great industrial use.

The project of protein-protein interactions can be further extended to humans. This will be a major step towards characterizing the interaction of all the proteins. Predicting protein-protein interactions will enhance the current understanding of diseases and will pave the way for novel therapeutic approaches.

[1] 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.

2 Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA, 2001; 10:4569-4574.

3 Utez P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerecisiae*. Nature 2000; 403:623-627.