

GRAPH-BASED PROTEIN-PROTEIN INTERACTION PREDICTION IN
SACCHAROMYCES CEREVISIAE

by

MARTIN SAMUEL RAO PARADESI

B.Tech, Jawaharlal Nehru Technological University, 2005

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Computing and Information Sciences
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2008

Approved by:

Co-Major Professor
William H. Hsu

Approved by:

Co-Major Professor
Doina Caragea

Abstract

The term ‘protein-protein interaction (PPI)’ refers to the study of associations between proteins as manifested through biochemical processes such as formation of structures, signal transduction, transport, and phosphorylation. PPI play an important role in the study of biological processes. Many PPI have been discovered over the years and several databases have been created to store the information about these interactions. von Mering (2002) states that about 80,000 interactions between yeast proteins are currently available from various high-throughput interaction detection methods. Determining PPI using high-throughput methods is not only expensive and time-consuming, but also generates a high number of false positives and false negatives. Therefore, there is a need for computational approaches that can help in the process of identifying real protein interactions.

Several methods have been designed to address the task of predicting protein-protein interactions using machine learning. Most of them use features extracted from protein sequences (e.g., amino acids composition) or associated with protein sequences directly (e.g., GO annotation). Others use relational and structural features extracted from the PPI network, along with the features related to the protein sequence. When using the PPI network to design features, several node and topological features can be extracted directly from the associated graph.

In this thesis, important graph features of a protein interaction network that help in predicting protein interactions are identified. Two previously published datasets are used in this study. A third dataset has been created by combining three PPI databases. Several classifiers are applied on the graph attributes extracted from protein interaction networks of these three datasets. A detailed study has been performed in this present work to determine if graph attributes extracted from a protein interaction network are more predictive than biological features of protein interactions. The results indicate that the performance criteria (such as Sensitivity, Specificity and AUC score) improve when graph features are combined with biological features.

Table of Contents

List of Figures	v
List of Tables	vi
Acknowledgements	vii
Dedication.....	viii
CHAPTER 1 - Introduction.....	1
1.1 Protein-protein interactions.....	1
1.1.1 Mass spectrometry based analysis of protein complexes.....	1
1.1.2 Protein microarrays.....	2
1.1.3 Yeast two hybrid (Y2H).....	2
1.2 Protein interaction databases.....	2
1.2.1 The Biomolecular Interaction Network Database (BIND)	2
1.2.2 The Database of Interacting Proteins (DIP)	3
1.2.3 IntAct	3
1.2.4 A Molecular INTERaction database (MINT)	3
1.2.5 The Munich Information Center for Protein Sequences (MIPS).....	4
1.3 Introduction to Machine Learning.....	4
1.3.1 Decision tree.....	5
1.3.2 Random Forest.....	6
1.3.3 Naïve Bayes.....	7
1.3.4 Support Vector Machine	7
1.3.5 K-Nearest Neighbors	7
1.3.6 Bagging	8
1.3.7 REPTree	8
1.4 Motivation.....	8
CHAPTER 2 - Background and significance	9
2.1 Qi, Bar-Joseph & Klein-Seetharaman	9
2.2. Licamele & Getoor.....	11
2.3 Paradesi, Caragea & Hsu	11

2.4 Chen & Liu	13
2.5 Advantages of graph-based PPI prediction.....	14
CHAPTER 3 - Experiments	14
3.1 Protein domain information	15
3.2 Biological features.....	15
3.3 Combining PPI databases	16
3.3.1 Protein domain information	16
3.3.2 Biological features	17
3.4 Features.....	17
CHAPTER 4 - Results.....	18
4.1 Protein domain information	19
4.2 Biological features.....	21
4.3 Combining PPI databases	23
4.3.1 Protein domain information	23
4.3.2 Biological features	24
CHAPTER 5 - Conclusion and Future Work.....	25
5.1 Conclusion	25
5.2 Future Work.....	25
5.2.1 Increase in quality and quantity of data	25
5.2.2 Improvement in classification algorithms.....	26
5.2.3 Use of protein interaction network analysis tools	26
5.2.4 Development of different approaches.....	26
Bibliography	28
Appendix A - Variance of the results across different experiments	35
Protein domain information.....	35
Biological features	37
Combining PPI databases	38
Protein domain information	38
Biological features.....	39

List of Figures

Figure 2-1 Comparison of results by Licamele & Getoor (2006) and Paradesi <i>et al.</i> (2007)	12
Figure 2-2 Comparison of results by Qi <i>et al.</i> (2006) and Paradesi <i>et al.</i> (2007).....	13
Figure 3-1 Graph features.....	18

List of Tables

Table 4-1 Results obtained from experiments using Chen & Liu (2005) dataset (5-fold).....	20
Table 4-2 Results obtained from experiments using Qi <i>et al.</i> (2006) dataset (5-fold)	22
Table 4-3 Results obtained from experiments using the Intersection dataset & Chen & Liu (2005) dataset (5-fold)	23
Table 4-4 Results obtained from experiments using the Intersection dataset & Qi <i>et al.</i> (2006) dataset (5-fold)	24
Table A-1 Variance of results obtained from experiments using Chen & Liu (2005) dataset (5-fold)	36
Table A-2 Variance of results obtained from experiments using Qi <i>et al.</i> (2006) dataset (5-fold)	37
Table A-3 Variance of results obtained from experiments using the Intersection dataset & Chen & Liu (2005) dataset (5-fold).....	38
Table A-4 Variance of results obtained from experiments using the Intersection dataset & Qi <i>et al.</i> (2006) dataset (5-fold)	39

Acknowledgements

I thank my co-advisors Dr. William H. Hsu and Dr. Doina Caragea for guiding me during my Master's program. I am grateful to them for their suggestions and comments. I also appreciate their help in guiding my research. I also thank Dr. Susan J. Brown for agreeing to serve on my Master's committee. I am grateful to Dr. Ruth Welti for funding me as a Graduate Research Assistant during my Master's program.

I thank my mother, father and sister for their love, support and encouragement. Above all, I thank my God for His grace, mercy, faithfulness and love. Last but not the least, I thank all my friends in the city of Manhattan, KS.

Dedication

Dedicated to my mother, Dr. Mary Prathibha Gollapalli, my father, Mr. Suresh Rao Paradesi and my sister Ms. Sharon Myrtle Paradesi.

CHAPTER 1 - Introduction

1.1 Protein-protein interactions

Proteins are groups of amino acids linked together by peptide bonds. They play a vital role in organisms and participate in many processes within cells. Proteins generally have at least one biological function. Protein-protein interactions (PPI) are the associations between proteins. Protein interactions are important in many biological processes. In my thesis, I investigate the protein interactions in *Saccharomyces cerevisiae* for two reasons. First, *Saccharomyces cerevisiae* has been firmly established as the most investigated eukaryotic organism (Mewes *et al.*, 2002). Second, *Saccharomyces cerevisiae* is a model system relevant to human biology (Gavin *et al.*, 2002).

There are several biological techniques to detect protein interactions in the yeast organism. Qi (2008) states that there are several *in vivo* methods and *in vitro* methods for identifying PPI. *in vivo* refers to a reaction that takes place inside an organism, while *in vitro* means performing a given experiment in a controlled environment outside of a living organism. Some experiments can identify interactions on a small scale, while others detect interactions on a large scale (also known as high-throughput methods of detecting protein-protein interactions). There are many experiments that provide physical interactions among proteins (at either binary or complex level) while few experiments provide functional associations among proteins. A few high-throughput methods that identify physical protein-protein interactions are listed below:

1.1.1 Mass spectrometry based analysis of protein complexes

Gavin *et al.* (2002) and Ho *et al.* (2002) develop a technique to identify protein interactions in *Saccharomyces cerevisiae* on a large scale. Qi (2008) states that this method involves the following steps:

1. A tag is attached to a target protein in order to capture bait proteins.
2. Bait proteins are systematically precipitated.
3. Purified protein complexes are separated according to mass.
4. Proteins are detected by mass spectrometry techniques.
5. Database-search algorithms are used to identify proteins.

1.1.2 Protein microarrays

MacBeath & Schreiber (2000) developed this technique to detect protein-protein interactions on a large scale. They state that this technique consists of three steps:

1. A high-precision contact-printing robot is used to deliver nanoliter volumes of protein samples to glass microscopic slides.
2. The proteins are immobilized because they are covalently attached to the slides but they still retain their functional properties on the slides.
3. The slides are probed with other proteins and the interactions are then identified.

1.1.3 Yeast two hybrid (Y2H)

Originally developed by Fields & Song (1989), this technique is one of the most widely used techniques for identifying protein-protein interactions. According to Shoemaker & Panchenko (2007), the Y2H method is based on the fact that many eukaryotic transcription activators have at least two distinct domains: one that directs binding to a promoter DNA sequence (BD) and another that activates transcription (AD). There are two Y2H approaches as described by Shoemaker & Panchenko (2007):

1. In the matrix approach, a matrix of prey clones is created where each clone expresses a particular prey protein in one well of a plate. Each bait strain is then mated with an array of prey strains. The diploids where two chimeric proteins interact are selected based on the expression of a reporter gene and the position on the plate.
2. In the library approach, each bait is screened against either an undefined prey library containing random cDNA fragments or open reading frames (ORFs). Diploid positives are selected based on their ability to grow on specific substrates. The interacting proteins are determined by DNA sequencing.

1.2 Protein interaction databases

1.2.1 The Biomolecular Interaction Network Database (BIND)

BIND (Bader *et al.*, 2003) stores information about interactions, complexes and pathways. It also contains a number of large scale interaction and complex mapping experiments using yeast two hybrid, mass spectrometry, genetic interactions and phage display. The group that maintains BIND has also developed a graphical analysis tool that provides users an

understanding of functional domains in protein interactions. They have also developed a clustering tool that allows users to divide the protein interaction network into specific regions of interest. BIND assumes that interactions can occur between two biological ‘objects’, which could be protein, RNA, DNA, molecular complex, small molecule, photon (light) or gene.

1.2.2 The Database of Interacting Proteins (DIP)

DIP (Salwinski *et al.*, 2004) is a database containing 18,343 interactions between 4,923 proteins validated from 23,366 experiments of the *Saccharomyces cerevisiae* organism. A few of the experiments from which they validate protein interactions are co-immunoprecipitation, yeast two-hybrid and in vitro binding assays. The group that maintains DIP has developed several quality assessment methods and uses them to identify the most reliable subset of the interactions that are inferred from high-throughput experiments. They also provide an online implementation of their evaluation methods that can be used to evaluate the reliability of new experimental and predicted interactions.

1.2.3 IntAct

IntAct (Kerrien *et al.*, 2007) contains data such as experimental methods, conditions and interacting domains that is extracted entirely from publications and is manually annotated by curators. It also formalizes the data by using a comprehensive set of controlled vocabularies in order to ensure data integrity. It is probably the only database that contains negative examples of protein interactions, i.e. they identify that two proteins do not interact. The database contains 169,792 interactions between 63,427 proteins. These interactions were obtained from 8,477 experiments that were performed on several organisms. The web site provides tools allowing users to search, visualize and download data from the repository.

1.2.4 A Molecular INTERaction database (MINT)

MINT (Chatr-aryamontri *et al.*, 2007) stores molecular interaction data extracted from several publications. Most of its curation work is focused on physical interactions, direct interactions and colocalizations between proteins. Genetic or computationally inferred interactions are not included in the database. It contains 42,044 interactions between 5,256 proteins of the *Saccharomyces cerevisiae* organism. An online graph visualization and editing

tool called "MINT Viewer" is available that allows users to view the interaction network and delete edges that are not of interest to the user.

1.2.5 The Munich Information Center for Protein Sequences (MIPS)

MIPS (Mewes *et al.*, 2002) provides information on Open Reading Frames (ORFs), RNA genes and other genetic elements. The research group that maintains MIPS has also applied techniques such as gene disruption in conjunction with powerful expression analysis and two-hybrid techniques as part of a systematic functional genome analysis. These methods generate information on how proteins cooperate in complexes, pathways and cellular networks. In addition, detailed information on transcription factors and their binding sites, transport proteins and metabolic pathways are being included or interlinked to the core data. The database also provides information on the molecular structure and the functional network of the yeast genome.

1.3 Introduction to Machine Learning

Machine learning algorithms (Mitchell, 1997) offer some of the most cost-effective approaches to automated knowledge discovery and data mining (discovery of features, correlations, and other complex relationships and hypotheses that describe potentially interesting regularities) from large data sets. In particular, machine learning algorithms have proven to be very successful for many bioinformatics problems, including protein-protein interaction prediction. The field of machine learning has developed terminology that is somewhat different from typical biological use. Here are some useful terms and definitions (Mitchell, 1997). A machine learning system is specified by the following components:

- A learner: An algorithm that can use experience to improve performance of some task. In our context, this is an algorithm that can predict if two given proteins interact or not.
- A task: In our case, the task is to predict protein interactions when provided with information about the PPI network.
- An experience source: For the PPI task, the source is a collection of the PPI databases that contain information about the protein interaction network.
- Background knowledge: It is the information the learner has about the task before the learning process. In our case, it is the list of known protein interactions that have been detected through high-throughput methods.

- Performance criteria: measures of the quality of the learning output in terms of accuracy, precision, efficiency, etc. In protein interaction prediction problem, performance is often measured in terms of the following criteria, where different importance may be placed on optimizing each probability:
 - sensitivity (Altman & Bland, 1994) : the probability of correctly predicting that a protein pair interacts,
 - specificity (Altman & Bland, 1994): the probability of correctly determining that there is no interaction between a protein pair,
 - Receiver Operating Characteristic (ROC) curve (Fawcett, 2004): a plot of the true positive rate versus the false positive rate, and
 - The area under the ROC curve (AUC) score.

In the problem of protein interaction prediction, we will focus on machine learning systems for classification tasks. These are tasks where the learner is provided with experience in the form of labeled examples (a.k.a., training data set or data source) and is asked to classify new unlabeled examples in one of several possible classes. In our case, the training examples are information about protein interactions. The variables used to encode an example are called attributes or features. The class label of an example is a special attribute representing the class to which that particular example belongs. The class label in protein interaction prediction problem is whether a protein pair interacts or not. The output of a learning algorithm is often termed a classifier, when the task considered is a classification task. Several strategies can be used to estimate the true error of a classifier. The simplest one is to divide the labeled data into a training set and a test set (a.k.a., validation set). The classifier is learned from the training set and its error is estimated using the test set. More commonly, the error is estimated by using a method called cross-validation. To use this method, the labeled data is divided into k folds. A classifier is learned from a training set consisting of $(k-1)$ folds and tested on the remaining k^{th} fold. The estimate for the true error is obtained by taking the average of the error of the k possible classifiers learned by leaving out one fold at a time.

1.3.1 Decision tree

Decision tree algorithms (Quinlan, 1986; Breiman *et al.*, 1984) are among some of the most widely used machine learning algorithms for building pattern classifiers from data. Their

popularity is due in part to their ability to: select from all attributes used to describe the data, a subset of attributes that are relevant for classification; identify complex predictive relations among attributes; and produce classifiers that are easy to comprehend for humans. The ID3 (Iterative Dichotomizer 3) algorithm proposed by Quinlan (1986) and its more recent variants such as C4.5 (Quinlan, 1993) represents a widely used family of decision tree learning algorithms. The ID3 algorithm searches in a greedy fashion, for attributes that yield the maximum amount of information for determining the class membership of instances in a training set D of labeled instances. The result is a decision tree that correctly assigns each instance in D to its respective class. The construction of the decision tree is accomplished by recursively partitioning D into subsets based on values of the chosen attribute until each resulting subset has instances that belong to exactly one of the m classes. The selection of an attribute at each stage of construction of the decision tree maximizes the estimated expected information gained from knowing the value of the attribute in question. C4.5 (Quinlan, 1993) is the most popular variant of the ID3 algorithm that has been implemented as the J48 classifier in WEKA (Waikato Environment for Knowledge Analysis - Witten, 2005), a popular machine learning toolkit. Some of the improvements that C4.5 has made over ID3 algorithm are: dealing with missing data, pruning the tree after creation and dealing with attributes of different costs.

1.3.2 Random Forest

Random Forest (Breiman, 2001) is known to produce highly accurate results in many problems. The algorithm involves the construction of multiple trees from the data and the trees vote for the class. Random Forest then chooses the class with the maximum number of votes. The method of constructing each tree is described by (Breiman, 2001) in the following steps:

1. If there are N examples in the training set, the tree will be built by sampling N examples at random with replacement,
2. If there are M input variables, a small subset of these examples m is chosen at each node to find the best split of the data at that node, and
3. There is no pruning of the trees that are constructed at each stage.

1.3.3 Naïve Bayes

Naïve Bayes is a highly practical learning algorithm (Mitchell, 1997), comparable to more powerful algorithms such as decision trees or neural networks in terms of performance in some domains. In the Naïve Bayes framework, each example x is described by a conjunction of attribute values, i.e. $x = \langle a_1, \dots, a_n \rangle$. The class label of an example can take any value from a finite set $C = \{c_1, \dots, c_m\}$. We assume that the attribute values are conditionally independent given the class label. A training set of labeled examples, $D = \{\langle x_1, y_1 \rangle, \dots, \langle x_t, y_t \rangle\}$, is presented to the algorithm. During the learning phase, a hypothesis h is learned from the training set. During the evaluation phase, the learner predicts the classification of new instances x as follows

$$c_{NB}(x) = \arg \max_{c_j \in C} \prod_{i=1}^n P(c_j) P(a_i | c_j)$$

1.3.4 Support Vector Machine

The Support Vector Machine (SVM) algorithm (Vapnik, 1998; Cortes & Vapnik, 1995; Scholkopf *et al.*, 1997; Cristianini & Shawe-Taylor, 2000) is a binary classification algorithm. If the data are linearly separable, it outputs a separating hyperplane, which maximizes the “margin” between classes. If data are not linearly separable, the algorithm works by implicitly mapping the data to a higher dimensional space, where the data become separable. A maximum margin separating hyperplane is found in this space. This hyperplane in the high dimensional space corresponds to a nonlinear surface in the original space. SVM classifiers are sometimes called “large margin classifiers” because they find a maximum margin separation. Large margin classifiers are very popular due to theoretical results that show that a large margin ensures a small generalization error bound (Vapnik, 1998) and also because they proved to be very effective in practice.

1.3.5 K-Nearest Neighbors

The K-Nearest Neighbors classifier (Cover & Hart, 1967; Mitchell, 1997) is a simple example of instance-based learning, also known as lazy learning. In the K-Nearest Neighbors algorithm, the nearest neighbors are defined in terms of a metric (a.k.a. distance) D between

instances. The class label for a new instance x is given by the most common class label among the k training examples nearest to x (according to the distance D).

1.3.6 Bagging

Bootstrap aggregating (Bagging) (Breiman, 1996) is an algorithm that helps improve the accuracy of a classifier. Bagging works by sampling examples from the training dataset D with replacement to create subsets of training data, which are called bootstrap samples. A classifier is applied to the different subsets and the output of these bootstrap samples is either averaged or they are allowed to vote for the class. Bagging is known to avoid overfitting and reduce variance in learning algorithms.

1.3.7 REPTree

REPTree is an implementation in WEKA (Witten & Frank, 2005) that builds a decision tree using information gain/variance reduction as the splitting criterion. It prunes the tree using reduced-error pruning (with backfitting). Missing values are dealt with by using fractional instances as in C4.5.

1.4 Motivation

Many PPI have been discovered over the years and several databases have been created to store the information about these interactions such as BIND (Bader *et al.*, 2003), DIP (Salwinski *et al.*, 2004), MIPS (Mewes *et al.*, 2002), IntAct (Kerrien *et al.*, 2007) and MINT (Chaturyamontri *et al.*, 2007). von Mering *et al.* (2002) states that about 80,000 interactions between yeast proteins are currently available from various high-throughput interaction detection methods. These methods detect if the interaction is either a physical binding between proteins or a functional association between proteins. The functional association between two proteins often leads to physical binding among them. Determining PPI using high-throughput methods is not only expensive and time consuming, but also generates a high number of false positives and false negatives. Therefore, there is a need for computational approaches that can help in the process of identifying real protein interactions.

Several methods have been designed to address the task of predicting protein-protein interactions using machine learning. Most of them use features extracted from protein sequences (e.g., amino acids composition) or associated with protein sequences directly (e.g., GO

annotation). Others use relational and structural features extracted from the PPI network, along with the features related to the protein sequence. When using the PPI network to design features, several node and topological features can be extracted directly from the associated graph. This thesis provides an overview of predicting PPI using the graph information extracted from a PPI network along with other available biological features of the proteins and their interactions.

CHAPTER 2 - Background and significance

Several graph-based approaches have been used to address the problem of predicting PPI. These approaches represent the PPI network as a graph and extract relational and structural features from it. These features are provided to machine learning algorithms, few of which were described in the previous section. The training dataset that contains protein pairs along with graph-based features are provided to classifiers. The testing dataset that contains protein pairs that are not present in the training dataset, are presented to the learning model generated by the classification algorithms. Statistical measures such as accuracy, sensitivity, specificity and AUC score are obtained from the learning algorithms. Several previous approaches to protein interaction prediction using graph-based features are detailed below.

2.1 Qi, Bar-Joseph & Klein-Seetharaman

Qi *et al.* (2006) divide the protein interaction prediction task into three sub-tasks: (1) prediction of physical (or actual) interaction among proteins, (2) prediction of proteins belonging to the same complex and (3) prediction of proteins belonging to the same pathway. They use different data sources for different subtasks: data from the MIPS (Mewes *et al.*, 2002) database for the first subtask, data from the DIP (Salwinski *et al.*, 2004) database for the second subtask and data from the KEGG (Kanehisa *et al.*, 2000) database for the third subtask. They assemble 162 features and vary their encoding to understand their effects on the protein interaction prediction subtasks. The categories that they divide their 162 features are:

- Gene expression: It contains 20 features (Pearson's correlation coefficient) calculated on 20 gene expression datasets that were recorded under more than 500 conditions (Bar-Joseph *et al.*, 2003).

- Gene Ontology (Molecular Function, Biological Process & Cellular Component): These 3 categories contain information of how many times a pair of protein occurs in the trees (Christie *et al.*, 2004).
- Protein Expression: It is the difference of the expression levels of the pair of proteins (Ghaemmaghami *et al.*, 2003).
- Essentiality: An essential gene cannot be made into a haploid or homozygous deletion strain. This feature contains the essentiality of the pair of proteins.
- High-throughput PPI datasets (HMS_PCI, TAP & Y2H): These 3 categories contain information extracted from several high-throughput protein interaction methods (Bader *et al.*, 2003; Gavin *et al.*, 2002; Ho *et al.*, 2002; Ito *et al.*, 2001; Uetz *et al.*, 2000).
- Synthetic Lethal: This feature was extracted by the union of Tong *et al.* (2001) and MIPS (Mewes *et al.*, 2002).
- Gene neighborhood / Gene Fusion / Gene Co-occur: This feature is the union of the three datasets described by von Mering *et al.* (2002).
- Sequence Similarity: It contains BLAST hit information of the query protein on SGD database (Christie *et al.*, 2004).
- Homology based PPI: This feature uses Sequence Similarity information to identify homology pairs. These pairs are then “BLAST”ed against NCBI non-redundant protein database and the count of their interactions was extracted.
- Domain-Domain Interaction: Deng *et al.* (2002) identify domain interactions based on sequence analysis. The value of this feature is the probability of interaction of a candidate protein pair.
- Protein-DNA TF group binding: Qi *et al.* (2006) group the TFs based on the MIPS protein class catalog into 16 TF groups. For each TF group, they counted the number TFs that bind to both genes, and used this number as one of their attributes.
- MIPS features (Protein Class and Mutant Phenotype): These 2 categories contain features that identify if the protein pair belongs to the same protein class and mutant phenotype. They apply several feature classifiers such as Random Forest (RF), RF similarity-based k-Nearest-Neighbor, Naive Bayes, Decision Tree, Logistic Regression, and Support Vector Machine (SVM) on their data and obtain reasonably good AUC scores. Their results show that

RandomForest is the one of the top two classifiers for all tasks; the other one is RandomForest similarity-based k-Nearest-Neighbor. They also observe that gene coexpression and few features extracted from the Gene Ontology were the best features for all three subtasks of the protein interaction prediction task.

2.2. Licamele & Getoor

Licamele & Getoor (2006) combine the link structure of the PPI graph with the information about proteins in order to predict the interactions in a yeast dataset. More specifically, they look at the shared neighborhood among proteins and calculate the clustering coefficient among the neighborhoods for the first-order and second-order protein relations. They also consider the Gene Ontology distance between proteins. However, they do not make a distinction between direct (physical interaction) and indirect (proteins belonging to the same complex) interactions in their study. They combine data from multiple data sources such as MIPS (Mewes *et al.*, 1999), BIND (Bader *et al.*, 2001), DIP (Xenarios *et al.*, 2002), yeast two-hybrid (Ito *et al.*, 2001; Uetz *et al.*, 2000) and In vivo pull-down (Gavin *et al.*, 2002; Ho *et al.*, 2002). They apply several classifiers such as Naive Bayes, kNN, Logistic Regression, C4.5, SVM, JRIP and Bagging with REPTrees on their data and obtain a reasonably good accuracy and AUC score when predicting new links from noisy high throughput data.

2.3 Paradesi, Caragea & Hsu

The above-mentioned approaches use relational data of the PPI network along with other biologically relevant information (such as, sequence, gene expression data, GO terms, etc.) to predict the protein interactions. Paradesi *et al.* (2007) address the problem of predicting protein-protein interactions based solely on the graph features of the PPI network. They identify nine structural features for *Saccharomyces cerevisiae* protein interaction network such as indegrees, outdegrees, mutual proteins and backward distance among proteins.

They also learn several classifiers such as Bagged Random Forest, Bagged REPTree, Random Tree, J48 and Classification via Regression on the data. They evaluated the learned models on the dataset of DIP (Salwinski *et al.*, 2004) and also that generated by Qi *et al.* (2006). The method developed by Paradesi *et al.* (2007) compares well with the existing methods for PPI

prediction, even though they used only the relational features of the network data in their study, and not the sequence information as used by Qi *et al.* (2006) and Licamele & Getoor (2006).

However, Paradesi *et al.* (2007) compare the results of different datasets with that obtained by the dataset of Licamele & Getoor (2006). Also, the method of generating negative examples using the dataset of Qi *et al.* (2006) does not provide the same negative examples as mentioned by Qi *et al.* (2006). These comparisons are made between algorithms applied to two different data sets. In order to make a fair comparison, the same dataset and features generated by Qi *et al.* (2006) are used. Due to the non-availability of the dataset by Licamele & Getoor (2006), the results obtained by using the graph features in this thesis cannot be compared with the results of Licamele & Getoor (2006). The comparisons of the previously published results are shown below:

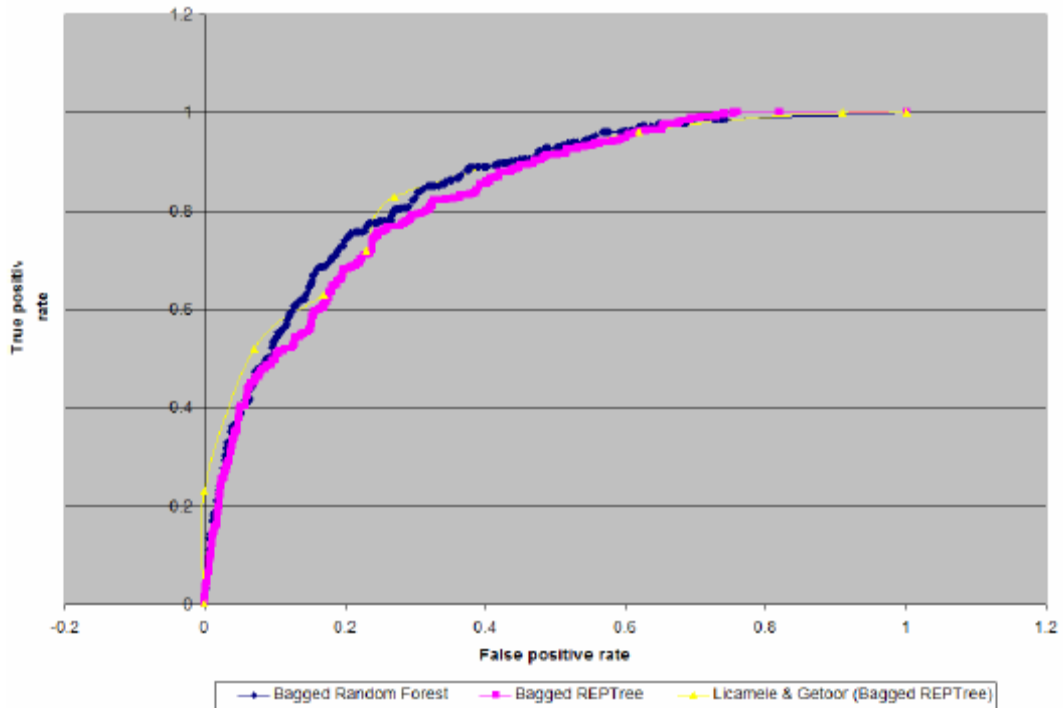


Figure 2-1 Comparison of results by Licamele & Getoor (2006) and Paradesi *et al.* (2007)

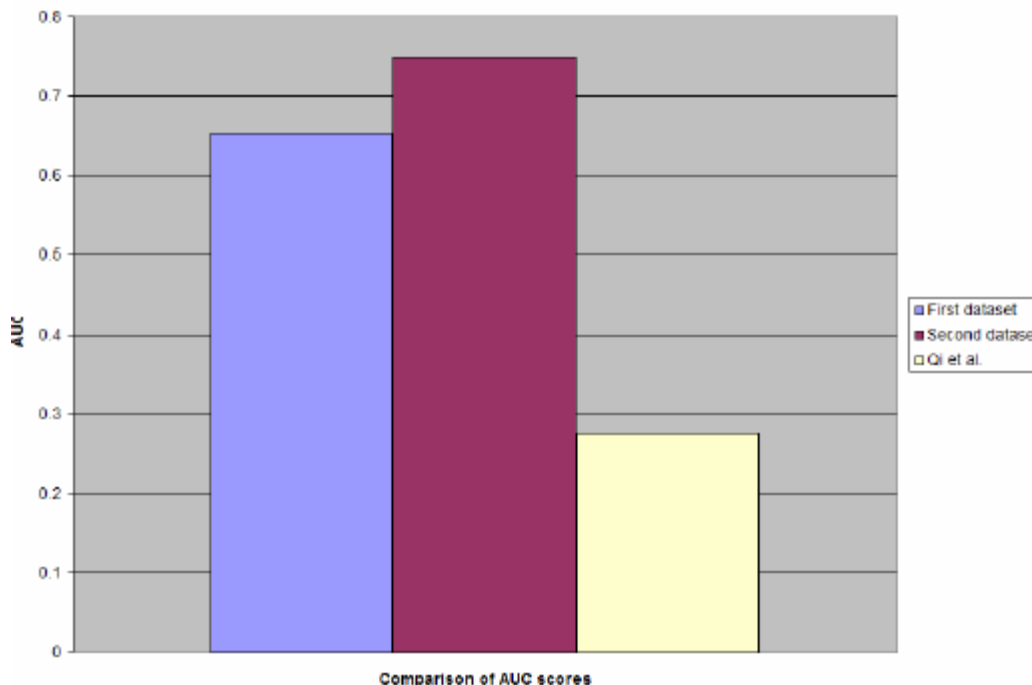


Figure 2-2 Comparison of results by Qi *et al.* (2006) and Paradisi *et al.* (2007)

Note: First dataset is DIP (Salwinski *et al.*, 2006) and second dataset is the dataset generated by Qi *et al.* (2006).

2.4 Chen & Liu

Protein interaction sites refer to the locations on the protein structures where one protein physically interacts with another protein. A protein domain is a functionally defined protein region. Chen & Liu (2005) predicts PPI using protein domain information. Many domain-based models for protein interaction prediction have been developed, and preliminary results have demonstrated their feasibility (Chen & Liu, 2005). Most of the existing domain-based methods, however, consider only single-domain pairs (one domain from one protein) and assume independence between domain–domain interactions. Chen & Liu (2005) introduced a new framework based on random forest for PPI prediction, which explores the contributions of all the possible domain combinations to predicting protein interactions. Furthermore, their model does not assume that domain pairs are independent of each other. They obtained the PPI data from DIP (Salwinski *et al.*, 2004; Deng *et al.*, 2002; Schwikowski *et al.*, 2000; Xenarios *et al.*, 2001). Chen & Liu (2005) extract the domain information for each protein and build a vector of the domain list of each candidate protein pair. The values in the vector are the number of

occurrences of the domain in both proteins. They obtain a better sensitivity and specificity when compared with a method that maximizes the likelihood of the observed protein interaction data and identifies domain interactions (Deng *et al.*, 2002). Due the availability of the dataset generated by Chen & Liu (2005), a comparison is done between the results obtained by them and the authors. This comparison uses the training and testing dataset generated by Chen & Liu (2005). The nine graph features mentioned in this thesis are extracted by the authors and the results of both approaches are compared.

2.5 Advantages of graph-based PPI prediction

Thus, as seen with previous approaches of predicting protein interactions, it is observed that graph-based features extracted from the PPI network are often more useful than biological features in predicting protein interactions. Protein interactions have generally been identified in laboratory experiments (these include small-scale and high-throughput experiments). Most of the predictions discovered in a laboratory might be false positives. Graph-based PPI prediction helps to identify new protein interactions based on the PPI network. If the graph-based PPI prediction can identify predictions that have been discovered by one experiment and not discovered by another experiment, one can assign higher confidence to the fact that the two proteins interact.

CHAPTER 3 - Experiments

In this thesis, several experiments have been performed on published datasets. This work presents a detailed study to determine if graph attributes extracted from a protein interaction network are more predictive than biological features of protein interactions. Another study has been performed to observe the variation in accuracy and AUC score by increasing the size of the protein interaction dataset. The datasets were parsed in order to construct directed networks of interacting protein pairs. The approach as described in Maslov & Sneppen (2002) has been applied, in which the authors represent the PPI network as a directed graph with a directed edge from a “bait” protein to a “prey” protein. A link is drawn between two proteins if and only if there exists an interaction between those two proteins. The absence of an interaction between two proteins results in not adding a link between those two proteins in the graph structure. Several methods of generating training and testing datasets are performed. The classification

problem reduces to the problem of classifying proteins within a distance $d(u, v)$ as either 1 (interacting) or greater than 1 (non-interacting).

3.1 Protein domain information

The dataset published by Chen & Liu (2005) contains 3,713 proteins and 9,832 protein interactions of the *Saccharomyces cerevisiae* organism. They built this dataset by combining data from the DIP database (Salwinski *et al.*, 2004), Deng *et al.* (2002) and Schwikowski *et al.* (2000). Although Chen & Liu (2005) claim that there are 9,834 protein interactions, it has been observed that two positive protein pairs have been repeated; one in the training dataset and the other in the testing dataset. Their dataset contains 4,293 protein domains that are present in the 3,713 proteins. The domain information was obtained from the Pfam database (Bateman *et al.*, 2004). Chen & Liu (2005) split the positive examples into training and testing datasets with 4,916 positive examples in each dataset. They also randomly sample 8,000 negative examples from a list of negative examples and store 4,000 negative examples in each of the training and testing dataset. Although this technique of data sampling is not accurate because there are more positive examples than negative examples in their dataset (In reality, there are many more negative examples than the positive examples in a protein interaction network), this technique has been followed in this thesis in order to compare with the published results of Chen & Liu (2005). In this study, graph features are added to the domain information to observe improved performance in predicting protein interactions.

3.2 Biological features

The dataset published by Qi *et al.* (2006) contains 6,270 proteins and 2,865 protein interactions that they retrieved from the DIP database (Xenarios *et al.*, 2002). There are 237,384 negative examples that are identified among all possible negative examples. Qi *et al.* (2006) encode 162 features for each protein pair in their positive and negative datasets as mentioned in Section 2.1. Qi *et al.* (2006) randomly sample 30,000 examples into each of the training and testing datasets. These 30,000 examples contain 50 positive examples and 29,950 negative examples so that there are 600 negative examples for every positive example. This ratio is a close approximation to the real ratio of negative examples to positive examples. However, sampling only 50 positive examples from 2,865 positive examples is not representative of all the

positive interactions. It would make an interesting study to sample at least around 390 positive examples, so that around 234,000 negative examples could be provided to the learning algorithm. However, despite the low sampling frequency of positive examples, the technique used by Qi *et al.* (2006) has been applied in this study in order to compare with their published results. Graph features are added to the 162 features to observe improved performance in predicting protein interactions.

3.3 Combining PPI databases

Section 2.3 describes a preliminary experiment comparing results from two different data sets. In order to make this comparison fair and rigorous, two new datasets are created that contain features from other approaches. Protein interactions that occur in DIP (Salwinski, 2004), IntAct (Kerrien, 2007) & MINT (Chatr-aryamontri, 2007) are combined to form a true positive dataset. The intersection of all these databases gives rise to protein interactions that are confirmed by several experiments. These databases share a common attribute – the UniprotID of the proteins – among them. However, most of the proteins have multiple UniprotIDs. A hashmap was used to identify unique proteins across the three datasets. This intersection dataset contains 956 proteins and 936 protein interactions. Thus, the dataset obtained by intersecting these three databases resulted in a sparser protein interaction network. In order to see if our features help in the prediction of protein interactions, the features of the intersection dataset are extracted from the published data of Chen & Liu (2005) and Qi *et al.* (2006).

3.3.1 Protein domain information

The positive protein pairs in the intersection data of DIP (Salwinski, 2004), IntAct (Kerrien, 2007) & MINT (Chatr-aryamontri, 2007) are identified in the published data by Chen & Liu (2005). 527 positive protein pairs of 571 proteins are present in both the datasets. There are 1,231 protein domains among these 527 interaction pairs. In order to determine the negative protein pairs, all possible protein pairs (571 x 571) are computed and subtracted from the positive protein pairs that are present in the published dataset by Chen & Liu (2005). 323,712 negative protein pairs are identified based on this technique. The ratio of positive to negative examples is approximately 1:614, which is a close approximation to the actual ratio.

3.3.2 Biological features

The positive protein pairs in the intersection data of DIP (Salwinski, 2004), IntAct (Kerrien, 2007) & MINT (Chatr-aryamontri, 2007) are identified in the published data by Qi *et al.* (2006). 317 positive protein pairs of 392 proteins are present in both the datasets. In order to determine the negative protein pairs, all possible protein pairs (392 x 392) are matched with the negative examples present in the published dataset by Qi *et al.* (2006). 936 negative protein pairs are identified based on this technique. This technique is used to generate negative protein pairs only because biological features have been identified for those protein pairs. The development time and computational time to compute biological features for all possible negative protein pairs is significantly large. However, this technique is a reasonably sound one to verify if there is an advantage of using graph-based features along with biological features in predicting PPI in a protein interaction network.

3.4 Features

The following graph features are extracted from the protein interaction network:

1. Indegree of the start node: Denotes the popularity (importance) of the start node (i.e., of the protein associated with the start node).
2. Indegree of the end node: Denotes the popularity (importance) of the end node (i.e., of the protein associated with the end node).
3. Outdegree of the start node: Denotes the number of proteins interacting with the protein at the start node.
4. Outdegree of the end node: Denotes the number of existing proteins interacting with the protein at the end node; correlates loosely with the likelihood of a reciprocal link.
5. Number of mutual proteins of a protein w , such that $u \rightarrow w \wedge w \rightarrow v$, for some proteins u and v .
6. Number of mutual proteins of a protein w , such that $v \rightarrow w \wedge w \rightarrow u$, for some proteins u and v .
7. Number of mutual proteins of a protein w , such that $u \rightarrow w \wedge v \rightarrow w$, for some proteins u and v .
8. Number of mutual proteins of a protein w , such that $w \rightarrow u \wedge w \rightarrow v$, for some proteins u and v .

9. Backward distance from v to u in the graph: identifies how far the protein v is from protein u.

The diagrammatic representations of the nine features considered are as shown in Figure 3-1 (a – i) below:

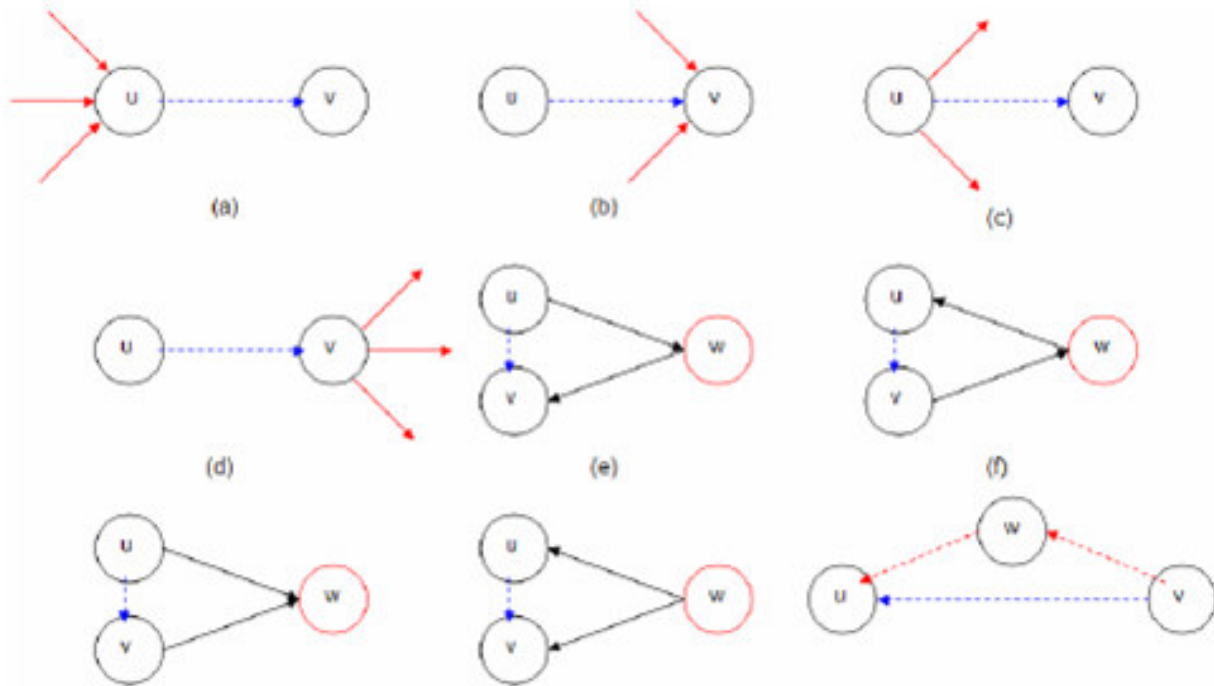


Figure 3-1 Graph features

Note: The objects in red denote the feature that we calculate. The dashed lines (in blue) above indicate that a link between two proteins u and v may be either present or absent, i.e. either u or v are directly connected or indirectly connected via another node w.

CHAPTER 4 - Results

The data is split into training and testing datasets according to the experiment setup described by Qi *et al.* (2006) and Chen & Liu (2005). Several classifiers such as RandomForest (RF), J48, ClassificationviaRegression (CVR), Naïve Bayes (NB) & Support Vector Machine (SVM) are applied to the training and testing datasets. It is important to note that the experiments performed using SVM classifier in WEKA use a linear kernel. The experiment is performed five times in order to negate the effect of randomly sampling examples. The average scores of the

performance criteria in the five experimental runs are reported. The performance criteria used in previous published works are used in related experiments of this study.

4.1 Protein domain information

As per the technique published by Chen & Liu (2005), 9,832 positive protein pairs are split into training and testing datasets. 8,000 negative protein pairs are randomly sampled and split into training and testing dataset. Chen & Liu (2005) calculate the sensitivity ($=\text{True positives}/(\text{True positives} + \text{False negatives})$) and specificity ($=\text{True negatives}/(\text{True negatives} + \text{False positives})$) performance measures in their study. The sensitivity (Se) and specificity (Sp) of the classifiers are shown:

Table 4-1 Results obtained from experiments using Chen & Liu (2005) dataset (5-fold)

	J48		NB		CVR		SVM	
	Se %	Sp %	Se %	Sp %	Se %	Sp %	Se %	Sp %
Domain	73.3	62.1	73.44	63.08	55.1	0	74.3	73.56
Degree	86.62	85.74	89.78	60.66	85.78	86.98	87.2	76.02
MutualProtein	96.68	59.3	97.5	57.72	55.1	0	98.96	55.94
BackwardDistance	99.52	65.8	99.52	65.8	99.52	65.8	99.52	65.8
Domain + Degree	86.3	86.08	88.94	62.52	85.92	86.7	80.7	77.76
Domain + MutualProtein	85.14	68.4	95.64	59.08	55.1	0	78.62	75.26
Domain + BackwardDistance	91.14	72.46	83.16	71.82	99.52	65.8	86.06	82.68
Degree + MutualProtein	87.6	86.56	93.96	65.24	86.72	87.14	89.26	77.44
Degree + BackwardDistance	92.56	91.54	93.1	71.88	92.02	92.14	93.3	83.54
MutualProtein + BackwardDistance	97.86	80.04	98.18	74.42	99.54	70.88	99.26	76.16
Domain + Degree + MutualProtein	87.8	85.88	93.16	65.94	86.86	86.84	83.18	78.76
Domain + MutualProtein + BackwardDistance	96.94	81.5	97.82	76.42	99.54	70.88	90.7	85
Domain + Degree + BackwardDistance	91.86	91.96	92.3	73.06	91.96	92.24	88.7	85.58
Degree + MutualProtein + BackwardDistance	93.04	93.1	94.68	69.4	92.94	93.14	95.74	85.68
Domain + Degree + MutualProtein + BackwardDistance	93.6	92.4	94.28	70.54	92.76	93.2	91.14	87.4

Table 4-1 indicates that the BackwardDistance attribute is the best attribute for predicting protein interactions along with protein domain information. The MutualProtein information is the next best feature for predicting protein interactions when using protein domain information. The

Degree feature provides more information about the protein interaction network to the classifiers than the Domain information alone. The highest sensitivity with respect to Domain information is obtained when both MutualProtein and BackwardDistance are added to the Domain information. The highest specificity with respect to Domain information is obtained when both Degree and BackwardDistance are added to the Domain information. It is interesting to observe that combining Degree, MutualProtein and BackwardDistance with the Domain information does not produce the highest sensitivity and specificity. This is observed despite the fact that all the three graph features perform better individually than the Domain information. Chen & Liu (2005) reported a sensitivity of 79.78% and specificity of 64.38% when they used domain information only. The sensitivity and specificity scores from the above table indicate that a higher sensitivity and specificity are obtained by using graph features along with the Domain information.

The results of RandomForest learning algorithm were not obtained due to memory constraints on the Beocat cluster. The experiments were run on the Beocat cluster with 16GB of RAM allotted to the program. The classifier built forests that exceeded the memory limit of 16GB.

4.2 Biological features

As per the technique published by Qi *et al.* (2006), 50 positive protein pairs are sampled and their features are stored in the training dataset. Similarly, another 50 positive protein pairs are sampled and their features are stored in the testing dataset. 29,950 negative protein pairs are randomly sampled and their features are stored in the training dataset. Similarly, 29,950 negative protein pairs are randomly sampled and their features are stored in the testing dataset. Qi *et al.* (2006) calculate the AUC (area under the ROC curve) score in their study. The AUC scores of the classifiers are shown:

Table 4-2 Results obtained from experiments using Qi *et al.* (2006) dataset (5-fold)

	J48	RF	NB	CVR	SVM
Feature	0.504	0.7052	0.7244	0.7466	0.504
Degree	0.5	0.7394	0.9442	0.9798	0.5
MutualProtein	0.61	0.806	0.826	0.5	0.622
BackwardDistance	0.79	0.79	0.79	0.5	0.73
Feature + Degree	0.57	0.7276	0.7378	0.9526	0.506
Feature + MutualProtein	0.5944	0.8172	0.737	0.8347	0.626
Feature + BackwardDistance	0.796	0.8438	0.7374	0.8634	0.736
Degree + MutualProtein	0.7052	0.8604	0.9632	0.9828	0.624
Degree + BackwardDistance	0.79	0.833	0.9646	0.9896	0.73
MutualProtein + BackwardDistance	0.79	0.94	0.948	0.5	0.758
Feature + Degree + MutualProtein	0.5784	0.8358	0.7428	0.9496	0.626
Feature + MutualProtein + BackwardDistance	0.796	0.9408	0.7438	0.9052	0.758
Feature + Degree + BackwardDistance	0.796	0.8678	0.7428	0.9768	0.736
Degree + MutualProtein + BackwardDistance	0.79	0.944	0.9774	0.9926	0.758
Feature + Degree + MutualProtein + BackwardDistance	0.796	0.927	0.7468	0.98	0.756

It is interesting to observe that different classifiers have different best attributes that predict protein interactions. However, it is not surprising to note that all the three graph features provide more information about the protein interaction network to the classifiers than the 162 biological features alone. The highest AUC score with respect to the biological features is obtained when Degree, MutualProtein and BackwardDistance are added to the biological features. Qi *et al.* (2006) reported an R50 AUC score of 0.25 when they used the biological features only. Due to the similar experiment design of this experiment and the experiment by Qi *et al.* (2006), it can be safely assumed that the AUC score obtained by Qi *et al.* (2006) using a RandomForest classifier could have been 0.7052. The AUC scores from the above table indicate that higher AUC scores are obtained by using graph features along with the biological features.

4.3 Combining PPI databases

4.3.1 Protein domain information

In order to maintain the same ratio of positive to negative examples from the original datasets, 100 positive protein pairs and 29,950 negative protein pairs are randomly sampled from the original datasets and split into training and testing datasets. The AUC scores of the classifiers are shown:

Table 4-3 Results obtained from experiments using the Intersection dataset & Chen & Liu (2005) dataset (5-fold)

	NB	CVR	SVM
Domain	0.4996	0.5	0.5
Degree	0.8128	0.8582	0.5
MutualProtein	0.7048	0.5	0.532
BackwardDistance	0.858	0.5	0.858
Domain + Degree	0.7082	0.81	0.5
Domain + MutualProtein	0.6652	0.5	0.6
Domain + BackwardDistance	0.841	0.5	0.85
Degree + MutualProtein	0.876	0.8862	0.546
Degree + BackwardDistance	0.9374	0.9278	0.858
MutualProtein + BackwardDistance	0.943	0.5	0.858
Domain + Degree + MutualProtein	0.791	0.8206	0.604
Domain + MutualProtein + BackwardDistance	0.9162	0.5	0.858
Domain + Degree + BackwardDistance	0.887	0.9326	0.85
Degree + MutualProtein + BackwardDistance	0.981	0.9164	0.858
Domain + Degree + MutualProtein + BackwardDistance	0.9402	0.919	0.858

Table 4-3 indicates that the BackwardDistance attribute is the best attribute for predicting protein interactions along with protein domain information. The Degree information is the next best feature for predicting protein interactions when using protein domain information. The MutualProtein feature provides more information about the protein interaction network to the classifiers than the Domain information alone. The highest AUC score with respect to Domain

information is obtained when Degree, MutualProtein and BackwardDistance are added to the Domain information. The AUC scores from the above table indicate that higher AUC scores are obtained by using graph features along with the Domain information.

4.3.2 Biological features

317 positive protein pairs and 936 negative protein pairs are split into training and testing datasets. The AUC scores of the classifiers are shown:

Table 4-4 Results obtained from experiments using the Intersection dataset & Qi *et al.* (2006) dataset (5-fold)

	J48	RF	NB	CVR	SVM
Feature	0.8248	0.9052	0.7062	0.7468	0.6398
Degree	0.7068	0.7928	0.759	0.8156	0.5048
MutualProtein	0.5	0.6096	0.6096	0.5	0.5664
BackwardDistance	0.85	0.85	0.8494	0.5	0.85
Feature + Degree	0.8748	0.9302	0.7448	0.863	0.6772
Feature + MutualProtein	0.8324	0.919	0.7374	0.7468	0.6898
Feature + BackwardDistance	0.9	0.9672	0.8112	0.8882	0.881
Degree + MutualProtein	0.7182	0.8058	0.7798	0.8264	0.5848
Degree + BackwardDistance	0.9114	0.9418	0.9398	0.958	0.85
MutualProtein + BackwardDistance	0.85	0.8872	0.8868	0.5	0.7318
Feature + Degree + MutualProtein	0.8742	0.934	0.7618	0.863	0.6952
Feature + MutualProtein + BackwardDistance	0.9014	0.967	0.8278	0.888	0.8898
Feature + Degree + BackwardDistance	0.9234	0.9754	0.8264	0.9458	0.881
Degree + MutualProtein + BackwardDistance	0.9328	0.9478	0.9482	0.9584	0.8684
Feature + Degree + MutualProtein + BackwardDistance	0.927	0.9706	0.8357	0.9464	0.8906

It is interesting to observe that different classifiers have different best attributes that predict protein interactions. It is also surprising to note that MutualProtein does provide more information about the protein interaction network to the classifiers than the 162 biological features alone. The highest AUC score with respect to the biological features is obtained when

MutualProtein and BackwardDistance are added to the biological features. It is interesting to observe that combining Degree, MutualProtein and BackwardDistance with the biological features does not produce the highest AUC score. The AUC scores from the above table indicate that higher AUC scores are obtained by using graph features along with the biological features.

CHAPTER 5 - Conclusion and Future Work

5.1 Conclusion

In this thesis, a comprehensive study about the importance of graph features in predicting protein interactions in a protein interaction network is performed. Two published datasets were used in this thesis. A third dataset was created by intersecting three protein interaction databases and combining the data from the other two published datasets. Nine graph features were extracted from the datasets and several learning algorithms were applied. The results indicate that graph features extracted from a protein interaction network are useful for predicting protein interactions.

5.2 Future Work

The future trends in the field of predicting protein interactions using graph-based features look very promising due to the following changes in the area of protein interaction prediction:

5.2.1 Increase in quality and quantity of data

There is a rapid increase in the discovery of new proteins and interactions based on the high-throughput methods. There is also a growth in techniques to identify actual protein interactions and eliminate false positive interactions. The research groups that maintain BIND (Bader *et al.*, 2003), MINT (Chatr-aryamontri *et al.*, 2007), DIP (Salwinski *et al.*, 2004), MPact (Güldener *et al.*, 2006) and IntAct (Kerrien *et al.*, 2007) have formed the IMEx consortium to build a large, consistent and non-redundant repository of protein interactions and information about the interactions. According to the IMEx consortium, the data they are gathering will be broader in scope and deeper in information than any individual effort (Kerrien *et al.*, 2007). The IMEx consortium will allow individual researchers and research groups to submit protein interaction information. This newly submitted data is run through several tests and manually

curated to ensure that the interaction is a true positive interaction. As the quality and quantity of data increases, the PPI network becomes more complete, thereby allowing highly predictive graph features to be extracted from the network. There is scope for researchers to work on increasing the quality and quantity of protein interactions by developing new computational techniques. A primary contribution of this work has been to identify graph features that will help in predicting protein interactions. Other feature construction and extraction algorithms (Liu & Motoda, 1998) can be added to this body of features.

5.2.2 Improvement in classification algorithms

There are also rapid advances in the machine learning and data mining algorithms. Most of the supervised and unsupervised learning algorithms used in the prediction of protein interactions were developed for tasks other than that. However, it has been observed that these algorithms have worked well for the protein interaction prediction task. There is a need for developing custom algorithms that can handle protein interaction data well.

5.2.3 Use of protein interaction network analysis tools

There are many protein interaction visualization tools such as ProViz (Iragne *et al.*, 2005), iPfam (Finn *et al.*, 2005), VisANT (Hu *et al.*, 2007), etc. and querying tools such as PathBLAST (Kelley *et al.*, 2004), APID (Prieto & De Las Rivas, 2006), etc. available. These tools allow users to view and search for proteins in any PPI network. They can be exploited to gather several graph-based features from the PPI network. The visualization and querying tools can also be used to split the PPI network into several overlapping sub-graphs. Interactions can be predicted at the sub-graph level and these predictions can be combined to identify protein interactions at the original graph level. A protein pair can be labeled as interacting if it is observed that the interaction between the protein pair appears in more than one sub-graph.

5.2.4 Development of different approaches

Although there have been continuing advances in data and improvement of algorithms and tools, attention must be paid to improve the approaches of solving the protein interactions prediction problem. There is no one-solution-solves-all-problems approach anymore. Instead, there is a need for developing approaches that solve the problem by applying an ensemble of various machine learning algorithms for different subgraphs of the PPI network. In other words,

one could extract different graph-based features from different subsets of the PPI network and run different machine learning algorithms on the features, depending on the data. The different machine learning classifiers could “vote” on the class, and the weighted average of the output could be assigned as the actual class.

Bibliography

- Altman, D.G., Bland, J.M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ*, 308(6943):1552.
- Albert, I., & Albert, R. (2004). Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*. 20(18):3346-52.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*.25(1):25-9.
- Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T., & Hogue, C.W. (2001). BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res*, 29(1):242-5.
- Bader, G.D., & Hogue, C.W. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*. 20(10):991-7.
- Bader, G.D., Betel, D., & Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31(1):248-50.
- Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A., Gifford, D.K. (2003). Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*. 21(11):1337-42.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., Eddy, S.R. (2004). The Pfam protein families database. *Nucleic Acids Res*;32(Database issue):D138-41.
- Breiman, L., Freidman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24 (2): 123-40.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45 (1), 5-32.

- Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G., & Chen, R. (2003). Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res*, 31(9):2443-50.
- Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., & Cesareni, G. (2007). MINT: the Molecular INTERaction database. *Nucleic Acids Res*, 35(Database issue): D572–D574.
- Chen, X.W., & Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework, *Bioinformatics*, 21(24):4394-4400.
- Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J.E., Hong, E.L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D., Cherry, J.M. (2004). Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res*. 32(Database issue):D311-4.
- Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273-297.
- Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, Vol. 13, No. 1, pp. 21-27.
- Cristianini, N. & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University.
- Deane, C.M., Salwinski, L., Xenarios, I. & Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1(5):349-56.
- Deng, M., Mehta, S., Sun, F., & Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 12(10):1540-8.
- Fawcett T. (2004). *ROC Graphs: Notes and Practical Considerations for Researchers*. Technical report, Palo Alto, USA: HP Laboratories.
- Fields, S., Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*; 340(6230):245-6.
- Finn, R.D., Marshall, M., & Bateman, A. (2005). iPfam: visualization of protein-protein

interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, 21(3):410-2.

- Frank, E., Wang, Y., Holmes, G., & Witten, I.H. (1998). Using model trees for classification. *Machine Learning*, 32 (1), 63-76.
- Gavin, A.C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., & Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141-7.
- Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S. (2003). Global analysis of protein expression in yeast. *Nature*, 425(6959):737-41.
- Güldener, U., Münsterkötter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.W., Stümpflen, V. (2006). MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*, 34(Database issue):D436-41.
- Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E., Young, R.A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99-104.
- Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sørensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D., & Tyers, M. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):180-3.
- Hu, Z., Ng, D.M., Yamada, T., Chen, C., Kawashima, S., Mellor, J., Linghu, B., Kanehisa,

- M., Stuart, J.M., & DeLisi, C. (2007). VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res*, 35(Web Server issue):W625-32.
- Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686-91.
 - Hsu, W.H., King, A.L., Paradesi, M.S.R., Pydimarri, T., & Weninger, T. (2006). Collaborative and Structural Recommendation of Friends using Weblog-based Social Network Analysis, *Proc. of Computational Approaches to Analyzing Weblogs - AAAI 2006 Technical Report SS-06-03*, 55-60.
 - Iragne, F., Nikolski, M., Mathieu, B., Auber, D., & Sherman, D. (2005). ProViz: protein interaction visualization and exploration. *Bioinformatics*, 21(2):272-4.
 - Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A.*, 98(8):4569-74.
 - Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28, 27-30.
 - Kelley, B.P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B.R., Ideker, T. (2004). PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res*, 32(Web Server issue):W83-8.
 - Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Liefstink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., & Hermjakob, H. (2007). IntAct--open source resource for molecular interaction data. *Nucleic Acids Res*, 35(Database issue):D561-5.
 - Licamele, K., & Getoor, L. (2006). Predicting Protein-Protein Interactions Using Relational Features, *Proc. of ICML Workshop on Statistical Network Analysis*.
 - Liu, H. & Motoda, H. (1998). Feature Selection for Knowledge Discovery and Data Mining. Kluwer.
 - MacBeath, G., & Schreiber, S.L. (2000). Printing proteins as microarrays for high-throughput function determination. *Science*; 289(5485):1760-3.

- Maslov, S., & Sneppen, K. (2002). Specificity and stability in topology of protein networks, *Science*, 296(5569):910-3.
- Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., & Frishman, D. (1999). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 27, 44–48.
- Mewes, H.W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., & Weil, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1):31-4.
- Mitchell T. (1997). *Machine Learning*, McGraw Hill.
- Mrowka, R., Patzak, A. & Herzel, H. (2001). Is there a bias in proteome research? *Genome Res.*, 11, 1971-73.
- Najafabadi, H.S, & Salavati, R. (2008). Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biology*, 9:R87
- Paradesi, M.S.R., Caragea, D., & Hsu, W.H. (2007). Structural Prediction of Protein-Protein Interactions in *Saccharomyces cerevisiae*, *Proc. of IEEE 7th International Symposium on Bioinformatics and BioEngineering*, vol. 2, pp. 1270-1274.
- Paradesi, M.S.R., Wang, L., Brown, S.J., & Hsu, W.H. (2006). Mining Domain Association Rules From Protein-Protein Interaction data, *Intelligent Engineering Systems through Artificial Neural Networks*, vol. 16, 213-218.
- Prieto ,C., & De Las Rivas, J. (2006). APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res*, 34(Web Server issue):W298-302.
- Qi, Y., Bar-Joseph, Z., & Klein-Seetharaman, J., (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Structure, Function, and Bioinformatics*, Volume 63, Issue 3, 490-500.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1, 81-106.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
- Salwinski, L. & Eisenberg, D. (2003). Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 13, 377-382.
- Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449-51.

- Sambrook, J. & Russell, D.W. (2006). Identification of Associated Proteins by Coimmunoprecipitation, *CSH Protocols*, doi:10.1101/pdb.prot3898.
- Schwikowski, B., Uetz, P., Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol*;18(12):1257-61.
- Scholkopf, B., Sung, K.-K., Burges, C. J. C., Girosi, F., Niyogi, P., Poggio, T., & Vapnik, V. (1997). Comparing support vector machines with gaussian kernels to radial basis function classifiers, *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2758-65.
- Shoemaker, B.A., & Panchenko, A.R. (2007). Deciphering Protein–Protein Interactions. Part I. Experimental Techniques and Databases. *PLoS Comput Biol* 3(3): e42.
- Shoemaker, B.A., & Panchenko, A.R. (2007). Deciphering Protein–Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. *PLoS Comput Biol* 3(4): e43.
- Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., Andrews, B., Tyers, M., Boone, C. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364-8.
- Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D.S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J.N., Lu, H., Ménard, P., Munyana, C., Parsons, A.B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A.M., Shapiro, J., Sheikh, B., Suter, B., Wong, S.L., Zhang, L.V., Zhu, H., Burd, C.G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F.P., Brown, G.W., Andrews, B., Bussey, H., Boone, C. (2004). Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808-13.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., 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, J.M. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623-7.
- Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

- von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417, 399-403.
- Wang, H., Segal, E., Ben-Hur, A., Koller, D., & Brutlag, D.L. (2005). Identifying Protein-Protein Interaction Sites on a Genome-Wide Scale. *In Advances in Neural Information Processing Systems 17*. Edited by: Saul L.K., Weiss Y, Bottou L. Cambridge, MA: MIT Press:1465-1472.
- Witten, I.H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, 2nd Edition. Morgan Kaufmann.
- Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30, 303–5.
- Young, K.H. (1998). Yeast two-hybrid: so many interactions, (in) so little time... *Biol Reprod*, 58(2):302-11.

Appendix A - Variance of the results across different experiments

This section presents the variance of results of all experiments conducted in this thesis.

Protein domain information

Table A-1 Variance of results obtained from experiments using Chen & Liu (2005) dataset (5-fold)

	J48		NB		CVR		SVM	
	Se	Sp	Se	Sp	Se	Sp	Se	Sp
Domain	1.76E-5	2.36E-5	6.24E-6	5.85E-5	0	0	2.0E-5	6.64E-6
Degree	4.57E-5	8.26E-5	4.57E-5	2.02E-5	7.76E-6	2.49E-5	1.56E-5	2.56E-6
MutualProtein	1.25E-5	3.9E-7	5.9E-6	1.59E-7	0	0	4.64E-6	2.4E-7
BackwardDistance	1.76E-6	0	1.76E-6	0	1.76E-6	0	1.76E-6	0
Domain + Degree	7.12E-5	1.06E-4	3.54E-5	1.05E-5	2.05E-5	3.91E-5	1.0E-5	1.74E-5
Domain + MutualProtein	6.34E-5	1.2E-6	1.33E-4	5.6E-7	0	0	1.61E-5	7.04E-6
Domain + BackwardDistance	4.91E-4	3.78E-5	7.44E-6	1.69E-5	1.76E-6	0	1.02E-5	7.36E-6
Degree + MutualProtein	2.04E-5	3.1E-5	9.84E-6	1.06E-5	7.75E-6	9.84E-6	2.3E-5	6.4E-7
Degree + BackwardDistance	7.9E-5	6.9E-5	7.16E-5	0.0021	5.25E-5	3.34E-5	1.16E-5	1.04E-6
MutualProtein + BackwardDistance	5.44E-6	2.4E-7	4.37E-5	0.0070	1.04E-6	8.56E-6	2.64E-6	2.4E-7
Domain + Degree + MutualProtein	2.72E-5	3.45E-5	2.5E-5	1.22E-5	7.04E-6	7.84E-6	9.36E-6	1.66E-5
Domain + MutualProtein + BackwardDistance	1.9E-5	2.8E-6	3.33E-5	0.0032	1.04E-6	8.56E-6	7.9E-6	4.79E-6
Domain + Degree + BackwardDistance	1.26E-5	1.38E-5	8.6E-5	0.0017	4.42E-5	2.94E-5	1.52E-5	5.36E-6
Degree + MutualProtein + BackwardDistance	7.02E-5	1.56E-5	1.77E-5	2.98E-4	9.04E-6	3.44E-6	1.1E-5	2.96E-6
Domain + Degree + MutualProtein + BackwardDistance	3.32E-5	1.32E-5	3.21E-5	3.7E-4	4.9E-5	1.4E-5	7.04E-6	9.6E-6

Biological features

Table A-2 Variance of results obtained from experiments using Qi *et al.* (2006) dataset (5-fold)

	J48	RF	NB	CVR	SVM
Feature	3.84E-4	3.24E-4	0.0120	0.0073	2.4E-5
Degree	0	3.45E-4	3.41E-5	2.34 E-4	0
MutualProtein	0.0081	0.0012	6.24E-4	0	4.96E-4
BackwardDistance	0.0018	0.0018	0.0018	0	0.015
Feature + Degree	0.0062	2.21E-4	0.0146	2.15E-4	6.4E-5
Feature + MutualProtein	0.0035	5.4E-4	0.0152	0.0043	4.24E-4
Feature + BackwardDistance	0.0012	0.0011	0.0157	9.79E-4	0.0123
Degree + MutualProtein	0.0228	3.3E-4	7.36E-6	1.11E-4	5.44E-4
Degree + BackwardDistance	3.6E-4	3.6E-4	3.6E-4	0	0.003
MutualProtein + BackwardDistance	0.0018	1.2E-4	2.16E-4	0	0.0087
Feature + Degree + MutualProtein	0.0017	6.19E-4	0.0152	7.9E-5	4.64E-4
Feature + MutualProtein + BackwardDistance	0.0012	1.73E-4	0.0172	0.0023	0.0071
Feature + Degree + BackwardDistance	0.0012	3.64E-4	0.0160	3.99E-4	0.0123
Degree + MutualProtein + BackwardDistance	0.0018	5.84E-4	5.44E-6	1.17E-4	0.0087
Feature + Degree + MutualProtein + BackwardDistance	0.0012	4.96E-4	0.0160	1.51E-4	0.0069

Combining PPI databases

Protein domain information

Table A-3 Variance of results obtained from experiments using the Intersection dataset & Chen & Liu (2005) dataset (5-fold)

	J48	RF	NB	CVR	SVM
Feature	6.47E-4	4.1E-4	5.37E-4	8.42E-4	2.78E-4
Degree	0.0023	2.6E-4	3.6E-4	7.38E-5	3.5E-5
MutualProtein	0	7.54E-5	7.54E-5	0	0.0014
BackwardDistance	1.10E-4	1.1E-4	1.08E-4	0	1.1E-4
Feature + Degree	6.85E-4	3.37E-5	5.25E-4	0.001	5.17E-4
Feature + MutualProtein	2.37E-4	1.50E-4	6.23E-4	8.42E-4	5.97E-4
Feature + BackwardDistance	9.76E-4	7.33E-5	2.01E-4	3.73E-4	4.28E-5
Degree + MutualProtein	0.0017	4.58E-4	3.81E-4	1.16E-4	3.43E-4
Degree + BackwardDistance	2.21E-5	2.21E-5	2.16E-5	0	2.21E-5
MutualProtein + BackwardDistance	1.1E-4	1.13E-4	1.1E-4	0	0.036
Feature + Degree + MutualProtein	6.67E-4	2.51E-4	5.22E-4	0.001	8.28E-4
Feature + MutualProtein + BackwardDistance	0.001	5.48E-5	2.61E-4	3.72E-4	4.37E-5
Feature + Degree + BackwardDistance	3.41E-4	1.58E-5	1.93E-4	7.41E-5	4.28E-5
Degree + MutualProtein + BackwardDistance	2.83E-4	1.12E-4	9.01E-5	1.94E-5	2.61E-4
Feature + Degree + MutualProtein + BackwardDistance	2.9E-4	4.02E-5	2.57E-4	7.58E-5	5.78E-5

Biological features

Table A-4 Variance of results obtained from experiments using the Intersection dataset & Qi *et al.* (2006) dataset (5-fold)

	NB	CVR	SVM
Feature	0.0013	0	0
Degree	0.0017	5.38E-4	0
MutualProtein	5.34E-4	0	0.0017
BackwardDistance	0.0015	0	0.0015
Feature + Degree	0.0013	0.002	0
Feature + MutualProtein	0.0023	0	2.8E-4
Feature + BackwardDistance	0.002	0	0.0015
Degree + MutualProtein	0.0012	3.05E-4	0.0027
Degree + BackwardDistance	2.99E-4	0	2.99E-4
MutualProtein + BackwardDistance	4.24E-4	0	0.0015
Feature + Degree + MutualProtein	0.0011	6.6E-4	5.84E-4
Feature + MutualProtein + BackwardDistance	0.001	0	0.0015
Feature + Degree + BackwardDistance	0.0016	0.003	0.0015
Degree + MutualProtein + BackwardDistance	1.4E-4	0.0027	0.0015
Feature + Degree + MutualProtein + BackwardDistance	2.53E-4	0.0037	0.0014