8-2011

Predicting Yeast Synthetic Lethal Genetic Interactions Using Protein Domains

Bo Li

Clemson University, bol@cs.clemson.edu

Follow this and additional works at: https://tigerprints.clemson.edu/all_dissertations

Part of the [Bioinformatics Commons](#)

Recommended Citation

Li, Bo, "Predicting Yeast Synthetic Lethal Genetic Interactions Using Protein Domains" (2011). *All Dissertations*. 752.
https://tigerprints.clemson.edu/all_dissertations/752

This Dissertation is brought to you for free and open access by the Dissertations at TigerPrints. It has been accepted for inclusion in All Dissertations by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.

PREDICTING YEAST SYNTHETIC LETHAL GENETIC INTERACTIONS USING
PROTEIN DOMAINS

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Computer Science

by
Bo Li
August 2011

Accepted by:
Dr. Feng Luo, Committee Chair
Dr. Brian C. Dean
Dr. David P. Jacobs
Dr. Pradip K. Srimani

ABSTRACT

Synthetic lethal genetic interaction (SLGI) is an important biological phenomenon. Such interactions are of interest as they can be used to predict the function of unknown proteins and find drug targets or drug combinations. High throughput biological experiments enhance the capability in identifying genetic interactions, but the large number of protein pairs still makes the task of genome-wide identification of genetic interactions overwhelming. Computation based prediction of SLGIs is promising but still hampered by the unclear molecular mechanism of SLGIs.

Protein domains with conserved functions serve as the building blocks of proteins. A genetic interaction that occurs between a pair of proteins could be essentially related to or even dominated by the domains underneath. We applied support vector machine (SVM) classifier and maximum likelihood estimation (MLE) method to predict SLGIs in yeast based on domain information in proteins. Our study demonstrates that yeast SLGIs could be explained by the genetic interactions between domains of those proteins. Moreover, we retrieved a set of polypeptide clusters and used them to predict SLGIs. Besides providing better performance, this approach allows us to predict genetic interactions in a more general fashion.

We propose a novel idea for the prediction of SLGIs, upon which multiple approaches were derived. This study helps the understanding of originality of functional relationships in SLGIs at the domain level that may significantly contribute to the biology community in further analysis of genetic interaction related studies.

DEDICATION

I dedicate this dissertation to my family. I couldn't have made it this far without their love and support over the years.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Feng Luo, for offering the precious opportunity working with him and providing guidance through the entire project.

I would like to thank Dr. Brian Dean, Dr. David Jacobs and Dr. Pradip Srimani for serving on my committee and giving me advice on research during my PhD study.

I would like to thank Dr. Weiguo Cao for assisting my project with the application in the area of DNA repair.

I would like to thank Thomas Grindinger, Xumeng Li, Pengfei Xuan and Yuehua Zhang for being great group mates.

Finally, I would like to thank the School of Computing and the Graduate School for providing me the wonderful learning experience.

.

TABLE OF CONTENTS

	Page
TITLE PAGE	i
ABSTRACT.....	ii
DEDICATION	iii
ACKNOWLEDGMENTS	iv
LIST OF TABLES.....	viii
LIST OF FIGURES	x
 CHAPTER	
1. INTRODUCTION	1
1.1. Problem Statement	1
1.2. Dissertation Summary.....	2
1.3. Research Contributions	4
1.4. Dissertation Outline.....	5
2. BACKGROUND	6
2.1. Genetic Interaction.....	6
2.2. Synthetic Lethal Genetic Interaction	8
2.3. Protein Domain.....	12
2.4. Classification and Prediction	12
2.5. Evaluation Criteria	14
2.6. Related Work.....	16
3. Predicting Yeast Synthetic Lethal Genetic Interaction Using Domain Features	28

Table of Contents (Continued)

	Page
3.1. Data Sources	28
3.2. Feature Encoding.....	29
3.3. SVM Classifier.....	29
3.4. Classification Results.....	31
3.4.1. Cross Validation Study	31
3.4.2. Prediction Performance.....	33
3.4.3. Oversampling the Negative Dataset	34
3.4.4. Updating the Negative Dataset	34
3.5. Updated Results	36
3.6. Discussion	37
4. Predicting Yeast Synthetic Lethal Genetic Interactions Using Domain Genetic Interactions.....	39
4.1. Estimating Probabilities of Domain Genetic Interactions with EM Algorithm.....	40
4.2. Domain Genetic Interactions Rarely Overlap Domain Physical Interactions.....	46
4.3. Domain Genetic Interaction Network.....	47
4.4. Domain Genetic Interactions in the SLGIs between Multi-domain Proteins	49
4.5. Prediction and Validation of Genome-wide SLGIs Using Domain Genetic Interactions	52
4.6. Novel SLGIs Predicted by MLE Approach	58
4.7. Compensatory Pathways from Predicted SLGIs.....	60
4.8. Discussion	65
5. Predicting Yeast Synthetic Lethal Genetic Interactions Using Short Polypeptide Clusters	67

Table of Contents (Continued)

	Page
5.1. Constructing the Short Polypeptide Clusters	68
5.1.1. Identifying Significant Short Polypeptide Sequence Pairs.....	68
5.1.2. Clustering Short Polypeptide Sequences.....	70
5.1.3. Post-processing Short Polypeptide Clusters.....	73
5.2. Using Short Polypeptide Clusters to Predict SLGIs.....	75
5.2.1. Cross Validation of SLGIs Based on Short Polypeptide Clusters Using SVM.....	76
5.2.2. Predicting SLGIs Based on Short Polypeptide Clusters Using MLE	78
5.3 Discussion	80
6. CONCLUSION AND FUTURE WORK.....	82
APPENDICES	84
A: Screenshots of www.genenetworks.net	85
REFERENCES	88

LIST OF TABLES

Table	Page
TABLE 2.1: OVERLAP BETWEEN GENETIC INTERACTIONS AND PROTEIN-PROTEIN INTERACTIONS	10
TABLE 3.1: RESULTS OF PROTEIN DOMAIN BASED SVM CLASSIFIER WITH VARIOUS SIZES OF NEGATIVE DATA.....	34
TABLE 3.2: CLASSIFICATION RESULTS OF PROTEIN DOMAIN BASED SVM CLASSIFIER WITH REPLACING MISCLASSIFIED NEGATIVE DATA	35
TABLE 3.3: CLASSIFICATION RESULTS OF PROTEIN DOMAIN BASED SVM CLASSIFIER WITH REMOVING MISCLASSIFIED NEGATIVE DATA	36
TABLE 3.4: PERFORMANCE OF PROTEIN DOMAIN BASED SVM CLASSIFIER WITH NEW DATA FROM SGD.....	37
TABLE 4.1: TOP TEN PAIRS OF DOMAIN GENETIC INTERACTIONS.....	46
TABLE 4.2: NUMBER OF PREDICTED SLGIs BY MLE APPROACH USING DIFFERENT POSTERIOR PROBABILITY THRESHOLDS.....	53
TABLE 4.3: SUMMARY STATISTICS OF DISTRIBUTION OF THE CORRELATION COEFFICIENT BETWEEN THE CELL CYCLE EXPRESSION PROFILES OF GENETIC INTERACTING PROTEINS	54
TABLE 4.4: SUMMARY STATISTICS OF DISTRIBUTION OF THE SIMILARITIES BETWEEN THE GO BIOLOGICAL PROCESSES OF GENETIC INTERACTING PROTEINS	56
TABLE 4.5: SUMMARY STATISTICS OF DISTRIBUTION OF THE SIMILARITIES BETWEEN THE GO CELLULAR COMPONENT OF GENETIC INTERACTING PROTEINS.....	57
TABLE 4.6: SUMMARY STATISTICS OF DISTRIBUTION OF THE SIMILARITIES BETWEEN THE GO MOLECULAR FUNCTION OF GENETIC INTERACTING PROTEINS	58
TABLE 4.7: PREDICTED NEW GENETIC INTERACTIONS WITH PROBABILITY GREATER THAN 0.9	60

List of Tables (Continued)

	Page
TABLE 5.1: THE NUMBER OF PROTEINS COVERED WITH VARIOUS THRESHOLDS OF HMMSEARCH AND POLYPEPTIDE CLUSTER FILTERS.	74
TABLE 5.2: THE NUMBER OF CLUSTERS OBTAINED WITH VARIOUS THRESHOLDS OF HMMSEARCH AND POLYPEPTIDE CLUSTER FILTERS.	75
TABLE 5.3: THE NUMBER OF SLGIs RETAINED WITH VARIOUS THRESHOLDS OF HMMSEARCH AND POLYPEPTIDE CLUSTER FILTERS.	75
TABLE 5.4: PERFORMANCE OF CROSS VALIDATION BY SVM CLASSIFIER BASED ON POLYPEPTIDE CLUSTERS WITH HMMSEARCH THRESHOLD SETTING TO 10^{-10}	77
TABLE 5.5: PERFORMANCE OF CROSS VALIDATION BY SVM CLASSIFIER BASED ON POLYPEPTIDE CLUSTERS WITH HMMSEARCH THRESHOLD SETTING TO 10^{-15}	77
TABLE 5.6: PERFORMANCE OF CROSS VALIDATION BY SVM CLASSIFIER BASED ON POLYPEPTIDE CLUSTERS WITH HMMSEARCH THRESHOLD SETTING TO 10^{-20}	77

LIST OF FIGURES

Figure	Page
2.1: ILLUSTRATION OF SYNTHETIC LETHAL GENETIC INTERACTION.....	9
3.1: ROC CURVE FOR SVM CLASSIFIER IN CROSS VALIDATION STUDY ON TRAINING DATA.....	32
3.2: ROC CURVE OF THE SVM CLASSIFIER IN PREDICTION	33
4.1: DOMAIN GENETIC INTERACTION COULD BE THE MECHANISM UNDERNEATH GENETIC INTERACTIONS	41
4.2: DOMAIN GENETIC NETWORK.....	48
4.3: THE LOG-LOG PLOT OF DEGREE DISTRIBUTION P(K) OF DOMAIN GENETIC INTERACTION NETWORK.....	49
4.4: THE DOMAIN GENETIC INTERACTIONS IN SLGIs BETWEEN MULTI-DOMAIN PROTEINS	50
4.5: THE SLGIs (PROBABILITIES>0.7) RELATED TO DNA REPAIR PROTEINS.	59
4.6: COMPENSATORY PATHWAYS IDENTIFIED FROM PREDICTED SLGIs.....	64
5.1: ILLUSTRATION OF MATCHING BETWEEN SIGNIFICANT SHORT POLYPEPTIDE SEQUENCE PAIRS	69
5.2: MULTIPLE SEQUENCE ALIGNMENT OF A TYPICAL POLYPEPTIDE CLUSTER	72
5.2: THE ROC CURVE OF FOR SVM CLASSIFIER USING POLYPEPTIDE CLUSTERS OBTAINED WITH 10^{-10} AS THE THRESHOLD OF HMMSEARCH AND 50 AS THE THRESHOLD OF POLYPEPTIDE CLUSTER FILTER.....	78
5.3: COMPARISON OF THE ROC CURVES OF MLE METHOD BASED ON SHORT POLYPEPTIDE CLUSTERS AND PROTEIN DOMAINS USING EXPERIMENTAL OBTAINED DATA.....	79
7.1: SEARCH THE PROBABILITY OF PREDICTED GENETIC INTERACTIONS USING DOMAIN GENETIC INTERACTIONS.	86
7.2: SEARCH THE PROBABILITY OF PREDITED DOMAIN GENETIC INTERACTIONS.	87

CHAPTER ONE

INTRODUCTION

1.1. Problem Statement

One of the important tasks in the post-genomic era is defining the functional relationships between proteins. A classical approach to understand gene functional relationships is to produce combinational mutant in two genes and observe genetic interaction [1]. Genetic interaction refers to the phenomenon in which the combined effect of mutations of two genes differs from individual effects of each mutation [2]. In extreme cases, mutations of two nonessential genes could lead to a lethal phenotype. This kind of genetic interaction is referred to as synthetic lethal genetic interaction (SLGI). As one of the major types of genetic interactions, synthetic lethal genetic interactions are of interest to many researchers because of their may be used to identify protein functions and find drug targets or drug combinations.

Genetic interactions can be identified by multiple biological experiments like mutant screens [3] or high throughput technologies such as synthetic genetic arrays (SGA) and synthetic lethal analysis by microarrays (SLAM) [4]. However, even high throughput methods will not be able to screen pair-wise combinations of all proteins in a genome with thousands of genes. Therefore, it is desirable to computationally predict genetic interactions.

Even though the computation-based approaches to predict SLGIs are promising, they are hampered by several issues. First, the molecular mechanism underlying genetic

interactions is not fully understood. This makes it difficult to select features for the prediction of SLGIs and understand how the features are related to SLGIs. Second, the number of positive SLGIs is limited and no reliable negative data is available. Researchers are usually interested in positive data and hence report only positive data. Also as a matter of fact, there are false positives and false negatives among data sets reported. For example, 1234 protein pairs are shared by the SLGIs from SGD (Saccharomyces Genome Database) and the non-genetic interaction dataset of Wong's [7] data. This situation adds difficulty to the prediction of SLGIs. For this reason, random datasets have been widely used to substitute negative datasets. Third, the class distribution of genetic interaction is highly skewed. There are far more non-genetic interactions than genetic interactions, which indicates that the cost of misclassifying interacting pairs differs from that of misclassifying non-interacting pairs. This greatly affects the performance of predicting SLGIs.

Several computational approaches have been proposed for the prediction of SLGIs [5-9] and various features, such as protein interactions, gene expression, functional annotation, gene location, protein network characteristics, and genetic phenotype, have been utilized by these methods. The big problem of previous study is that they all require features from high throughput experiments, which are either not quite available or noisy.

1.2. Dissertation Summary

This dissertation is dedicated to the prediction and understanding of yeast SLGIs. We establish new features based on characteristics of sequences that are usually available

for all organisms. Also, we explore new computational approaches for the prediction of yeast SLGIs.

Representing the structures and function of proteins, protein domains are usually regarded as building blocks of proteins and are conserved during evolution. Many studies have used domain information to predict protein-protein interactions [10-12] based on the assumption that proteins interact through domains. We hypothesize that there is a strong correlation between the protein-domain relationship and genetic interactions. The genetic interaction that occurs between a pair of proteins could be essentially related to or even dominated by the domains underneath. Multiple approaches were derived from this idea in the study. This research project started by predicting SLGIs with domain features using a machine-learning model called Support Vector Machine (SVM). The performance obtained indicated that there is strong correlation between protein domains and genetic interactions. To further investigate how domains contribute to genetic interactions, domain genetic interaction was used to represent the domain pairs contained by the protein pair of a genetic interaction. The significant domain genetic interaction and their probabilities were explored using a maximum likelihood estimation (MLE) algorithm from known SLGIs, and further utilized to predict novel SLGIs from all protein pairs in which the domain genetic interactions reside. We also identified novel compensatory pathways from the predicted SLGIs. Moreover, we retrieved a set of short polypeptide sequence clusters and used them for the prediction of SLGIs.

1.3. Research Contributions

Our study demonstrates a novel idea for the prediction of SLGIs using protein domains. Multiple approaches were derived from this idea (domain features, domain genetic interactions and the short polypeptide clusters) with convincing performance. Our results indicate that there is strong correlation between domain information and genetic interactions and that such correlation can be utilized for the genome-wide prediction of SLGIs. A new concept termed "domain genetic interaction" is defined corresponding to the term "genetic interaction" as there is no physical connection between items described in either term, namely domains in domain genetic interaction and proteins in genetic interaction. The domain genetic interactions help more precisely define the specific function related to different functionalities of multi-domain proteins. This identification of domain genetic interactions helps the understanding of originality of functional relationship in SLGIs at the domain level which can significantly contribute to the further analysis of genetic interaction related studies. Our study significantly improves the understanding of multidomain proteins in yeast, the SLGIs, and the functional relationships between proteins and pathways. It also shows the efficiency of classifying genetic interaction with SVM and MLE by using characteristics of sequences as predictors. Moreover, we created a framework to retrieve short polypeptide sequence clusters in yeast and used them for the prediction of SLGIs. Besides allowing us to predict genetic interactions in a more general method, this part of the work also investigates the potential of finding super families and functional sites with this approach.

Essentially, our study contributes a new view of viewing the underlying mechanism of genetic interactions to the biological community.

1.4. Dissertation Outline

The rest of the dissertation is organized as follows. In chapter 2, the background of genetic interaction and related information are presented, along with a discussion of previous work. In chapter 3, the method and results of cross validation for predicting SLGIs based on domain features using SVM is presented. In chapter 4, an MLE method is introduced to predict SLGIs from significant domain genetic interactions. The MLE problem is solved by expectation maximization (EM) algorithm. In chapter 5, retrieval of polypeptide sequence clusters is introduced to substitute protein domains for the prediction of SLGIs. By providing higher coverage among proteins, this method provides a more general method compared to the protein domain based approach. Finally, conclusions and future work are presented in chapter 6.

CHAPTER TWO

BACKGROUND

2.1. Genetic Interaction

Genetic interaction is a phenomenon in which the combined effect of mutations of two genes differs from individual effects of each mutation [2]. A gene mutation is a change in the DNA sequence in which the gene is composed. Mutations can occur in the range from one single DNA base to a large number of DNA bases. Gene mutations can be inherited from a parent or acquired during an organism's lifetime. Acquired mutations can be caused by environmental factors such as ultraviolet radiation from the sun, or by a mistake made in DNA duplication. Mutations can lead to various types of results: they can have no effect, alter the protein produced by a gene or prevent the gene from functioning. It is also important to understand the robustness of an organism against random mutation.

Genetic interaction can be defined as the difference ϵ between the combined effect of two mutations on the fitness of an organism and the product of their individual effects:

$$\epsilon = W_{ab} - W_a \times W_b$$

The fitness represents the growth rates as decimal fractions. Here, W_a , W_b and W_{ab} represent the fitness of organisms with mutation a , mutation b and double mutants respectively. Non-interacting gene pairs have ϵ close to zero. Genetic interactions have ϵ greater or less than zero. Synergistic gene pairs inclusive of synthetic sick and

synthetic lethal gene pairs have $\epsilon < 0$ and antagonistic gene pairs have $\epsilon > 0$ meaning that the growth rate is higher for double mutants than that of each individual mutation. The synergistic gene pairs are also called aggravating interactions. The antagonistic gene pairs are also called alleviating interactions. The genes involved in synergistic interactions may act at the same step of a pathway or in redundant pathways to enhance each other. One gene of the antagonistic gene pairs maybe defined as suppressor, as the mutation of this gene can correct the phenotypic defects of another mutation.

Differs from protein-protein interaction, genes involved in genetic interaction usually don't have physical connection. From this perspective, genetic interaction can be better understood as a "genetic relation".

Genetic interaction is also termed as epistasis by population geneticists. William Bateson defined epistasis in 1907. It was described as the phenomenon in which one mutation masks or suppresses the effect of another allele at another locus. Classical genetics supports this definition. The population geneticists extended the definition of epistasis by including aggravating interaction in which two mutations together yield an unexpected deleterious phenotype [13]. The definition of epistasis was further extended to encompass alleviating interactions recently.

Genetic interactions can be identified by multiple biological experiments like mutant screens [3]. High throughput technologies such as synthetic genetic arrays (SGA) and synthetic lethal analysis by microarrays (SLAM) [4] have been developed for parallel and massive detection of genetic interactions. As an adaptation of SGA, E-MAP (Epistatic Miniarray Profile) was recently proposed to quantitatively measure both

positive and negative genetic interactions among hundreds of genes. The usage of these quantitative genetic interactions significantly improved the performance of gene function prediction.

2.2. Synthetic Lethal Genetic Interaction

In the extreme cases, mutation of two nonessential genes could lead to a lethal phenotype. This kind of genetic interaction is referred as synthetic lethal genetic interaction (SLGI). Two synthetic lethal genes can either reside on the same pathway or on parallel pathways. These genes can compensate for one another in case of a mutation. This concept of genetic buffering is supported by various studies systematically implementing double mutations of large amount of gene pairs [4]. For this reason, we can say that genetic interactions define overlapping functions and compensatory pathways. The studies also indicate that only a very small percentage (0.5%) of protein pairs have synthetic sick and lethal interactions [5]. Figure 2.1 illustrates one such synthetic lethal genetic interaction.

Figure 2.1 Illustration of synthetic lethal genetic interaction. A) and B) the cell is still alive after knocking out one gene; C) the cell died after knocking out both genes.

A synthetic lethal screen can identify any gene that, if mutated, causes cell death with a nonlethal single mutation. Through the use of synthetic lethal screening, the entire genome of an organism can be scanned to identify mutations in related pathways or proteins with redundant functions.

Synthetic lethal interactions in the yeast, *Saccharomyces Cerevisiae*, have been studied for a long time [9, 14-16] due to its simplicity of having a single cell. Synthetic sick and lethal interactions compose a network that is denser than protein interaction network, but largely non-overlapping over protein pairs. So, genetic and protein interaction networks provides complementary information.

As one of major types of genetic interaction, SLGIs are of interest to many researchers because they can be used for understanding protein functions and have the potential for finding drug target or drug combinations [7, 8, 17, 18]. For example, if we know the function of a protein, the genetic partner of this protein should also have the

same function. Cancer is caused by the loss of functions of some genes, killing the genetic partners of impaired genes can get rid of the cancer cell while keeping the normal cells. SLGI is also the topic of our project because its special characteristic and significant usage.

It is of interest to us to examine the overlap between protein interactions and genetic interactions, especially SLGIs. Hence, we compared our 11145 synthetic lethal interactions with protein-protein interaction datasets of yeast from the Database of Interacting Proteins (DIP) [19] and the Munich Information Center of Protein Sequences (MIPS) [20]. There are 6264 and 7406 protein interactions in the DIP core and the MIPS network datasets, among 2724 and 4146 proteins respectively. The result shows that even though our synthetic lethal interaction dataset shares a large amount (about 50%) of proteins with protein-protein interaction datasets, only a small portion (less than 14%) of protein pairs exists in both synthetic lethal interactions and protein-protein interactions (Table 2.1). It also indicates that a pair of proteins having protein-protein interaction can also serve as a buffer of each other because of the complicated nature of organisms.

Table 2.1 Overlap between genetic interactions and protein-protein interactions

	SLGI	DIP	MIPS
Protein	11145	2724	4146
Protein Pair	2640	6264	7406
Shared Protein	N/A	1610 (59.10%)	1950 (47.03%)
Shared Protein Pair	N/A	356 (13.07%)	227 (5.48%)

In human beings, SLGIs are involved in many complex phenotypes and are the defining basis of multigenetic disease. The study of the yeast SLGIs may help us understand complex human disease. The prediction of human SSL genetic interactions

could be valuable in discovering drug targets for cancer therapy, and in identifying genes that are responsible for multigenic diseases. The implications of synthetic lethal screens have been discussed in the context of drug development as synthetic lethal pairs could be used to selectively kill cancer cells, but leave normal cells relatively unharmed. Reliable prediction of genetic interaction in higher organisms has the potential to impact pharmacies. As tumor suppressor genes are frequently inactivated in cancer, using a drug to inhibit their synthetic sick and lethal partners may selectively impair cancer cells while normal tissue persists [86].

Studies have shown that human diseases caused by Mendelian single gene effects are sensitive to mutations of some other genes. Essentially, the mutations of the genetic interaction partners of disease genes enhance the severity of some Mendelian diseases. Diseases caused by the combination of two mutations are called digenic diseases. The prediction of SLGIs can be used to explore such diseases, as the effect of such diseases could be similar to that of genetic interactions and genetic interactions derived from model organisms could be conserved.

Reliable prediction could also dramatically improve the statistical efficiency of multigenic disease mapping. Testing for association between a disease and all possible combinations of genes requires an unreasonable number of tests, severely diminishing the usage of statistical methods in association studies. The prediction of SLGIs allows a candidate gene pair approach, in which only combinations of genes most likely to genetically interact are tested, hence making multigenic disease mapping more feasible using reduced patient populations [21].

2.3. Protein Domain

A protein domain is the building block of proteins and is evolutionarily conserved. It is part of the protein sequence that can fold, function, evolve and exist independently compared to the rest of a protein chain. Obviously, multiple proteins from either the same organism or different organisms can contain the same protein domain. A protein may contain multiple domains with each has its own function. Identification of protein domains has been conducted for many years using both biological and computational methods. Some available databases of protein domains include Pfam [22], DOMINE [23], InterDOM [24], and ProDOM [25].

Various studies related to protein domains also have been conducted subject to protein-protein interaction. However, no work has been done inferring genetic interactions from domain information. The phenotype produced by the mutation of a gene is caused by the loss of function of its protein product, which is mainly due to the loss of protein domains in the protein product. Then, the effect of the mutation of two genes is caused by the loss of protein domain combinations in both protein products. To the best of our knowledge, we developed the first domain-based study for the prediction of genetic interactions. Our study uses PfamA domains of yeast proteins, as they are mainly referred by the community.

2.4. Classification and Prediction

Classification is the task of finding a model from training data and predicting the class of testing data. The model deduced from training data should be able to characterize and distinguish data classes and concepts. The derived model could be rules,

decision tree or mathematical functions. Classification is also called supervised learning. Training data are usually observations or measurements with known label attributes. Each instance in the training data is accompanied by a label indicating the class that the sample belongs to. Test data are usually samples without predefined classes. Both training data and test data are represented by a series of attributes called features. Features selected for the samples should relate to the classes of samples. The labels predicted by the classifier are usually categorical. For models generating continuous numerical values, the labels can be obtained by discretization. Some typical classification algorithms are decision tree, random forest, support vector machine and naïve Bayesian.

To validate the prediction result, researchers have been trying various approaches. One of the most utilized approaches is cross validation. In cross validation, the sample data is randomly split to n sets. During each iteration, one set is used for testing and the other $n - 1$ sets are used for training. Totally n iterations of training and testing will be conducted in which each set gets a chance to be tested. The result of each iteration will be summed up and averages of all metrics will be calculated to represent the performance. The number of sets can be an arbitrary number. If we split the sample to two sets with one set for training and one set for testing, it is called two-fold cross validation. By applying the same logic, there are 4-fold cross-validation, ten-fold cross-validation and so on.

2.5. Evaluation Criteria

We have employed multiple criteria to evaluate our classification results, which include sensitivity, specificity, precision, F-measure and accuracy. In the content of this paper, the abbreviations TP, TN, FP and FN represent the number of true positive, true negative, false positive and false negative predictions, respectively. The value of true positive is equal to the number of correctly predicted positive data. The value of true negative is equal to the number of correctly predicted negative data. The value of false positive is the number of predicted positive data that actually belong to the negative category. The value of false negative is the number of predicted negative data that actually belong to the positive category.

The sensitivity is defined as the percentage of correctly predicted positive data over the total number of positive data.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Sensitivity is identical to recall in the context of classification. In our study, the sensitivity is the percentage of correctly predicted synthetic lethal interactions over the total number of synthetic lethal interactions in the test dataset, which is the sum of true positive and false negative.

The specificity is defined as the percentage of correctly identified negative data over the total number of negative data, which is the sum of true negative and false positive. It is the percentage of correctly identified non-synthetic lethal interactions over the total number of non-synthetic lethal interactions included in the test dataset in our case.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

The precision is defined as the percentage of correctly predicted positive data over the total number of predicted positive data, which is the sum of true positive and false positive. In our study, the precision is the percentage of correctly predicted synthetic lethal interactions over the total number of predicted interactions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The F-measure is also called F-score. As a weighted average of the precision and recall, it considers both the precision and recall of the test to compute the score. The best F-measure score is 1 and the worst F-measure score is 0.

$$F = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

The accuracy is defined as the percentage of correctly predicted positive and negative data over the sum of positive and negative data. In this paper, the accuracy reflects the percentage of all correctly identified interaction and non-interactions over the size of the test dataset.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

We also use the ROC curve to evaluate the classifier in the cross validation study and the prediction performance. ROC stands for receiver operating characteristics. The ROC curve plots the fraction of true positives against the false positive rate as the threshold of prediction varies. The ROC evaluates the performance of classifiers based

on the tradeoff between specificity and sensitivity. While the ROC curve provides a visualization method of evaluating a classifier, the AUC (area under the curve) score is widely used by providing a numeric value for the comparison of prediction performance. The ROC curve and AUC value serve to evaluate the discriminate power of models, and can be used to select the optimal models. While an AUC value of 1 means perfect prediction, an area of 0.5 indicates random prediction. Most common models should have AUC values between 0.5 and 1.0. The higher the AUC value, the better the model. High AUC value means that lowering the threshold for the prediction only brings in limited false positive samples.

2.6. Related Work

Many computational approaches like decision tree, support vector machine and naïve Bayesian have been proposed for the prediction of genetic interactions especially SLGIs in *Saccharomyces cerevisiae* and *Caenorhabditis elegans* [5-9]. Researchers also have been trying various kinds of features, such as protein-protein interactions, gene expression, functional annotation, gene location, protein network characteristics, and genetic phenotype for the prediction of genetic interaction. As well, multiple approaches inclusive of biological experiments and various kinds of cross validation were used to validate the prediction results. Constrained by its own requirement, only limited groups have the condition to conduct web lab experiments to verify the *in silicon* results.

Wong et al. [7] applied decision tree to genome-wide prediction of synthetic lethal interactions in *Saccharomyces cerevisiae* by integrating multiple features including physical interactions, mRNA expression, functional annotation, gene location and

characteristic of protein network topology. As arguably the most comprehensive work in predicting genetic interaction, the study represents a rigorous demonstration that genetic interactions can be predicted. Most features describe how the protein pair of a genetic interaction behaves in the context. For example, are the genes orthologs? Are the genes located in certain range of distance among the genome? Are they physical interaction partners? Specially, they presented some 2Hop features. The 2Hop features describe the topology around a pair of genes in various networks. For example, if protein *A* physically interacts with protein *C*, and protein *B* genetically interacts with protein *C*, then gene pair *A-B* possesses the 2Hop physical-genetic characteristics. Five types of characteristics were utilized to compose following 2Hop characteristics: H-S, P-S, S-S, S-X, H-H, H-P, H-R, H-X, P-P, P-R and X-X, where the H denotes sequence homology with BLAST E values less than 10^{-3} , the S denotes synthetic sick or lethal, the X denotes correlated mRNA expression with Pearson correlation coefficients greater than 0.7, the R denotes common upstream regulator with P value less than 0.001, the P denotes physical interaction. Decision trees were constructed in the study based on feature vectors of the training data. Only binary characteristics were used to construct the decision trees. While some features are inherently binary, continuous characteristics were mapped to several binary characteristics with alternative thresholds. At the beginning, all gene pairs of the training set were included in the root node. Gene pairs of each node *N* were recursively partitioned into children nodes based on the feature yielding the highest conditional information gain. Each node is partitioned into two children nodes, as all features are binary. For any single node *N*, let $Y_c(t)$ denotes the value of a gene pair for

screened attribute c , all gene pairs of this node is distributed between nodes $N_0 = \{t \in N, Y_c(t) = 0\}$ and $N_1 = \{t \in N, Y_c(t) = 1\}$. The conditional information gain for the class X of the training data was calculated by:

$$H_N(X) - \sum_{a=0,1} \frac{|N_a|}{|N|} H_{N_a}(X)$$

where $H_N(X)$ is the entropy of X at node N . $H_N(X)$ is defined as

$$-p_N \log(p_N) - (1 - p_N) \log(1 - p_N)$$

and P_N is the probability that a gene pair $t \in N$ is synthetic sick and lethal. To avoid overfitting the training data, "early stopping" criterion is enforced based on the Bayesian Information Criterion (Minimum Description Length). A node is made to be a decision node if the maximal conditional information gain doesn't exceed $\frac{\log(|T|)}{2|N|}$. Here, $|N|$ is

the number of gene pairs in node N while $|T|$ is the number of pairs in the entire training set. Instead of assigning a class like positive or negative to the decision node, each leaf received a score equal to the fraction of its genes pairs that were synthetic sick and lethal (SSL). To compensate for small sample size, a pseudo count 1 was added to the number of SSL and non-SSL. The denominator was added by 2. The highest scored protein pairs became the predicted SSL pairs. The reliability of Wong et al. [7] method was demonstrated by both experimental evidence and cross validation. The study demonstrated that genetic interactions could be predicted. The performance of their prediction was assessed by 4-fold cross-validation on 692,865 gene pairs, of which 3868 (0.56%) were SSL. The sensitivity was 80%, while the false positive was 18%. The

decision trees also indicated that the top predictors were 2hop SSL-SSL and 2Hop physical-SSL. These two predictors were consistent with the finding that SSL partners of a gene tend to interact with each other in the genetic network and compensating pathways of gene products. Experimental validation of some SSL predictions on 35,996 new gene pairs was conducted in Wong et al.'s work. The predicted results were also screened by SGA experiments with performance similar to that of cross-validation.

Paladugu et al. [8] proposed an approach based on graph-theoretic properties of two proteins in a protein interaction network as features for the prediction of synthetic sick and lethal genetic interactions. Many studies had been done on using graph properties of protein interaction network to infer gene and protein function in yeast. The rationale behind this idea is that the local connectivity and global position of a protein in a protein interaction network correlate with its functional properties like its essentiality or dispensability. The network properties of two proteins can be considered simultaneously, which possibly could determine their joint dispensability. To abstract the protein interaction network as a graph, nodes were denoted as proteins while edges were represented as their interactions. Correlations between network properties and aspects of biological functions were found in several reports [26, 27]. Examples of network properties include the pair-wise correlation between protein degree, protein centrality, gene essentiality and evolutionary rate. Graph-theoretic data was also combined with other data sources, such as co-expression data and functional annotation of neighboring proteins in the protein interaction network. The principal observation of these studies is that essential proteins tend to be over-represented among proteins with high degrees and

high centralities in a protein interaction network. An essential protein can produce a lethal phenotype when deleted. Some graph properties of a single node in a network used are its degree, clustering coefficient, closeness centrality etc. To construct the feature vector of protein pairs by these single node properties, either average value or the absolute difference were computed. These features along with two-nodes properties like number of mutual neighbors, shared synthetic lethal partner, shared physical interaction partner and inverse of shortest distance were fed into the SVM model for training and testing. Ten-fold cross validation were conducted to evaluate the performance of Paladugu et al.'s [8] SVM classifier. The accuracies without 2Hop features for positive and negative data were 0.7960 and 0.7550 respectively. The overall accuracy with all features was 90.4%. The results indicated the importance of 2Hop features, which was consistent with Wong's work.

Zhong et al. [9] predicted genome-wide genetic interactions in *Caenorhabditis elegans* by constructing a probability models using five features, such as interaction data, gene expression data, phenotype data and functional annotation data from three model organisms: *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*. The positive training data included 1,816 genetic interaction pairs curated from the literature and 2,878 physical interaction pairs from yeast two-hybrid experiments. The negative training data included 3,296 pairs for linked *cis* markers used in genetic mapping experiments. These gene pairs had double mutant scores but were not reported as genetic interactions. The test data set was constructed independent of the training set. 5,515 gene pairs from KEGG pathway database were extracted and 5000

random gene pairs with a minimal distance of six or above in the positive training set. Multiple data sets were collected and categorized to fourteen predictors. GO annotations generated electronically were excluded in order to maintain the data independency since they were converted from the phenotypes. Orthologs of *Caenorhabditis elegans* genes were identified with BLAST, Cluster of Orthologous Groups (COG) and InParanoid. For each gene in a *Caenorhabditis elegans* gene pairs, a group of orthologs in other organisms were identified. The authors used logistic regression and naive Bayesian network as classifiers and achieved similar performance. The first method was naive Bayesian network. This model assumes that all predictors are independent. Thus, the overall likelihood of genetic interaction L is:

$$L = \prod_{i=1}^n L_i$$

where L_i is the likelihood of a feature defined as

$$L_i = \frac{P(v_i | pos)}{P(v_i | neg)}$$

The $P(v_i | pos)$ and $P(v_i | neg)$ are frequencies of gene pairs having the feature v_i in the positives and the negatives respectively. The value of L indicates the strength of predictive power of the features. A value of L greater than 1 means that the feature is enriched in genetic interaction pairs. The other method is logistic regression. This model does not require the data independency. Let O be the odds of a gene pair being an interacting pair for predictor v_i , following relationship L_i between and O can be derived.

$$\begin{aligned}
O &= \frac{P(pos | v_i)}{P(neg | v_i)} \\
&= \frac{P(v_i | pos) \times P(pos) / P(v_i)}{P(v_i | neg) \times P(neg) / P(v_i)} \\
&= L_i \times \frac{P(pos)}{P(neg)}
\end{aligned}$$

where $P(pos)$ and $P(neg)$ are the frequencies of interacting pairs and non-interacting pairs among all possible gene combinations which can be considered constants. Logistic regression uses a weighted sum to integrate scores into overall odds:

$$\ln O = \ln \frac{p}{1-p} = c + \sum_{i=1}^n a_i \ln L_i$$

where c and a_i are constants determined by fitting of the training set and p is the final output. They validated the predicted interactions of two human disease-related genes by wet lab experiments. Excluding untested predictions, the accuracy is 44% for *let-60* and 60% for *itr-1*.

By applying random walks on biological networks, Chipman et al. [5] were able to predict synthetic lethal interactions with both true positive rates and true negative rates of more than 90% in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. Instead of using all the features along, they suggested organizing the features as biological networks. The biological networks came from protein-protein interaction (PPI), Homologs, Co-Expression network, and GO similarities data. Protein pairs were connected if their relationship exceeded certain threshold. Based on the 2Hop idea from Wong et al.'s work, these biological networks were integrated with the genetic

interactions used for training data to compose the feature vector. The 2Hop network motifs capture the relationship between a pair of genes. If genes A and B share a protein interaction, while genes A and C have synthetic lethal genetic interaction, the 2Hop scheme suggests that genes B and C might also be synthetic lethal genetic interaction. Researches also suggest that genes tested to interact physically are more likely to form a stable complex if they share common genetic interactions and hence physical interaction can be used to model genetic interactions. A random walk on undirected graph can be regarded as a time-reversible Markov chain. The usage of random walks theory had been demonstrated in seeds expansion, graph partition and genetic interaction prediction [28], [29], [30], [31]. Provided a graph and a starting point, we can choose a neighbor of the starting point at random and move to this neighbor. We can select a neighbor for this new point and move to it. After perform this process repeated, we can obtain a sequence of points. The sequence of points is a random walk of the graph. The input of the random walk algorithm was the biological network G , a start node s , and restart probability c . The output was the proximity matrix P . This stationary vector could be obtained by either solving the dominant eigenvector or by running equation

$$\vec{p}_s = (1 - c)A\vec{p}_s(V) + c\vec{r}_s(V)$$

iteratively until convergence where A was the column normalized adjacency matrix and $\vec{r}_s(V)$ was the restart vector with value 0 for all of its entries except a 1 for the entry denoted as node s . The random walk method is great for organisms partially mapped or for lesser-studied organisms for which homologous information from a close relative is available. Chipman et al. applied random walks on biological networks especially 2Hop

networks to predict SLGIs. The restart probability c was set to 0.2 in the study. The interaction probability obtained from the random walk algorithm between gene pairs from multiple biological networks were fed into a decision tree for modeling and prediction. By using the probability of each leaf in the decision tree, they were able to generate ROC curves with the threshold varying by a factor of 0.0001 over a range of 0 to 1. This random walks algorithm was applied to both genetic interaction dataset and non-genetic interaction dataset. Proximities between genes of each protein pair were calculated for both cases. These two variables were considered as input to build the classifier that was a decision tree. ROC curves were compared between random walk and Wong et al. 2Hop for *Saccharomyces cerevisiae* and among random walks, Wong et al. 2Hop and Zhong et al. for *Caenorhabditis elegans*. Also, the ROC curves were compared among human orthologs, yeast orthologs, worm interactome, co-expressions, genetic interaction and integrated datasets for *Caenorhabditis elegans*, among GO, PPI, Genetic Interactions, and Integrated Datasets for *Saccharomyces cerevisiae* and among SSL interactions, Non-Interactions and SSL+Non-Interactions for *Caenorhabditis elegans* and *Saccharomyces cerevisiae*. The authors also compared the function of information between random walk and Wong et al. 2Hop. Also, the ROC curves were plotted for prediction of Collins et al. data and Tong et al. data by cross validation. The AUC values were both more than 0.9 that is just a little lower than the composite dataset. The random walk feature works well by using a decision tree classifier, and the performance sustained with other classifiers like logistic regression.

For some reason, E-MAPs contain many missing values. For example, over 34% of interactions in the Chromosome Biology E-MAP were not measured. Recently, Igor et al. [32] proposed to quantitatively predict both the positive and negative genetic interactions not measured in E-MAP based on E-MAP information and other genomic data. The authors also tested multiple methods to predict genetic interaction with cross-validation. By using data from three E-MAPs [33-35], they predicted about 190k genetic interactions. A variety of genomic features like the existence of a physical interaction and co-expression were used as input for the classifiers. These 167 features were divided into four groups. Network group: features based physical and GI networks. Instead of encoding gene pairs sharing any complex as an indicator like previous studies, individual features were constructed for each complex in this work. 15 features were considered in this group. Genomic group: features based on various genomic characteristics. Separate individual features were also defined for each phenotype and localization. 117 features were considered in this group. GSG group: information on genetically similar genes. A set of genes were called GSGs of gene A if the GI profiles are the most similar to those of A among all the genes in the E-MAP. This feature captures measured GIs between A and genes similar to B given the proteins A and B of a genetic interaction. 10 features were considered in this group. GSG-Matrix group: the S-scores between GSGs of each gene in a gene pair. This feature captures measured GIs between genes similar to A and genes similar to B . 25 features were considered in this group. For the task of estimating quantitative S-scores between genes, four regressors and k-nearest-neighbors classifier were compared using various groups and group combinations such as GSG only,

GSG+Matrix, Network+Genomic and combinations of all four groups. These four regressors include linear regression, M5 (a decision tree with regression models at the leaves), least median squared linear regression and Gaussian radial basis function network. The procedure of ten-fold cross-validation was applied while computing the correlation between the predicted and actual S-scores. Linear regression with all features achieved the best performance. Linear regression based on GSD+Matrix features yielded comparable result. It indicates the strong predictive power of these two novel features proposed by the authors. For the prediction of GI classes, five classifiers were evaluated inclusive of naive Bayesian, random forest, J48 decision tree, logistic regression, and discretized linear regression. The diffusion kernel proposed by Qi et al. [6] was also compared. The G^- diffusion kernel based on the number of odd-length paths between the two genes was used to predict negative genetic interactions, and the G^+ diffusion kernel based on the number of even-length paths was used to predict positive genetic interactions. ROC curves and precision recall curves were plotted to assess the performance of classifiers. The area under the precision-recall curve (AUPR) provided a better quantitative assessment of the performance when the data is skewed. The best performance was archived by logistic regression and naive Bayesian classifiers with all the features. Using GSG or GSG+Matrix features obtained near-optimal classification accuracy again as did in quantitative estimation of genetic interactions. The difficulty for prediction of positive interactions was aggravated by the selection of S-score threshold. By using about 4.2% of the interactions, the best recall obtained is 40%, and the precision obtained is 61%. All the features and the logistic regression were used to predict the

RNA E-MAP by ChromBio E-MAP, the accuracy for the prediction of negative interactions was 52.5% while the accuracy for that of positive interactions was 37.5%.

CHAPTER THREE

Predicting Yeast Synthetic Lethal Genetic Interaction Using Domain Features

Multiple computational approaches have been reported for the prediction of genetic interactions [5-9] and various features, such as protein-protein interactions, gene expression, functional annotation, gene location, protein network characteristics, and genetic phenotype, had been utilized by these methods. However, no study has associated protein domains with genetic interactions. Genetic interactions define function relationships between genes while protein domain is the building block of genes. The mutation of a gene causes the loss of function of its protein product, which may accredit to the loss of protein domains in the protein product. We hypothesize that there is strong correlation between protein domain relationship and genetic interaction. We applied support vector machine (SVM) to predict SLGIs in *Saccharomyces cerevisiae* based on domain information in proteins.

3.1. Data Sources

The protein domain data was collected from Pfam (Protein families database) [22]. The Pfam database provides two types of protein family data. PfamA domains are manually curated while PfamB domains are automatically generated. In our study, only PfamA domains are considered. The total number of PfamA domains selected is 2289.

Genetic interactions of yeast were downloaded from the *Saccharomyces* Genome Database (SGD) [36], which were compiled from different biological experiments. Synthetic lethal interaction data set was extracted from the file containing all the genetic

interactions information. Initially, there were totally 14248 pairs of synthetic lethal interactions. We removed protein pairs from study if either protein in the pair does not contain any domain. Eventually we obtained 7435 synthetic lethal interactions among 2000 proteins. We randomly selected half (3717 pairs) of them for training and rest half for testing. Then, we generated equal size of negative pairs for both training and testing. The negative pairs were randomly sampled from those 2000 proteins without overlapping to the positive dataset.

3.2. Feature Encoding

The domain information was widely used to predict protein-protein interactions. We encoded the protein pairs using the method proposed by Chen et al. [37]. Each protein pair is represented by a protein domain feature vector that includes all 2289 unique PfamA domains. Each domain feature has a possible value of 0, 1 or 2 in feature vector. The value is 0 if none of the proteins in the pair contains the domain. The value is 1 if one protein of the protein pair contains the domain. The value is 2 if both proteins of the protein pair contain the domain. The coded data is a big matrix with the number of columns equal to the number of features and number of rows equal to the number of samples.

3.3. SVM Classifier

The Support Vector Machine (SVM) [38] is a computational method with learning capability. The goal of the SVM is to generate a model that can be used to predict class labels of data instances in the test data set with attributes only. The SVM creates a hyperplane that best separates the data sets from different classes. It does this

by maximizing the distance between the sets of data from the hyperplane. A buffer called "soft margin" is used to allow certain errors during the training stage in order to avoid overfitting. Currently, various implementations of SVM are available. Some frequently mentioned implementations in the community include LibSVM [39], SVMlight [40] and TinySVM [41]. SVMs can only handle two-class classification problems. To solve a multi-class problem, one can train a SVM model for each class. The positive training data can be selected from instances belong to that class and negative training data can be selected from instances not belong to that class. All the SVM models classify new input and the prediction with furthest distance to the margin is accepted.

The prediction of synthetic lethal interaction can be formulated as a two-class classification problem. A protein pair is either synthetic lethal or non-synthetic lethal. We use 1 to represent synthetic lethal class and 0 to represent non-synthetic lethal class. Support Vector Machine (SVM) is a widely used method for the binary classification problem. We chose the LibSVM tool provided by Chang et al. [42]. The C-Support Vector Classification in LibSVM is utilized in our application.

As there is no trace that indicates that our data can be linearly separated, the Gaussian radial basis function (RBF) kernel is recommended for the SVM. The RBF can be formally described as:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$$

The kernel function is also used in the decision function to predict unknown instances. This kernel function calculates the similarity between two vectors. If the similarity between the unknown vector and the support vector is very small, the value

returned by the decision function will be curved to the side represented by the support vector.

There are two parameters considered by RBF kernel. The C parameter (cost) denotes the penalty enforced to the support vectors. A small cost would result in high misclassification rate. The other parameter γ (gamma) denotes the shape of the decision boundaries. A greater gamma value produces highly irregular decision boundaries in which case the model adapts perfectly to the training data while the model may not have good results for unseen data samples as it is too adapted to the training samples. A smaller gamma value can be used to avoid noise and overfitting problem brought by greater gamma value.

We used the radial basis function kernel. The parameters (cost and gamma) were estimated with five-fold cross validation by following the guide of LibSVM [42], which were selected based on the highest accuracy that the trained SVM classifier achieved. In our study, the selected cost and gamma is 8 and 0.125 respectively.

3.4. Classification Results

We conducted multiple experiments for this study inclusive of cross validation on training data, cross validation with both training data and testing data. As well, we tried to oversample negative dataset and update negative dataset to improve the performance. The results were also reported in our conference paper [43] in 2009.

3.4.1. Cross Validation Study

To evaluate the performance of SVM, we have carried a five-fold cross-validation study on the training data for the SVM classifier. Among all these protein pairs of the

training data, interaction pairs were randomly divided to five groups and same number of non-interaction pairs were randomly assigned to each group. Eventually five groups with equal number of interaction pairs and non-interaction pairs were generated. The SVM classifier was trained and tested for five times. During each run, four groups were selected as training data while the remaining groups served as testing data. These five groups took turns to be tested and eventually each group tested once. The ROC curve was generated based on the cross validation results (Figure 3.1). The AUC of our five-fold cross validation study is 0.925 that indicates the remarkable performance of the SVM classifier.

Figure 3.1 ROC curve for SVM classifier in cross validation study on training data

3.4.2. Prediction Performance

After training the SVM classifier with our training dataset, we applied the model to classify the test dataset. We achieved 84.4% in accuracy, 85.61% in sensitivity, 83.19% in specificity, 83.58% in precision and 84.58% in F-score. Table 3.1 (negative data size 1) shows the result in detail. These results indicated that using domain information alone could achieve high performance in the prediction of synthetic lethal genetic interactions.

The classifier was also tested with ROC analysis based on the prediction results. The AUC value reached 0.9272 that is even higher than the AUC of cross validation study on training data. The ROC curve is showed in Figure 3.2.

Figure 3.2 ROC curve of the SVM classifier in prediction

Table 3.1 Results of protein domain based SVM classifier with various sizes of negative data

	1	2	4	10
TP	3182	2770	2234	1771
FP	625	635	643	796
TN	3092	6799	14225	36374
FN	535	947	1483	1946
Sensitivity	85.61%	74.52%	60.10%	48.00%
Specificity	83.19%	91.46%	95.68%	98.00%
Precision	83.58%	81.35%	77.65%	69.00%
Accuracy	84.40%	85.81%	88.56%	93.00%
F-score	84.58%	77.79%	67.76%	56.62%

3.4.3. Oversampling the Negative Dataset

Because the number of non-interaction protein pairs is much higher than the number of genetic interaction protein pairs, we tested the effect of the size of negative data on our classification results. We trained our SVM classifier with various sizes of the negative dataset. The sizes of selected negative dataset were 2, 4 and 10 times of the size of positive dataset in our experiment. The result showed that the more the negative data, the better the specificity (Table 3.1). However, the sensitivity decreased along with the increasing of negative data size. Meanwhile, the result showed that the accuracy increases along with the increasing of negative data size due to the increasing of specificity. The F-scores, which combines the effect of specificity and sensitivity, decreased along the increasing of negative data size.

3.4.4. Updating the Negative Dataset

As the negative data were randomly sampled from non-positive data, it is possible that some protein pairs in the negative dataset could be positive data. In order to improve

the performance of our SVM model, we examined two approaches to update the negative dataset to reduce the possibility of including positive data in it. The first approach was to replace misclassified negative protein pairs with new randomly generated protein pairs during the training process. After each run of training and testing, misclassified non-interaction protein pairs in the training dataset were replaced with new randomly generated non-interaction protein pairs. Parameters for the SVM classifier were also re-estimated at each iteration step. After three iterations, the change of F-score became trivial. The prediction results for the test data reaches 85.18% in accuracy, 88.35% in sensitivity, 82.00% in specificity and 83.08% in precision. The improvement of accuracy and sensitivity indicate that by replacing misclassified negative data in the training dataset, the classifier is capable of better modeling the real genetic interactions (see Table 3.2 for detailed results).

Table 3.2 Classification results of protein domain based SVM classifier with replacing misclassified negative data

	TP	FP	TN	FN	Sensitivity	Specificity	Precision	Accuracy	F-score
Iteration 0	3182	625	3092	535	85.61%	83.19%	83.58%	84.40%	84.58%
Iteration 1	3264	653	3064	453	87.81%	82.43%	83.33%	85.12%	85.51%
Iteration 2	3284	669	3048	433	88.35%	82.00%	83.08%	85.18%	85.63%

The second approach is to remove misclassified negative data. After each run of training, misclassified negative protein pairs were removed from the training data while the test data remains the same all the time. At the initial stage, the number of negative protein pairs is equal to the size of the positive protein pairs for both the training data and test data. After three iterations, the change of F-score also became trivial. The prediction

result showed performance improvement as the test data reaches 85.01% in accuracy, 88.43% in sensitivity, 81.57% in specificity and 82.75% in precision (Table 3.3). Along with the iterations, the number of negative data decreased from 3717 to 3642 in the training dataset that means seventy-five non-interaction protein pairs were removed after three iterations.

Table 3.3 Classification results of protein domain based SVM classifier with removing misclassified negative data

	TP	FP	TN	FN	Sensitivity	Specificity	Precision	Accuracy	F-score
Iteration 0	3182	625	3092	535	85.61%	83.19%	83.58%	84.40%	84.58%
Iteration 1	3284	684	3033	433	88.35%	81.60%	82.76%	84.97%	85.46%
Iteration 2	3287	685	3032	430	88.43%	81.57%	82.75%	85.01%	85.50%

3.5. Updated Results

We obtained a newer version of SLGIs from Saccharomyces Genome Database (SGD) (July 2009 version) recently and kept the same Pfam domain data since there is no update from Pfam database. There are 7475 SLGIs with PfamA domains in this new data set which 40 more than it was in the old version. We used the same approach to build the training and testing data sets. For example, half SLGIs were used for training and the other half SLGIs were used for testing. Also, same amount of random protein pairs were used as negative data set. We re-ran the SVM classification based method on the new data set 10 times and summarized the average performance with standard deviation in Table 3.4. We didn't oversample the negative dataset, nor did we update the negative dataset for this new study. The performance archived by this new data set is better than that of the old data set in all criteria. The comparison indicates the feasibility of our

method with more accurate GI data set available from biological community. It also reveals the potential of improvement in the performance with more accurate training data in the future.

Table 3.4 Performance of protein domain based SVM classifier with new data from SGD

	Sensitivity	Specificity	Precision	Accuracy
Average	0.8802	0.8409	0.8476	0.8583
(Std)	(0.0059)	(0.0045)	(0.0034)	(0.0028)

3.6. Discussion

Our novel approach of genetic interaction prediction based on domain information achieves high sensitivity (88.35%) and specificity (82.00%). This study clearly demonstrated that there is strong correlation between protein domain relationship and genetic interactions, especially synthetic lethal genetic interactions. Besides serving as features for the prediction of protein-protein interactions which is demonstrated by other studies, domain information can also be used to predict genetic interactions as indicated by our results. As the building blocks of proteins, we believe the essential status of protein domains decides their importance in the prediction of both protein and genetic interactions.

As the uncertainty of negative data, we have employed two approaches, replacing misclassified negative data and removing misclassified negative data, to improve the performance of SVM classifier. Both approaches have improved the accuracy of the prediction. The replacing misclassified negative data approach with new randomly

generated negative data improves the accuracy of prediction from 84.40% to 85.18% with certain increase in sensitivity and little decrease in specificity. The removing misclassified negative data approach improves the accuracy from 84.40% to 85.01%.

We also examined the effect of oversampling the non-interaction pairs. Although increasing the size of negative data improved the specificity and the accuracy, it dramatically reduced the sensitivity of the prediction. One possible reason is that the parameters of the classifier are obtained by optimizing the accuracy of the classifier. Our results imply that this approach did not handle the problem of unbalanced dataset well.

CHAPTER FOUR

Predicting Yeast Synthetic Lethal Genetic Interactions Using Domain Genetic Interactions

In previous chapter, we performed a cross validation study on predicting SLGIs using protein domains as features. Our support vector machine (SVM) classifiers were able to achieve high performance with AUC (The area under the ROC curve) score as 0.9272. These results suggested that using domain information might catch the genetic relationships between proteins. However, we still don't understand the mechanism between protein domains and SLGIs. Here, we define a new concept called domain genetic interaction. Representing the structures and functions of proteins, protein domains are usually regarded as building blocks of proteins and are conserved during evolution. The mutation of a gene causes the loss of function of its protein product, which may accredit to the loss of protein domains in the protein product. Then, the effect of the mutation of two genes may be caused by the loss of protein domain combinations in both protein products. We refer the phenomenon in which combined effect of mutations of two domains in two proteins differs from individual effects of mutation of each domain as domain genetic interactions. The domain genetic interactions may correlate to SLGIs. We may use the domain genetic interactions to explain and predict the SLGIs. Furthermore, in multidomain proteins, different domains may fulfill different functions independently or collaboratively. Although genetic interaction analysis provides a promising method to understand the functional relationship between proteins, we cannot tell the contributions of different domains to certain functionality of

multidomain proteins based on their genetic interactions. Studying the domain genetic interactions may help elucidate the SLGIs between multidomain proteins at domain level. Recently, the MLE method had been used to infer domain physical interactions from protein physical interactions [10, 45]. To further explore the study, we applied the Maximum Likelihood estimation (MLE) approach to estimate the probabilities of domain genetic interactions from yeast SLGIs. The EM algorithm developed by Dempster et al. [44] had been used to iteratively compute the maximum likelihood. We identified significant domain genetic interactions, which rarely overlap with the domain physical interactions from iPfam database [46]. We also showed that domains in multidomain yeast proteins contribute to their genetic interactions differently. Then, we were able to predict novel yeast SLGIs using probabilities of domain genetic interactions, which demonstrate the ability of protein domains in predicting SLGIs. We were also able to apply our predicted yeast SLGIs to understand the compensatory pathways. A complete description of our results and methods is given in the sections below.

4.1. Estimating Probabilities of Domain Genetic Interactions with EM Algorithm

We believe there are some significant domain pairs underneath the synthetic lethal genetic interactions as showing in (Figure 4.1). By discovering these significant domain pairs, we can try to predict new SLGIs from all protein pairs including these domain pairs. We name these significant domain pairs as significant domain genetic interactions. We tried to identify these significant domain genetic interactions with expectation maximization (EM) algorithm. The probability of a domain pair indicates its propensity to genetically interact. The high probability of a domain pair imply that the

number of SLGIs including the domain interaction is high comparing to the number of protein pairs including the domain interaction that are not genetically interacted. Maximum likelihood method has been previously explored by Deng et al. [10] to inferring domain interactions from yeast two-hybrid protein interaction data.

Figure 4.1 Domain genetic interaction could be the mechanism underneath genetic interactions

We treated the protein genetic interactions $P_{m,n}$ and domain genetic interaction $D_{i,j}$ as random variables. The $P_{m,n}=1$ if two proteins i and j genetically interact and $P_{m,n}=0$ otherwise. The $D_{i,j}=1$ if two domain i and j genetically interact and $D_{i,j}=0$ otherwise. We estimated the probabilities of potential domain genetic interactions $\Pr(D_{i,j}=1)$ by maximizing the likelihood of observed genetic interactions using the Expectation-

Maximization (EM) algorithm [10, 44, 45]. The EM algorithm iteratively estimates the maximum likelihood of the ‘complete data’ that combine the observed data and unobserved data. Here, the protein genetic interactions and the domain information of proteins are our observed data and the domain genetic interactions are our unobserved data. By assuming the independences among domain genetic interactions, the likelihood for the observation of protein genetic interactions based on domain genetic interactions can be obtained as:

$$L = \prod_{i,j} \Pr(D_{i,j} = 1)^{M_{i,j}+a} (1 - \Pr(D_{i,j} = 1))^{N_{i,j}+K_{i,j}+b}$$

where $M_{i,j}$ be the number of genetic interacting pairs between domain i and j in all protein genetic interactions; $N_{i,j}$ be the number of non genetic interacting domain pairs between domain i and j in protein genetic interactions; and $K_{i,j}$ be the number of non genetic interacting protein pairs including domain i in one protein and j in the other one. The value of $K_{i,j}$ is counted from all possible protein pairs with domain i in one protein and j in the other one with excluding the known genetic interacting protein pairs. The $K_{i,j}$ will keep unchanged during EM computation. The a and b are pseudo counts to avoid the $\Pr(D_{i,j} = 1)$ or $\Pr(D_{i,j} = 0)$ to be zero when instances of domains i and j are rare. We set both a and b to 1 in our calculation. Initially, the $M_{i,j}$ was set to the number of genetic interactions between domain i and j in experimental genetic interactions; $N_{i,j}$ is set to 0. And $\Pr(D_{i,j} = 1)$ was initialized as following:

$$\Pr(D_{i,j} = 1) = \frac{M_{i,j}}{M_{i,j} + N_{i,j} + K_{i,j}}$$

In each Expectation step of EM algorithm, we first estimated the expected values of $E[M_{i,j}]$ and $E[N_{i,j}]$ using the current $\Pr(D_{i,j}=1)$:

$$E[M_{i,j}] = \sum_{m,n} \frac{\Pr(D_{i,j} = 1)}{1 - \prod_{i \in D(m), j \in D(n)} (1 - \Pr(D_{i,j} = 1))}$$

$$E[N_{i,j}] = \sum_{m,n} \left(1 - \frac{\Pr(D_{i,j} = 1)}{1 - \prod_{i \in D(m), j \in D(n)} (1 - \Pr(D_{i,j} = 1))}\right)$$

Then, we calculate the $\Pr(D_{i,j}=1)$ using the $E[M_{i,j}]$ and $E[N_{i,j}]$ as following (Maximization step):

$$\Pr(D_{i,j} = 1) = \frac{E[M_{i,j}] + a}{E[M_{i,j}] + E[N_{i,j}] + K_{i,j} + a + b}$$

The EM algorithm was iterated the above Expectation and Maximization steps till the change of likelihood L is less than a pre-defined small value.

Then we calculated an evidence score $E_{i,j}$ [45] for each domain genetic interaction to estimate its relative significance. The evidence score $E_{i,j}$ of domain pair i and j is defined as the ratio between the probability that a pair of proteins, m and n , genetically interact given that the pair of domains, i and j , genetically interact and the probability that a pair of proteins, m and n , genetically interact given that the pair of domains, i and j , do not genetically interact:

$$E_{i,j} = \sum_{\substack{m:i \in D(m) \\ n:j \in D(n)}} \log \frac{1 - \prod_{k \in D(m), l \in D(n)} (1 - \Pr(D_{k,l} = 1))}{1 - \prod_{k \in D(m), l \in D(n)} (1 - \Pr(D_{k,l}^{i,j} = 1))}$$

The $\overline{\Pr(D_{k,l}^{i,j} = 1)}$ denotes the probability of genetic interacting between domains k and l given that the domains i and j do not genetically interact.

Essentially, the numerator computes the probability that domain pair i and j is related; the denominator computes the probability that domain i and j is not related. The E score is calculated based on the likelihood of with and without the domain pair. More genetic interactions with domain pair i and j can lead to higher E_{ij} . This feature is consistent with the intuition that more observation of a domain pair should increase the confidence of that domain genetic interaction. The E scores are not the standard likelihood ratio test as they are calculated using only the positive data [45]. The E scores indicate the relative significances of the domain pairs in the SLGIs. The probability and the E score give different view about the significances of domain genetic interactions. The high probabilities do not always correlate to high E score. For example, there are a large number of protein pairs including a single domain pair. If very few of them are SLGIs, the probability of the domain pair will be low. However, if that domain pair is the only domain pair in those SLGIs, it will have high E score. On the other hand, in case there are a large number of protein pairs containing the same multiple domain pairs, if most of the protein pairs are SLGIs, the probabilities of those domain pairs will be high. However, the E scores of those domain pairs will be relatively low as removing any of those domain pairs will have limited effect on the probabilities of SLGIs. Therefore, we used both the probability and E score to select significant domain genetic interactions.

We first applied the MLE approach to the 7475 genetic interactions with both proteins having protein domains. We were able to assign probabilities and E scores to

11,189 domain pairs. Those domain genetic interactions are available on our website, www.genenetworks.net, for searching and downloading as the screenshots (Figure 7.1, Figure 7.2) show in the appendix. We first selected the significant domain genetic interactions with E scores greater than or equal to 2.0, which corresponds to an approximate seven fold drop of the probability of SLGIs if this domain genetic interactions is excluded. Then, we selected the domain genetic interactions with probability greater than 0.5 even though they have low E score values. Totally, we obtained 3848 domain genetic interactions of 1027 domains. Table 4.1 lists top ten domain genetic interactions with highest E scores. It showed that the prefoldin domain dominates in the top ten domain genetic interactions (7 of 10). Our results are consistent with the findings of Meur and Gentlman [47], in which they showed that the prefoldin complex is in 9 of their top 10 pairs synthetic multi-protein complexes genetic interactions.

Table 4.1 Top ten pairs of domain genetic interactions

E score	Prob.	Domain_a	Domain_a Pfam ID	# of Yeast proteins with domain_a	Domain_b	Domain_b Pfam ID	# of Yeast proteins with domain_b	# of GI
67.48	0.46	PF00225	Kinesin	6	PF01920	Prefoldin_2	4	11
65.59	0.72	PF01302	CAP_GLY	4	PF01920	Prefoldin_2	4	12
63.31	0.32	PF00022	Actin	9	PF01920	Prefoldin_2	4	11
41.73	0.025	PF00071	Ras	24	PF00071	Ras	24	13
39.53	0.036	PF00400	WD40	87	PF01920	Prefoldin_2	4	14
38.41	0.12	PF00071	Ras	24	PF04893	Yip1	3	8
36.53	0.027	PF00069	Pkinase	114	PF01920	Prefoldin_2	4	17
35.15	0.35	PF00225	Kinesin	6	PF02996	Prefoldin	3	6
33.16	0.8	PF01920	Prefoldin_2	4	PF03114	BAR	2	7
31.20	0.013	PF00022	Actin	9	PF00069	Pkinase	114	17

4.2. Domain Genetic Interactions Rarely Overlap Domain Physical Interactions

To investigate the relationship between domain genetic interactions and the domain physical interactions, we compared the 3848 significant domain genetic interactions with the domain physical interactions from the iPfam database [46]. There are 4030 domain physical interactions of 1867 domains in iPfam database (2008 version). Among them, 1556 domain physical interactions of 1048 domains exist in yeast proteins we studied. There are 422 domains in both the genetic and physical interactions. We applied the Fisher's exact test in R to examine if domains participating genetic interactions are independent from domains participating in physical interactions. We obtained a P-value of 0.00025, which indicated a significant difference between two sets of protein domains. Furthermore, there are only 70 domain pairs overlapped between the

1556 domain physical interactions and 3848 domain genetic interactions. The Fisher's exact test obtained a p-value of 0, which showed that the domain genetic interactions significantly differ from the domain physical interactions. Thus, the domain genetic interaction is a new type of relationship among domains and can provide a complementary view about the relationships between domains.

4.3. Domain Genetic Interaction Network

To obtain an overview of the domain genetic interactions, we modeled the domain genetic interactions as a network (Figure 4.2), in which each node represents a domain and each link represents a genetic interaction between two domains. Then, we examined the properties of this domain genetic interaction network. The average connectivity of the network is 7.392. The domain PF00022 (Actin) has the highest connections of 186. The average shortest path among all nodes is 3.31. Comparing the size of nodes, this number is very small. Furthermore, the average node clustering coefficient of the network is 0.159. These properties indicated that the domain genetic interaction network has the small world property [48].

The analysis of the connectivity distribution of this network showed a power-law distribution with an exponent degree of 1.45 (Figure 4.3) 1), which indicated that the domain genetic interaction network is a scale free network [49]. These results demonstrated that the domain genetic interaction network follows the common principles of biological networks [88].

Figure 4.2 Domain genetic network

Figure 4.3 The log-log plot of degree distribution $P(K)$ of domain genetic interaction network. The linear characteristics indicated by red line ($y=x^{-1.45}$) imply that $P(K)$ follows a power law.

4.4. Domain Genetic Interactions in the SLGIs between Multi-domain Proteins

Most proteins are multidomain proteins, which are created as a result of different genetic events, such as insertions and duplications [50, 51]. Multidomain proteins may have different functionalities due to different domains. Our identification of domain genetic interactions helps understand how the domains contribute to functionality defined by the SLGIs, and then help elucidate the functional relationships between proteins at domain level from their genetic interactions.

Figure 4.4 The domain genetic interactions in SLGIs between multi-domain proteins. The thick lines indicate significant domain genetic interactions. The thin lines indicate the domain genetic interactions with low probability. The probabilities of domain genetic interactions are labeled besides the line. Those examples show different domain genetic interaction architectures in SLGIs between yeast multidomaon proteins.

Figure 4.4 shows examples of domain genetic interactions in three SLGIs. Figure 4.4A shows the domain genetic interactions between SGS1 and TOP1. The SGS1 has three domains and TOP1 has two domains. Only one of the three domains of SGS1, PF00570 (HRDC domain), has high probabilities to interact with two domains of TOP1: PF01028 (Eukaryotic DNA topoisomerase I catalytic core domain) and PF02919 (Eukaryotic DNA topoisomerase I DNA binding domain). The other two domains: PF00270 (DEAD/DEAH domain) and PF00271 (Helicase conserved C-terminal domain) of SGS1 have no genetic interaction with two domains of TOP1 (very low probabilities). Figure 4.4B shows the domain genetic interactions between TOP3 and TOP1. Both domains of TOP3: PF01131 (DNA topoisomerase domain) and PF01751 (Toprim domain) show high probabilities of genetic interaction with two domains of TOP1. Figure 4.4C shows the domain genetic interactions between RAD5 and RAD50. The RAD5 has four domains and RAD50 has two domains. The genetic interaction between RAD5 and RAD50 is mainly due to a single domain genetic interaction between PF08797 (HIRAN domain) of RAD5 and PF04423 (zinc hook domain) of RAD50. The PF00097 (Zinc finger domain), PF00176 (SNF2 family N-terminal domain) and PF00271 (Helicase conserved C-terminal domain) of RAD5 and PF02463 (RecF/RecN/SMC N terminal domain) of RAD50 have low contribution to the genetic interaction between RAD5 and RAD50. Those examples imply different domain genetic interaction architectures in SLGIs between yeast multidomaon proteins. The domain genetic interactions may exist between all domains of two proteins (TOP3 and TOP1), or between part of domains of one protein and all domains of the other protein (SGS1 and

TOP1), or between part of domains of one protein and part of domains of the other protein (RAD5 and RAD50). The domain genetic interactions are able to help understand functional relationships between multidomain proteins at domain level.

We also investigated the domain genetic interactions of SLGIs between SGS1 and other proteins. We found that only the PF00570 (HRDC domain) of SGS1 has significant genetic interactions with other domains. The results implied that certain functionality of SGS1 may be only due to its HRDC domain, rather than its DEAD/DEAH domain and helicase conserved Cterminal domain. Previous study showed that the HRDC domain of SGS1 is required for its cellular functions involving topoisomerases [52]. Thus, the domain genetic interaction analysis can help understand how domains contribute to the different functionalities of multidomain proteins.

4.5. Prediction and Validation of Genome-wide SLGIs Using Domain Genetic Interactions

Having established that there is a strong correlation between domain genetic interactions and SLGIs, we explored to predict the probabilities of protein pairs to be SLGIs using probabilities of domain genetic interactions calculated from MLE. We assumed that two proteins genetically interact ($P_{m,n}=1$) if and only if at least one domain pair from the two proteins genetically interact ($D_{i,j}=1$). Then, we calculated the probability of two proteins genetically interacting $\Pr(P_{m,n}=1)$ as following:

$$\Pr(P_{m,n} = 1) = 1.0 - \prod_{i \in D(m), j \in D(n)} (1 - \Pr(D_{i,j} = 1))$$

We were able to assign 599,752 protein pairs with probabilities greater than 0. Table 4.2 lists the number of SLGIs predicted by MLE approach at different probability cutoffs. All predicted SLGIs are hosted on our website for searching and downloading.

Table 4.2 Number of predicted SLGIs by MLE approach using different posterior probability thresholds

Threshold	Number of SLGIs
0.9	82
0.8	383
0.7	792
0.6	1821
0.5	3267
0.4	5143
0.3	7767
0.2	14096
0.1	32881
0.09	35702
0.08	39785
0.07	44989
0.06	51564
0.05	60807
0.04	77050
0.03	94460
0.02	131199
0.01	205758
0	599752

Table 4.3 Summary statistics of distribution of the correlation coefficient between the cell cycle expression profiles of genetic interacting proteins

Threshold	T-score (vs. original GI)	P value (vs. original GI)	T-score (vs. All pairs)	P value (vs. All pairs)
0.95	-1.2311	0.2292	-0.0525	0.9586
0.9	-2.0549	0.0442	-0.3759	0.7083
0.85	-1.1770	0.2417	0.9884	0.3251
0.8	0.5674	0.5710	3.8704	1.32E-04
0.75	0.2503	0.8024	4.3109	1.94E-05
0.7	0.3446	0.7310	5.0998	4.35E-07
0.65	1.6739	0.0943	8.0533	1.79E-15
0.6	1.4894	0.1370	8.6120	1.63E-17
0.55	1.3339	0.1824	8.9319	9.58E-19
0.5	0.7308	0.4650	10.1916	5.34E-24
0.45	0.6319	0.5275	10.6774	3.35E-26
0.4	0.4825	0.6290	12.2157	8.43E-34
0.35	0.0290	0.9769	12.8720	2.16E-37
0.3	-0.7195	0.4720	13.4718	7.30E-41
0.25	-2.3167	0.0205	13.4981	3.80E-41
0.2	-3.6548	2.58E-04	14.1176	6.34E-45
0.15	-4.6716	3.01E-06	15.7401	1.85E-55
0.1	-7.2015	6.20E-13	15.9569	4.35E-57
0.05	-9.6593	5.26E-22	16.5797	1.36E-61
0	-13.9797	5.11E-44	24.3758	3.57E-131

We then compared the correlation coefficients of gene expressions of predicted SLGIs to those of known SLGIs and those of all possible protein pairs. We used a yeast cell cycle gene expression data [53], which contains 77 data points. We calculated the T-score and P-value for the null hypothesis that there is no difference between the means of predicted SLGIs and the means of known SLGIs and the null hypothesis that there is no difference between the means of predicted SLGIs and the means of all pairs. The results are shown in Table 4.3. The correlation coefficients of gene expression of predicted SLGIs with threshold greater than 0.3 are similar to those of known SLGIs. The gene expression correlation coefficients of predicted SLGIs using different thresholds

significantly differ from those of all pairs except for SLGIs with probabilities greater than 0.85, which have only small numbers of SLGIs. Those results indicated that the correlation coefficients of gene expressions of predicted SLGIs are similar to those of known SLGIs, rather than to those of random pairs. Recently, it was reported that the SLGIs are likely to have similar GO annotations [54]. We studied the distribution of similarities of Gene Ontology (GO) annotations between predicted SLGIs and also compared them to those of known SLGIs and all possible protein pairs. As shown in Table 4.4, Table 4.5 and Table 4.6, the mean similarities of GO annotations of predicted SLGIs always significantly differ from those of all pairs. At certain probability thresholds, the mean similarities of GO annotations of predicted SLGIs show no significant differences from those of known SLGIs. Those thresholds are 0.35 for biological process and cellular component and 0.25 for molecular function. As the probability thresholds increase, the mean similarities of GO annotations also increase, which will make them differ from those of known SLGIs. The studies of GO annotations similarities and expression correlation coefficients showed that the predicted SLGIs at probability threshold around 0.3 are similar to experimentally obtained SLGIs.

Table 4.4 Summary statistics of distribution of the similarities between the GO biological processes of genetic interacting proteins

Threshold	T-score (vs. original GI)	P value (vs. original GI)	T-score (vs. All pairs)	P value (vs. All pairs)
0.95	2.8616	8.18E-03	7.0986	1.54E-07
0.9	4.1388	1.09E-04	9.6944	6.83E-14
0.85	4.5079	1.58E-05	11.5580	7.19E-21
0.8	3.5569	4.27E-04	15.3384	1.62E-40
0.75	5.1504	3.51E-07	20.2872	6.66E-69
0.7	4.6786	3.36E-06	22.3357	5.43E-85
0.65	6.2352	5.66E-10	29.9002	4.46E-152
0.6	5.4971	4.28E-08	31.8686	9.99E-177
0.55	5.0716	4.21E-07	33.2063	1.48E-193
0.5	2.8398	4.53E-03	39.0308	5.67E-272
0.45	1.8511	0.064	40.8923	1.03E-300
0.4	0.5845	0.559	46.6709	0.00
0.35	-0.6147	0.539	50.3936	0.00
0.3	-3.5334	4.11E-04	53.5202	0.00
0.25	-6.3149	2.76E-10	60.1044	0.00
0.2	-11.6473	2.92E-31	63.5508	0.00
0.15	-16.2162	8.97E-59	70.9904	0.00
0.1	-23.2261	1.81E-117	82.4446	0.00
0.05	-31.1076	2.91E-205	99.6669	0.00
0	-49.0786	0.00	185.9402	0.00

Note: the Resnik method is used to calculate the similarity between the biological process terms

Table 4.5 Summary statistics of distribution of the similarities between the GO cellular component of genetic interacting proteins

Threshold	T-score (vs. original GI)	P value (vs. original GI)	T-score (vs. All pairs)	P value (vs. All pairs)
0.95	2.4239	0.0226	5.4338	1.07E-05
0.9	4.0408	1.53E-04	7.6053	2.31E-10
0.85	4.9111	3.03E-06	9.8846	5.60E-17
0.8	4.8803	1.62E-06	13.7995	1.64E-34
0.75	5.5758	3.74E-08	17.6206	1.89E-55
0.7	5.6365	2.39E-08	20.1136	5.25E-72
0.65	6.9807	4.23E-12	26.7665	3.22E-127
0.6	5.4308	6.21E-08	27.8059	2.90E-141
0.55	4.8200	1.52E-06	29.0693	2.93E-155
0.5	2.8470	4.43E-03	33.9117	1.57E-214
0.45	2.0025	0.0453	35.5617	3.54E-237
0.4	0.2845	0.7760	39.9349	3.08E-301
0.35	-1.7901	0.0735	41.7910	0.00
0.3	-4.7863	1.71E-06	44.2991	0.00
0.25	-8.2880	1.22E-16	48.8099	0.00
0.2	-12.9517	3.19E-38	53.1532	0.00
0.15	-16.7738	1.04E-62	59.9635	0.00
0.1	-24.262	1.07E-127	67.8925	0.00
0.05	-32.1225	6.40E-218	77.9517	0.00
0	-55.8164	0.00	43.4725	0.00

Note: the Resnik method is used to calculate the similarity between the cellular component terms

Table 4.6 Summary statistics of distribution of the similarities between the GO molecular function of genetic interacting proteins

Threshold	T-score (vs. original GI)	P value (vs. original GI)	T-score (vs. All pairs)	P value (vs. All pairs)
0.95	2.4895	0.0204	3.7151	1.14E-03
0.9	4.1629	1.15E-04	5.8254	3.43E-07
0.85	2.7182	7.78E-03	5.2245	1.03E-06
0.8	3.3616	8.92E-04	7.5871	6.66E-13
0.75	3.9179	1.03E-04	9.3705	4.77E-19
0.7	4.6401	4.27E-06	10.8787	4.29E-25
0.65	7.5061	1.25E-13	15.3320	1.44E-47
0.6	7.7377	1.82E-14	16.9870	2.03E-58
0.55	7.1848	9.68E-13	17.4099	7.03E-62
0.5	5.4622	5.05E-08	18.8874	5.92E-74
0.45	4.7483	2.12E-06	19.5387	1.88E-79
0.4	2.9593	0.0031	21.2428	1.80E-94
0.35	3.0087	0.0026	23.4789	4.78E-115
0.3	2.1694	0.0301	25.6045	6.41E-137
0.25	-0.0979	0.9220	27.3813	1.31E-157
0.2	-2.4949	0.0126	28.9940	7.19E-178
0.15	-6.1706	6.99E-10	30.3384	3.11E-196
0.1	-9.5693	1.30E-21	35.3075	1.82E-266
0.05	-12.8642	1.49E-37	42.3348	0.00
0	-17.3589	2.94E-66	110.0738	0.00

Note: the Resnik method is used to calculate the similarity between the cellular molecular function terms

4.6. Novel SLGIs Predicted by MLE Approach

The MLE approach was able to predict novel SLGIs. Table 4.7 lists 17 novel SLGIs (not included in our training data) with probability >0.9. We predicted the MYO4/DYN1 pair to be SLGI with the highest probability of 0.9895. The MYO4 is one of two type V myosin motors. The other one is MYO2, which is known to genetically interact with DYN1 [54]. Among 17 SLGIs, 12 SLGIs are between transcription initiation factor genes and genes from RNA polymerase complex. Previously, many SLGIs between transcription initiation factor genes and RNA polymerase genes have

already been reported [55-57]. We expected our novel SLGIs to help further elucidate the transcription machine. We then investigated genes involved in cellular response to stresses caused by DNA damage. We downloaded a list of 116 DNA repair and recombination genes from Kyoto Encyclopedia of Genes and Genomes (KEGG) database [58]. Then, we extracted SLGIs in which at least one protein of SLGI pair are related to DNA repair. Figure 4.5 shows the DNA repair related SLGIs with probability>0.7. Of a total of 133 SLGIs, 22 SLGIs are novel. Some of new predicted SLGIs are supported by previous studies. For example, the TOP3 and RAD1 double mutant has shown extreme synergistic growth defects in a previous study [59]. A recent study showed that the RTT109 and YKu70 double mutant exhibits synergistic defects under hydroxyurea treatment [60]. The PAP2 and POL2 were also shown to genetically interact at high temperatures [61].

Figure 4.5 The SLGIs (probabilities>0.7) related to DNA repair proteins. The cycles indicate DNA repair proteins and the triangle indicate non DNA repair proteins. Each genetic interaction involves at least one DNA repair protein. The wide links represent new genetic interactions. The figure was produced using Cytoscape [62].

Table 4.7 Predicted new genetic interactions with probability greater than 0.9

Protein 1		Protein 2		Probability
Name	Annotation	Name	Annotation	
MYO4	type V myosin motors	DYN1	Cytoplasmic heavy chain dynein	0.9895
RPO21	RNA polymerase II largest subunit	BRF1	TFIIIB B-related factor	0.9816
DST1	General transcription elongation factor TFIIS	RET1	Second-largest subunit of RNA polymerase III	0.9797
STO1	Large subunit of the nuclear mRNA cap-binding protein complex	RPA190	RNA polymerase I subunit	0.9586
STO1	Large subunit of the nuclear mRNA cap-binding protein complex	RPO21	RNA polymerase II largest subunit	0.9586
DST1	General transcription elongation factor TFIIS	RPA135	RNA polymerase I subunit A135	0.9570
BNR1	Formin, nucleates the formation of linear actin filaments	DYN1	Cytoplasmic heavy chain dynein	0.9441
CEG1	Alpha (guanylyltransferase) subunit of the mRNA capping enzyme	RPA190	RNA polymerase I subunit	0.9437
CEG1	Alpha (guanylyltransferase) subunit of the mRNA capping enzyme	RPO31	RNA polymerase III subunit	0.9437
RPB5	RNA polymerase subunit	RPA190	RNA polymerase I subunit	0.9437
RPB5	RNA polymerase subunit	RPO31	RNA polymerase III subunit	0.9437
DST1	General transcription elongation factor TFIIS	RPA190	RNA polymerase I subunit	0.9265
DST1	General transcription elongation factor TFIIS	RPO31	RNA polymerase III subunit	0.9265
ARO1	Pentafunctional arom protein	SPT16	Subunit of the heterodimeric FACT complex	0.9198
RET1	Second-largest subunit of RNA polymerase III	SUA7	Transcription factor TFIIB	0.9035
BRF1	TFIIIB B-related factor	RET1	Second-largest subunit of RNA polymerase III	0.9035
BRF1	TFIIIB B-related factor	RPB2	RNA polymerase II second largest subunit	0.9035

4.7. Compensatory Pathways from Predicted SLGIs

Protein pathways are a part of gene network in the cell that can accomplish certain functionality. The SLGIs have been proposed to have high probability of occurrence in compensatory pathways [4]. Thus, the SLGIs within the pathways are rare and the SLGIs between pathways are significantly abundant. Identification compensatory pathways

from synthetic lethal genetic interactions can be a powerful way to understanding cellular functional relationships. We expected our new predictions will increase the ability of understanding compensatory pathways. We applied the algorithm of Ma [63] to identify compensatory pathways from 7583 predicted SLGIs with probability higher than 0.3. Among 7583 SLGIs, 4497 SLGIs are novel predictions. Although Ma et al. have shown that physical interactions are enriched in discovered pathways [63], there is no assumption that proteins in those pathways are physically interacting. Totally, we obtained 167 pairs of compensatory pathways, which include 638 proteins and 3535 SLGIs. Then, we examined the GO term co-occurrences in each pathway using the SGD GO Term Finder [36]. The GO Term Finder calculated the p-values that reflect the probability of observing the co-occurrence of proteins with a given GO term in a certain pathway by chance based on a binomial distribution. Among 167 compensatory pathway pairs, 153 pairs are found significantly enriched GO terms with p-value less than 0.05. All 167 pairs of compensatory pathways and their most significant GO terms are listed on our www.genenetworks.net website.

Figure 4.6A and Figure 4.6B list two pairs of compensatory pathways related to DNA double strand break (DSB) repair. Among total 28 SLGIs between those two compensatory pathways, only 12 of them are known. The DSB is a kind of lethal DNA damage in which both strands of double helix are cleaved. The cell maintains multiple mechanisms to repair double strand breaks with the homologous recombination (HR) and non-homologous end joining (NHEJ) as two major mechanisms. Pathways involving those two mechanisms had shown to be compensatory to each other in *Drosophila* [40].

The pathways on the left of two compensatory pairs involve in NHEJ. The KU complex (YKU80 and YKU70) is the damage detector of NHEJ [64]. Mutation of nucleoporins NUP84 and NUP133 was reported to be hypersensitive to DNA damage [65]. Furthermore, the nucleoporins (NUP84 and NUP133) had been reported to colocalize and coimmunoprecipitate with Slx5/Slx8 [66], which regulate a DNA repair pathway that counteract Rad5-independent HR [67]. Our discovery suggested that the nucleoporins (NUP84 and NUP133) may regulate the NHEJ pathway through Ku complex for double strand breaks repair. The RAD24 is a DNA damage checkpoint protein [68]. Studies also had shown that the YKU80 and RAD24 are in the same NHEJ pathways [68] to repair irradiation and methylmethanesulphonate (MMS) damages. The CSM3 is a DNA replication checkpoint protein. Genetic interaction between RAD24 and CSM3 may indicate the pathway on the right side of Figure 4.6A can actually become two pathways: one is YKU80, RAD24, NUP84; the other is YKU80, CSM3 and NUP84.

The pathways on the right of those two compensatory pairs of Figure 4.6A and Figure 4.6B involve in HR. On the right of Figure 4.6A, the pathway involves three proteins: RAD57, RAD51 and DMC1. The DMC1 and RAD51 are known to form a complex [69] and have roles in recombination [70]. RAD51 and RAD57 are in the same protein family and it has been shown that RAD57/RAD55 bind with RAD51 [71]. On the right of Figure 4B, the pathway involves four proteins: SGS1, RRP6, MRE11 and RAD52. The MRE11 and SGS1 are part of a two-step mechanism to initial HR [72]. The RAD52 plays a major role in the single strand annealing and strand exchange.

Figure 4.6C also shows two compensatory DNA repair pathways. Among 18 SLGIs between two pathways, 11 are novel predictions. Many previous studies have supported our prediction of compensatory functionalities of those two pathways. Guillet and Boiteus [73] reported that the APN2 and MUS81/MMS4 have overlapping function to repair 3'-blocked single strand breaks (SSBs). Vance and Wilson [74] showed that TDP1 and RAD1 function as redundant pathways. SGS1/TOP3 had also been reported to have overlapping functionality with the MMS4/MUS81 [75]. Proteins in these compensatory pathways involve in many DNA-repair pathways, such as base excision repair (BER), nucleotide excision repair (NER) and HR. Studies already showed that BER, NER and HR pathways have overlapping specificities [76]. Our prediction of these compensatory pathways may help to understand the overlapping functionalities among BER, NER and HR pathways.

Another interesting compensatory pathway pair is related to hydroperoxides response in the cell. As shown in Figure 4.6D, 11 of 15 SLGIs connecting two pathways are novel predictions. The three proteins (TSA1, TSA2 and PRX1) on the right are all thioredoxin peroxidase. They play a role in reduction of hydroperoxides in cytoplasm (TSA1 and TSA2) and mitochondrion (PRX1). Meanwhile, the five proteins (RAD52, RAD5, MDM31, MDM32, and MRE11) in the left pathway are related to DNA repair. RAD52, RAD5 and MRE11 are known related to double strand break repair [77-79]. The MDM31 and MDM32 relate the stability of mitochondrial DNA [80]. Those five proteins may relate to repair the damage created by the hydroperoxides. This

compensatory pathway pair implies that removing hydroperoxides or repairing damage is alternative strategies for cells to survive from hydroperoxides.

Figure 4.6 Compensatory pathways identified from predicted SLGIs. (A, B) two compensatory pathways related to DNA double strand breaks repairs. Pathways on the left of A and B belong to non-homologous end joining. Pathways on the right of A and B belong to homologous recombination. (C) A compensatory pathways related to DNA repairs. (D) A compensatory pathways related to hydroperoxides response in the cell. Dashed lines indicate known SLGIs and solid lines indicated novel predicted SLGIs. The figure was produced using Cytoscape [62].

4.8. Discussion

We demonstrated that the synthetic lethal genetic interactions between two proteins could be described by the genetic interactions between domains of those proteins. We applied a probabilistic model to successfully identify significant domain genetic interactions. The significant domain genetic interactions rarely overlap with the domain physical interactions from iPfam [20], which suggests that the domain genetic interactions and domain physical interactions are complementary to each other. The domain genetic interactions offer a new understanding of the relationship among domains, and then among proteins. Furthermore, with the identified domain genetic interactions, we showed that the contributions of domains in a multi-domain protein to its genetic interactions are significantly different. The domain genetic interactions will help to decipher the domains that perform the function related to the genetic interaction. For example, analyzing the domain genetic interactions confirms that only the PF00570 (HRDC domain) of SGS1 involve in the cellular functions that are compensatory to topoisomerases.

Our identification of domain genetic interactions and prediction of SLGIs is not complete. First, our training data is limited. It is believed that the available genetic interactions are just a small fraction of the whole genetic interactions [58]. Second, the training data largely come from several high-throughput genetic interaction screens, which emphasized certain biological processes, such as DNA repair. Thus, the probability of domain genetic interaction may be effected by the future available SLGIs. Moreover, we should keep in mind that our predicted SLGIs require further validation to

exclude false positives. The MLE approach assumes the independence of domain genetic interactions. However, there may be dependence between domain genetic interactions. An apparent extension of this work is to consider the dependences among cooperative domains in multi-domain proteins. Due to those limitations, the performance of predicting genome-wide SLGIs using protein domains may be not as good as other classifiers [11].

This study significantly improved the understanding of different domains in mulitdomain proteins. The identification of domain genetic interactions helps the understanding of originality of functional relationship in SLGIs at domain level. Furthermore, our prediction of SLGIs expanded the ability to elucidate the functional relationships between proteins and pathways. This part of the study was published recently in a journal [84].

CHAPTER FIVE

Predicting Yeast Synthetic Lethal Genetic Interactions Using Short Polypeptide Clusters

Our study showed that domain genetic interactions could be used to characterize and predict SLGIs. However, the feasibility of domain-based prediction of SLGIs is limited by the coverage of protein domains. For example, among more than 6500 yeast proteins, only 4480 of them contain PfamA domains. It is known as a virtual axiom in biology that the “sequence specifies structure and structure determines functionality” [87]. The functionalities of a protein are conserved in parts of its sequence like super families and function sites. We can use the short polypeptide sequences, which are typically shorter than the classically defined protein domains, to characterize the functionalities of proteins. The short polypeptide sequences have previously been used to predict protein-protein interactions in yeast [81]. Here, we hypothesize that the genetic interaction between a pair of proteins can be determined by the genetic interactions between the short polypeptide sequences. Using short polypeptide sequences, rather than protein domains, will allow us to predict genetic interactions with much higher coverage. We developed a framework to obtain short polypeptide clusters and further used them to predict SLGIs using polypeptide cluster information between genetic interactions. We demonstrated that the genetic interaction between a pair of proteins could be determined by the genetic interactions between the short polypeptide clusters of those proteins. Using short polypeptide clusters as features, we can not only increase the prediction coverage, but also improve the prediction performance.

5.1. Constructing the Short Polypeptide Clusters

We constructed the short polypeptide clusters in three steps. First, we identified significant short polypeptide sequence pairs from yeast proteins based on the similarities of local alignments. Then, we developed a new clustering algorithm to group the short polypeptide sequences. Finally, we post-process the short polypeptide clusters to construct the features to predict SLGIs.

5.1.1. Identifying Significant Short Polypeptide Sequence Pairs

We designed a method to identify significant short polypeptide pairs based on their similarity. For each protein A , we first chopped the protein sequence into short polypeptides with length of L sequentially in a moving window size w . Then, we used Smith-Waterman algorithm [82] to search all local matches with similarity scores beyond a predefined threshold for each short polypeptide in protein A from the sequences of all other available proteins (Figure 5.1). We used the BLOSUM62 to score the similarity. If we find a significant local match between a short polypeptide a_i from protein A and a short polypeptide b_j from protein B , we will obtain a peptide pair a_i and b_j . To assure strong similarity between this pair of peptides, a significant local match between the polypeptide b_j from protein B and a short polypeptide a_i from protein A also needs to be obtained in order to keep this peptide pair (a_i, b_j) . This peptide pair retained is identified as significant polypeptide pair.

Figure 5.1 Illustration of matching between significant short polypeptide sequence pairs.

The significances of local matching of short polypeptide sequences are determined using a P-value threshold. For each protein, we obtained a series of similarity scores by aligning each polypeptide chunk to all other proteins. We modeled those scores using a Gaussian distribution. Then, the significant local matching will have a score beyond: $z \times \sigma + m$, where the m and σ are the mean and standard deviation of all scores of each protein. The standard score z can be determined by the P-value threshold we chosen. For example, the z -score is 4.75 for P-value 10^{-6} and the z -score is 5.61 for P-value 10^{-8} . Since m and σ values are different for each protein, the thresholds of similarities used to determine the significant local matching are different for each protein.

We could optimize following parameters of the method: 1) the size of the short polypeptide sequence L ; 2) the moving window w ; 3) the penalty for gap and mismatch in the alignment; 4) the p-value for the threshold of similarity score. In this study, we chose the size of each short polypeptide sequence L to be 25 and the window w is set to 5. The penalty for gap and mismatch was chosen to be 14. And we have tested P-value equal to 10^{-8} for the threshold of similarities. We eventually obtained 1,193,806 polypeptide pairs covering 6260 proteins from yeast. And there are totally 1,587,944 unique polypeptides involved in these polypeptide pairs.

5.1.2. Clustering Short Polypeptide Sequences

After identifying significant polypeptide sequence pairs, we developed a clustering algorithm to group similar short polypeptide sequences into clusters, which may contain more than two short polypeptide sequences. Initially, each significant polypeptide sequence pair was considered as a short polypeptide cluster. We first align the significant polypeptide sequence pair using ClustalW [85]. Then, we built a hidden Markov (HMM) model using the output of the multiple sequence alignment as the seed. The HMM model was constructed by the HMMbuild tool from HMMER [83]. After that, we obtained a short polypeptide cluster from the polypeptide sequence pair by searching relevant polypeptide sequences using HMMsearch against all significant polypeptide sequences. Using a strict threshold, HMMsearch can screen the polypeptide sequences consist of all the polypeptide sequence pairs with the HMM model. Polypeptide sequences significantly similar to the members of the seed were added to the cluster. For each polypeptide sequence pair, this process was repeated iteratively so relevant polypeptide sequences could be continuously added to the polypeptide cluster until no new polypeptide sequence could be enrolled. The short polypeptide sequence clustering method implemented is summarized as follows:

Input: a pair of short polypeptide sequences

Initialization: add the short polypeptide sequence pair into a polypeptide cluster

Step 1. Conduct multiple sequence alignment (MSA) for the members of the polypeptide cluster using ClustalW;

Step 2. Build a HMM model using HMMbuild from the output of MSA conducted in step 1;

Step 3. Search all relevant polypeptide sequences using HMMsearch and add them to the polypeptide cluster;

Step 4. Go back to step 1 until no new polypeptide sequences can be recruited.

Several stringent thresholds (10^{-10} , 10^{-15} and 10^{-20}) were tested for HMMsearch in order to include all similar short polypeptide sequences and reduce false cluster members. Due to the large size of the short polypeptide sequence pairs, we first obtained short polypeptide cluster from each sequence pair as the seed. Then, we post-processed the short polypeptide clusters. This strategy allowed us to easily run the clustering algorithm on a computer cluster.

```

YNL339C-190-214      --FNRKDYDTLSLFFYLNRGYYNELSFR----- 25
YGR296W-190-214      --FNRKDYDTLSLFFYLNRGYYNELSFR----- 25
YBL111C-126-150      --FNRKDYDTLSLFFYLNRGYYNELSFR----- 25
YBL111C-131-155      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YGR296W-195-219      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YNL339C-195-219      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YPL283C-195-219      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YDR545W-131-155      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YER190W-17-41        -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YHL049C-131-155      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YHR218W-67-91        -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YIL177C-131-155      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YJL225C-131-155      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YLR462W-131-155      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YLR467W-131-155      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YML133C-131-155      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YOR396W-131-155      -----YDTLSLFFYLNRGYYNELSFRVLERC 25
YBL111C-130-154      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YDR545W-130-154      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YER190W-16-40        -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YGR296W-194-218      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YHL049C-130-154      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YHR218W-66-90        -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YIL177C-130-154      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YJL225C-130-154      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YLR462W-130-154      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YLR467W-130-154      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YML133C-130-154      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YNL339C-194-218      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YOR396W-130-154      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YPL283C-194-218      -----DYDTLSLFFYLNRGYYNELSFRVLER- 25
YBL111C-129-153      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YER190W-15-39        -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YHR218W-65-89        -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YDR545W-129-153      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YGR296W-193-217      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YHL049C-129-153      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YIL177C-129-153      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YJL225C-129-153      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YLR462W-129-153      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YLR467W-129-153      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YML133C-129-153      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YNL339C-193-217      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YOR396W-129-153      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
YPL283C-193-217      -----KDYDTLSLFFYLNRGYYNELSFRVLE-- 25
*****

```

Figure 5.2 Multiple sequence alignment of a typical short polypeptide cluster

5.1.3. Post-processing Short Polypeptide Clusters

Our goal is to use the short polypeptide clusters to represent the functionalities of proteins. Like protein domains are not allowed to overlap with existing Pfam families [22], we didn't allow overlapping short polypeptide clusters. Thus, short polypeptide clusters were merged together if the clusters share polypeptide sequence. So eventually no polypeptide overlap can be found among polypeptide clusters. Although we used a loose merging criterion, the stringent thresholds used by HMMsearch still allowed the short polypeptide sequences in each cluster to have high similarity. As well, we filtered out polypeptide clusters with only two polypeptides, as those clusters will have no power for prediction of SLGIs. Also, we filtered out short polypeptide clusters contained by too many proteins. Those common short polypeptide clusters exist in both positive and negative data and also do not provide prediction power. For example, polypeptide clusters contained by more than 20, 50 or 100 proteins were removed for further analysis. Figure 5.2 shows the output of multiple sequence alignment of a typical polypeptide cluster obtained after the step of merging. The cluster contains 212 polypeptide sequences but only part of the output was provided in the figure considering the dimension of the page. We can observe that those polypeptide sequences contained by the cluster are highly conserved with significant number of identical amino acids.

Next, we explored how the choosing of thresholds for HMMsearch and for filtering clusters will affect the short polypeptide clusters. We tested the different threshold configurations that combined one of three E-value thresholds for HMMsearch: 10^{-10} , 10^{-15} , 10^{-20} ; and one of four thresholds for filtering out short polypeptide clusters:

20, 50, 100, No Filter. The number of proteins involved, the number of SLGIs covered and the number of polypeptide clusters for all available configurations are listed in Table 5.1, Table 5.2 and Table 5.3 respectively. In these three tables, the first column lists the thresholds for filtering out short polypeptide clusters and the first row lists the E-value thresholds used by HMMsearch. The results showed that the E-value threshold affects the number of retrained short polypeptide clusters, and then affects the number of proteins and the number of SLGIs covered. On the other hand, changing the threshold for filtering out short polypeptide clusters has limited effect. Comparing to 7475 SLGIs covered by PfamA domains, our polypeptide clusters cover much more SLGIs, which is 10,129 when the HMMsearch threshold is chosen to be 10^{-10} and clusters contained by more than 100 proteins are removed. Also 5966 proteins are covered in this configuration comparing to 4480 proteins covered by PfamA dataset. The maximum number of polypeptide clusters contained by a protein is 123 (YIR023W) when the HMMsearch threshold is chosen to be 10^{-10} and clusters contained by more than 100 proteins are removed. The results indicate that polypeptide clusters provide better coverage than that of PfamA domains.

Table 5.1 The number of proteins covered with various thresholds of HMMsearch and polypeptide cluster filters

	E-10	E-15	E-20
20	5926	5215	4984
50	5963	5239	4996
100	5966	5255	5004
N/A	6260	5345	5104

Table 5.2 The number of clusters obtained with various thresholds of HMMsearch and polypeptide cluster filters.

	E-10	E-15	E-20
20	16414	33385	51740
50	16538	33692	53415
100	16581	33829	53755
N/A	16594	33854	53778

Table 5.3 The number of SLGIs retained with various thresholds of HMMsearch and polypeptide cluster filters.

	E-10	E-15	E-20
20	10084	8693	8285
50	10129	8725	8304
100	10129	8749	8309
N/A	10786	8947	8579

5.2. Using Short Polypeptide Clusters to Predict SLGIs

We then used the short polypeptide clusters as features to predict SLGIs. We conducted two studies. In first study, we performed a two-fold cross-validation using support vector machine (SVM) as the classifier to predict SLGIs. In the second study, we evaluate the maximum likelihood estimation (MLE) based classifier on an experimentally obtained genetic interactions and non-genetic interactions.

5.2.1. Cross Validation of SLGIs Based on Short Polypeptide Clusters Using SVM

The polypeptide clusters could be used to substitute protein domains in the context of predicting SLGIs. We applied those clusters to predict SLGIs using support vector machine. The SVM models are trained from existing SLGIs with encoding of cluster information. Instead of using those genetic interactions with Pfam domains, we can select all genetic interactions with cluster information now. We performed the two-fold cross-validation using short polypeptide clusters obtained by different threshold configuration. Same number of random protein pairs were generated to serve as negative data set. We achieved better performance than protein domain based prediction by cross-validation. We obtained the accuracies for models of all HMMsearch thresholds with various cutoffs of filter configuration. We observe that the average accuracy is the highest (89.77%) while using 10^{-10} as the threshold of HMMsearch and removing clusters contained by more than 50 proteins. Other configurations achieved slightly lower but similar performance. The cost was set to 32 and the gamma was set to 0.0078125 during the training of SVM models. The performance of various configurations is also available with multiple criteria listed in Table 5.4, Table 5.5 and Table 5.6 with the HMMsearch threshold setting to 10^{-10} , 10^{-15} and 10^{-20} respectively where the first column of each row describes the threshold used by cluster filters.

We also evaluated the SVM classifier using the ROC curve. Figure 5.3 shows the ROC curve for the SVM classifier using short polypeptide clusters obtained by E-value setting to 10^{-10} as and 50 as cluster filter. The AUC score of this ROC curve is 0.9618, which indicates great performance of the SVM classifier.

Table 5.4 Performance of polypeptide clusters based cross validation by SVM classifier. The short polypeptide clusters were obtained using 10^{-10} as the threshold for HMMsearch and different thresholds for filtering.

	TP	FP	TN	FN	Sensitivity	Specificity	Precision	Accuracy	F-score
20	4610 (27.6)	642 (18.71)	4386.67 (13.12)	432 (27.36)	0.9143 (0.0054)	0.8700 (0.0026)	0.8756 (0.0016)	0.8922 (0.0016)	0.8945 (0.0019)
50	4637 (3.27)	609.33 (53.27)	4454.67 (53.27)	427 (3.27)	0.9157 (0.0007)	0.8796 (0.0105)	0.8840 (0.0090)	0.8977 (0.0050)	0.8995 (0.0044)
100	4638 (19.87)	633.67 (25.72)	4430.33 (25.72)	426 (19.87)	0.9158 (0.0039)	0.8749 (0.0051)	0.8798 (0.0039)	0.8954 (0.0011)	0.8975 (0.0009)

Table 5.5 Performance of polypeptide clusters based cross validation by SVM classifier. The short polypeptide clusters were obtained using 10^{-15} as the threshold for HMMsearch and different thresholds for filtering.

	TP	FP	TN	FN	Sensitivity	Specificity	Precision	Accuracy	F-score
20	3950.33 (19.75)	571 (22.99)	3775 (22.99)	395.67 (19.75)	0.9090 (0.0046)	0.8686 (0.0053)	0.8737 (0.0048)	0.8888 (0.0043)	0.8910 (0.0041)
50	3960.33 (27.15)	590 (32.07)	3770.67 (33.51)	435 (30.24)	0.9003 (0.0069)	0.8644 (0.0077)	0.8692 (0.0056)	0.8823 (0.0009)	0.8844 (0.0009)
100	3961.33 (23.30)	593 (26.77)	3781 (26.77)	412.67 (23.30)	0.9056 (0.0054)	0.8644 (0.0061)	0.8698 (0.0058)	0.8850 (0.0057)	0.8874 (0.0056)

Table 5.6 Performance of polypeptide clusters based cross validation by SVM classifier. The short polypeptide clusters were obtained using 10^{-20} as the threshold for HMMsearch and different thresholds for filtering.

	TP	FP	TN	FN	Sensitivity	Specificity	Precision	Accuracy	F-score
20	3678.33 (14.61)	591.67 (14.84)	3550.33 (14.84)	463.67 (14.61)	0.8881 (0.0035)	0.8571 (0.0036)	0.8614 (0.0032)	0.8726 (0.0028)	0.8746 (0.0028)
50	3665.33 (22.01)	590.67 (4.50)	3561.33 (4.50)	486.67 (22.01)	0.8828 (0.0053)	0.8578 (0.0011)	0.8612 (0.0013)	0.8703 (0.0029)	0.8719 (0.0031)
100	3699.67 (31.66)	625 (46.70)	3529 (38.13)	458 (21.23)	0.8897 (0.0051)	0.8495 (0.0092)	0.8554 (0.0068)	0.8696 (0.0020)	0.8722 (0.0012)

Figure 5.3 The ROC curve for SVM classifier using polypeptide clusters obtained with 10^{-10} as the threshold of HMMsearch and 50 as the threshold of polypeptide cluster filter. The AUC is 0.9618.

5.2.2. Predicting SLGIs Based on Short Polypeptide Clusters Using MLE

In our previous study, we used the MLE method to infer domain genetic interactions from SLGIs. To further demonstrate the prediction of SLGIs using short polypeptide clusters, we applied the Maximum Likelihood estimation (MLE) approach to estimate the probabilities of genetic interactions between short polypeptide clusters. We also evaluated the MLE methods on an experimental obtained genetic interactions and non-genetic interactions from Wong's study. Besides providing higher coverage, polypeptide clusters based approach also achieved better performance as indicated by the ROC curves (Figure 5.4). For the data from Wong's study, the number of SLGIs is 3771

and the number of non-SLGIs is 688045. The MLE algorithm based on short polypeptide clusters predicted the probabilities of genetic interaction for 4,872,245 protein pairs with the threshold of HMMsearch setting to 10^{-10} and cluster filter setting to 50. This method covers 1580 SLGIs from Wong's data and the AUC is 0.8866. While, the MLE algorithm based on protein domain has predicted the probabilities of genetic interaction for 599752 protein pairs. As limited by both the coverage of proteins and domain pairs, the protein domain-based approach only covers 45 SLGIs from Wong's study and the AUC is 0.7544. The ROC curves show that, at the same false positive rate level, the true positive rate is always higher for short polypeptide clusters based MLE method than that for protein domain based MLE method. Intuitively, the specificity is 54.30% while choosing the sensitivity to 86.67% using protein domain based method. While the specificity is 85.08% by choosing the sensitivity to 87.47% using polypeptide cluster based method.

Figure 5.4 Comparison of the ROC curves of MLE method based on short polypeptide clusters and protein domains using an experimental obtained genetic interactions and non-genetic interactions. Polypeptide clusters based prediction has threshold of HMMsearch setting to 10^{-10} and polypeptide cluster filter setting to 50 and the AUC is 0.8943. Protein domains based prediction has the AUC equal to 0.7544.

5.3. Discussion

In this study, we created a framework to retrieve short polypeptide clusters in yeast. We hypothesized that those short polypeptide cluster represent the functionalities of proteins, like the protein domains. The framework consists of multiple steps including identifying matching polypeptide sequences, clustering short polypeptides and post-processing polypeptide clusters. Various tools and algorithms were used for local alignment (Waterman-Smith), multiple sequence alignment (ClustalW) and Hidden Markov Model (HMMER). Polypeptide clusters obtained provide better coverage than

PfamA protein domains from both the perspective of proteins and SLGIs. Considering this is a computational intensive project, most of the work was conducted on the distributed system called Palmetto in Clemson University. Tasks were submitted to multiple machines with each machine responding on a section of the work. We also used these polypeptide clusters for the prediction of SLGIs using both SVM and MLE method. Both cross-validation and evaluation on an experimental data sets showed that the short polypeptide clusters based approach is superior to the previous protein domain based approach. Moreover, the short polypeptide clusters based approach provides much higher coverage for predicting SLGIs.

CHAPTER SIX

CONCLUSION AND FUTURE WORK

In this dissertation, we established sequence features inclusive of protein domains, domain genetic interactions and short polypeptide clusters to characterize yeast SLGIs, which help understand the molecular mechanisms of SLGIs. Our study helps more precisely define the functional relationship between yeast proteins. We also provided novel approaches to predict new SLGIs.

Our approach of genetic interaction prediction based on domain features using SVM achieves high sensitivity and specificity. As the building blocks of proteins, we believe the essential status of protein domains decides their importance in the prediction of both protein and genetic interactions. The results clearly demonstrated that there is strong correlation between protein domain relationship and genetic interactions, especially synthetic lethal genetic interactions. The study also indicates the efficiency of SVM in predicting yeast SLGIs using protein domains as features.

As well, we demonstrated that the SLGIs between two proteins could be described by the genetic interactions between domains of those proteins. We applied a MLE model to successfully identify significant domain genetic interactions. The domain genetic interactions offer a new understanding of the relationship among domains, and then among proteins with domains contributing differently in the context of genetic interaction. Our study significantly improved the understanding of different domains in

multi-domain proteins and expanded the ability to elucidate the functional relationships between proteins and pathways.

Furthermore, we created a framework to obtain conserved short polypeptide clusters. Our study indicates that short polypeptide clusters can achieve performance comparable to that of protein domains for the prediction of yeast SLGIs. Besides that, polypeptide clusters provide higher coverage than protein domains, which allow us to predict SLGIs in a general method. They also have the potential of identifying function sites like motifs in further study.

Though our study has much to contribute to the community especially in the area of understanding and predicting SLGIs, there is a lot space to improve and extend. One direction is to focus on certain types of protein domains while investigating genetic interactions. Along with the growing understanding of functions of protein domains, we can focus on protein domains with recognizable molecular functions. The true positive would be potentially improved with the cost from coverage, which is still significant especially in assisting small-scale experimental study. Our polypeptide based prediction achieved impressive performance while applied to predict SLGIs. They also possess the potential of predict protein-protein interactions when appropriately applied. As these polypeptide clusters were obtained in a way similar to that of protein domains, they can be further annotated and serve as a start point of discovering super families and function sites. It is also promising if we can extend our approach to other organisms to understand the evolutionary conservation of domain genetic interactions and short polypeptide cluster interactions.

APPENDICES

Appendix A

Screenshots of www.genenetworks.net

Search Predicted Genetic Interaction

Genetic Interaction probability of this protein pair:

Protein 1	Protein 2	Probability
YDL140C (RPB1)	YGL043W (DST1)	0.99861 *

Domains contained by these proteins:

YDL140C (RPB1)	YGL043W (DST1)
PF00623	PF01096
PF04983	PF07500
PF04990	PF08711
PF04992	
PF04997	
PF04998	
PF05000	
PF05001	

The probabilities between domains:

	PF00623	PF04983	PF04990	PF04992	PF04997	PF04998	PF05000	PF05001
PF01096	0.076931	0.076931	0.200056	0.200056	0.076931	0.076931	0.076931	0.200056
PF07500	0.142886	0.142886	0.333488	0.333488	0.142886	0.142886	0.142886	0.333488
PF08711	0.250087	0.250087	0.500349	0.500349	0.250087	0.250087	0.250087	0.500349

The Scores between domains:

	PF00623	PF04983	PF04990	PF04992	PF04997	PF04998	PF05000	PF05001
PF01096	0.000116	0.000116	0.000347	0.000347	0.000116	0.000116	0.000116	0.000347
PF07500	0.000231	0.000231	0.000693	0.000693	0.000231	0.000231	0.000231	0.000693
PF08711	0.000462	0.000462	0.001386	0.001386	0.000462	0.000462	0.000462	0.001386

Select an organism :

Saccharomyces cerevisiae

Protein 1: *

RPB1

The first protein of the genetic interaction. Both gene ID and name are acceptable, for example, YDL140C or RP021.

Protein 2: *

DST1

The second protein of the genetic interaction. Both gene ID and name are acceptable, for example, YGL130W or CEG1.

SEARCH RESULT

Figure 7.1 Search the probability of predicted genetic interactions using domain genetic interactions

Search Predicted Domain Genetic Interactions

Genetic Interaction probability of this domain pair:

Domain 1	Domain 2	Probability	E-score
PF07500	PF00623	0.142886	0.000231

Proteins containing these domains:

PF07500	PF00623
YGL043W (DST1)	YDL140C (RPB1)
YKL005C (BYE1)	YOR116C (RPC1)
	YOR341W (RPA190)

The probabilities between proteins (* indicates known GI):

	YGL043W (DST1)	YKL005C (BYE1)
YDL140C (RPB1)	0.99861*	0.863032
YOR116C (RPC1)	0.926478	0.537412
YOR341W (RPA190)	0.926478	0.537412

Domain 1: *

The name of first domain of the genetic interaction, for example PF01920

Domain 2: *

The name of second domain of the genetic interaction, for example PF07572

SEARCH RESULT

Figure 7.2 Search the probability of predicted domain genetic interactions

Bibliography

1. Hartman JLt, Garvik B, Hartwell L: **Principles for the buffering of genetic variation.** *Science* 2001, **291**(5506):1001-1004.
2. Mani R, St Onge RP, Hartman JL, Giaever G, Roth FP: **Defining genetic interaction.** *Proceedings of the National Academy of Sciences* 2008, **105**(9):3461-3461.
3. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al*: **Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis.** *Science* 1999, **285**(5429):901-906.
4. Kelley R, Ideker T: **Systematic interpretation of genetic interactions using protein networks.** *Nature biotechnology* 2005, **23**(5):561-566.
5. Chipman KC, Singh AK: **Predicting genetic interactions with random walks on biological networks.** *BMC Bioinformatics* 2009, **10**:17.
6. Qi Y, Suhail Y, Lin YY, Boeke JD, Bader JS: **Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions.** *Genome Res* 2008, **18**(12):1991-2004.
7. Wong SL, Zhang LV, Tong AHY, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey H: **Combining biological networks to predict genetic interactions.** *Proceedings of the National Academy of Sciences* 2004, **101**(44):15682-15687.
8. Paladugu S, Zhao S, Ray A, Raval A: **Mining protein networks for synthetic genetic interactions.** *BMC Bioinformatics* 2008, **9**(1):426-426.
9. Zhong W, Sternberg PW: **Genome-wide prediction of *C. elegans* genetic interactions.** *Science's STKE* 2006, **311**:1481.
10. Deng M, Mehta S, Sun F, Chen T: **Inferring domain-domain interactions from protein-protein interactions.** *Genome Res* 2002, **12**(10):1540-1548.
11. Huang C, Morcos F, Kanaan SP, Wuchty S, Chen DZ, Izaguirre JA: **Predicting protein-protein interactions from protein domains using a set cover approach.** *IEEE ACM Transactions on Computational Biology and Bioinformatics* 2007, **4**(1):78-78.

12. Sprinzak E, Margalit H: **Correlated sequence-signatures as markers of protein-protein interaction.** *J Mol Biol* 2001, **311**(4):681-692.
13. Roth F, Lipshitz H, Andrews B: **Q&A: Epistasis.** *Journal of Biology* 2009, **8**(4):35.
14. Boone C, Bussey H, Andrews BJ: **Exploring genetic interactions and networks with yeast.** *Nature Reviews Genetics* 2007, **8**(6):437-449.
15. Carter GW, Galas DJ, Galitski T, Bader JS: **Maximal Extraction of Biological Information from Genetic Interaction Data.** *PLoS Comput Biol* 2009, **5**(4):e1000347-e1000347.
16. Järvinen AP, Hiissa J, Elo LL, Aittokallio T: **Predicting Quantitative Genetic Interactions by Means of Sequential Matrix Approximation.** *PLoS ONE* 2008, **3**(9).
17. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS: **Gene function prediction from congruent synthetic lethal interactions in yeast.** *Molecular Systems Biology* 2005, **1**(1).
18. Dixon SJ, Fedyshyn Y, Koh JLY, Prasad TSK, Chahwan C, Chua G, Toufighi K, Baryshnikova A, Hayles J, Hoe K-L *et al*: **Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(43):16653-16658.
19. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.** *Nucleic acids research* 2002, **30**(1):303-303.
20. Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, Frishman D: **MIPS: a database for genomes and protein sequences.** *Nucleic acids research* 1999, **27**(1):44-44.
21. Hoh J, Ott J: **Mathematical multi-locus approaches to localizing complex human trait genes.** *Nat Rev Genet* 2003, **4**(9):701-709.
22. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL: **The Pfam protein families database.** *Nucleic acids research* 2002, **30**(1):276-276.

23. Raghavachari B, Tasneem A, Przytycka TM, Jothi R: **DOMINE: a database of protein domain interactions**. *Nucleic Acids Res* 2008, **36**(Database issue):D656-661.
24. Ng SK, Zhang Z, Tan SH: **Integrative approach for computationally inferring protein domain interactions**. *Bioinformatics* 2003, **19**(8):923-929.
25. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D: **ProDom: automated clustering of homologous domains**. *Brief Bioinform* 2002, **3**(3):246-251.
26. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks**. *Nature* 2001, **411**(6833):41-42.
27. Hahn MW, Kern AD: **Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks**. *Molecular Biology and Evolution* 2005, **22**(4):803-806.
28. Spielman DA, Teng SH: **A local clustering algorithm for massive graphs and its application to nearly-linear time graph partitioning**. *CoRR* 2008.
29. Can T, Singh AK: **Analysis of protein-protein interaction networks using random walks**. In: 2005; 2005: 68.
30. Anderson R, Lang KJ: **Communities from seed sets**. In: *Proceedings of the 15th international conference on World Wide Web: 2006*; 2006: 232.
31. Harel D, Koren Y: **On clustering using random walks**. *Lecture Notes in Computer Science* 2001.
32. Ulitsky I, Krogan NJ, Shamir R: **Towards accurate imputation of quantitative genetic interactions**. *Genome Biol* 2009, **10**(12):R140.
33. Wilmes GM, Bergkessel M, Bandyopadhyay S, Shales M, Braberg H, Cagney G, Collins SR, Whitworth GB, Kress TL, Weissman JS *et al*: **A genetic interaction map of RNA-processing factors reveals links between Sem1/Dss1-containing complexes and mRNA export and splicing**. *Mol Cell* 2008, **32**(5):735-746.
34. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF *et al*: **Exploration of the Function and Organization of the Yeast Early Secretory Pathway through an Epistatic Miniarray Profile**. 2005, **123**(3):507-519.

35. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M *et al*: **Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map.** *Nature* 2007, **446**(7137):806-810.
36. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M: **SGD: Saccharomyces genome database.** *Nucleic Acids Research* 1998, **26**(1):73-73.
37. Chen XW, Liu M: **Prediction of protein-protein interactions using random decision forest framework.** *Bioinformatics* 2005, **21**(24):4394–4400-4394–4400.
38. Cortes C, Vapnik V: **Support-Vector Networks.** *Machine Learning* 1995, **20**(3):273-297.
39. Fan RE, Chen PH, Lin CJ: **Working set selection using the second order information for training SVM.** *Journal of Machine Learning Research* 2005, **6**:1889-1918.
40. Joachims T: **Making Large-Scale SVM Learning Practical.** In: *Advances in Kernel Methods - Support Vector Learning.* Edited by B.Scholkopf, C.Burges. Cambridge: MIT press; 1999.
41. **TinySVM** [<http://www.chasen.org/~taku/software/TinySVM/>]
42. Chang CC, Lin CJ: **LIBSVM: a library for support vector machines, 2001.** *Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>* 2001.
43. Li B, Luo F: **Predicting yeast synthetic lethal genetic interactions using protein domains.** In: *IEEE International Conference on Bioinformatics and Biomedicine: 2009; Washington DC.; 2009:* 43-47.
44. Dempster AP, Laird NM, Rubin DB: **Maximum Likelihood from Incomplete Data Via Em Algorithm.** *Journal of the Royal Statistical Society Series B-Methodological* 1977, **39**(1):1-38.
45. Riley R, Lee C, Sabatti C, Eisenberg D: **Inferring protein domain interactions from databases of interacting proteins.** *Genome Biol* 2005, **6**(10):R89.
46. Finn RD, Marshall M, Bateman A: **iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions.** *Bioinformatics* 2005, **21**(3):410-412.

47. Le Meur N, Gentleman R: **Modeling synthetic lethality.** *Genome Biol* 2008, **9**(9):R135.
48. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393**(6684):440-442.
49. Barabasi AL, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286**(5439):509-512.
50. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes.** *J Mol Biol* 2001, **310**(2):311-325.
51. Vogel C, Teichmann SA, Pereira-Leal J: **The relationship between domain duplication and recombination.** *Journal of Molecular Biology* 2005, **346**(1):355-365.
52. Mullen JR, Kaliraman V, Brill SJ: **Bipartite structure of the SGS1 DNA helicase in *Saccharomyces cerevisiae*.** *Genetics* 2000, **154**(3):1101-1114.
53. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9**(12):3273-3297.
54. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M *et al*: **Global mapping of the yeast genetic interaction network.** *Science* 2004, **303**(5659):808–813-808–813.
55. Wilcox CB, Rossetini A, Hanes SD: **Genetic interactions with C-terminal domain (CTD) kinases and the CTD of RNA Pol II suggest a role for ESS1 in transcription initiation and elongation in *Saccharomyces cerevisiae*.** *Genetics* 2004, **167**(1):93-105.
56. Malagon F, Tong AH, Shafer BK, Strathern JN: **Genetic interactions of DST1 in *Saccharomyces cerevisiae* suggest a role of TFIIS in the initiation-elongation transition.** *Genetics* 2004, **166**(3):1215-1227.
57. Archambault J, Lacroute F, Ruet A, Friesen JD: **Genetic Interaction between Transcription Elongation-Factor Tfiis and Rna Polymerase-Ii.** *Molecular and Cellular Biology* 1992, **12**(9):4142-4152.
58. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32**(Database issue):D277-280.

59. Shor E, Gangloff S, Wagner M, Weinstein J, Price G, Rothstein R: **Mutations in homologous recombination genes rescue top3 slow growth in *Saccharomyces cerevisiae*.** *Genetics* 2002, **162**(2):647-662.
60. Jessulat M, Alamgir M, Salsali H, Greenblatt J, Xu J, Golshani A: **Interacting proteins Rtt109 and Vps75 affect the efficiency of non-homologous end-joining in *Saccharomyces cerevisiae*.** *Arch Biochem Biophys* 2008, **469**(2):157-164.
61. Edwards S, Li CM, Levy DL, Brown J, Snow PM, Campbell JL: ***Saccharomyces cerevisiae* DNA polymerase epsilon and polymerase sigma interact physically and functionally, suggesting a role for polymerase epsilon in sister chromatid cohesion.** *Mol Cell Biol* 2003, **23**(8):2733-2748.
62. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**(11):2498-2504.
63. Ma X, Tarone AM, Li W: **Mapping genetically compensatory pathways from synthetic lethal interactions in yeast.** *PLoS ONE* 2008, **3**(4):e1922.
64. Pastwa E, Blasiak J: **Non-homologous DNA end joining.** *Acta Biochimica Polonica* 2003, **50**(4):891-908.
65. Therizols P, Fairhead C, Cabal GG, Genovesio A, Olivo-Marin JC, Dujon B, Fabre E: **Telomere tethering at the nuclear periphery is essential for efficient DNA double strand break repair in subtelomeric region (vol 172, pg 189, 2006).** *Journal of Cell Biology* 2006, **172**(6):951-951.
66. Nagai S, Dubrana K, Tsai-Pflugfelder M, Davidson MB, Roberts TM, Brown GW, Varela E, Hediger F, Gasser SM, Krogan NJ: **Functional targeting of DNA damage to a nuclear pore-associated SUMO-dependent ubiquitin ligase.** *Science* 2008, **322**(5901):597-602.
67. Burgess RC, Rahman S, Lisby M, Rothstein R, Zhao XL: **The slx5-slx8 complex affects sumoylation of DNA repair proteins and negatively regulates recombination.** *Molecular and Cellular Biology* 2007, **27**(17):6153-6162.
68. de la Torre-Ruiz MA, Lowndes NF: **The *Saccharomyces cerevisiae* DNA damage checkpoint is required for efficient repair of double strand breaks by non-homologous end joining.** *Febs Letters* 2000, **467**(2-3):311-315.

69. Bishop DK: **RecA Homologs Dmc1 and Rad51 Interact to Form Multiple Nuclear-Complexes Prior to Meiotic Chromosome Synapsis.** *Cell* 1994, **79**(6):1081-1092.
70. Masson JY, West SC: **The Rad51 and Dmc1 recombinases: a non-identical twin relationship.** *Trends in Biochemical Sciences* 2001, **26**(2):131-136.
71. Sung P: **Yeast Rad55 and Rad57 proteins form a heterodimer that functions with replication protein a to promote DNA strand exchange by Rad51 recombinase.** *Genes & Development* 1997, **11**(9):1111-1121.
72. Mimitou EP, Symington LS: **Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing.** *Nature* 2008, **455**(7214):770-U773.
73. Guillet M, Boiteux S: **Endogenous DNA abasic sites cause cell death in the absence of Apn1, Apn2 and Rad1/Rad10 in Saccharomyces cerevisiae.** *Embo Journal* 2002, **21**(11):2833-2841.
74. Vance JR, Wilson TE: **Yeast Tdp1 and Rad1-Rad10 function as redundant pathways for repairing Top1 replicative damage.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(21):13669-13674.
75. Kaliraman V, Mullen JR, Fricke WM, Bastin-Shanower SA, Brill SJ: **Functional overlap between Sgs1-Top3 and the Mms4-Mus81 endonuclease.** *Genes & Development* 2001, **15**(20):2730-2740.
76. Swanson RL, Morey NJ, Doetsch PW, Jinks-Robertson S: **Overlapping specificities of base excision repair, nucleotide excision repair, recombination, and translesion synthesis pathways for DNA base damage in Saccharomyces cerevisiae.** *Molecular and Cellular Biology* 1999, **19**(4):2929-2935.
77. D'amours D, Jackson SP: **The Mre11 complex: At the crossroads of DNA repair and checkpoint signalling.** *Nature Reviews Molecular Cell Biology* 2002, **3**(5):317-327.
78. Lisby M, Rothstein R, Mortensen UH: **Rad52 forms DNA repair and recombination centers during S phase.** *Proc Natl Acad Sci U S A* 2001, **98**(15):8276-8282.
79. Johnson RE, Prakash S, Prakash L: **Yeast DNA repair protein RAD5 that promotes instability of simple repetitive sequences is a DNA-dependent ATPase.** *J Biol Chem* 1994, **269**(45):28259-28262.

80. Dimmer KS, Jakobs S, Vogel F, Altmann K, Westermann B: **Mdm31 and Mdm32 are inner membrane proteins required for maintenance of mitochondrial shape and stability of mitochondrial DNA nucleoids in yeast.** *Journal of Cell Biology* 2005, **168**(1):103-115.
81. Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N *et al*: **PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.** *BMC Bioinformatics* 2006, **7**:365.
82. Waterman MS, Eggert M: **A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons.** *J Mol Biol* 1987, **197**(4):723-728.
83. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**(9):755-763.
84. Li B, Cao W, Zhou J, Luo F: **Understanding and predicting synthetic lethal genetic interactions in *Saccharomyces cerevisiae* using domain genetic interactions.** *BMC Syst Biol*, **5**:73.
85. Thompson JD, Gibson TJ, Higgins DG: **Multiple sequence alignment using ClustalW and ClustalX.** *Curr Protoc Bioinformatics* 2002, **Chapter 2**:Unit 2 3.
86. Kaelin WG, Jr.: **The concept of synthetic lethality in the context of anticancer therapy.** *Nat Rev Cancer* 2005, **5**(9):689-698.
87. Anfinsen CB: **Principles That Govern Folding of Protein Chains.** *Science* 1973, **181**(4096):223-230.
88. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**(2):101-113.