Predicting yeast synthetic lethal genetic interactions using protein domains

Bo Li, Feng Luo
School of Computing
Clemson University
Clemson, SC, USA
e-mail: bol, luofeng@clemson.edu

***Abstract*—Synthetic lethal genetic interactions are of interest as they can be used to predict function of unknown proteins and find drug target or drug combinations. In this study, we applied support vector machine (SVM) classifier to predict synthetic lethal genetic interactions in Saccharomyces cerevisiae based on domain information in proteins. We found that our method can predict synthetic lethal genetic interactions with high sensitivity (88.35%) and specificity (82.00%). To the best of our knowledge, the work reported in this paper is the first domain-based model for the prediction of genetic interactions. Our study indicates that there is strong correlation between protein domain relationship and synthetic lethal genetic interactions.**

Keywords-Genetic interactions; SVM; protein domains; prediction

I. INTRODUCTION

Genetic interaction is a phenomenon in which the combined effect of mutations of two genes differs from individual effects of each mutation [1]. In the extreme cases, mutation of two nonessential genes could lead to lethal phenotype. This kind of genetic interactions is referred as synthetic lethal interactions. Two synthetic lethal genes can either reside on the same pathway or on parallel pathways [2]. As one of major types of genetic interactions, synthetic lethal interactions are of interest to many researchers because they are able to be used for understanding protein functions and have potential for finding drug target or drug combinations [3-6]. Synthetic lethal interactions in the yeast, Saccharomyces cerevisiae, have been studies for a long time [3, 4, 6-10] due to its simplicity of having a single cell.

Genetic interactions can be identified by mutant screens, synthetic genetic arrays (SGA) or synthetic lethal analysis by microarrays (SLAM) [2]. However, even high throughput methods, such as SLAM, will not be able to screen pairwise combinations of all proteins in a genome with thousands of genes. Therefore, it is of interest to computationally predict genetic interactions. Recently, many computational approaches have been proposed for the prediction of genetic interactions, especially synthetic lethal interactions. Various feature information are utilized by these methods. Wong et al. [6] applied decision tree for genome-wide prediction of synthetic lethal interactions by utilizing features including protein interactions, gene expression, and functional annotation, gene location and protein network characteristics. The method can be used to discover synthetic sick and lethal pairs with 80% precision from 20% of protein pairs of yeast. Paladugu et al. [4] used the graph theoretic properties of protein network to predict synthetic sick and lethal interactions. They were able to predict synthetic genetic interactions with sensitivity and specificity exceeding 85%. Zhong et al. [10] 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. They had also experimentally tested the predicted interactions of two human disease-related genes.

Many studies have used domain information to predict protein-protein interactions [11-13]. These approaches are based on the assumption that proteins interact though domains. Representing the structures and functions of proteins, protein domains are usually regarded as building blocks of proteins and conserved during evolution. 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. Here, we hypothesize that there is a strong correlation between protein domain relationship and genetic interactions.

To the best of our knowledge, there is no report on genetic interaction prediction using domain information. In this paper, we conducted a novel study of predicting synthetic lethal interactions in yeast based on protein domain information using SVM.

II. MATERIAL AND METHODS

A. Data Sources

The protein domain data was collected from Pfam (Protein families database) [14]. The Pfam database provides two types of protein family data. Pfam-A domains are manually curated while Pfam-B domains are automatically generated. In our study, only Pfam-A domains are considered. The total number of PfamA domains selected is 2289.

Genetic interactions of yeast were downloaded from the Saccharomyces Genome Database (SGD) [15]. 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,

978-0-7695-3885-3/09 $26.00 © 2009 IEEE
DOI 10.1109/BIBM.2009.37

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.

The domain information was widely used to predict protein-protein interaction. It is of interest to us to examining the overlap between protein interactions and genetic interactions. Hence, we compared our 7435 synthetic lethal interactions with protein-protein interaction datasets of yeast from Database of Interacting proteins (DIP) [16] and Munich Information Center of Protein Sequences (MIPS) [17]. There are 6264 and 7406 protein interactions in DIP and MIPS datasets among 2724 and 4146 proteins respectively. The result shows that even though our synthetic lethal interaction dataset shares a large amount (more than 20%) of proteins with protein-protein interaction datasets, only a small portion (about 5%) of protein pairs exists in both synthetic lethal interactions and protein-protein interactions (Table 1).

TABLE I. OVERLAP BETWEEN GENETIC INTERACTIONS AND PROTEIN-PROTEIN INTERACTIONS

	DIP	MIPS
Overlap protein (percentage)	1277 (20.39%)	1492 (35.89%)
Overlap interactions (percentage)	144 (5.29%)	106 (0.0143%)

B. *Feature Encoding*

We encoded the protein pairs using the method proposed by Chen et al. [18] . 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.

C. *Support Verctor Machine*

The prediction of synthetic lethal interaction is 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. [19]. The C-Support Vector Classification in LibSVM is utilized in our application. We used the radial basis function kernel. The parameters (cost and gamma) were estimated by following the guide of LibSVM [20], which were selected based on the highest accuracy that the trained SVM classifier achieved.

D. *Evaluation of Results*

We have employed multiple criteria to evaluate our classification results, which include sensitivity, specificity, precision, F-measure and accuracy. In the content of paper, the abbreviations of TP, TN, FP and FNrepresent the number of true positive, true negative, false positive and false negative predictions, respectively. The sensitivity is defined as the percentage of correctly predicted positive data over the total number of positive data.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (1)$$

Sensitivity is identical to recall in the classification context. 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.

The specificity is defined as the percentage of correctly identified negative data over the total number of negative data. 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.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad (2)$$

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (3)$$

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 computer the score. The best F-measure score is 1 and the worst F-measure score is 0.

$$\text{F} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \qquad (4)$$

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 percentage of all correctly identified interaction and non-interactions over the size of the test dataset.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

We also used ROC (receiver operating characteristics) to evaluate the classifier in the cross validation study and the prediction performance. The ROC evaluates the performance of classifiers based on the tradeoff between specificity and sensitivity. The area under the ROC curve (AUC) can be used to compare the prediction performance. While an area of 1 means perfect prediction, an area of 0.5 indicates random prediction.

III. EXPERIMENTAL RESULTS

A. *Crsos Validataion 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 equal 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 1). The AUC of our five-fold cross validation study is 0.925 which indicates the remarkable performance of the SVM classifier.

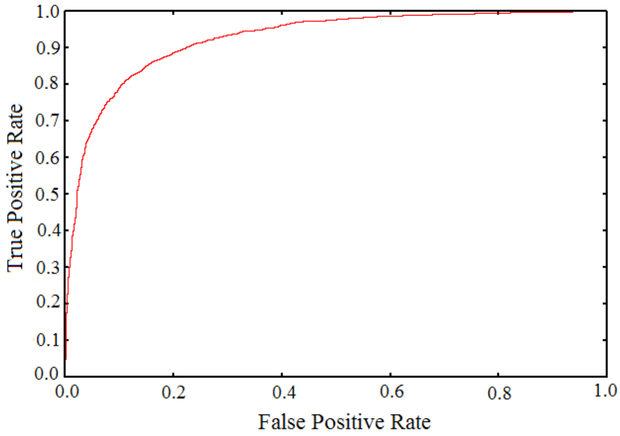

Figure 1. ROC curve for SVM classifier in cross validation study on training data.

B. *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 2 (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 which is even higher than the AUC of cross validation study on training data. The ROC curve is showed in Figure 2.

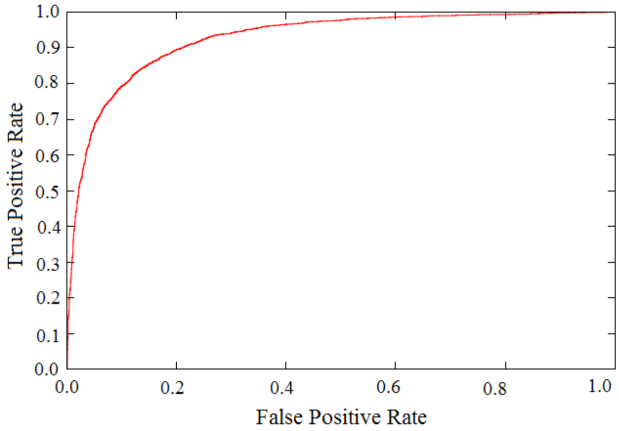

Figure 2. ROC curve of the SVM classifier in prediction

TABLE II. RESULT 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%

C. *Oversampling the Negative Dataset*

Because the number of non-interaction protein pairs is much higher than the number of genetic interaction protein pairs, we test 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 2). 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.

D. *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 is 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 85.12% in accuracy, 87.79% in sensitivity, 82.46% in specificity and 83.35% 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 for detailed results).

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 4). Along with the iterations, the number of negative data decreased from 3717 to 3642 in the training dataset which means seventy five non-interaction protein pairs were removed after three iterations.

TABLE III. CLASSIFICATION RESULTS 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%

TABLE IV. CLASSIFICATION RESULTS 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%

IV. CONCLUSION AND DISCUSSION

Impressively, our novel approach of genetic interaction prediction based on domain information achieves high sensitivity and specificity. Our study clearly demonstrated that there is the strong correlation between protein domain relationship and genetic interactions, especially synthetic lethal 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 85.12% 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.

In the future, we would like to explore other classification methods, like random forest, to improve the prediction of genetic interactions. Moreover, it is worthwhile to study new methods to select parameters of SVM for unbalanced data

ACKNOWLEDGMENT

Bo Li and Feng Luo are supported by NSF EPSCoR grants EPS-0447660. Feng Luo would like to acknowledge financial support from the Institute for Modeling and Simulation Applications at Clemson University.

REFERENCES

[1] R. Mani, R. P. St Onge, J. L. Hartman, et al., "Defining genetic interaction," *Proceedings of the National Academy of Sciences,* vol. 105, pp. 3461-3461, 2008.

[2] R. Kelley and T. Ideker, "Systematic interpretation of genetic interactions using protein networks," *Nature biotechnology,* vol. 23, pp. 561-566, 2005.

[3] [3] P. Ye, B. D. Peyser, X. Pan, et al., "Gene function prediction from congruent synthetic lethal interactions in yeast," *Molecular Systems Biology,* vol. 1, 2005.

[4] S. Paladugu, S. Zhao, A. Ray, et al., "Mining protein networks for synthetic genetic interactions," *BMC Bioinformatics,* vol. 9, pp. 426-426, 2008.

[5] S. J. Dixon, Y. Fedyshyn, J. L. Y. Koh, 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,* vol. 105, pp. 16653-16658, 2008.

[6] S. L. Wong, L. V. Zhang, A. H. Y. Tong, et al., "Combining biological networks to predict genetic interactions," *Proceedings of the National Academy of Sciences,* vol. 101, pp. 15682-15687, 2004.

[7] C. Boone, H. Bussey, and B. J. Andrews, "Exploring genetic interactions and networks with yeast," *Nature Reviews Genetics,* vol. 8, pp. 437-449, 2007.

[8] G. W. Carter, D. J. Galas, T. Galitski, et al., "Maximal Extraction of Biological Information from Genetic Interaction Data," *PLoS Comput Biol,* vol. 5, pp. e1000347-e1000347, 2009.

[9] A. P. Järvinen, J. Hiissa, L. L. Elo, et al., "Predicting Quantitative Genetic Interactions by Means of Sequential Matrix Approximation," *PLoS ONE,* vol. 3, 2008.

[10] W. Zhong and P. W. Sternberg, *Genome-wide prediction of C. elegans genetic interactions*: American Association for the Advancement of Science, 2006.

[11] M. Deng, S. Mehta, F. Sun, et al., "Inferring domain-domain interactions from protein-protein interactions," in *Proceedings of the sixth annual international conference on Computational biology*, Washington, DC, USA, 2002, pp. 117-126.

[12] C. Huang, F. Morcos, S. P. Kanaan, et al., "Predicting protein-protein interactions from protein domains using a set cover approach," *IEEE ACM Transactions on Computational Biology and Bioinformatics,* vol. 4, pp. 78-78, 2007.

[13] E. Sprinzak and H. Margalit, "Correlated sequence-signatures as markers of protein-protein interaction," *Journal of Molecular Biology,* vol. 311, pp. 681-692, 2001.

[14] A. Bateman, E. Birney, L. Cerruti, et al., "The Pfam protein families database," *Nucleic acids research,* vol. 30, pp. 276-276, 2002.

[15] J. M. Cherry, C. Adler, C. Ball, et al., "SGD: Saccharomyces genome database," *Nucleic Acids Research,* vol. 26, pp. 73-73, 1998.

[16] I. Xenarios, L. Salwinski, X. J. Duan, et al., "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks

of protein interactions," *Nucleic acids research,* vol. 30, pp. 303-303, 2002.

[17] H. W. Mewes, K. Heumann, A. Kaps, et al., "MIPS: a database for genomes and protein sequences," *Nucleic acids research,* vol. 27, pp. 44-44, 1999.

[18] X. W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Bioinformatics,* vol. 21, pp. 4394–4400-4394–4400, 2005.

[19] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines, 2001," *Software available at* http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001.

[20] H. Chih-Wei, C. Chih-Chung, and L. Chih-Jen, "A practical guide to support vector classification," *National Taiwan University,* pp. 1-12, 2004.