

Meta-analysis of Genomic and Proteomic Features to Predict Synthetic Lethality of Yeast and Human Cancer

Min Wu¹, Xuejuan Li¹, Fan Zhang¹, Xiaoli Li², Chee-Keong Kwoh¹, and Jie Zheng^{1,3,*}

¹ School of Computer Engineering, Nanyang Technological University, Singapore 639798

² Data Analytic Department, Institute for Infocomm Research, A*STAR, Singapore 138632

³ Genome Institute of Singapore, A*STAR, Singapore 138672

{wumin, lixj, asckkowh}@ntu.edu.sg, fzhang005@e.ntu.edu.sg, xlli@i2r.a-star.edu.sg,

*Corresponding author: zhengjie@ntu.edu.sg

ABSTRACT

A major goal in cancer medicine is to find selective drugs with reduced side-effect. A pair of genes is called synthetic lethality (SL) if mutations of both genes will kill a cell while mutation of either gene alone will not. Hence, a gene in SL interactions with a cancer-specific mutated gene will be a promising drug target with anti-cancer selectivity. Wet-lab screening approach is still so costly that even for yeast only a small fraction of gene pairs has been covered. Computational methods are therefore important for large-scale discovery of SL interactions. Most existing approaches focus on individual features or machine learning methods, which are prone to noise or overfitting. In this paper, we propose an approach of meta-analysis that integrates 17 genomic and proteomic features and the outputs of 10 classification methods. It thus combines the strengths of existing methods. It also adjusts relative contributions of multiple methods with weights learned from the training data. Running on a dataset of the yeast strain of *S. cerevisiae*, our method achieves AUC (area under ROC curve) of 87.2% the highest among all competitors. Moreover, through orthologous mapping from yeast to human genes, we predicted a list of SL pairs in human that contain top mutated genes in lung and breast cancers recently reported by The Cancer Genome Atlas (TCGA). Our method and predictions would shed light on mechanisms of SL and lead to discovery of novel anti-cancer drugs.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: [General]; J.3 [Life And Medical Sciences]: [Biology and Genetics]

General Terms

Algorithms

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

BCB'13, September 22 - 25 2013, Washington, DC, USA
Copyright ©2013 ACM 978-1-4503-2434-2/13/09...\$15.00.

Keywords

Synthetic lethality, Cancer, Classification, Meta-analysis, Comparative genomics, TCGA

1. INTRODUCTION

The current anti-cancer drug development is faced with multiple issues, such as low approval rate of new drugs despite enormous amounts of money and time invested in the drug discovery; emergence of drug-resistance; side-effects of single-target drugs [3, 13]. Recently, a novel anti-cancer strategy called “Synthetic Lethality” has shown great potential to address these issues. A pair of genes is defined synthetic lethality (SL) if mutation of either gene alone has little effect on the cell but mutations of both genes would lead to cell death [3, 13]. While DNA mutation rate is extremely low in normal cells, there are abundant somatic mutations in cancer cells. Thus, a drug that targets the SL partner gene of a cancer-specific mutated gene will kill cancer cells, but spare normal cells. Originally discovered in genetic experiments of yeast and fruit fly [11], SL was proposed by Hartwell *et al.* as a new framework for anti-cancer therapies in 1997 [12], and since then has been under intense research. Recently, clinical success for breast cancer therapy has been achieved by an SL based drug, namely the inhibitor of PARPs (Poly ADP-ribose polymerases), which has SL interactions with the BRCA1 and BRCA2, two well-known genes for DNA repair whose mutations lead to breast cancer [3].

The dominant approach to discovery of SL is high throughput screening using RNA interference (RNAi) or compound libraries. For instance, Tong *et al.* developed a genome-wide strategy for the construction of double mutants named synthetic genetic array (SGA) analysis [26]; Ooi *et al.* introduced the genomic approach of synthetic lethality analysis by microarray (SLAM) [18]. In addition, a variant of the SGA method, called Epistatic Miniarray Profiling (E-MAP), was developed to quantify the synthetic effects [7, 6]. However, the screening based approach has limitations, e.g. high cost, false positive, lack of mechanistic interpretation, inconsistency among cell lines. As a result, very few SL pairs have been discovered in human cancer [8]. With abundant data of genetic interactions including SL, yeast (*S. cerevisiae*) is a popular model organism for cancer research. It is also because several pathways critical for cancer (e.g. DNA damage, cell cycle) are highly conserved between human and yeast [8]. Nevertheless, even for yeast, the number of known SL pairs is low compared with all possible genetic

interactions. Many potential SL candidates remain to be discovered for yeast as well as other model organisms (e.g. *C. elegans*, zebrafish). To this end, computational prediction can play important roles, as a cheap and efficient approach complementary to the wet-lab screening. Moreover, computational methods of systems biology based on pathway modelling and functional analysis could shed light on mechanisms of SL interactions.

Recently, several machine learning methods have been proposed and tested on the benchmark dataset of yeast, which show computational methods have great potential to analyse and predict SL [22, 19, 15, 20]. Qi *et al.* applied diffusion kernels defined on the network of yeast SL interactions in a support vector machine (SVM) classifier [22]. Paladugu *et al.* extracted multiple features from protein-protein interaction networks, which were used in an SVM to predict new SL interactions [19]. Li *et al.* used protein domain as the main type of features to achieve high performance of SL prediction [15]. These methods tend to focus on a particular type of features and to use a single machine learning method. However, as a highly complex cellular phenomenon, SL interaction is likely to be caused by different mechanisms. Thus, integrative analysis of multiple features would be desirable. Pandey *et al.* proposed a method called “MNMC” (multi-network multi-classifier) that integrates results of multiple predictive methods into one system [20]. However, MNMC combines the predictions of multiple classifiers without considering their difference in predictive performances. As such, if any classifier employed makes poor predictions, the overall performance of MNMC may be affected.

In this paper, we propose a meta-analysis approach called “MetaSL” which integrates multiple features into multiple predictive models. In contrast to MNMC, MetaSL assigns different weights to the predictions from various models, according to their performances (measured by AUC) during training process. In other words, the final decisions will be made based on a weighted consensus derived from votes of the participating classifiers. Running on yeast benchmark data, MetaSL is able to achieve an AUC of 87.2%, better than MNMC and other methods. Moreover, we conducted analysis of feature ranking output by MetaSL, which provides biological insights into the observed SL of yeast. Mapping yeast SL to orthologous human genes frequently mutated in cancer patients, using new data released by TCGA recently [17, 16], this paper reports human SL candidates that may lead to novel drug targets for lung and breast cancers.

2. METHODS

2.1 Features from multiple data sources

Synthetic lethality (SL) means that two non-essential genes result in a lethal phenotype. Therefore, two genes with a SL relationship generally have back-up functions. As such, we collect a category of features to measure the similarity between two genes, including GO semantic similarity, topological similarity in PPI networks, gene expression correlation and so on. We denote these features as similarity-based features (S features in short). In addition, each gene in the SL pairs is non-essential. We thus collect features for individual genes to reflect their propensity to be non-essential and these features are denoted as lethality-based features (L fea-

tures in short). All these features are summarized in Table 1.

Next, we briefly introduce the coding of features from various data sources. We calculate the semantic similarity between genes based on the GO term similarity that is defined in [28]. As we know, GO has 3 sub-ontologies (biological process, molecular function and cellular component) and we are able to calculate a semantic similarity for two genes in each sub-ontology of GO. Therefore, we have 3 features for GO semantic similarity between genes. For two genes in a PPI network, the number of their common neighbours can be utilized to measure their similarity. We employ a simplified variant of FSweight [5, 29] to show the topological similarity between two genes. In Tandem Affinity Purification with Mass Spectrometry (TAP-MS) experiments, two proteins occurring more frequently in the same purifications (i.e., bait-prey and prey-prey relationships) tend to have a higher similarity. Here, we utilize a recently-developed method called C2S [31] to calculate the similarity from TAP-MS data. For two genes, the Pearson Correlation Coefficient between their expression profiles is applied to measure their similarity. In addition, similarity-based features for two genes in this paper include their co-complex membership, co-pathway membership, whether or not they are paralogs as well as the number of their common or interacting domains.

For each gene, the degree (i.e., the number of incident edges) in a PPI network, the number of paralogs and the number of domains are used as lethality-based features. In total, 17 features are used to predict SL pairs, consisting of 11 similarity-based features and 6 lethality-based features.

2.2 Individual classifiers and the meta-classifier

Once we collected the features for gene pairs, various classifiers can be applied for predicting whether a given pair is a SL pair or not. In this paper, 8 classifiers from the WEKA machine learning suite [10] were used, namely, random forest, J48 (a type of decision tree), Bayesian logistic regression, Bayesian network, PART (a rule-based classifier), RBFNetwork, bagging (bootstrap aggregating) and classification via regression. Among the 8 classifiers, random forest is a well-known ensemble classifier. A random forest is a set of decision trees such that each tree is built from a random subset of features [2]. In addition, support vector machines (SVM) is a state-of-the-art classification technique and it has been proven to one of the best classifiers in many application domains [27]. SVM will find a maximum-margin hyperplane for classification by solving a convex optimization problem. Thus, we also explored SVM with linear and gaussian RBF kernels (using SVM^{light} software [14]) for predicting SL pairs. With SVM (2 kernels) and the above 8 classifiers from WEKA, we have 10 individual classifiers in all.

Given a gene pair x , assume that $p_i(x)$ is the probability of x to be SL as predicted by the i^{th} classifier ($1 \leq i \leq N$, N is the number of classifiers and is 10 in this paper). The MNMC method [20] combined the results from the above 10 classifiers in Equation 1. Here, $\prod_{i=1}^N p_i(x)$ and $\prod_{i=1}^N (1 - p_i(x))$ are the products of the probabilities of the instance x to be SL and non-SL respectively. The score $p(x)$ as their difference will thus provide a more accurate estimate of the likelihood of x to be true SL.

Table 1: Data sources and features for predicting SL pairs.

Data sources	Features	# of features	Remark	Category
Gene Ontology	Semantic similarity	3	3 sub-ontologies	S
PPI network	Topological similarity	1	FS-weight	S
	Degree in PPI network	2	for individual protein	L
TAP-MS	Similarity based on purifications	1	C2S scores	S
Protein complexes	Co-complex membership	2	real and predicted complexes	S
Pathways	Co-pathway membership	1		S
Gene expression	Gene expression correlation	1	Pearson Correlation	S
Paralog	Paralog pair	1		S
	The number of paralogs	2	for individual protein	L
Domain	Common/interacting domains	1		S
	The number of domains	2	for individual protein	L

$$p(x) = \prod_{i=1}^N p_i(x) - \prod_{i=1}^N (1 - p_i(x)) \quad (1)$$

As we know, the above individual classifiers may have different performance for classification. However, the MNMC method treats them equally when combining them in Equation 1 and does not take their relative importance into account. In this work, we apply the following weighted sum in Equation 2 to combine the individual classifiers. We assign the weight w_i to the classifier i based on its classification performance during the training process, e.g., a classifier with higher performance will be assigned with a larger weight. Here, we measure the classification performance for classifiers using the AUC. It is the area under the Receiver Operating Characteristics (ROC) curve, which is a graphical plot of the sensitivity vs. 1-specificity for a classifiers as the decision threshold varies.

$$p(x) = \sum_{i=1}^N w_i \times p_i(x) \quad (2)$$

2.3 Prediction of SL in human cancers

After individual classifiers were trained from the benchmark data, we are thus able to conduct prediction for novel gene pairs in yeast based on Equation 2. However, the classifiers can not be trained directly for human due to the limited number of known human SL pairs. Alternatively, given two human genes h_i and h_j , their propensity score to be SL, $p(h_i, h_j)$ can be inferred from $p(y_i, y_j)$ where y_i and y_j are the yeast orthologs of h_i and h_j , respectively. As such, we can still conduct *de novo* prediction for any two human genes when they both have orthologs in yeast.

In particular, we focus on the prediction of SL in human cancers. First, we collect significantly mutated genes in breast and lung cancers recently reported by The Cancer Genome Atlas (TCGA) [16, 17]. Second, two human genes are considered as a candidate SL pair when they satisfy the following two conditions, (1) both have orthologs in yeast and (2) one of them is a mutated gene prevalent among cancer patients.

3. RESULTS

3.1 Experimental data

Yeast SL data were downloaded from BioGRID [24]. Originally, there are 10,885 SL pairs in total. However, some of them contain essential genes, which should be excluded because by definition of SL each single gene in a SL is non-essential. With the list of essential genes downloaded from <http://bioinfo.mbb.yale.edu/genome/yeast/cluster/essential/>, we collected 7,347 SL pairs where every gene is non-essential. To train various classifiers, we considered these 7,347 SL pairs as positive data and generated the same number of random pairs (they are not involved in the positive data and have no essential genes) as negative data.

Gene ontology (GO) data were downloaded from <http://www.geneontology.org/>. The yeast PPI data (e.g., DIP data [23]), gene expression profiles for yeast genes and their domain information were downloaded from [29]. The real complexes were downloaded from Wodak’s lab [21] and the predicted complexes were generated by the COACH algorithm [30] from DIP data. The pathways were collected from SGD database. The orthologs between yeast and human genes, and the paralogs for yeast genes were downloaded from the Ensembl database. Frequently mutated genes in human breast and lung cancers were obtained from recent reports of TCGA [16, 17].

3.2 Feature importance analysis

In our dataset, 17 features are used for predicting SL pairs as shown in Table 1. Next, we aim to answer the question—which features are most correlated with SL prediction?

After training the linear SVM, the absolute values of the feature weights or coefficients show the importance of these features [4], i.e., the larger $|c_j|$ is (c_j is the coefficient for the j^{th} feature), the more important is the j^{th} feature in SL prediction. In addition, the coefficients from LASSO [25, 32] also indicate the importance of individual features. Table 2 shows the feature importance indicated by both SVM and LASSO coefficients, in which the first column shows individual features. For instance, GO_BP_Sim, GO_CC_Sim and GO_MF_Sim represent the semantic similarity between two genes based on the three GO sub-ontologies — biological process (BP), cellular component(CC) and molecular function (MF), respectively. In addition, Paralog_A and Paralog_B are the numbers of paralogs of two genes, while Paralog_AB represents whether these two genes themselves are paralogs of each other. The second and fourth columns are

the coefficients from linear SVM and LASSO respectively.

For these two feature rankings in Table 2, their Spearman Correlation Coefficient is 0.75 with p-value 0.00079, demonstrating that they are quite consistent. For example, the C2S scores from TAP-MS data are both ranked 1st by the two methods. In addition, the lethality-based features, such as degree, paralog and domain for individual genes, have similar importance indicated by both methods.

Table 2: Feature importance indicated by SVM and LASSO coefficients.

Features	SVM	Rank	LASSO	Rank
C2S_Sim	4.513	1	4.76	1
Degree_A	4.111	2	1.2319	4
Degree_B	3.606	3	1.3284	3
Paralog_A	-2.497	4	-0.7769	5
Paralog_B	-2.257	5	-0.6273	8
Paralog_AB	1.949	6	0.2494	11
Domain_B	1.917	7	0.7575	6
GO_CC_Sim	1.421	8	0.653	7
Domain_A	1.129	9	0.4055	9
GO_BP_Sim	0.991	10	0.082	13
GO_MF_Sim	-0.814	11	-0.3235	10
Domain_AB	0.713	12	-0.0181	17
PPI_FSweight	0.678	13	-2.7049	2
Co-Pathway	0.579	14	0.1236	12
GeneExpression	0.548	15	-0.0737	14
Co-Complex-Real	0.453	16	-0.0585	15
Co-Complex-Pre	0.404	17	0.0551	16

The C2S score [31] was originally proposed to measure the co-complex membership between two proteins. It is ranked as the most important feature, indicating that co-complex information are highly correlated with SL prediction. However, two features based on the co-complex membership in both real and predicted complexes have low importance demonstrated by both SVM and LASSO coefficients. The reason could be that only a small number of positive and negative SL pairs are co-complex pairs, e.g., 167 out of 7,347 positive SL pairs and 5 out of 7,347 negative SL pairs are co-complex pairs in real complexes. Interestingly, the significant difference between the numbers of positive and negative co-complex SL pairs (167 vs. 5 in real complexes and 154 vs. 4 in predicted complexes) leads to high gain ratio scores (another feature importance indicator) for these co-complex based features. In particular, two features Co-Complex-Real and Co-Complex-Pre have gain ratio scores 0.104 (Ranked 4th) and 0.105 (Ranked 3rd). Therefore, the two features based on co-complex memberships have high gain ratio scores, which is consistent with the high rank of C2S scores measured by SVM and LASSO coefficients.

As described above, our 17 features can be divided into two categories, i.e., similarity-based features (S features) and lethality-based features (L features). Figure 1 shows the performance of individual classifiers on the S and L features respectively. We can thus make the following two observations. First, various classifiers achieve significantly higher performance (i.e., AUC) on S features than L features, implicating S features are more important for the prediction of SL pairs. Second, individual classifiers generally achieve better performance after we combined both S and L fea-

tures (except for two classifiers J48 and RBFNetwork). This demonstrates that L features are also helpful although S features are relatively more important.

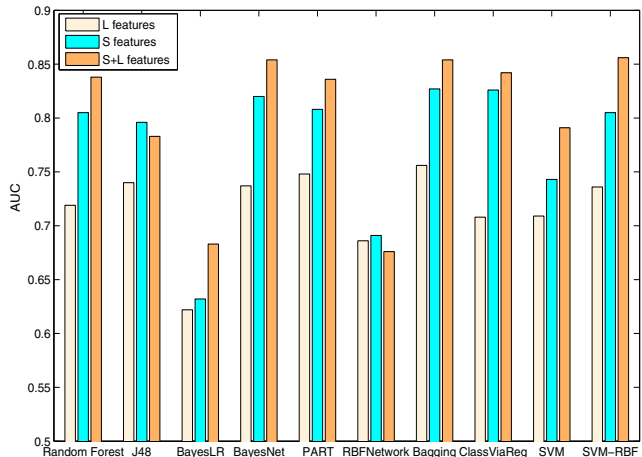


Figure 1: The performance of various classifiers across different feature sets.

3.3 Performance of individual classifiers and MetaSL

We divided our data into two parts, i.e., two-thirds of the data for training and one-third for testing. On the training data, we perform 5-fold cross validation and then obtain the AUC for individual classifiers. With the weights for classifiers based on their AUC, we are able to combine the results of various classifiers on the test set using Equation 2. Table 3 shows the AUC of various classifiers. It is obvious that Bagging, SVM with RBF kernel and Bayesian Network (BayesNet) achieve higher AUC than other classifiers. Assume the i^{th} classifier achieves an AUC of x_i , its weight for MetaSL, w_i as shown in the third column in Table 3, is scaled by $(x_i - min)/(max - min)$, where $max = 0.854$ (achieved by both Bagging and SVM with RBF kernel) and $min = 0.679$ (achieved by Bayesian Logistic Regression, i.e., BayesLR in Table 3).

Table 3: AUC for various classifiers on the training data as well as their weights for MetaSL.

Classifiers	AUC	weights
Random forest	0.841	0.926
J48	0.772	0.531
BayesLR	0.679	0
BayesNet	0.853	0.994
PART	0.836	0.897
RBFNetwork	0.699	0.114
Bagging	0.854	1
ClassificationViaRegression	0.837	0.903
SVM	0.794	0.657
SVM-RBF	0.854	1

With the above weights in Table 3, we show the performance of various classifiers as well as MNMC and our

MetaSL on the test data in Figure 2. Bayesian network (BayesNet) with an AUC of 0.853 is the best performer among the 10 individual classifiers. Meanwhile, MNMC achieves an AUC 0.841. With an AUC of 0.872, MetaSL outperforms MNMC and the 10 individual classifiers as shown in Figure 2. In addition, we also tried different scaling for the weights of classifiers, i.e., $w_i = \alpha * (x_i - min)/(max - min) + 1 - \alpha$. For $\alpha = 0.7, 0.8$ and 0.9 , the AUC values for our MetaSL are 0.870, 0.871 and 0.871, respectively. The result shows that the performance of MetaSL is robust to different scaling of the weights.

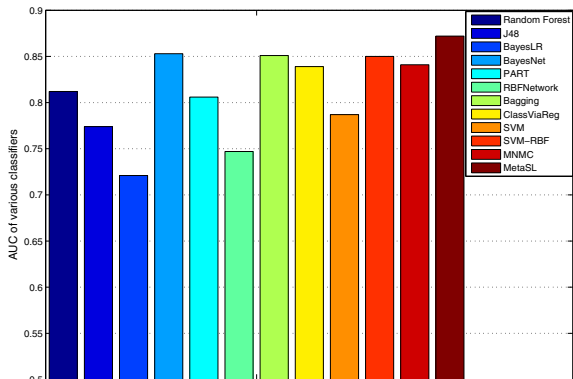


Figure 2: AUC of various classifiers including MNMC and MetaSL on the test data.

3.4 Results for predicted yeast SL pairs

We introduced the results on the training and testing data in the above subsections. Next, we show the results of novel yeast SL pairs predicted by MetaSL. In our experiments, there are 5,504 non-essential yeast genes. The number of non-essential genes in reality may be less than 5,504 because the list of 694 essential genes here is still far from complete. We aim to make predictions for all the gene pairs, i.e., $\frac{5504 * 5503}{2} - 2 * 7347 = 15,129,562$ pairs. However, due to this large number of gene pairs, the computation is extremely time- and space-consuming. Therefore, we only selected a subset of 100,000 candidate pairs with the highest C2S scores. As C2S score is the most important feature for predicting SL pairs from the benchmark dataset as shown in Table 2, these 100,000 pairs are more likely to be true SL interactions than the remaining 15,029,562 pairs.

Table 4 shows the top 10 yeast SL interactions predicted by MetaSL. We observe that in these gene pairs the two genes tend to have high functional similarity, e.g., 9 out of these 10 pairs have GO similarity higher than 0.65 as shown in the third column. In particular, it is interesting to notice that the following three pairs, (YNL104C, YOR108W), (YLR186W, YPL217C) and (YBR009C, YNL030W), have already been reported as genetic interactions in BioGRID [24]. Moreover, the two genes in the pair (YNL104C, YOR108W), which is ranked as 2nd in Table 4, has synthetic growth defect as validated by experiments [9]. In addition, all the three gene pairs (YNL104C, YOR108W), (YBR009C, YNL030W) and (YLR270W, YOR173W) share common protein domains. Therefore, we believe that gene pairs top-ranked by MetaSL

provide good candidates for experimental screening in the future.

3.5 Predicted SL pairs in human cancers

With the frequently mutated genes collected from TCGA [17, 16], we generated 34,841 candidate SL pairs in human cancers and ranked them based on the scores that were assigned to their yeast ortholog pairs by MetaSL. Table 5 shows the top 14 potential SL pairs in human cancers.

In Table 5, GATA3 is a frequently mutated gene in breast cancer. It is likely to form potential SL pairs with other GATA family members, e.g., GATA1, GATA2, GATA4, GATA5 and GATA6, for the following potential reasons. First, two yeast candidate SL pairs, (YFL021W, YJL110C) and (YFL021W, YER040W), are both top ranked by MetaSL. YFL021W and GATA3 are orthologous of each other while YJL110C and YER040W are both orthologous to the GATA family. The SL interactions between GATA3 and other GATA family members are thus supported by *multiple* yeast SL pairs. Second, they are in the same family and thus have similar functions, for instance, GATA3 and GATA1 have 8 GO annotations in common while GATA3 and GATA2 have 6. Lastly, members of GATA family are transcription factors and they are all annotated with the function “regulation of transcription from RNA polymerase II promoter”. Many SL interactions involving transcription initiation factors and RNA polymerase II genes have been reported both experimentally and computationally in previous studies [1, 15].

4. DISCUSSION AND CONCLUSIONS

Synthetic lethality based anti-cancer treatment is an emerging strategy that targets critical difference between normal and tumor cells, thereby killing tumor cells selectively. The sequencing technologies have provided new data about somatic mutations and other alterations in cancer. Finding SL partners of these cancer-specific mutated genes would provide candidates of drug targets. However, due to high cost of wet-lab screening of genetic interactions, there is a dearth of confirmed SL in human cancer. With abundant benchmark data, yeast is a good model organism for the study of SL. But even for yeast, the number of benchmark SL may be still low. Thus, computational methods play important roles for large-scale discovery of SL.

In this paper, we proposed an integrative approach that combines multiple genomic and proteomic features, and multiple machine learning methods into one meta-analysis system, called MetaSL. Our features consist of those depicting similarity between two genes and the lethality of single genes. As far as we know, only one previous method (called “MNMC”) combines multiple classifiers to predict SL [20]. However, MNMC treats the results of different methods equally, despite their different performances. By contrast, our method of MetaSL takes into account the differences of predictive methods, using AUC-based weights learned from the training data of yeast SL. Note that the different weights are not specific to the yeast dataset, but largely due to inherent difference of strengths among the classifiers. We have also analysed the relative importance of features to the predictive performance, which sheds lights on causal factors of SL interactions. Testing on the SL benchmark data of *S. cerevisiae*, MetaSL achieved an AUC of 87.2% the highest among all methods of SL prediction. Furthermore, through orthologous mapping between human and yeast genes, we

Table 4: Top 10 predicted yeast SL pairs and their GO term similarity.

Rank	Gene A	Gene B	GO similarity	Common GO terms
1	YMR128W	YPL217C	0.787	
2	YNL104C	YOR108W	1	leucine biosynthetic process
3	YLR186W	YNL075W	0.652	
4	YHR148W	YOR310C	1	ribosome biogenesis and assembly
5	YLR186W	YPL217C	0.525	
6	YKL172W	YLR276C	0.691	ribosome biogenesis and assembly
7	YLR270W	YOR173W	1	deadenylation-dependent decapping
8	YBR065C	YPL151C	0.707	nuclear mRNA splicing, via spliceosome
9	YBR009C	YNL030W	1	chromatin assembly or disassembly
10	YNL075W	YPR144C	0.771	ribosome biogenesis and assembly

identified SL partners of the most prevalent mutated genes in lung and breast cancers using TCGA data.

In spite of promising performance of MetaSL, we have noticed its limitations which point to future work. First, although our feature weight ranking and GO analysis provide some clues about causal factors of SL, the underlying mechanisms of SL remain unclear. To address this issue in future, we will add pathway analysis to interpret the discovered SL pairs. Using additional post-processing, we hope to filter out false positives, and select top reliable SL candidates for experimental study. Second, the number of features we have used here is still low, and there are inter-dependence among features. In future, we will collect a comprehensive set of features (e.g. considering epigenetic features of histone modifications) and conduct feature selection before training our model. Third, the inference of SL in human cancer genes from yeast SL candidates is hindered by the scarcity of known orthologous relation between the two species. For instance, no yeast orthologous gene is known for p53, a critical gene for human cancer. Moreover, some yeast genes have multiple orthologous genes in human, and vice versa. In the future, we will also consider functional and structural homologs between human and yeast, and design a reliable algorithm to map genes between the two species.

Overall, synthetic lethality based cancer medicine is still in its infancy. Our method of MetaSL combines strengths of previous computational methods by meta-analysis of genome-wide features of yeast genes. By integrating additional data and knowledge, we will not only improve the predictive performance, but also gain mechanistic understanding of synthetic lethality, which will contribute to the design of next-generation anti-cancer therapies.

Acknowledgement

This research has been supported by Tier 1 AcRF Grant RG32/11 (M4010977.020) from Ministry of Education, Singapore and NTU Start up grant (CoE_SUG/RSS_1FEB11.1/8), Singapore.

5. REFERENCES

- [1] J. Archambault, F. Lacroute, A. Ruet, and J.D. Friesen. Genetic interaction between transcription elongation factor *tfiis* and rna polymerase ii. *Mol Cell Biol*, 12(9):4142–52, 1992.
- [2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [3] D.A. Chan and A.J. Giaccia. Harnessing synthetic lethal interactions in anticancer drug discovery. *Nat Rev Drug Discov*, 10(5):351–364, 2011.
- [4] Yin-Wen Chang and Chih-Jen Lin. Feature ranking using linear svm. *Journal of Machine Learning Research - Proceedings Track*, 3:53–64, 2008.
- [5] Hon Nian Chua, Wing-Kin Sung, and Limsoon Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.
- [6] S. R. Collins, K. M. Miller, N. L. Maas, A. Roguev, J. Fillingham, C. S. Chu, and *et al.* Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, 446(7137):806–810, 2007.
- [7] S. R. Collins, M. Schuldiner, N. J. Krogan, and J. S. Weissman. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome biology*, 7(7):R63, 2006.
- [8] N. Conde-Pueyo, A. Munteanu, and *et al.* Human synthetic lethal inference as potential anti-cancer target gene detection. *BMC Syst Biol*, 3:116, 2009.
- [9] A. DeLuna, K. Vetsigian, N. Shoresh, M. Hegreness, M. Colon-Gonzalez, and *et al.* Exposing the fitness contribution of duplicated genes. *Nat Genet*, 40(5):676–681, 2008.
- [10] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.
- [11] J. L. Hartman, B. Garvik, and L. H. Hartwell. Principles for the buffering of genetic variation. *Science*, 291(5506):1001–4, 2001.
- [12] L.H. Hartwell, P. Szankasi, C.J. Roberts, A.W. Murray, and S.H. Friend. Integrating genetic approaches into the discovery of anticancer drugs. *Science*, 278(5340):1064–1068, 1997.
- [13] J. D. Iglehart and D. P. Silver. Synthetic lethality—a new direction in cancer-drug development. *N Engl J Med*, 361(2):189–91, 2009.
- [14] T. Joachims. Making large-scale svm learning practical. *Advances in Kernel Methods: Support Vector Machines*, 1999.
- [15] B. Li, W. Cao, J. Zhou, and F. Luo. Understanding and predicting synthetic lethal genetic interactions in *saccharomyces cerevisiae* using domain genetic

Table 5: Top predicted SL pairs in human cancers. The genes in bold are significantly mutated in cancers.

Gene A	Gene B	Common GO terms
PRPF6	SF3B1	RNA splicing, via transesterification reactions nuclear mRNA splicing, via spliceosome RNA splicing
TBL1XR1	ING1	positive regulation of transcription, DNA-dependent transcription, DNA-dependent
TBL1XR1	ING2	chromatin modification positive regulation of transcription, DNA-dependent
TBL1XR1	ING3	transcription, DNA-dependent
TBL1XR1	ING4	negative regulation of transcription, DNA-dependent
TBL1XR1	ING5	transcription, DNA-dependent positive regulation of transcription, DNA-dependent
PRPF3	SF3B1	RNA splicing, via transesterification reactions nuclear mRNA splicing, via spliceosome RNA splicing
GATA3	GATA1	in utero embryonic development transcription from RNA polymerase II promoter blood coagulation negative regulation of cell proliferation male gonad development erythrocyte differentiation embryonic hemopoiesis positive regulation of transcription from RNA polymerase II promoter
GATA3	GATA2	cell fate determination blood coagulation inner ear morphogenesis negative regulation of fat cell differentiation positive regulation of transcription from RNA polymerase II promoter cell maturation
GATA3	GATA4	in utero embryonic development transcription from RNA polymerase II promoter blood coagulation male gonad development response to drug response to estrogen stimulus positive regulation of transcription, DNA-dependent positive regulation of transcription from RNA polymerase II promoter
GATA3	GATA5	blood coagulation positive regulation of transcription from RNA polymerase II promoter
GATA3	GATA6	in utero embryonic development transcription from RNA polymerase II promoter blood coagulation male gonad development response to drug response to estrogen stimulus negative regulation of transcription, DNA-dependent positive regulation of transcription from RNA polymerase II promoter
NCOR1	SIN3A	negative regulation of transcription from RNA polymerase II promoter cellular lipid metabolic process
NCOR1	SIN3B	transcription, DNA-dependent cellular lipid metabolic process

- interactions. *BMC Syst Biol*, 5:73, 2011.
- [16] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.
- [17] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012.
- [18] S.L. Ooi, D.D. Shoemaker, and J.D. Boeke. Dna helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nat Genet*, 35(3):277–286, 2003.
- [19] S. Paladugu, S. Zhao, A. Ray, and A. Raval. Mining protein networks for synthetic genetic interactions. *BMC Bioinformatics*, 9(1):426, 2008.
- [20] Gaurav Pandey, Bin Zhang, Aaron N. Chang, Chad L. Myers, Jun Zhu, Vipin Kumar, and Eric E. Schadt. An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Computational Biology*, 6(9), 2010.
- [21] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 37(3):825–831, 2009.
- [22] Y. Qi, Y. Suhail, Y.Y. Lin, J.D. Boeke, and J.S. Bader. 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*, 18(12):1991–2004, 2008.
- [23] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 30(1):449–451, 2004.
- [24] C. Stark, B. Breitkreutz, A. Chatr-aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, and *et al*. The biogrid interaction database: 2011 update. *Nucleic Acids Research*, 39(Database-Issue):698–704, 2011.
- [25] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. of the Royal Statistical Society. Series B*, 58:267–288, 1996.
- [26] A.H. Tong, M. Evangelista, A.B. Parsons, H. Xu, G.D. Bader, and *et al*. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–2368, 2001.
- [27] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc. New York, NY, USA, 1995.
- [28] J.Z. Wang, Z. Du, R. Payattakool, P. Yu, and C. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.
- [29] M. Wu, X. Li, H. Chua, C.-K. Kwoh, and S.-K. Ng. Integrating diverse biological and computational sources for reliable protein-protein interactions. *BMC Bioinformatics*, 11(S7):S8, 2010.
- [30] M. Wu, X. Li, C. K. Kwoh, and S.-K. Ng. A core-attachment based method to detect protein complexes in ppi networks. *BMC Bioinformatics*, 10, 2009.
- [31] Z. Xie, C. K. Kwoh, X. Li, and M. Wu. Construction of co-complex score matrix for protein complex prediction from ap-ms data. *Bioinformatics*, 27(13):i159–i166, 2011.
- [32] Y. Yuan, Y. Xu, J. Xu, R. L. Ball, and H. Liang. Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data. *Bioinformatics*, 28(9):1246–1252, 2012.