USING RELATIVE IMPORTANCE METHODS TO MODEL HIGH-THROUGHPUT GENE PERTURBATION SCREENS

Ying Jin*, Naren Ramakrishnan, and Lenwood S. Heath
Department of Computer Science, Virginia Tech,
Blacksburg, VA 24061, U.S.A.
Email: {jiny,naren,heath}@cs.vt.edu.

Richard F. Helm
Department of Biochemistry, Virginia Tech,
Blacksburg, VA 24061, U.S.A.
Email: helmrf@vt.edu.

With the advent of high-throughput gene perturbation screens (e.g., RNAi assays, genome-wide deletion mutants), modeling the complex relationship between genes and phenotypes has become a paramount problem. One broad class of methods uses 'guilt by association' methods to impute phenotypes to genes based on the interactions between the given gene and other genes with known phenotypes. But these methods are inadequate for genes that have no cataloged interactions but which nevertheless are known to result in important phenotypes. In this paper, we present an approach to first model relationships between phenotypes using the notion of 'relative importance' and subsequently use these derived relationships to make phenotype predictions. Besides improved accuracy on *S. cerevisiae* deletion mutants and *C. elegans* knock-down datasets, we show how our approach sheds insight into relations between phenotypes.

1. INTRODUCTION

There are now a variety of mechanisms to study loss of function phenotypes in specific cell types or at different stages of development in an organism. Genome wide deletion mutants, e.g., for *Saccharomyces cerevisiae* [1, 2], use homologous recombination to replace genes with targeted cassettes so that the resulting strain can be screened for specific phenotypes (or lack thereof). RNA interference methodologies, in organisms such as *Caenorhabditis elegans* [3, 4], use post-transcriptional gene silencing to degrade specific RNA molecules, thus causing a drastic attenuation of gene expression. Since RNAi may not completely deplete the expressed RNA molecules, its use is referred to as a 'knock-down', in contrast to a complete 'knockout' exhibited by a deletion mutant. Through the use of high-throughput screens, both these techniques now support large scale phenotypical studies.

A central goal of bioinformatics research is to model the phenotype effects of gene perturbations. The mapping between gene function and expressed phenotype is complex. A single gene perturbation (through deletion or RNAi interference) can lead to a cascade of changes in transcription or post-transcriptional pathways. It is impractical to make a comprehensive empirical analysis when there is a large number of candidate genes. An emerging area of interest therefore is to use diverse, highly heterogeneous, data (e.g., microarrays, RNAi studies, protein-protein interaction assays) to computationally model phenotype effects for mutations.

Previous studies have shown that by considering interactions between candidate genes and target genes (which have been known to result in a desired phenotype) the accuracy of phenotype prediction can be improved. Examples of interactions that have been considered by such works include physical interactions between proteins [5], interactions underlying protein complexes [6], and integrated gene networks constructed from multiple data sources [7]. Most of these methods can be classified as 'direct' methods since they require a direct interaction between a gene and another gene with the target phenotype in order to predict the phenotype for the given gene.

Statistical and computational methods to prioritizing genes by using combinations of gene expression and protein interaction data have also been pro-

*Corresponding author.

posed, e.g., CGI [8] and GeneRank [9]. In addition to direct interactions, these methods take into account indirect interactions, i.e., links from genes to target genes through other intermediate genes. However, these approaches assume that there is at least one path from a candidate gene to some target gene(s). Since many genes do not have any catalogued interactions, this limits their applicability.

Markowetz *et al.* [10] proposed the NEM (nested effects models) approach to rank genes according to subset relations between phenotypes. NEM uses phenotype profiles only, i.e., it does not consider any protein-protein interactions. While this overcomes the limitations mentioned previously, NEM has shortcomings in scalability with respect to the number of phenotypes and to overcome the increased computational cost, NEM focuses on inference only from pairwise and triple relations.

Contributions: We propose a new graph theoretic approach to predicting phenotype effects of gene perturbation using phenotype relations (P^3). Our approach focuses on relative importance methods to infer relations between phenotypes and uses these relations to predict phenotype effects. We integrate phenotype profiles with the gene network to derive phenotype relations. It is assumed that genes tightly connected are likely to share the same phenotypes. We use a weighted directed graph to model the relations between phenotypes such that more complicated relations can be illustrated and interpreted instead of just subset relations. Since predictions are carried out purely based on the phenotype relations derived, there is no requirement for known interaction paths from candidate genes to target genes. Furthermore, once the relations between phenotypes are derived, they can be used repetitively in the prediction process. In particular, complete perturbation effects across all phenotypes can be predicted simultaneously from the relations between known phenotypes and others. Therefore, P^3 is more effective for large-scale phenotype prediction than previous methods that rank genes for each phenotype, one at a time. Experimental results on *S. cerevisiae* and *C. elegans* also show that our approach outperforms the direct and GeneRank methods consistently. In particular, for genes without any interactions in *S. cerevisiae*, we show that our method can predict 96% of their phenotypes with AUC (area under ROC curve) greater than 0.8, and 60% of the phenotypes in *C. elegans*.

2. WORKING EXAMPLE

Table 1 describes an example of phenotype profiles resulting from many gene perturbations. Each row represents a phenotype and each column a gene. The cell value indicates whether the gene perturbation exhibits the corresponding phenotype, e.g., g_1 gives rise to p_1 but not p_2 and p_3. A second form of data available is a gene network as shown in Figure 1 (left), that shows interactions between genes. For ease of interpretation, genes that result in the same phenotype as shown in Table 1 are also grouped in Figure 1 (left). Suppose that the only information about g_7 that we are given is that it results in phenotype p_3 and we desire to use computational methods to predict that it will also cause p_2 but not p_1 (see last column of Table 1).

Table 1. Example phenotype profiles.

	g_1	g_2	g_3	g_4	g_5	g_6	g_7
p_1	1	1	0	0	0	1	0
p_2	0	0	1	1	1	0	1
p_3	0	0	0	1	0	1	1

- **Using phenotype profiles:** If we were to use only Table 1 to make a prediction, it is not clear whether g_7 should result in p_1 or p_2. p_1 and p_2 involve three genes each, and p_3 has (exactly) one gene in common with both sets. Obviously, p_1 and p_2 have an equal chance to be predicted, no matter what association measure is used.
- **Using network information:** If we assume that all links in Figure 1 (left) have the same weight, then in fact the prediction result will be p_1. To see this, observe that g_7 has only one interaction partner g_2, and it is known that g_2 contributes to p_1 only. And there are no paths from g_7 to any genes resulting in phenotype p_2. Hence, no matter what graph theoretic methods are used, p_1 has a better chance of being predicted.

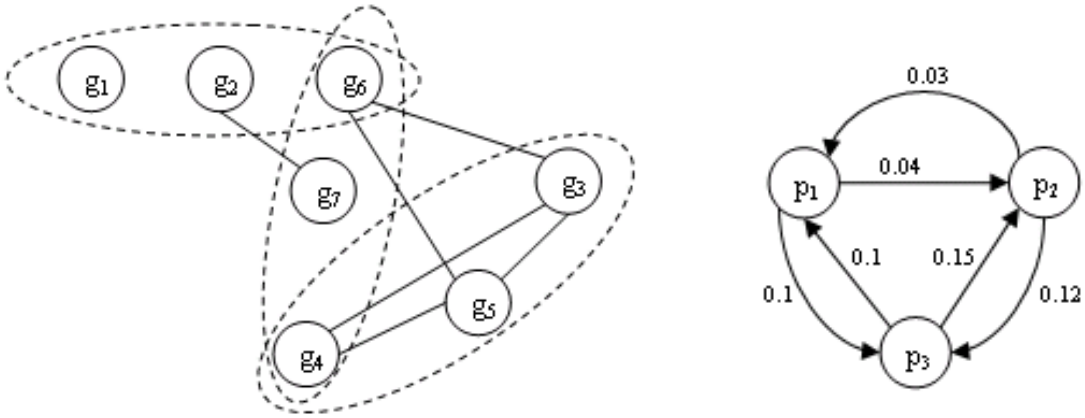

Fig. 1. (left) Example of gene network. (right) Induced relationships between phenotypes.

We propose to combine the selective superiorities of the two methods to model phenotypes. In this process, we develop a method that resembles a collaborative filtering algorithm [11] used in recommender systems research. First, we derive relationships between phenotypes from Table 1 and Figure 1 (left). Figure 1 (right) demonstrates the relationships between phenotypes obtained by applying our algorithm presented in the following section. The value on the arrow from phenotype p_i to phenotype p_j denotes the tendency that a gene perturbation causing p_i also causes p_j. From such a relation, we can predict that if a gene perturbation results in p_3, then it is more likely to result in p_2 rather than p_1. Some characteristics of existing methods and our approach are listed in Table 2.

3. METHODS

3.1. Inferring Relations Between Phenotypes

As stated earlier, inferring relations between phenotypes is a one-time cost and can be amortized over the prediction process. Our method is motivated by the study of relative importance in networks [12]. Original link analysis methods, such as PageRank [13] and HITS [14], rank nodes according to their "importance" or "authority" in the entire network. Relative importance analysis focuses on determining the importance of nodes with respect to a subset of the network, called the "root set." Multiple algorithms have been proposed for relative importance computation, such as k-short path, k-short node-disjoint paths, PageRank with priors, HITS with priors, and the k-step Markov approach, which are all surveyed by White and Smyth [12].

Suppose that there are n genes $G = \{g_i | 1 \leq i \leq n\}$, and m phenotypes $P = \{p_i | 1 \leq i \leq m\}$ in a study. Let $W_{n \times n}$ denote the connection matrix of the network, where $w_{i,j}$ denotes the weight of the connection between gene g_i and gene g_j. W is required to be a symmetric matrix whose diagonal is uniformly 0. For each phenotype p_j, there is a corresponding vector $\overline{p}_j = \langle v_1, v_2, ..., v_n \rangle$, where $v_i = 1$ indicates that gene g_i is known to result in p_j, otherwise $v_i = 0$. These vectors are grouped together to form a gene phenotype matrix $V_{m \times n}$, where rows are phenotypes and columns are genes. Given a phenotype p, genes resulting in this phenotype form a root set R. Similar to PageRank with priors, each gene is assigned a prior rank score, as shown in Equation 1. Observe that the sum of all initial rank scores is 1.

$$r^0_{g_i} = \begin{cases} \frac{1}{\|R\|} & \text{if } g_i \in R, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Let $N(g_i) = \{g_j | w_{i,j} > 0, i \neq j, \text{and } g_j \in G\}$ denote the set of all other genes that interact with g_i. Define parameter β, $0 \leq \beta \leq 1$, to be the relative weight between the original score of a gene and the score that results through the influence of its neighbors. The formula for iteratively computing gene rank scores is shown in Equation 2.

Table 2. Comparison of P^3 to other methods for phenotype prediction.

Method	Use phenotype profiles?	Use gene interactions?	Ability to rank phenotypes?	Induce phenotype relations?
Wormnet (Direct)		√		
GeneRank		√		
NEM	√		√	√
P^3	√	√	√	√

$$r_{g_i}^{k+1} = \beta r_{g_i}^0 + (1-\beta)\left(\sum_{g_j \in N(g_i)} \frac{w_{i,j}}{\pi_{g_i}} r_{g_j}^k\right). \quad (2)$$

Here, $\pi_{g_i} = \sum_{j=1}^{n} w_{i,j}$ is the total weight of interactions involving gene g_i. k in the equation indicates the number of iterations. After convergence, we obtain rank scores of all genes with respect to phenotype p. The above procedure can be repeated for every phenotype to obtain the corresponding list of rank scores of all genes. The list of rank scores of genes to a phenotype corresponds to a vector $\overline{\mathcal{R}_{p_i}} =< r_{g_1}, ..., r_{g_n} >$, where r_{g_k} is the rank score of g_k. Let $C_{m\times m}$ denote a "closeness" matrix of phenotypes, where both rows and columns are phenotypes, and each entry $c_{i,j}$ stores the closeness value from phenotype p_j to p_i. It is defined as the p_i's average rank scores of genes causing p_j. The formula is given in Equation 3, where $\overline{p_j}^T$ is the transpose of $\overline{p}_j$.

$$c_{i,j} = \overline{p}_j^T \times \overline{\mathcal{R}_{p_i}} \quad (3)$$

Note that this matrix is not necessarily symmetric, since the rank score of a gene to a phenotype depends on the scores of its neighbors, but for two phenotypes p and q, genes involved in phenotype p may not have the same neighbors as genes involved in phenotype q. For simplicity, the diagonal of the matrix is set to 0, because the closeness of a phenotype to itself is not of interest. This matrix thus maps to a weighted directed graph, such as seen in Figure 1 (right), where nodes are phenotypes, and the weight of the directed edge from phenotype p_i to phenotype p_j is $c_{i,j}$. After the whole matrix C is computed, prediction is carried out using this matrix.

3.2. Predicting Phenotype Effects of Gene Perturbations

Algorithms for ranking genes to a phenotype and ranking phenotypes for a gene using the phenotype graph are described below.

3.2.1. *Ranking Genes for a Phenotype*

Given a phenotype p, suppose that there is a gene g which is known to result in phenotypes $\{q_1, ...q_k\}$. The closeness of phenotype q_i, $1 \le i \le k$, to p is the weight of the edge from p to q_i in the phenotype graph. There are multiple ways to define the rank score of a gene g to the phenotype p, for example, we can utilize the maximum closeness from q_i, $1 \le i \le k$, to p. Here, we used the average closeness from known phenotypes of the gene to the target phenotype. The rank scores of all genes to all target phenotypes can be calculated simultaneously by a simple matrix computation, as shown in Equation 4.

$$RG = V' \times C \quad (4)$$

V', with entries $v'_{i,j} = \frac{v_{j,i}}{\sum_{k=1}^{m} v_{k,i}}$, is obtained by transposing the phenotype-gene matrix V and dividing each entry by the number of 1s in the corresponding row. RG is thus an $n \times m$ matrix, where rows are genes and columns are phenotypes, and the value of each cell is the rank score of the gene to the corresponding phenotype.

3.2.2. *Ranking Phenotypes for a Gene*

Given a gene g, assume that it is known to result in phenotypes $\{q_1, ..., q_k\}$. For any other phenotype p in the phenotype graph, the closeness from p to phenotype q_i, $1 \le i \le k$ is the weight of the edge from q_i to p. The method of ranking phenotypes to a gene is very similar to ranking genes for a phenotype, described above. In ranking genes, the weights

on the edges incident on phenotypes $\{q_1, ..., q_k\}$ are used, but in ranking phenotypes, the edges outgoing from phenotypes $\{q_1, .., q_k\}$ are considered. The rank score of phenotype p to gene g is the average of the closeness values from p to phenotypes $\{q_1, ..., q_k\}$. Analogously as stated earlier, rank scores of all phenotypes to all genes can be computed at the same time. Equation 5 describes the method, where RG is the resulting rank score matrix.

$$RP = V' \times C^T \tag{5}$$

The only difference from the method to ranking genes is that the transpose of the closeness matrix is used here.

4. EXPERIMENTAL RESULTS

We illustrate the effectiveness of our methodology by comparing it to the Direct method (as used in Lee et al. [7]) and GeneRank [9] on two real datasets: deletion mutants on yeast and an RNAi study of early embryogenesis in the *C. elegans* nematode. We further analyze the phenotype graphs derived by clustering phenotypes with high closeness values and present a biological interpretation.

4.1. Data

Two datasets are used in this study: the dataset of *C. elegans* RNAi induced early embryo defects [4] and the yeast knockout dataset from the Munich Information Center for Protein sequences (MIPS) database [15].

We focus on 45 RNAi induced defect categories in the *C.elegans* early embryo (data available in [16]) and use an interaction network extracted from Wormnet [7]. The original core Wormnet contains 113,829 connections and 12,357 genes. To compare with the Direct and GeneRank methods, we select genes resulting in at least two early embryo defects and interacting with at least one other gene, and retain all interactions between them in Wormnet. To evaluate the applicability of P^3 on predicting phenotypes for genes without interactions, we prepare another gene set that retains genes without any interactions.

In the yeast data, the underlying network involves protein-protein interactions, and is built by combining the yeast protein interaction data from several sources (CYGD [17], SGD [18], and BioGrid [19]). Phenotypes and genes are selected according to the same criteria as above. The statistics of these datasets are listed in Table 3.

4.2. Experiment Setup

We implement the Direct method and use the log-likelihood value of each interaction published with Wormnet as the edge weights for the *C. elegans* network. For a given phenotype, genes known to result in that phenotype are considered as the seed set. The rank score of other genes are the sum of the log-likelihoods of interactions to the seed set. In the case of yeast, we simply set the same weight on all interactions.

Table 3. Statistics of datasets used in this work.

Organism	Genes	Interactions	Phenotypes
Caenorhabditis elegans	420	6677	45
Saccharomyces cerevisiae	1232	13228	72

In addition to the connectivity matrix of the network, GeneRank has another input, namely the expression changes vector, which is used to set initial ranks. In our case, we use a binary phenotype signature vector, where 1 means that the corresponding gene is known to show that phenotype, 0 otherwise. There is also a parameter d that determines relative weights of expression changes and connectivity information to the rank value. We tried multiple values on d from 0.1 to 0.9 with interval 0.1, and chose the one gives optimal prediction results in performance comparison (0.1). The implementation published with the original paper is used.

To compare with the above methods, the algorithm for ranking genes for a given phenotype is applied. Another algorithm to ranking phenotypes for a given gene is used to predict phenotypes for genes without any interactions. There is one parameter β in P^3 to derive relations between phenotypes. We studied different values on β from 0.1 to 0.9 with step 0.1, and found that 0.6 gives the best performance. We used 0.6 in all the experiments described below.

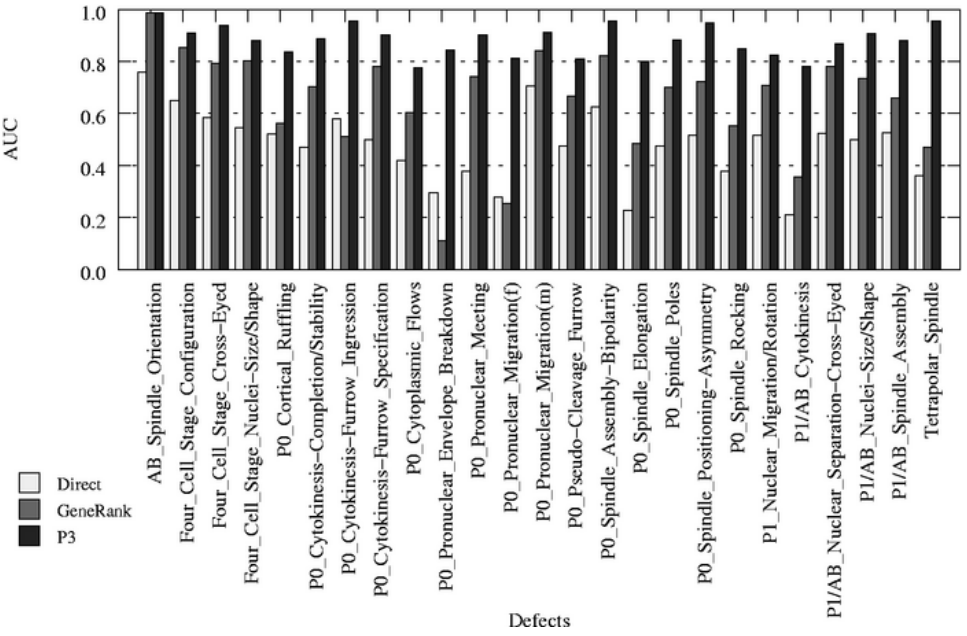

Fig. 2. Overall performance comparison on the *C. elegans* dataset. Direct : ranking genes using the interaction network only; GeneRank : $d = 0.1$; P^3 : $\beta = 0.6$

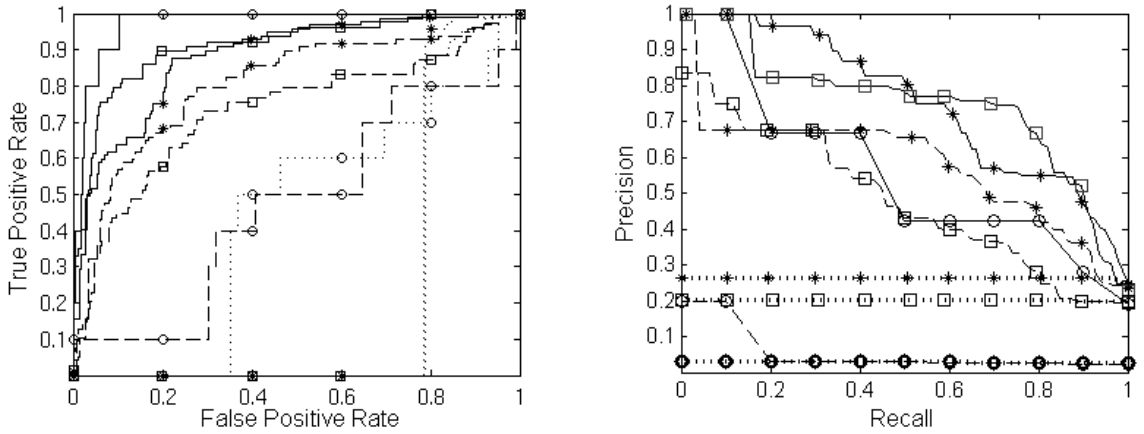

Fig. 3. (left) ROC curves on *C. elegans*. (right) Precision vs. Recall on *C. elegans*. Direct: points, GeneRank: dashed line, P^3: solid line; square: P1-AB Nuclear-Size-Shape, star: Four Cell Stage Nuclei-Size-Shape, circle: Tetrapolar Spindle.

4.3. Results

To evaluate the prediction performance for each phenotype, we used the leave-one-out and k-fold cross validation approaches. For the leave-one-out approach, one gene/phenotype pair is ignored from the original dataset each time, and the prediction algorithm is applied on the remaining dataset to see if that gene/phenotype pair is predicted correctly. Results show that our method outperforms the direct method and GeneRank method almost in all cases. We compared the Area Under the Receiver Operating Characteristic (AUC ROC) curve for each phenotype and plot the ROC curve and Precision-Recall curves for some phenotypes for further performance comparison. For k-fold cross validation, the original gene/phenotype pairs are separated into k groups, 10 in *C. elegans* and 5 in yeast, one of them is selected as test data and the remaining are used as training data. The distributions of AUC were compared. P^3 outperforms other methods in all cases. In the exper-

iment of predicting phenotypes to genes without any interactions, results show that P^3 is able to predict a majority of these phenotypes with high accuracy.

4.3.1. *Leave-one-out*

C. elegans: For each phenotype prediction, we computed true-positive rate versus false-negative rate to measure the recovery of genes with the given phenotype. The comparison of the area under the Receiver Operating Characteristic curve for each phenotype is shown in Figure 2. For visualization purpose, 20 defects are randomly selected for discussion here. The defect "AB Spindle Orientation" shows the highest AUC in the results of all three methods, with values of 0.99 in P^3 and GeneRank, and 0.76 in Direct method. P^3 is always better than the Direct method and outperforms the GeneRank method in most cases. The AUCs of P^3 are greater than those of Direct method and GeneRank by 0.37 and 0.2, in average respectively, and the maximum differences are 0.6 and 0.73, respectively. Only three defects, "Egg Size/Shape", "AB Spindle Orientation" and "P1/AB Cortical Activity" show that GeneRank method is slightly better than P^3, with the maximum difference of AUC as 0.028. Three phenotypes, "Tetrapolar Spindle", "Four Cell Stage Nuclei-Size-Shape", and "P1-AB Nuclear-Size-Shape", that have both high AUC and precision-recall for P^3 were chosen for further comparison. Figure 3 (left) shows their ROC curves, and the corresponding precision-recall curves are shown in Figure 3 (right).

Yeast: Similar to the study in *C. elegans*, we computed true-positive rate versus false-negative rate and precisions at certain recall levels. The comparison of the area under the Receiver Operating Characteristic curve for each phenotype is shown in Figure 4. For simplicity, we show the results for 28 phenotypes among the 72 examined phenotypes. The highest AUC in the selected results of P^3 is 0.98, from "Cell wall-Hygromycin B", that of the direct method is about 0.81, from "Peroxisomal mutants", and GeneRank has the highest AUC value about 0.88, from "Sensitivity to immunosuppressants". P^3 outperforms GeneRank and Direct method in most cases. The AUCs of P^3 are greater than those of Direct method and GeneRank by 0.4 and 0.2 in average respectively, and the maximum differences are 0.6 and 0.8 respectively. Three phenotypes that have both high AUC values and precisions among the result of P^3 method were chosen for further comparison. They are "Conditional phenotypes", "Carbon utilization", and "Cell morphology and organelle mutants". Figure 5 (left) shows their ROC curves, and the corresponding precision-recall curves are shown in Figure 5 (right).

4.3.2. *k-fold cross validation*

C. elegans. 10-fold cross validation was carried out on *C. elegans* data. Figure 6 shows the distributions of AUC values of each method. The median, lower quantile and upper quantile of each group is plotted. As is evident, the performance is considerably improved by using P^3 for phenotype prediction.

Yeast. 5-fold cross validation was carried out on the yeast data. Figure 7 shows the comparison of distributions of AUC. The median, lower quantile and upper quantile of each group is plotted. P^3 outperforms the other two methods in all cases.

4.3.3. *Predicting Phenotypes to genes without any interactions*

To evaluate our approach in predicting phenotypes for genes without any interaction information, we identified those genes that have at least two phenotypes but without interactions in both datasets. We used the phenotype graphs obtained in the leave-one-out experiment, that were derived without any information about the test genes. The target gene/phenotype pairs are separated almost equally into two groups: one for training and another for testing. For example, for each gene, if it has two phenotypes then one is in the training group and another is in the test group. Results show that P^3 can predict most of the phenotypes successfully. Table 4 presents the characteristics of the data and results.

Table 4. Predicting phenotypes for genes without interactions.

Organism	Genes	Predicted with AUC $\geq$ 0.8
Caenorhabditis elegans	42	24
Saccharomyces cerevisiae	48	46

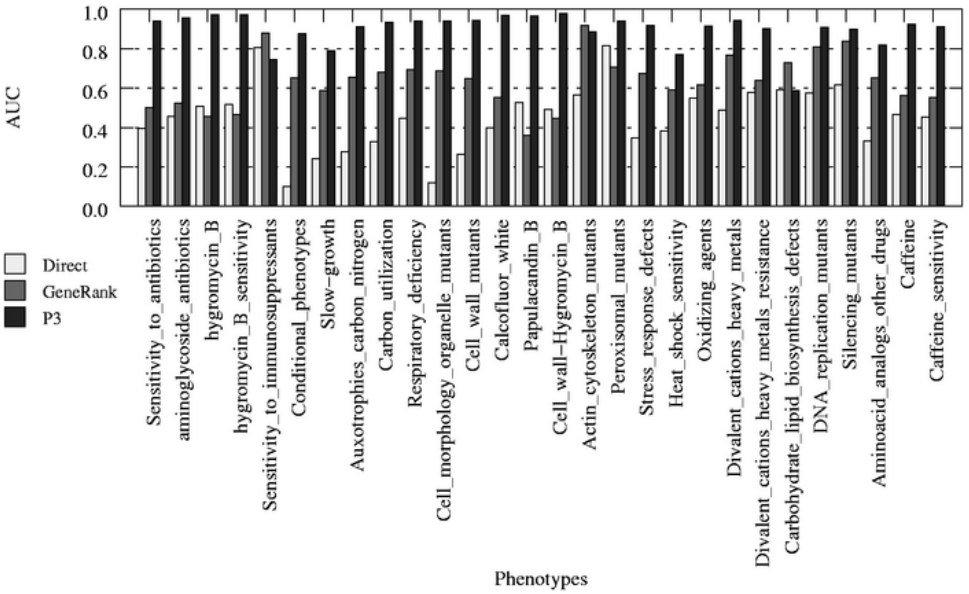

Fig. 4. Overall performance comparison on yeast phenotype dataset. Direct : ranking genes using the interaction network only; GeneRank : $d = 0.1$; P^3 : $\beta = 0.6$

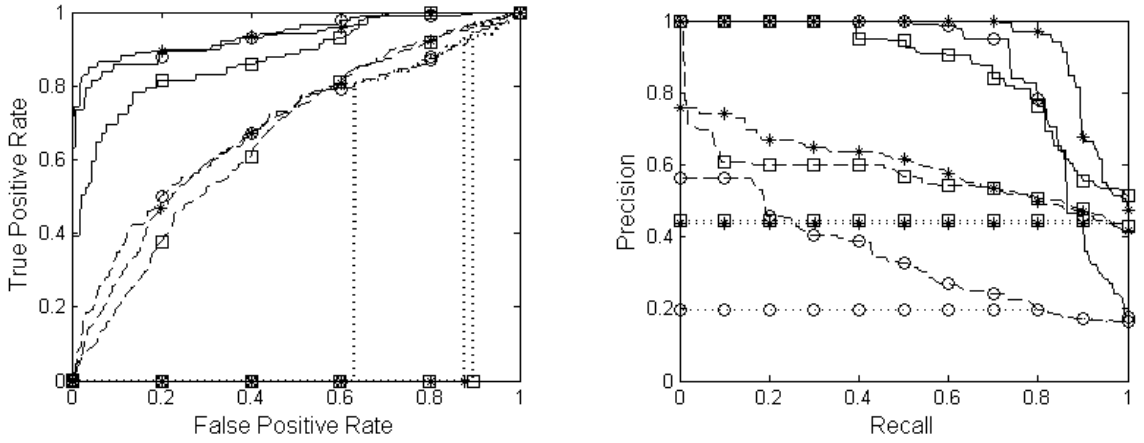

Fig. 5. (left) ROC curves on yeast. (right) Precision vs. Recall on yeast. Direct: points, GeneRank: dashed line, P^3: solid line; circle: Carbon utilization , square: Conditional phenotypes, star: Cell morphology and organelle mutants.

4.4. Phenotype relations

The complete directed graph of phenotypes is too complex to describe in detail here. Therefore, we partition the graph into several highly connected subgraphs by using the CAST [20] algorithm. CAST is a heuristic approach for solving the 'corrupted cliques' problem. It transforms an undirected graph into a set of cliques or almost cliques by repetitively adding nodes having maximum average similarity to the current clique, as long as the similarity is above a threshold λ, and removing nodes with minimum average similarity to the clique, when the similarity is less than the threshold. The process stops when there are no more nodes to add or remove. First, directions are removed from the edges in the original phenotype graph. For each pair of phenotypes, two directed edges are merged into one undirected edge. Every new edge is assigned a new weight that is the average of weights of the original two edges. The graph is further adjusted by deletions of "weak" connections between phenotypes. For example, if the

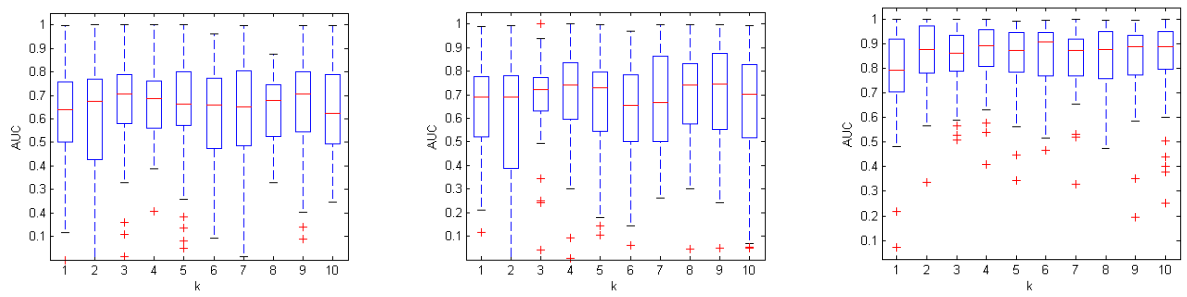

Fig. 6. AUC distributions on *C. elegans*. Direct method (left), GeneRank method (middle), and P^3 (right).

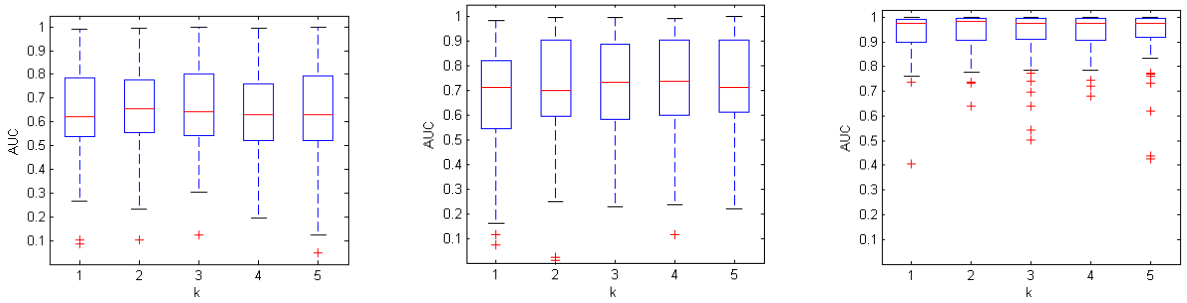

Fig. 7. AUC distributions on yeast. Direct method (left), GeneRank method (middle), and P^3 (right).

weight of the connection between phenotype p to q is less than a threshold t, then the corresponding edge is removed. We run the CAST algorithm on this simplified graph. A set of cliques and almost cliques are obtained. Each clique/almost clique is a cluster of single or a set of highly related phenotypes. Genes causing these phenotypes tend to interact or function together. Figure 8 and Figure 9 show some of the phenotype cliques obtained. The thickness of links represents the closeness between phenotypes. Multiple values are used for parameter t and λ. As t and λ decrease, the number of cliques decreases and the size of the maximum clique increases. We choose the parameter values that give small cliques so that they are relatively easy to interpret biologically. In *C. elegans*, there are 23 cliques/almost cliques, and the largest clique contains 11 nodes, one clique with 5 nodes, 4 nodes, and 3 nodes respectively, three cliques with 2 nodes, and the rest are singletons. In yeast, there are 41 cliques/almost cliques, and the largest clique contains 11 nodes, one clique with 4 nodes, six with 3 nodes, and six with 2 nodes, the remaining are singletons.

The *C. elegans* phenotypes identified in Figure 8 are all related to cell division. The edges suggest that there are distinct relationships between the formation and behavior of the nuclei, indicative of a functional role for structural proteins. The role of structural proteins, acting as conduits for macromolecular and organellar movement can also be seen in the largest clique where cytokinesis (splitting of the cytoplasm to form two cells) and furrow formation (where the cells are divided in half) are related.

The larger yeast clique in Figure 9 pertain to drug sensitivities, including antibiotics. Such associations could potentially be reflective of the role of the extracellular domain in resistance or non-resistance to select antibiotics. Inasmuch, caffeine sensitivity has been related to the synthesis of phospholipids (cell membrane components) and changes in calcium flux. Indeed, the smaller clique relates all of these concepts through sensitivity to immunosuppresants, a sensitivity that is related to phosphorylation-based signal transduction cascades.

5. DISCUSSION

In this paper, we have presented an approach to modeling phenotype relations and using these rela-

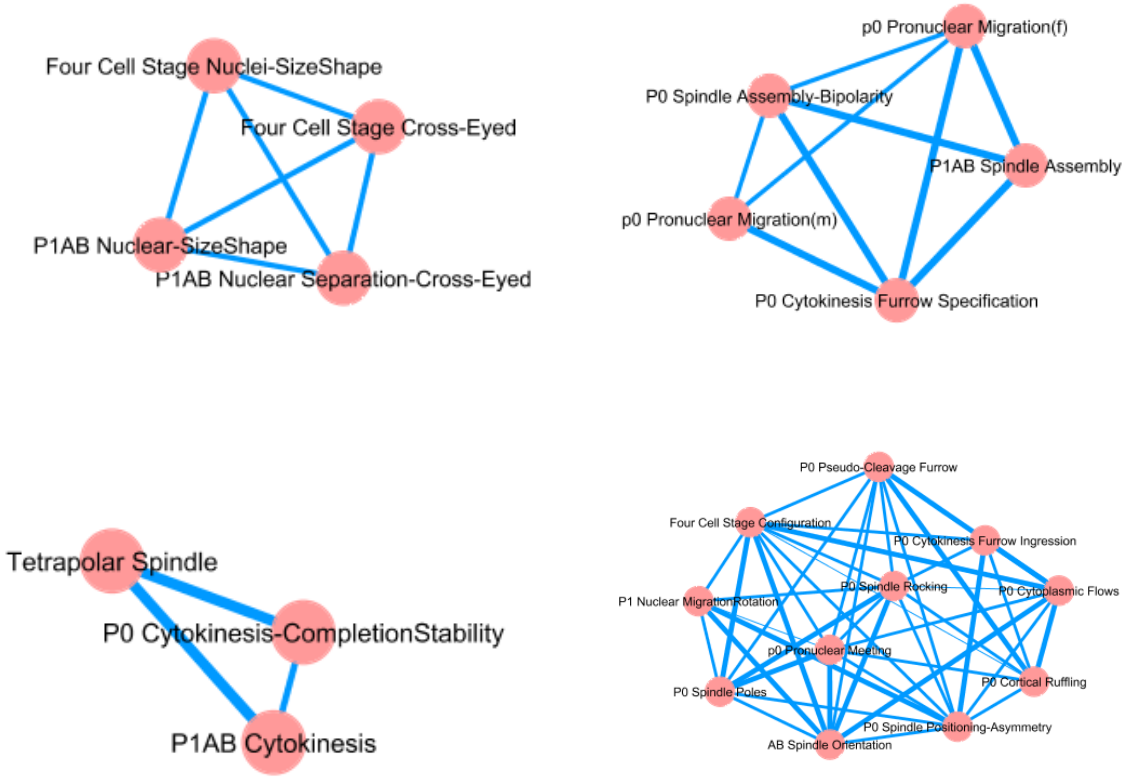

Fig. 8. Phenotype cliques in the *C. elegans* dataset derived from P^3.

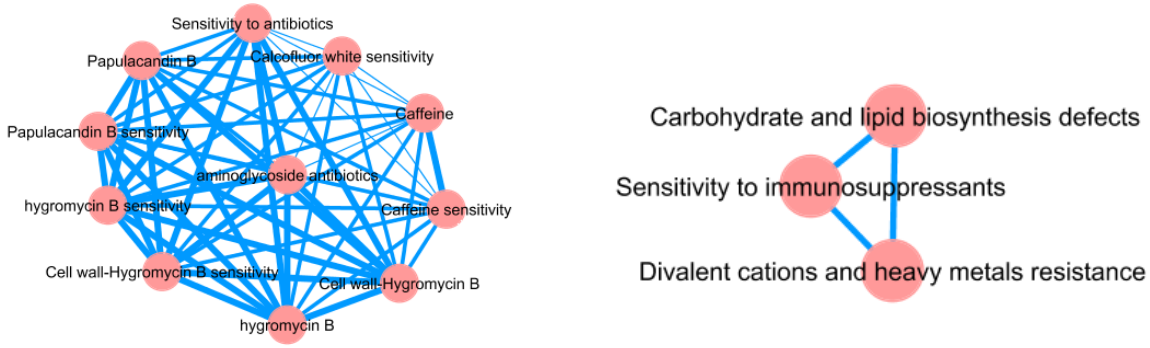

Fig. 9. Phenotype cliques in the *S. cerevisiae* dataset derived from P^3.

tionships to predict phenotypes for uncharacterized genes. The strong results indicate that the combination of gene networks and phenotype profiles provides a powerful synergy that is not obtainable with either method alone. One limitation is that to be able to make predictions, a gene should have at least one known phenotype. In future work, we seek to capture more complex many-many effects between

genes and phenotypes and design new experiments to validate the predictions made.

Acknowledgments

This work is supported in part by US NSF grant ITR - 0428344.

References

1. Scherens, B., Goffeau, A.: The uses of genome-wide yeast mutant collections. Genome Biol 5(7) (2004)
2. Ohya, Y., *et al.*: High-dimensional and large-scale phenotyping of yeast mutants. PNAS 102(52) (December 2005) 19015–19020
3. Piano, F., *et al.*: Gene clustering based on RNAi phenotypes of ovary-enriched genes in C. elegans. Curr Biol 12(22) (November 2002) 1959–1964
4. Sonnichsen, B., *et al.*: Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. Nature 434(7032) (2005) 462–469
5. Oti, M., Snel, B., Huynen, M.A., Brunner, H.G.: Predicting disease genes using protein-protein interactions. J Med Genet 43(8) (2006) 691–698
6. Lage, K., *et al.*: A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature Biotechnology 25(3) (2007) 309–316
7. Lee, I., *et al.*: A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nature Genetics 40(2) (2008) 181–188
8. Ma, X., Lee, H., Sun, F.: CGI: a new approach for prioritizing genes by combining gene expression and proteinprotein interaction data. Bioinformatics 23(2) (2007) 215–221
9. Morrison, J., Breitling, R., Higham, D., Gilbert, D.: GeneRank: Using search engine technology for the analysis of microarray experiments. BMC Bioinformatics 6(1) (2005)
10. Markowetz, F., Kostka, D., Troyanskaya, O.G., Spang, R.: Nested effects models for high-dimensional phenotyping screens. Bioinformatics 23(13) (2007) i305–312
11. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press (1999) 230–237
12. White, S., Smyth, P.: Algorithms for estimating relative importance in networks. In Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press (2003) 266–275
13. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project (1998)
14. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. JACM 46(5) (1999) 604–632
15. Mewes, H. W., *et al.*: MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res 34(Database issue) (2006)
16. Pati, A., Jin, Y., Klage, K., Helm, R.F., Heath, L.S., Ramakrishnan, N.: CMGSDB: integrating heterogeneous Caenorhabditis elegans data sources using compositional data mining. Nucleic Acids Res 36(Database issue) (2008)
17. Morrison, J., Breitling, R., Higham, D., Gilbert, D.: CYGD: Comprehensive Yeast Genome Database. BMC Bioinformatics 6(1) (2005)
18. Saccharomyces Genome Database, http://www.yeastgenome.org/
19. BioGrid, http://www.thebiogrid.org
20. Ben-Dor, A., Shamir, R., Yakhini, Z.: Clustering Gene Expression Patterns. J Comput Biol. 6(3/4) (1999) 281–297