BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 2 2007, pages 215–221
doi:10.1093/bioinformatics/btl569

Systems biology

CGI: a new approach for prioritizing genes by combining gene expression and protein–protein interaction data

Xiaotu Ma[1,†], Hyunju Lee[1,2,¶,†], Li Wang[1] and Fengzhu Sun[1,*]

[1]Molecular and Computational Biology Program, Department of Biological Sciences and
[2]Department of Computer Science, University of Southern California, Los Angeles, CA 90089-2910, USA

Received on July 7, 2006; revised on October 23, 2006; accepted on November 3, 2006
Advance Access publication November 10, 2006
Associate Editor: Trey Ideker

ABSTRACT

Motivation: Identifying candidate genes associated with a given phenotype or trait is an important problem in biological and biomedical studies. Prioritizing genes based on the accumulated information from several data sources is of fundamental importance. Several integrative methods have been developed when a set of candidate genes for the phenotype is available. However, how to prioritize genes for phenotypes when no candidates are available is still a challenging problem.

Results: We develop a new method for prioritizing genes associated with a phenotype by Combining Gene expression and protein Interaction data (CGI). The method is applied to yeast gene expression data sets in combination with protein interaction data sets of varying reliability. We found that our method outperforms the intuitive prioritizing method of using either gene expression data or protein interaction data only and a recent gene ranking algorithm GeneRank. We then apply our method to prioritize genes for Alzheimer's disease.

Availability: The code in this paper is available upon request.

Contact: fsun@usc.edu

Supplementary data: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

With the rapid development of high-throughput biotechnologies, biologists have amassed a large amount of data at various levels such as gene expression profiles, (Cho *et al*., 1998; Spellman *et al*., 1998; Hughes *et al*., 2000; Gasch *et al*., 2000), protein–protein interactions (Ito *et al*., 2000, 2001; Uetz *et al*., 2000; Gavin *et al*., 2002; Ho *et al*., 2002), single nucleotide polymorphisms, transcription regulation networks, etc. These resources give us insight into the underlying mechanisms of basic biological processes, which can lead to improvements in public health. A typical problem in biological and biomedical studies is to identify genes responsible for a phenotype (e.g. disease status, quantitative trait values, gene expression values, etc.). How to integrate evidences from different resources to assist biologists on this task remains a challenge for bioinformaticians.

Prioritizing genes by combining genetic study results and other molecular level data has attracted much attention recently, since the resolution of genetic studies are low and the number of candidate genes could be potentially large (Maraganore *et al*., 2005). Franke *et al*. (2006) proposed to rank genes in candidate regions by their connectivity with the genes in other linked regions. The intuition behind the study is that the disease associated genes could be a set of genes interacting with each other, for example from a particular pathway. However, this method depends strongly on the availability of results from genetics studies. Under similar rationale, Aerts *et al*. (2006) developed a Bayesian model to identify new genes (in addition to the already known genes) involved in a given disease, using many currently available data types. However, they assumed that a set of genes responsible for the disease is already known, which limits the applicability of their method.

Gene expression data and protein interaction data have been integrated for gene function prediction. For example, Ideker *et al*. (2002) used protein interaction data and gene expression data to screen for differentially expressed subnetworks between different conditions. In Tornow and Mewes (2003) and Segal *et al*. (2003), gene expression data and protein interactions are used to group genes into functional modules. These methods provide insights into the regulatory modules of the whole networks at the systems biology level. However, it is not clear how to adapt their methods to identify genes contributing to the phenotype of interest. Morrison *et al*. (2005) adapted the Google search engine to prioritize genes for a phenotype by integrating gene expression profiles and protein interaction data. However, the algorithm ignores the information from proteins linked to the target protein through other intermediate proteins, referred to in the rest of this paper as indirect neighbors.

Here we propose an approach motivated from Markov Random Field theory to prioritize genes using high-throughput data, including gene expression profiling and protein interaction mapping. Our approach focuses on relating (ranking) genes to phenotypes without requiring any known candidate genes. We study the effect of different data integration methods and different definitions of the neighborhood of the protein interaction network. We show that the performance of our approach outperforms that of using gene expression data only or using protein interaction data only, as well as that of the existing method GeneRank, which is a modified version of Google search engine PageRank. We also study the performance of our method with respect to noise in protein interaction network. Finally, we apply our approach on data from human Alzheimer's disease.

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[¶]Present address: Harvard-Partners Center for Genetics and Genomics, 77 Avenue Louis Pasteur Boston, MA 02115, USA

2 MATERIALS AND METHODS

2.1 Materials

Three gene expression data sets are used in this study: the Yeast Compendium Knockout (KO) data (Hughes *et al*., 2000), Stress Response (SR) data (Gasch *et al*., 2000) and Cell Cycle (CC) data (Cho *et al*., 1998; Spellman *et al*., 1998).

Several protein interaction databases for yeast are available, including data generated using the yeast two-hybrid method (Ito *et al*., 2000, 2001; Uetz *et al*., 2000) and the mass spectrometric analysis of protein complexes (Gavin *et al*., 2002; Ho *et al*., 2002). The fractions of true interactions among the different observed interaction data sets, referred to as reliability, have been extensively studied (Mrowka *et al*., 2001; Deane *et al*., 2002; Deng *et al*., 2003). We choose to use the highly reliable MIPS (Munich Information Center for Protein Sequences) physical interaction data set (Mewes *et al*., 2002), which includes interactions collected from small-scale experiments and the core data of Ito *et al*. (2000, 2001). The other protein interaction data sets such as DIP (Database of Interacting Proteins) core (Xenarios *et al*., 2002), Uetz (Uetz *et al*., 2000) and Ito (Ito *et al*., 2000, 2001) are also used to evaluate the robustness of our approach with respect to noise in protein–protein interaction data.

In order to assess the usefulness of various prioritizing methods, we use the functional annotation from the Gene Ontology (GO, The Gene Ontology Consortium 2001). GO is a rooted directed acyclic graph (DAG) and the nodes close to the root are more abstract than the nodes far away from the root. To avoid the problem of too broad or too specific functional categories in GO, we use the concept of informative GO nodes defined as those containing at least 40 genes and at most 200 genes, without any of its offspring having the same number of genes as itself, similar to the definition presented in Zhou *et al*. (2002).

2.2 Gene Expression Profiles and Association Measures

Suppose that there are m subjects (or individuals/conditions) in a study. Let φ_i be the phenotype value of the i-th subject. For each subject, the expression values of n genes are measured. Let l_{ij} be the log-expression value of the j-th gene for the i-th subject. The phenotype values and the gene expression data are organized as in Table 1. For simplicity of presentation, we assume that the phenotype takes real continuous values (see Dudoit *et al*., 2002 for qualitative phenotype values). The Pearson correlation coefficient between the phenotype values and the expression values of each gene can be a measure of the association between the phenotype and genes. Since too many missing values will make the correlation estimate unstable, we eliminate genes with >5 missing values and then standardize the remaining data on the rows in Table 1, by subtracting the mean expression measurements and dividing by the standard deviation. In general, the phenotype values or gene expression values may not have a normal distribution. The following modified correlation coefficient (MCC) can be used to avoid this problem.

Without loss of generality, we assume that both the phenotype values φ and the expression values of gene g are available for the first $m-k$ subjects. Let $\mathrm{RP}=(rp_1, rp_2, \ldots, rp_{m-k})$ and $\mathrm{RG}=(rg_1, rg_2, \ldots, rg_{m-k})$ be the rank of the values of phenotype φ and the gene expression values of gene g across the first $m-k$ subjects. In case of ties, we randomly assign ranks to these tie-subjects. An inverse normal transform (Li, 2002) is applied to the rank vectors RP and RG:

$$x_i = \Phi^{-1}((rp_i + 0.5)/(m-k+1)),$$

$$y_i = \Phi^{-1}((rg_i + 0.5)/(m-k+1)),$$

where the fraction 0.5 in the numerator and 1 in the denominator are introduced to prevent the corresponding item from being 0 or 1. Φ is the cumulative distribution function of standard normal. After the transformation, we use the Pearson correlation coefficient between $(x_1, x_2, \cdots, x_{m-k})$ and $(y_1, y_2, \cdots, y_{m-k})$ to measure the association between phenotype φ and gene g:

$$\mathrm{MCC}(\varphi, g) = \frac{1}{m-k}\sum_{i=1}^{m-k} x_i y_i. \tag{1}$$

Table 1. The data structure for the phenotype value and the gene expression values for the studied subjects

Subject	Phenotype value	Gene expression			
		Gene 1	Gene 2	$\cdots$	Gene n
1	φ_1	l_{11}	l_{12}	$\cdots$	l_{1n}
2	φ_2	l_{21}	l_{22}	$\cdots$	l_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
m	φ_m	l_{m1}	l_{m2}	$\cdots$	l_{mn}

To make our efforts immediately applicable for robust estimation of association scores between the phenotype and gene expression profiles using protein interaction data, we apply the Fisher's transformation (David, 1949) on the modified correlation coefficient to obtain:

$$O_g = \frac{\sqrt{m-k-3}}{2}\ln\frac{1+\mathrm{MCC}(\varphi, g)}{1-\mathrm{MCC}(\varphi, g)}, \tag{2}$$

where O_g has approximate standard normal distribution N(0,1). We sort the genes according to O_g. Note that whether or not we apply Fisher's transformation does not affect the prioritizing result when using expression data only. However, it does affect the prioritizing result when integrating gene expression and protein interaction data. We emphasize that O_g is used here to indicate that the value in Equation (2) is the observed value, not necessarily the true underlying association score R_g.

2.3 Prioritizing genes by combining gene expression profiles and protein interaction data

The association scores between the phenotype value φ and the expression profiles of interacting genes are correlated. Therefore, we can use O_g and $O_{g'}$, where g' are the interaction partners of gene g, to obtain a more accurate estimation of the association between the phenotype and gene g.

The rationale behind our approach is that (1) the gene expression profiles measured by microarray are noisy and thus the derived association score is also noisy; (2) a protein is likely to be co-expressed with its interaction partners (Jansen *et al*., 2002); (3) estimation of the association between a protein and the phenotype can be calibrated by considering the association between the protein's interaction partners and the phenotype. Since indirect interaction partners may contribute to the accurate estimation in (3), we also take them into consideration. The basic idea of our approach, CGI, is shown in Figure 1.

Two issues need to be clarified. One is the definition of neighborhood system in the protein interaction network and the other is the method for data integration. We consider the following neighborhood systems:

(1) *Direct neighbors*. For a given protein g, $\mathcal{N}_g = \{h \mid h$ *interacts with* g and $h \neq g\}$ is the direct neighborhood of g. A similarity measure S_{gh} is defined as 1 if $h \in \mathcal{N}_g$ and 0, otherwise. S is the adjacency matrix in graph theory. The direct neighborhood system cannot capture information from indirect neighbors.

(2) *Shortest path*. Let d_{gh} be the graph-theory shortest distance between proteins g and h in the protein interaction network. The similarity between them is defined as $S_{gh} = 1/(d_{gh}+1)$ (Krauthammer *et al*., 2004). Shortest distance neighborhood system captures information from indirect neighbors, and gives indirect neighbors lower weight than direct neighbors. However, this similarity matrix may not utilize the information contained in the neighbor proteins efficiently since in general d_{gh} is small.

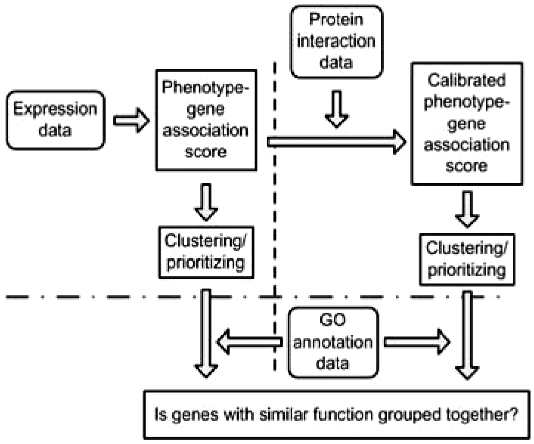
Protein interaction data
Expression data
Phenotype-gene association score
Calibrated phenotype-gene association score
Clustering/ prioritizing
Clustering/ prioritizing
GO annotation data
Is genes with similar function grouped together?

Fig. 1. Schematic flowchart of our approach. Shown left to the vertical dashed line is the intuitive gene-prioritizing method using the gene expression data only. Our approach, CGI, integrates the gene expression and protein interaction data to obtain a calibrated association score between the phenotype and the gene. The GO functional annotation is used to evaluate the prioritizing result of the intuitive method and our approach.

(3) *Diffusion kernel*. The diffusion kernel is defined as $K = \exp(\tau H)$ (Kondor and Lafferty 2002), where H is defined as:

$$H_{gh} = \begin{cases} 1 & \text{if protein } h \in \mathcal{N}_g, \\ -\sum_{g':g' \in \mathcal{N}_g} 1 & \text{if } h = g, \\ 0 & \text{otherwise.} \end{cases}$$

The similarity score between two proteins g and h are defined as $S_{gh} = K_{gh}/\sqrt{K_{gg} \times K_{hh}}$, for $h \in \mathcal{N}_g$ (so that S has unit diagonal elements).

To integrate protein interaction data and gene expression data, we consider the following probabilistic model. Let R_i be the underlying true association score between the phenotype and the i-th gene of interest and $R = \{R_i, i = 1,2,\ldots,n\}$, where n is the number of genes. We treat R as a random vector and model R by Markov Random Field (MRF) theory (for more detail on MRF and Gibbs distribution see Geman and Geman, 1984). We model the probability density function of R to be proportional to $\exp(-U_0(R)/T)$, where T is called temperature in statistical physics and $U_0(R)$ is referred as the potential function. The potential function $U_0(R)$ defines a global Gibbs distribution of the entire configuration of random vector R:

$$\Pr(R) = \frac{1}{Z} \exp\left(-U_0(R)/T\right),$$

where $U_0(R) = \sum_{i \neq j} S_{ij}(R_i - R_j)^2$ and $Z = \sum_R \exp\left(-U_0(R)/T\right)$ is referred as the *partition* function. We also assume that the observed association O_i is a summation of R_i and random noise ε_i, i.e. $O_i = R_i + \varepsilon_i, i = 1, 2, \ldots n$, where ε_i are independent and identically distributed random variables with normal distribution N(0, σ^2). The likelihood of the observed association scores $(O_1, O_2, \cdots, O_n)$ can be written as

$$\Pr(O) \sim \exp\left(-U(R)/2\sigma^2\right),$$

where $U(R)$ can be written as

$$U(R) = \sum_i (O_i - R_i)^2 + \lambda \sum_{i \neq j} S_{ij}(R_i - R_j)^2, \tag{3}$$

where $\lambda = 2\sigma^2/T$. The first term represents the contribution of the observed association score. The second term represents the constraint by pairwise interactions. When no noise is present, $\lambda = 0$, only the first term is effective and the maximum a posteriori (MAP) estimation for R is exactly the same as the observed data, $R = \text{O}$. As λ becomes larger, the solution is more influenced by the second term. Our goal here is to obtain the least square estimation of the underlying association scores $\hat{R}$ by Equation (3). We refer to the ranking result using $\hat{R}$ as CGI-1.

In addition to CGI-1, we also consider two other relatively simple methods for integrating gene expression profiles and protein interactions. The first method is based on the assumption that the association scores between the phenotype values and the gene expression profiles of interacting genes are positively correlated. Therefore, we update the association score between the phenotype and the i-th gene by the weighted average of the observed association scores for the neighboring genes,

$$\hat{R}_i = \frac{O_i + \lambda \sum_{k:k \neq i} S_{ik} O_k}{1 + \lambda \sum_{k:k \neq i} S_{ik}}, \; i = 1, 2, \cdots, n. \tag{4}$$

We can then sort the genes according to the values of $\hat{R}_i$. We refer to this approach as CGI-2.

Although several studies have shown that interacting proteins are more likely to be positively correlated, some interacting gene pairs can also be negatively correlated (Deng *et al*., 2003). By using CGI-2, the neighbors with positive association score and negative association score may counteract the significance of their contribution. Thus, we introduce a modified version by

$$\hat{R}_i = \frac{O_i + \lambda \sum_{k:k \neq i} S_{ik} \,|\, O_k \,|}{1 + \lambda \sum_{k:k \neq i} S_{ik}}, \quad i = 1, 2, \cdots, n. \tag{5}$$

This approach is referred as CGI-3. We will show in section 3.1.3 the reason of the usefulness of Equation (5).

2.4 Evaluation of gene prioritizing methods

One difficulty in evaluating the gene prioritization approaches is that there are not many microarray data sets with a clearly defined phenotype in yeast. On the other hand, highly reliable and abundant protein interaction data are available for yeast and thus it is an ideal model organism to evaluate the various approaches. Considering that the phenotypes of interest (e.g. disease status in human) are generally the functional consequence of the expression of many genes, we choose to simply take the expression levels of one (called simple phenotype) or the average of the expression levels of two (called complex phenotype) genes as phenotypes (we do not consider noise for simplicity). Note that gene expression levels are frequently used as quantitative traits to locate quantitative trait loci (QTL) (Brem, 2002, 2005; Morley *et al*., 2004). This strategy also allows us to empirically estimate the performance of different prioritizing methods using the external criteria of GO (The Gene Ontology Consortium 2001). For convenience of notation, the gene whose expression level is treated as phenotype is referred as target gene. The other genes can be regarded as genotypes as usual. The goal of our prioritization is to rank genes with the same function (GO annotation) as the target gene(s) on the top. We first describe how we evaluate prioritization results using simple phenotypes. The evaluation of prioritization results using complex phenotypes are illustrated in the Results section.

For each target gene φ and its functional annotation F (informative nodes as defined in the Materials subsection), we test the hypothesis that the genes with functional annotation F are ranked higher than the other genes using one-sided Wilcoxon rank sum test. Denote the resulting P-value as $\text{p}_{\varphi,F}$. We repeat this procedure for many gene–function pairs (φ, F) and obtain P-values for these pairs. Note that a gene can be counted multiple times if it belongs to multiple functional categories (Deng *et al*., 2003). Denote the empirical cumulative distribution function of the P-values of these gene–function pairs as $G(p)$, $p \in [0,1]$. Since it is anticipated that genes with the same function as the target gene should have higher rank than other genes, many gene-function pairs (φ, F) will have small P-values and $G(p)$ increases very fast when p is close to 0. The faster $G(p)$ increases, the better the corresponding prioritizing method is. We define the performance measure as follows:

$$\text{performance} = \int_0^1 G(p)d(p), \tag{6}$$

i.e. the area under the cumulative distribution function $G(p)$.

Note that if a set of genes are unrelated to the target genes, the resulting P-values would follow a uniform distribution in [0,1] and the performance measure defined above will be 0.5. Higher performance value means better performance of the gene prioritizing method.

3 RESULTS

In this section, we first study the effect of the three definitions of neighborhood systems (direct neighbors, shortest path and diffusion kernel) on the performance of the different integration approaches. It is shown that CGI-3 combined with the diffusion kernel neighborhood system consistently outperforms the other combinations. CGI-3 combined with the diffusion kernel also outperforms the GeneRank algorithm. The robustness of the performance of CGI-3 with respect to noise in the protein interaction network is also studied. We then study the performance of CGI-3 for complex phenotypes. Finally, we apply CGI-3 to the Alzheimer's disease data set. Due to the incompleteness of the protein interaction data set, we focus on the 4136 genes having at least one interacting partners in the physical interaction data set (Mewes *et al*., 2002).

3.1 Effect of neighborhood systems

3.1.1 Direct neighbor and shortest path kernels Note that in this case only one parameter λ is involved. We first explore empirically the effect of the parameter λ under the direct neighborhood system, using prioritizing methods CGI-1, CGI-2 and CGI-3, respectvely. We tried several values for λ. The optimal estimation $\hat{\lambda}$ is chosen as the one maximizing Equation (6) for each data set (listed in 'Direct Neighbor' column of Table 2). Our study shows that CGI-3 (Equation 5) outperforms the other prioritizing methods under direct neighborhood kernel. Therefore, we first present the effect of different values of λ with CGI-3. The results of CGI-1 and CGI-2 are shown later (see Fig. 4).

Figure 2 shows the relationship between the performance index and λ for different gene expression data sets, using CGI-3. When $\lambda = 0$, only the gene expression information is used and the performance index is low. The performance index first increases as λ reaches an optimal value and decreases thereafter. The optimal values for λ are consistent across the different data sets. For example, when $\lambda = 0$, i.e. using only microarray data, the performance index is 0.74, 0.74 and 0.80, for the CC, KO and SR data, respectively. The corresponding performance index increases to 0.83, 0.82 and 0.85, when λ is 0.5.

Similar results are observed for shortest path kernel. But in this case CGI-3 is not always better than the other prioritizing methods. Since the prioritizing method CGI-3 with direct neighbor kernel outperforms that using shortest path kernel (see Fig. 4), we continue on studying the diffusion kernel.

3.1.2 Diffusion kernel neighborhood system Here two parameters λ and τ are involved. Again the prioritizing method CGI-3 (Equation 5) outperforms the other two methods. By sampling λ and τ from 2D space $[0, 6] \times [0, 2\ ]$. We found that the optimal value for λ is always close to 1 and thus we fix $\lambda = 1$. The relationship between the performance index and τ is given in Figure 3 for different gene expression data sets. The performance index stays high for a large range of values of τ. Comparison with the result of direct neighborhood system also shows that the neighborhood system defined by the diffusion kernel outperforms the directed neighbors. This phenomenon has been observed to hold for protein function prediction using protein interaction networks (Lee *et al*., 2006). The empirical estimation of the optimal τ is listed in Table 2.

Table 2. Empirical optimal parameters for λ (CGI-3 with direction neighbors) and τ (CGI-3 with diffusion kernel neighborhood system). CC: Cell cycle; SR: Stress response; KO: Knockout

	Direct neighbor ($\hat{\lambda}$)	Diffusion kernel ($\hat{\lambda}$, $\hat{\tau}$)
CC	0.50	(1, 0.70)
SR	0.50	(1, 0.65)
KO	0.55	(1, 0.50)

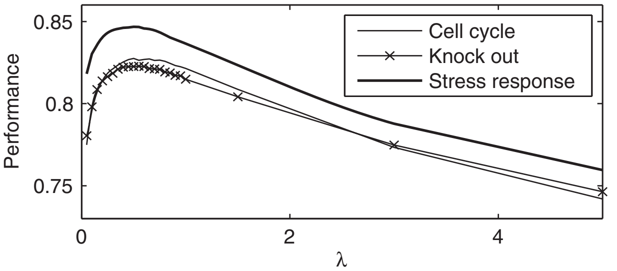

Fig. 2. The performance of CGI-3 with direct neighborhood kernel for different λ_s on three data sets.

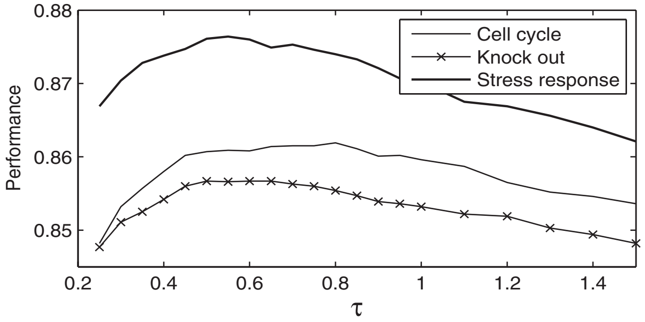

Fig. 3. The performance of CGI-3 with diffusion kernel for different $\tau(\lambda = 1)$.

3.1.3 Comparison of the performance of CGI-1, CGI-2, CGI-3 and GeneRank We compare the performance of different combinations of neighborhood systems: direct neighbors, shortest path and diffusion kernel with different prioritizing methods: CGI-1, CGI-2, and CGI-3. The values of $\hat{\lambda}$ and $\hat{\tau}$ given in Table 2 are used. In Figure 4, the naive prioritizing method of using microarray data only (Equation 1) is also presented as open bars (denoted by MCC). It is clear that MCC has the lowest performance index. In addition, we rank proteins according to the number of their interacting partners (denoted as PPI in Fig. 4). This procedure always gives highly connected proteins higher rank and does not depend on the phenotype. Despite this drawback, it can be seen that the result

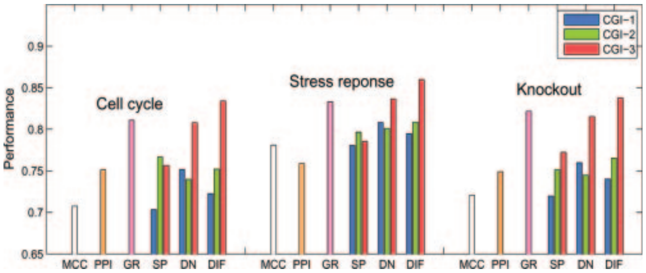

Fig. 4. The performance of CGI-1, CGI-2 and CGI-3 by different neighborhood systems. MCC (open bars): prioritizing genes with expression data only; PPI (purple bars): prioritizing genes with protein-interaction data only; SP: shortest path neighborhood system; DN: direct neighborhood system; DIF: diffusion kernel neighborhood system. GR: GeneRank algorithm. The optimal empirical $\hat{\lambda}$ and $\hat{\tau}$ values are used.

of prioritizing genes by protein interaction data outperforms that by gene expression data only.

With the direct neighborhood system, the performance of CGI-3 is much better than that of CGI-1 and CGI-2, whereas the performance of CGI-2 is comparable with the performance of CGI-1. In case of diffusion kernel, the performance of CGI-3 is again much better than that of CGI-1 and CGI-2 and the performance of CGI-1 is also comparable with CGI-2. With the shortest path neighbor system, CGI-3 is comparable with CGI-2 and better than CGI-1. Moreover, CGI-1, CGI-2 and CGI-3 with the shortest path kernel all perform worse than that of CGI-3 with direct neighborhood system and the diffusion kernel system. As a conclusion, the integration method CGI-3 with the diffusion kernel performs much better in all three microarray data sets.

We implemented the GeneRank algorithm (Morrison *et al*., 2005) and evaluated its performance using our criteria. Specifically, we used the GeneRank algorithm to recursively update the association score between the expression profile of each gene and the phenotype of interest, where the protein interaction data was used as the network for this algorithm. We tried different values for the free parameter d in GeneRank from 0.5 to 1 with step 0.1. The performance increases as d increases and reaches the optimum at $d = 0.9$, the same as recommended. The performance of GeneRank for different data sets is shown in Figure 4 as purple bars. It can be seen that CGI-3 with direct neighbor performs similarly with GeneRank and CGI-3 with diffusion kernel outperforms GeneRank.

The improvement by CGI-3 with the direct neighbors and the diffusion kernel may be attributed to the following reasons. First, it borrows strength from both positive and negative associations between interacting proteins. On the other hand, CGI-1 and CGI-2 only borrow strength from positive association between interacting proteins. Second, by using absolute values for the association scores for the neighboring proteins, CGI-3 gives high weights to proteins having large number of interaction neighbors highly associated with the phenotype. To illustrate this point, we use the expression values of gene YIL138C as a phenotype (TPM2/YIL138C is a gene directing polarized cell growth in yeast by binding to and stabilizing actin cables and filaments (Evangelista *et al*., 1997), thus our phenotype is cell polarity-related). Among the genotypes, YLR319C (a protein involved in the organization of the actin cytoskeleton (Pruyne and Bretscher, 2000)) has the same function as YIL138C. With the cc gene expression data, the MCC for the expression profiles of YLR319C and YIL138C is only 0.62 with rank 1214 (out of 3202 genes). YLR319C has eight direct neighbors with MCC scores (1.23, −2.26, −2.65, −0.96, −3.06, −0.35, 2.36 and −1.54, note that the null distribution of the scores are standard normal), respectively. Four out of the eight neighbors have absolute MCC values at least 2.26. Both strong positive and negative association scores are present among the neighbors. CGI-3 assigns it a rank of 312, an increase of 902 in ranking.

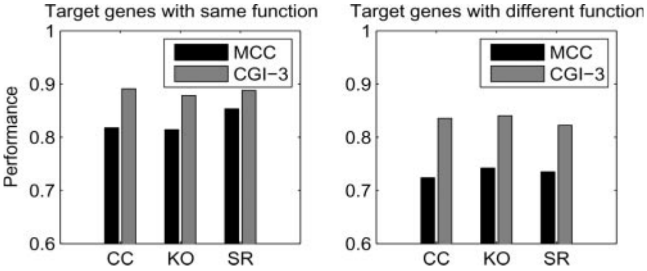

Fig. 5. Performance of CGI-3 with diffusion kernel for complex phenotypes. SR: stress response, CC: cell cycle, KO: knock out. Left panel: the target genes have the same function; Right panel: the target genes have different functions.

3.1.4 Robustness of CGI-3 with diffusion kernel with respect to noise in the protein interaction data We study the robustness of CGI-3 with the diffusion kernel when the reliability of protein interaction data sets is lower than that of MIPS. We choose to test our method using other protein interaction data sets including DIP-core, Ito's, and Uetz's protein interaction data and we found that the performance of CGI-3 with diffusion kernel decreases as the reliability of the protein interaction data decreases. The detailed results are presented in Supplementary materials. In addition, we also randomly add noise into the most reliable MIPS protein interaction data. It is found that the performance of CGI-3 with diffusion kernel decreases as the noise level increases, which is also provided in Supplementary results.

3.2 Prioritization of complex phenotypes

Since genes interweave as a network, the observed phenotype in general is the functional consequence of the expression values of a set of genes, rather than any given single gene. A phenotype can in theory be modeled as a combination of the expression levels of several genes. For simplicity, we assume that the phenotype is a linear combination of the expression values of two genes. In particular, for genes f and g, where $f \in F$ and $g \in G$, the phenotype is simulated as $\varphi = f + g$, where f and g are also used to represent their expression levels (Again, no noise is introduced to avoid complexity). Our objective is to rank genes with function F or G on the top after prioritizing genes with respect to φ. The P-values are summarized and the performance (Equation 6) is displayed in Figure 5. We first study the scenario where $F = G$ [to avoid double counting, our statistics are on different combinations of triplets (f, g, F)]. This case corresponds to the basic assumptions that disease associated genes tend to be in the same pathway in recent studies (Franke *et al*., 2006; Aerts *et al*., 2006). As can be seen from Figure 5 (left panel), when the two target genes are from the same functional category, the prioritizing result by CGI-3 is much better than that by expression data alone. One interesting observation is that in this case both prioritizing methods (by expression data only and by CGI-3) have potential improvement (Fig. 5 left panel) relative to simple phenotypes. This suggests that there is more chance to discover the genes related to a given phenotype if the phenotype is determined by several genes with similar functions rather than by only one of them individually. However, it should be noted that in reality the phenotypes are much more complicated than our ‘complex phenotypes’, both in terms of the number of causal genes and in terms of the dependency between causal genes.

We then study the scenario that $F \neq G$ are distinct GO nodes far away from each other ($\geq$9 in the GO annotation hierarchy). In Figure 5 right panel, it can also be seen that prioritizing result by CGI-3 is much better than using expression data only. However, in this case the overall performance is slightly worse than that for phenotypes caused by genes of similar function (Fig. 5, left panel), suggesting that the prioritizing task is harder when the phenotype is caused by functionally unrelated genes.

3.3 Application to Alzheimer’s disease data

We applied our approach to the data on Alzheimer’s disease by Blalock *et al*. (2004). In this data set, the gene expression levels are measured on 31 subjects. Also, a reliable clinical phenotype, Mini-Mental Status Examination (MMSE), and a neurofibrillary tangle (NFT) score across these 31 subjects are provided. The human protein interaction pairs are available from database HPRD (Peri *et al*., 2003). A total of 4703 genes appear in both Blalock *et al*. (2004) and HPRD. Four genes (*APP*, *APOE*, *PSEN*1 and *PSEN*2) are known to be associated with Alzheimer’s disease (Krauthammer *et al*., 2004). sixty additional expert selected genes are provided in Krauthammer *et al*. (A total of 34 such genes are included in the 4,703 genes). We first ranked the genes according to the association score between the gene expression profiles and the phenotype MMSE (our approach does not result in significant improvement for phenotype NFT, data not shown). We also ranked the genes according to the updated association score by Equation (5). It turns out that the four known genes are ranked significantly higher when the updated association score by CGI-3 with diffusion kernel is used (Table 3, one-sided Wilcox rank sum test *P*-value changed from 0.52 to 0.06). If we pool the expert selected candidate genes together, the *P*-value of the one-sided Wilcox rank sum test for the 34 genes changed from 0.097 for using expression data to 0.0051 by CGI-3. However, we also found that most of the 34 candidate genes are highly connected in the protein interaction network (Wilcox rank sum test *P*-value $<10^{-4}$). Thus, it suggests that the disease related genes are the ‘important genes’ which are highly connected with other genes. On the other hand, it is also possible that the known candidate genes are biased to the highly connected genes since they tend to be more well-studied. To overcome the potential biases, large-scale unbiased protein interactions from Y2H (Stelzl *et al*., 2005; Rual *et al*., 2005; Lim *et al*., 2006) may be used. Unfortunately, this data set does not include the four known Alzheimer’s disease genes. Therefore our method can not be evaluated with it.

Table 3. Rank of the known genes of Alzheimer’s disease (smaller means better)

Gene name	Expression data only	CGI-3
APOE	3203	1568
PSEN1	2558	1345
PSEN2	2732	1313
APP	1379	1010

In the top 50 genes ranked by Equation (5), we found 17 mitochondrial genes and additional 6 ATP-related genes. The enrichment (hypergeometric *P*-value $<10^{-15}$) of 108 mitochondrial genes in the top 50 genes agrees with the observation that Alzheimer’s disease might be related with mitochondrial function (Castellani *et al*., 2002). On the contrary, there are only two mitochondrial genes and six ATP-related genes in the top 50 genes ranked by expression data only, which is no longer significant. In addition, the connectivity of mitochondrial genes in the protein interaction network is significantly less than the non-mitochondrial genes (Wilcox rank sum test *P*-value $<2 \times 10^{-6}$). Although the connectivity of the 17 top-ranked mitochondrial genes is higher than that of other mitochondrial genes, their connectivity is still similar to non-mitochondrial genes (Wilcox rank sum test *P*-value $= 0.98$).

4 DISCUSSION

Genome-wide expression profiling and protein interaction mapping studies allow researchers to discover disease genes systematically. Clustering analysis is the most widely used data exploration method. Although there are many studies on clustering algorithms and similarity metrics, efforts to integrate gene expression with other information such as protein interactions for prioritizing genes are limited, to our best knowledge. In this paper, we studied several approaches for prioritizing genes related to a phenotype by integrating gene expression profiles and protein interactions.

We studied the effect of neighborhood systems and different data integration approaches, in particular CGI-1, CGI-2 and CGI-3. It is found that CGI-3 together with the diffusion kernel is the best approach for prioritizing genes with respect to a phenotype. We also studied the robustness of our approach in terms of noise in the protein interaction data using DIP core, Uetz’s and Ito’s interaction data. The performance of our approach increases as the reliability of the protein interaction data set increases. Since a phenotype in general is the functional consequence of many genes, we also investigate the prioritizing performance of our approach by simulating ‘complex phenotypes’. The result shows that our approach performs even better when the contributing genes of the phenotype are from the same functional category. In the case where the phenotype results from genes of different functional categories, our study suggests that the prioritizing result is only slightly worse than the case of ‘simple’ phenotypes.

With the experience on yeast, we applied our approach to the data from human Alzheimer’s disease. It turns out that the four known genes as well as the 30 expert-selected genes related to Alzheimer’s disease are ranked significantly higher by our approach. We also found that the mitochondrial genes are significantly enriched in the top 50 genes by our approach, which deserves more attention.

The CGI-3 developed in this paper should be able to find genes positively associated with the phenotype. To find negatively associated genes, CGI-3 should be changed to

$$\hat{R}_i = \frac{O_i - \lambda \sum_{k:k\neq i} S_{ik} \mid O_k \mid}{1 + \lambda \sum_{k:k\neq i} S_{ik}}, \quad i = 1, 2, \cdots, n. \tag{7}$$

and we choose the genes ranked at the bottom.

In this paper we treat the interaction between proteins as a binary variable. However, it is quite possible that in some circumstances there is only a value describing the confidence of the existence of interaction between proteins. How to efficiently utilize this kind of information is a topic of future research.

Due to space limitation, we did not study the possibility of prioritizing genes within a subset of all the collected conditions (Getz *et al*., 2000) by integrating protein interaction data, which is clearly of much interest when the experiment is specially designed, e.g., for some developmental studies. We will pursue these issues in the future.

ACKNOWLEDGEMENTS

The authors are grateful to Mat Lebo for critical readings of the manuscript. This research is supported by NIH/NSF joint mathematical biology initiative DMS-0241102 and by NIH P50 HG 002790.

Conflict of Interest: none declared.

REFERENCES

Aerts,S. *et al*. (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol*., **24**, 537–544.

Blalock,E.M. *et al*. (2004) Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc Natl. Acad. Sci. USA*, **101**, 2173–2178.

Brem,R.B. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Brem,R.B. *et al*. (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, **436**, 701–703.

Castellani,R. *et al*. (2002) Role of mitochondrial dysfunction in Alzheimer's disease. *J. Neurosci. Res*., **70**, 357–360.

Cho,R.J. *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

David,F.N. (1949) The moments of the *z* and *F* distributions. *Biometrika*, **36**, 394–403.

Deane,C.M. *et al*. (2002) Protein interactions: Two methods for assessment of the reliability of high-throughput observation. *Mol. Cell. Proteom*., **1**, 349–356.

Deng,M.H. *et al*. (2003) Assessment of the reliability of protein–protein interactions and protein function prediction. *Pac. Symp. Biocomput*. (PSB2003), 140–151.

Dudoit,S. *et al*. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Evangelista,M. *et al*. (1997) Bni1p, a yeast formin linking cdc42p and the actin cytoskeleton during polarized morphogenesis. *Science*, **276**, 118–122.

Franke,L. *et al*. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet*., **78**, 1011–1025.

Gasch,A.P. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*., **11**, 4241–4257.

Gavin,A. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Geman,S. and Geman,D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Machine Intell*., **6**, 721–741.

Getz,G. *et al*. (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12084.

Ho,Y. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Hughes,T.R. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ideker,T. *et al*. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, S233–S240.

Ito,T. *et al*. (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA*, 1143–1147.

Ito,T. *et al*. (2001) A comprehensive two hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jansen,R. *et al*. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res*., **12**, 37–46.

Kondor,R.I. and Lafferty,J. (2002) Diffusion kernels on graphs and other discrete inpute spaces. *Proc. Int. Mach. Learn*., 315–322.

Krauthammer,M. *et al*. (2004) Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl Acad. Sci. USA*, **101**, 15148–15153.

Lee,H.J. *et al*. (2006) Diffusion kernel based logistic regression models for protein function prediction. *Omics: J. Integr. Biol*., **10**, 40–55.

Li,K.C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875–16880.

Lim,J. *et al*. (2006) A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, **125**, 801–814.

Maraganore,D.M. *et al*. (2005) High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet*., **77**, 685–693.

Mewes,H.W. *et al*. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*., **30**, 31–34.

Morley,M. *et al*. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.

Morrison,J.L. *et al*. (2005) GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, **6**, 233.

Mrowka,R. *et al*. (2001) Is there a bias in proteome research? *Genome Res*., **11**, 1971–1973.

Peri,S. *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*., **13**, 2363–2371.

Pruyne,D. and Bretscher,A. (2000) Polarization of cell growth in yeast. *J. Cell Sci*., **113(Pt 4)**, 571–585.

Rual,J.F. *et al*. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.

Segal,E. *et al*. (2003) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, **19**, 264–272.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stelzl,U. *et al*. (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.

The Gene Ontology Consortium. (2001) Creating the gene ontology resource: design and implementation. *Genome Res*., **11**, 1425–1433.

Tornow,S. and Mewes,H.W. (2003) Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res*., **31**, 6283–6289.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Xenarios,I. *et al*. (2002) DIP: The database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*., **30**, 303–305.

Zhou,X. *et al*. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.