BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 16 2007, pages 2121–2128
doi:10.1093/bioinformatics/btm294

*Systems biology*

# Edge-based scoring and searching method for identifying condition-responsive protein–protein interaction sub-network

Zheng Guo[1,2,*,†], Yongjin Li[1,†], Xue Gong[1], Chen Yao[1], Wencai Ma[1], Dong Wang[1], Yanhui Li[1], Jing Zhu[1], Min Zhang[1], Da Yang[1] and Jing Wang[1]

[1]Department of Bioinformatics, Bio-pharmaceutical Key Laboratory of Heilongjiang Province-Incubator of State Key Laboratory, Harbin Medical University, Harbin 150086 and [2]Bioinformatics Centre and School of Life Science, University of Electronic Science and Technology of China, Chengdu, 610054, China

Received on January 14, 2007; revised on May 19, 2007; accepted on May 25, 2007

Advance Access publication June 1, 2007

**ABSTRACT**

**Motivation:** Current high-throughput protein–protein interaction (PPI) data do not provide information about the condition(s) under which the interactions occur. Thus, the identification of condition-responsive PPI sub-networks is of great importance for investigating how a living cell adapts to changing environments.

**Results:** In this article, we propose a novel edge-based scoring and searching approach to extract a PPI sub-network responsive to conditions related to some investigated gene expression profiles. Using this approach, what we constructed is a sub-network connected by the selected edges (interactions), instead of only a set of vertices (proteins) as in previous works. Furthermore, we suggest a systematic approach to evaluate the biological relevance of the identified responsive sub-network by its ability of capturing condition-relevant functional modules. We apply the proposed method to analyze a human prostate cancer dataset and a yeast cell cycle dataset. The results demonstrate that the edge-based method is able to efficiently capture relevant protein interaction behaviors under the investigated conditions.

**Contact:** guoz@ems.hrbmu.edu.cn

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

## 1 INTRODUCTION

No protein performs its function in isolation. Instead, its 'functional expression', involved in regulating a cellular activity, is realized through interacting with other proteins (Barabasi and Oltvai, 2004). In recent years, high-throughput experiments have populated the public databases with thousands of protein–protein interaction (PPI) data (Uetz *et al.*, 2000), and PPI networks have been widely used for constructing biological pathways (Segal *et al.*, 2003) or identifying protein complexes (Spirin and Mirny, 2003). However, one weakness of the high-throughput PPI data is that it contains no information about the conditions under which the interactions may take place. In other words, the PPI network is not a real snapshot of the interactions *in vivo*, but a union of the interactions activated under various conditions. Therefore, for a given set of proteins, the interactions among them may possibly be involved in several different pathways responding to different environments or conditions (e.g. a disease state).

It is a popular way to use the gene expression information to measure the 'activity' of a molecular network or pathway in response to the investigated condition. From the entire interaction networks, some groups identified responsive PPI sub-networks based on the significant changes of gene expressions over a particular condition(s) (Ideker *et al.*, 2002; Scott *et al.*, 2005; Sohler *et al.*, 2004). Other groups ranked the activity of protein interactions, protein complexes or molecular pathways based on co-expression of the involved genes (Han *et al.*, 2004; Jansen *et al.*, 2002; Rahnenfuhrer *et al.*, 2004), under the hypothesis that higher expression correlation of the genes implies genuine interactions of the proteins under the investigated conditions. For example, Jansen *et al.* (Jansen *et al.*, 2002) distinguished condition-relevant protein complexes by the co-expression of the genes encoding the subunits of the complexes. Han *et al.* (Han *et al.*, 2004) calculated the average Pearson correlation coefficients of hubs (i.e. proteins having many interaction partners) with their partners in the PPI network, and then by the bimodal distribution of correlation coefficient, they found one type of hubs as 'date hubs' which interact with their different partners under different conditions.

Aiming at finding the sub-network of interactions responding to a particular condition, researchers have proposed some methods to extract the condition-relevant PPI sub-network by resorting to the gene expression profiles corresponding to the investigated condition. Here, we refer to such a condition-relevant PPI sub-network as a responsive sub-network, which reflects the intricate interplay between the genes (thereafter proteins) responding to the specific condition. Ideker *et al.* (Ideker *et al.*, 2002) were one of the first groups who attempted to identify such a sub-network (that they call an active sub-network), based on the speculation that the majority of genes encoding the proteins in the responsive sub-network are likely to be differentially expressed. Sohler *et al.* (Sohler *et al.*, 2004) and Scott *et al.* (Scott *et al.*, 2005) searched for significant area of a PPI network by spanning the network starting with a

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be declared as joint First Authors.

 

given set of seed vertices (proteins). These proposed methods scored and searched sub-networks by the differential expression of genes, and took all known interactions among the identified proteins as the edges of the responsive sub-network. We refer to such approaches as the vertex-based methods, which usually do not further select the active interaction relationships among the identified proteins. However, simply taking all the interactions among the 'active' proteins as the edges of the sub-network is inadequate, because under a particular condition, only a part of the interactions among a set of proteins may be active.

In this article, to identify the responsive sub-network under a particular condition, we proposed a novel edge-based scoring and searching method, using interactions (edges) between protein pairs as basic units to measure the overall activity of a sub-network. By virtue of the scoring way, we quantify the response of a sub-network from more comprehensive aspects: variation of gene expression level, gene co-expression between directly connected proteins and the topology of the sub-network, whereas many other existing methods only take a part of these factors into account. By the edge-based searching procedure utilizing connected edges, instead of vertices, what we construct is a sub-network with the topology structure of condition-relevant interactions, making an improvement over the previous vertex-based methods.

We applied the proposed method to the human PPI network from HPRD (Peri *et al.*, 2004) using a gene expression dataset of prostate cancer (Lapointe *et al.*, 2004) and the yeast PPI network from DIP (Salwinski *et al.*, 2004) using a gene expression dataset of cell cycle (Spellman *et al.*, 1998). Results demonstrated that our method was of improved efficiency in capturing relevant interaction behaviors under the investigated conditions. For the prostate cancer dataset, by taking prostate cancer related genes (Li *et al.*, 2003) as seeds, we combined the edge-based seed expansion approach (Chen *et al.*, 2006) to explore the network in more detail. The advantage for the seed expansion approach is that it can directly use prior knowledge of known disease proteins.

## 2 METHODS

### 2.1 Protein interaction and gene expression data

The PPI data was derived from the physical PPI dataset of DIP (2006 release) (Salwinski *et al.*, 2004) and HPRD (Peri *et al.*, 2004) (Release 6). We processed the data as follows: (i) removing self-interactions; (ii) removing reduplicate interactions.

The prostate dataset (Lapointe *et al.*, 2004) consists of about 26 000 genes measured in 71 prostate tumors as well as 41 normal prostate specimens. The expression dataset for cell cycle (Spellman *et al.*, 1998) contains the relative expression changes of yeast genes during the cell cycle measured in 77 different time points. For each of the above cDNA microarray datasets, we screened out genes with missing data in more than 10% of arrays and applied a base-2 logarithmic transformation(Wang *et al.*, 2006). Then, we carried out data normalization so that the observations had the mean 0 SD 1 in every array.

By integrating the processed PPI and expression data, we constructed the entire network to be searched for the condition-responsive sub-network. Briefly, from a PPI network with proteins as vertices and interactions as edges, we deleted the vertices without gene expression data. Finally, the entire network to be searched contained 6509 vertices with 23 157 edges for the prostate cancer dataset, while the entire network contained 3619 vertices with 11 083 edges for the cell cycle dataset.

### 2.2 Responsive score for a given sub-network

In this work, the responsive score of an interaction (edge) is defined by the covariation of the expression levels of interaction partners, which accounts for not only the co-expression between the directly connected proteins but also differential expressions of the genes.

Let $e_{(x,y)}$ denotes the edge between two directly connected proteins, $x$ and $y$, in the entire network. Then, the edge score is defined as

$$Score(e_{(x,y)}) = Cov(X,Y) = Corr(X,Y)std(X)std(Y)$$

Here, $Corr(X,Y)$ is the Pearson correlation coefficient of the expressions of the genes $x$ and y. The differential expressions of the genes are measured as the overall expression variation ($std(X)$ and $std(Y)$), which has been adopted by several researchers for selecting differentially expressed genes (Ding, 2003; Dudoit and Fridlyand, 2003; Xu *et al.*, 2006; Zien *et al.*, 2000).

Based on the edge scores, the score for a connected sub-network $G=(V, E)$ is defined as

$$T(\mathrm{G}) = \sum_{e \in E} Score(e).$$

Obviously, $T$ is affected by the number of the edges in the network. To eliminate this effect, for a sub-network with $k$ edges, we randomly sampled 10 000 edge sets of size $k$ from the entire PPI network, and calculate $T(G_{\mathrm{rand}})$ for each edge set. Under the null hypothesis that a sub-network is not responsive to the investigated condition, the expressions of its edges (interaction pairs) are randomly related, and therefore, the standardized score of the sub-network will not be affected by its connection structure.

Then, we estimate the mean $avg_k$ and SD $std_k$ of $T(G_{\mathrm{rand}})$. The standardized overall score of the sub-network with $k$ edges is defined as follows. The standardized $Score(G)$ of random sub-networks are guaranteed to have mean $\mu=0$ and SD $\sigma=1$.

$$Score(\mathrm{G}) = \frac{\mathrm{T(G)} - \mathrm{avg}_k}{std_k}$$

To be comparable, $Score(G)$ of sub-networks with different $k$ edges should have the same distribution. We empirically compared their scores distribution. Taking $k$ from 50 to 5000 by the step of 50, for each $k$, we randomly sample 10 000 sub-networks from the entire network and generated a score population with 10 000 scores. Then, we compared every two populations, from the generated 100 score populations, by the two-sample Kolmogorov–Smirnov test for distribution goodness-of-fit at the significant level 0.1. The null hypothesis is that the two populations of samples come from the same population. The larger the $P$-value at which the null hypothesis cannot be rejected, the more likely that they come from the same population. For both datasets, more than 99% of all 4950 $\left(C_n^2\right)$ pair-wise comparisons indicated that two populations follow the same distribution. Thus, the scores of networks with different number of edges roughly follow the same distribution and are generally comparable. Therefore, a higher-scoring sub-network, as a whole, indicates its statistical significance and biological activity in response to a specific condition.

### 2.3 Searching for responsive sub-network

Because finding the highest-scoring sub-network in the entire network is a NP-hard problem (Ideker *et al.*, 2002), we implemented the searching procedure based on simulated annealing, which has the advantage of being capable of jumping out from local optimization (Kirkpatrick *et al.*, 1983). In each iteration step, it tests whether the addition or removal of an edge will increase the score. If the score increases, the try

will be accepted, otherwise, it will be accepted with a certain probability. The pseudocode of the edge-based simulated annealing algorithm is described as below.

Input: entire PPI network $G_0 = (V, E)$; edge-score array; a set of parameters for running simulated annealing: start temperature $T_{\text{start}}$, end temperature $T_{\text{end}}$, number of iterations $N$.

Output: the connected sub-network with the highest score.

(1) Initialize $G_{RS}$ by setting each edge $e \in E$ to active/inactive with probability 0.5; Calculate scores for all connected components (sub-networks) of $G_{RS}$ and get its score vector $V_{RS}$;

(2) For $i = 1$ to $N$, Do

(3) Calculate the current temperature $T_i = T_{start} \bullet \left(\frac{T_{end}}{T_{start}}\right)^{\frac{i}{N}}$;

(4) $G_{try} \leftarrow G_{RS}$;

(5) Randomly pick an edge $e \in E$

IF ($e \in G_{try}$), remove $e$ from $G_{try}$;
ELSE add $e$ to $G_{try}$;

(6) Calculate scores for all connected components of $G_{try}$ and get their score vector $V_{trv}$;

(7) Calculate $\Delta = V_{try} - V_{RS}$ (see the rule in the Supplementary Material)

IF $\Delta > 0$, then $G_{RS} \leftarrow G_{try}$;
ELSE, accept $G_{RS} \leftarrow G_{try}$ with the probability $p = e^{\Delta/T_i}$;

(8) END (end for)

(9) Output the connected sub-network with the highest score in $G_{RS}$.

In order to avoid the possible influence on results by insufficiency of iteration, after annealing, we traversed every edge of the entire network at temperature = 0, so that every edge had been tried at least one time. As described in the pseudocode, the algorithm only outputs the connected sub-network with the highest score. The program was implemented in Matlab and Java. The program and data are available on request.

### 2.4 Evaluating the responsive sub-network by functional enrichment analysis

The identification of the responsive sub-network is essentially unsupervised, so it is a difficult problem to find an objective criterion (i.e. 'gold standard') to evaluate the biological relevance of the results to the investigated condition. Here, we empirically investigated the modular behavior of the proteins in the responsive sub-networks.

Based on the hypergeometric distribution statistics (Draghici *et al.*, 2003), we calculated the probability value of a Gene Ontology (GO) (Harris *et al.*, 2004) biological process category having at least the number of the annotated proteins in the responsive sub-network by random chance. Owing to the hierarchical nature of the GO categories, there are some redundancies in the selected categories, e.g. parent–child relationship between the categories. In such a case, only the child category was analyzed, because its functional description is more specifically defined. To concentrate on specific functions, we removed the GO categories at level 5 and above.

In addition, for the prostate cancer data, we used the hypergeometric distribution statistics to calculate the enrichment of prostate cancer genes in the responsive sub-network with respect to the entire network. The prostate cancer related genes were obtained from Prostate Gene Database (PGDB) (Li *et al.*, 2003), which covers genes, as published in the literature, involved in many molecular and genetic events of the prostate cancer, including gene amplification, mutation, gross deletion, methylation, polymorphism, linkage and over-expression. Currently, there are 175 prostate cancer genes in PGDB, 118 of which are included in the entire network.

### 2.5 Comparison with the vertex-based algorithm

For comparison, we used jActiveModules to identify the responsive sub-network based on the vertex-based algorithm proposed by Ideker *et al.* (Ideker *et al.*, 2002). The score of a sub-network was calculated based on the *P*-values (significance levels) of the differential expression of the genes corresponding to the proteins included in the sub-network. For the prostate cancer data, the *P*-values were obtained by the two-tailed *t*-test. While in the case of cell cycle data, for each time point, we assigned a *P*-value as the two-sided cumulative probability of the standardized expression levels, $p_i = 2*(1 - \text{normcdf}(x))$, where $x$ is the standardized expression level of gene $i$. The assumption to do this is that the standardized gene expression values in one array follow the standard normal distribution.

For each dataset, we compared the *P*-values of the functional categories in the two sub-networks identified by the edge-based and vertex-based methods respectively. A smaller *P*-value of a category in a sub-network indicates that the corresponding method is more efficient in capturing proteins relevant to the function described by the category.

## 3 RESULTS

To find whether the algorithm reaches convergence, we investigated how the score variation (described as $\Delta$ in the pseudocode) changed with the iterations. For each dataset, as shown in Figure 1, after iterating 30 000 times, $\Delta$ nearly came to zero, suggesting that the score in general had reached convergence. Then, we changed the starting temperature and initial active/inactive probability, and found that the obtained top scores only had little difference (Table 1 and Supplementary Table S1). Therefore, we ran simulated annealing with parameters $N = 30\,000$, $T_{\text{start}} = 1$, $T_{\text{end}} = 0.01$, for further analysis. We note that this algorithm, as others, cannot insure that the score is globally optimal. However, even a sub-network with high-score nearly optimal is still of biological interest (Ideker *et al.*, 2002).

In the prostate cancer data, the responsive sub-network identified by the edge-based method contained 2181 vertices and 3200 edges. Among the 2181 proteins in the sub-network, there were 10 157 interactions in the original entire network and 6957 interactions were filtered out of the sub-network. The responsive sub-network identified by the vertex-based method contained 1493 vertices, and all the 2430 edges among these proteins were included in the sub-network. In the yeast cell cycle data, the responsive sub-network identified by the edge-based method contained 1511 vertices and 2616 edges. Among the 1511 proteins in the sub-network, there were 5797 interactions in the original network and 3181 interactions were filtered out of the sub-network. The responsive sub-network identified by the vertex-based method contained 2726 vertices, and all the 7563 edges among these proteins were included in the sub-network. The resulting large sub-networks indicate that a large portion of proteins may be interaction-connected for coordinately carrying out the affected functions.

One of the advantages of the edge-based method is that, among the proteins in a sub-network, many less relevant interactions are thinned out. Unfortunately, because biological

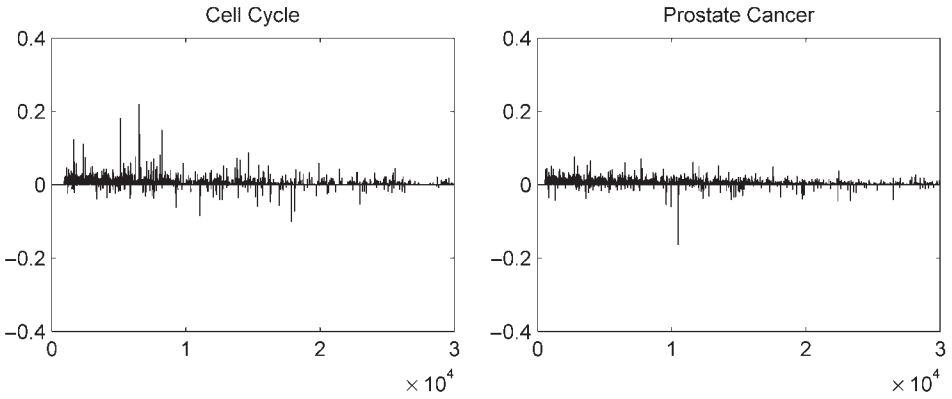


**Fig. 1.** The variation of the score during annealing process in searching the responsive sub-networks to human prostate cancer and yeast cell cycle.

**Table 1.** Effect of starting temperature on responsive scores

| $T_{start}$ | Cell cycle | | Prostate cancer | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| 1 | 45.9 | 0.64 | 69.7 | 2.08 |
| 2 | 47.5 | 0.83 | 71.8 | 2.17 |
| 5 | 46.9 | 0.34 | 69.3 | 1.60 |
| 10 | 47.5 | 0.24 | 69.8 | 1.80 |

*Note*: Each row summarizes results from five annealing runs starting from different random states of the entire network. $T_{start}$ is the starting annealing temperature. For all runs, $T_{end} = 0.01$, $N = 30\,000$.

studies are biased towards reporting positive results, it is difficult to find negative examples in literatures to justify the filtered interactions to be inactive under the investigated condition. To help interpret the results according to the common hypothesis that higher degree of gene co-expression implies genuine interactions of the proteins (Han *et al*., 2004; Jansen *et al*., 2002; Rahnenfuhrer *et al*., 2004), we calculated Pearson correlation coefficients of the expression values of protein pairs for the interactions included in the responsive sub-network and the filtered ones, respectively. As shown in Figure 2, for each dataset, the filtered edges had much lower expression correlation than the edges included in the responsive sub-network.

### 3.1 Analysis of the responsive sub-network of the human prostate cancer

The responsive sub-network identified by the edge-based method covered 74 prostate cancer genes obtained from PGDB (Li *et al*., 2003). Based on the hypergeometric distribution statistics, the random probability that the sub-network includes 74 of the 118 prostate cancer genes included in the original network was $P = 1.04\ E{-}11$. In contrast, there were only 38 prostate cancer proteins captured in the responsive sub-network based on the vertex-based method, and the random enrichment probability was $p = 7.3E{-}03$.

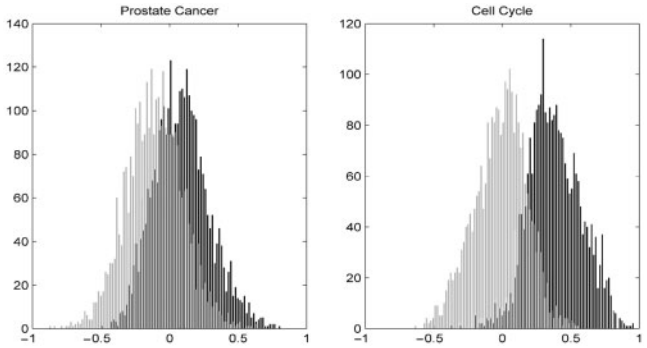


**Fig. 2.** The distribution of the co-expression data. For the active (black area) and inactive (gray area) edges, respectively, we divided the correlation coefficient values into 100 intervals and counted the frequency of the values in each interval. The *Y* axis represents the frequency of the Pearson correlation coefficients in each interval (*X* axis).

Then, we investigated the functional modules enriched with the proteins in the responsive sub-networks identified by either the edge-based or the vertex-based algorithm. All the identified modules were listed in Supplementary Table S2. After removing redundancy (see details in Methods section), the significant modules identified by at least one method were listed in Table 2. Generally, as described in Table 2, according to the *p*-values of the functional categories identified by both methods, the responsive sub-network extracted by the edge-based method was lightly more efficient in capturing proteins in the functional categories (with smaller *P*-values), which are relevant to prostate cancer, as described below. Additionally, when simply using differentially expressed genes selected by *t*-test at a given *P*-value cutoff, we found fewer significant categories, as described in supplementary Table S4.

Oncogenesis and tumor progression are relevant to alterations in cell signaling, and blocking the activation of upstream signal transduction proteins is a promising approach to cancer therapy (Hudes, 2002). Some prostate cancer relevant pathways, such as 'I-kappaB kinase/NF-kappaB cascade' (GO:0007249), 'integrin-mediated signaling pathway' (GO:0007229), 'activation of MAPK activity' (GO:0000187),

**Table 2.** Enriched GO categories for human prostate cancer

| GO ID | GO categories | Edge-based | Vertex-based |
|---|---|---|---|
| 0006796 | Phosphate metabolism | 6.96*E*−10 | 4.72*E*−04 |
| 0006917 | Induction of apoptosis | 1.55*E*−06 | 2.86*E*−01 |
| 0007229 | Integrin-mediated signaling pathway | 3.64*E*−06 | 9.42*E*−01 |
| 0008286 | Insulin receptor signaling pathway | 6.86*E*−06 | 4.36*E*−01 |
| 0006916 | Anti-apoptosis | 9.88*E*−06 | 6.83*E*−01 |
| 0019221 | Cytokine and chemokine mediated signaling pathway | 2.78*E*−05 | 9.57*E*−01 |
| 0045944 | Positive regulation of transcription from RNA polymerase II promoter | 2.88*E*−05 | 1.32*E*−01 |
| 0001525 | Angiogenesis | 3.83*E*−05 | 6.48*E*−01 |
| 0007409 | Axonogenesis | 4.04*E*−05 | 1.29*E*−01 |
| 0050870 | Positive regulation of T cell activation | 8.03*E*−05 | 6.81*E*−01 |
| 0006261 | DNA-dependent DNA replication | 1.01*E*−04 | 7.90*E*−02 |
| 0000079 | regulation of cyclin-dependent protein kinase activity | 1.05*E*−04 | 2.09*E*−01 |
| 0006469 | Negative regulation of protein kinase activity | 1.05*E*−04 | 2.87*E*−02 |
| 0008285 | Negative regulation of cell proliferation | 1.12*E*−04 | 1.24*E*−01 |
| 0007010 | Cytoskeleton organization and biogenesis | 1.60*E*−04 | 6.64*E*−01 |
| 0007178 | Transmembrane receptor protein serine/threonine kinase signaling pathway | 1.84*E*−04 | 4.86*E*−03 |
| 0007188 | G-protein signaling, coupled to cAMP nucleotide second messenger | 1.99*E*−04 | 2.99*E*−01 |
| 0008544 | Epidermis development | 2.01*E*−04 | 7.78*E*−01 |
| 0042475 | Odontogenesis (sensu Vertebrata) | 2.08*E*−04 | 4.04*E*−01 |
| 0030097 | Hemopoiesis | 2.71*E*−04 | 2.31*E*−02 |
| 0000187 | Activation of MAPK activity | 3.46*E*−04 | 6.81*E*−01 |
| 0046777 | Protein amino acid autophosphorylation | 3.69*E*−04 | 1.13*E*−01 |
| 0050671 | Positive regulation of lymphocyte proliferation | 4.00*E*−04 | 8.16*E*−01 |
| 0050730 | Regulation of peptidyl-tyrosine phosphorylation | 5.50*E*−04 | 1.00*E*+00 |
| 0048002 | Antigen processing and presentation of peptide antigen | 6.95*E*−04 | 1.00*E*+00 |
| 0042994 | Cytoplasmic sequestering of transcription factor | 8.27*E*−04 | 2.01*E*−01 |
| 0007249 | I-kappaB kinase/ NF-kappaB cascade | 8.48*E*−04 | 6.67*E*−01 |
| 0050803 | Regulation of synapse structure and function | 8.69*E*−04 | 3.46*E*−01 |
| 0007507 | Heart development | 9.13*E*−04 | 5.73*E*−01 |
| 0002443 | Leukocyte mediated immunity | 9.41*E*−04 | 5.23*E*−01 |
| 0007265 | Ras protein signal transduction | 1.66*E*−02 | 7.00*E*−04 |
| 0051052 | Regulation of DNA metabolism | 1.45*E*−01 | 7.00*E*−04 |

were enriched with proteins in our responsive sub-network. For example, I-kappaB kinase can regulate transcription factor nuclear factor-kappaB (NF-kappaB), which is a key anti-apoptotic factor in mammalian cell. It was found that inhibition of NF-kappaB anti-apoptotic and proliferative activation pathways could inhibit prostate cancer cell growth (Gasparian *et al*., 2002). Some integrin-activated signaling pathways such as FAK (focal adhesion kinase) and PI3-kinase (phosphatidylinositol 3-kinase) are involved in controlling proliferation, survival and migration of prostate cancer cells (Fornaro *et al*., 2001). Activation of MAPK activity is an important portion in prostate cancer progression. Prostate cancer cells achieve the transition from androgen-sensitivity to androgen-independence by different multistep routes, including adapting the AR (androgen receptor) pathway via MAPK (Edwards and Bartlett, 2005). In addition to signal transduction pathways, many other enriched categories were also relevant to prostate cancer, such as angiogenesis (GO:0001525) which is highly relevant to metastatic potential of prostate cancer cells (Aalinkeel *et al*., 2004).

It is often believed that hub proteins play central roles in both cellular processes and PPI network topology (Han *et al*., 2004; Jeong *et al*., 2001). The top 20 hub proteins with the highest degrees in the responsive sub-network were listed in supplementary Table S5. The hub protein with the highest degree in the responsive sub-network was TP53, which is a tumor suppressor protein inactivated in many cancers, including prostate cancer. The second highest degree hub was SRC which is a coactivator of androgen receptor, playing a role in androgen-independent prostate cancer (Agoulnik *et al*., 2005). The third hub FYN was a member of Src family kinases essential for many cell functions (Cohen, 2005). It has been reported that FYN was down-regulated in prostate cancer cell lines, which might be a new tumor suppressor of prostate cancer (Sorensen *et al*., 2006).

By combining some other biology knowledge and methods, we can explore the responsive sub-network for some more detailed biological results. For example, because disease proteins tend to interact with each other (Gandhi *et al*., 2006; Xu and Li, 2006), proteins directly interact with many disease proteins may suggest valuable information to biologists, especially when the proteins are in response to the disease condition.

Here, restricted to the responsive sub-network identified by the edge-based method for prostate cancer, we applied the method suggested by Chen *et al*. (Chen *et al*., 2006) to find a sub-region closely related to disease proteins, by connecting the proteins directly interacting with at least one of the 118 disease proteins included in both the PGDB and the responsive sub-network. As shown in Figure 3, the largest connected sub-region contained 123 interactions and 109 proteins, including 40 disease-related proteins. In the sub-region, there were 17 proteins interacting with at least two prostate cancer proteins. Eight of them (SRC, STAT3, CREBBP,FOS, JUN, PRKCA, IRS1 and FYN) were reported to be related to prostate cancer in recent published literatures (Agoulnik *et al*., 2005; Azare *et al*., 2007; Comuzzi *et al*., 2004; Chen *et al*., 2006; Neuhausen *et al*., 2005; Sorensen *et al*., 2006; Stewart and O'Brian, 2005).

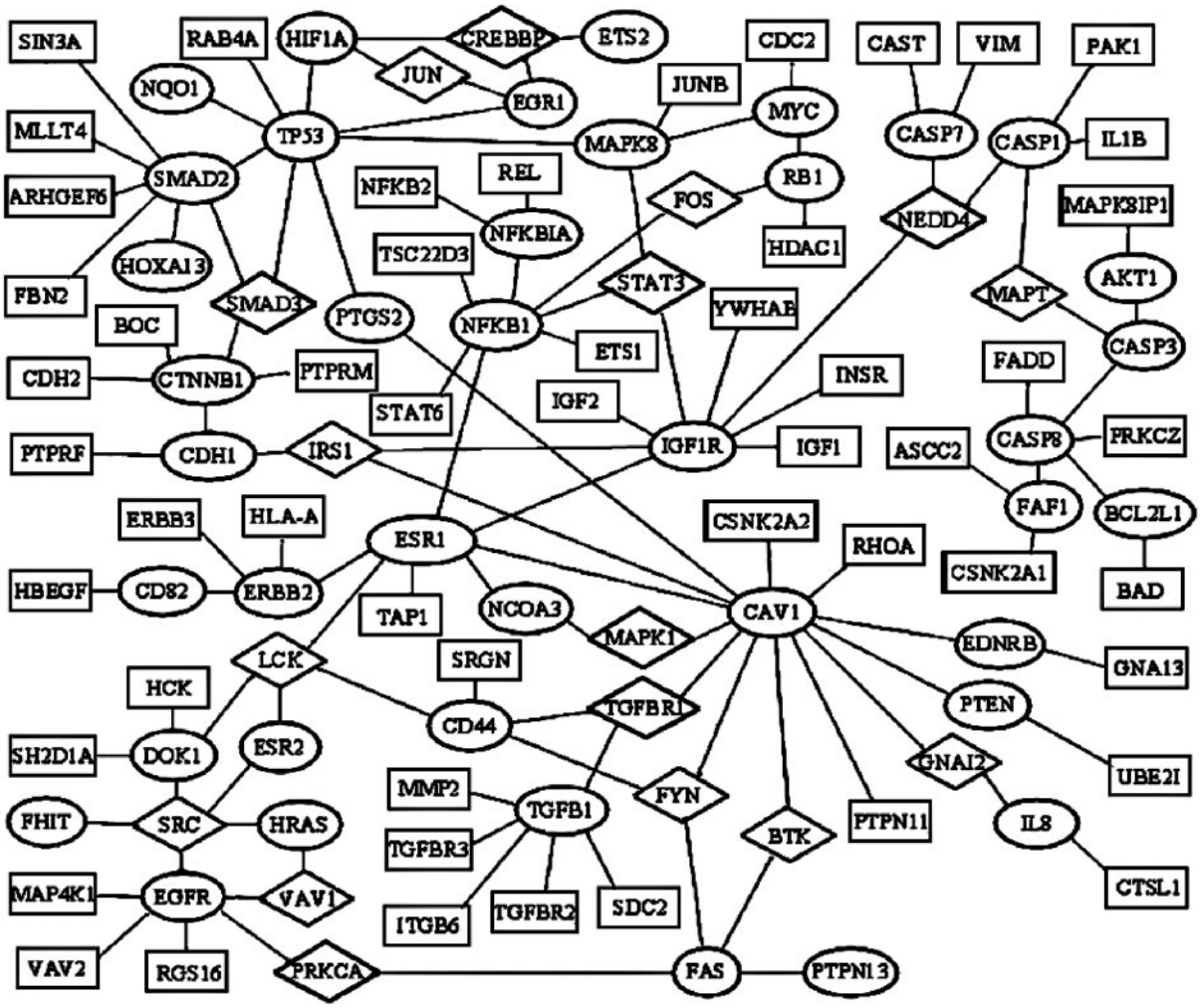


**Fig. 3.** Ellipses denote disease proteins, and diamonds indicate proteins interacting with more than two disease proteins.

### 3.2 Analysis of the responsive sub-network of the yeast cell cycle

After removing redundancy, the significant modules identified by at least one method were listed in Table 3. All the identified modules were listed in Supplementary Table S3. Similar to the results of the prostate cancer, generally, when a category was identified by both methods, the *P*-value for the edge-based method was smaller.

Yeast cell cycle is a tightly regulated process during which the replication of DNA, transcription of RNA and translation of protein are indispensable (Sobel, 1997). The enrichment analysis suggest that the edge-based method was more efficient in identifying relevant categories such as 'G1/S transition of mitotic cell cycle' and 'G2/M transition of mitotic cell cycle', which are key regulatory points of the cell cycle. Protein synthesis has the capacity of restricting a cell to progress past the cell cycle (Yu *et al*., 2006).

Ribosomes are 'factories' of protein synthesis and synthesis of ribosomes, which is a key control point for the regulation of cell growth and division (Dez and Tollervey, 2004). As shown in Table 3, we were able to identify ribosomal biogenesis related categories. Some genes in these categories are related to cell cycle. For instance, over-expression of YLR197W (*SIK1*), shortens the G1 phase of the yeast cell cycle and causes spindle orientation defects (Bogomolnaya *et al*., 2004). Depletion of YNL110C (*NOP15*) leads to a cell cycle defect (Oeffinger and Tollervey, 2003). And YGR103W (*YPH1*) is pivotal for yeast cells to exit G0 and initiate a cell cycle, whose depletion will lead to cells arrest in G1 or G2 (Du and Stillman, 2002).

The top 20 hubs in the responsive sub-network for the yeast cell cycle data were listed in Supplementary Table S6. Also, we found that, some key hub proteins in the responsive sub-network were involved in the functions related to the cell cycle process. The hub protein with the highest–degree was YBR160W (*CDC28*), which is the catalytic subunit of the main cell cycle cyclin-dependent kinase (CDK) (Mendenhall and Hodge, 1998). CDK activity drives events of the cell cycle through phosphorylation of key substrates (Loog and Morgan, 2005; Ubersax *et al*., 2003; Wittenberg, 2005). The hub protein with the second highest-degree in the responsive sub-network

**Table 3.** Enriched GO categories for yeast cell cycle

| GO ID | GO categories | Edge-based | Vertex-based |
|---|---|---|---|
| 0006365 | 35S primary transcript processing | 1.35*E*−10 | 5.63*E*−02 |
| 0030490 | Processing of 20S pre-rRNA | 4.15*E*−10 | 5.50*E*−01 |
| 0006511 | Ubiquitin-dependent protein catabolism | 3.68*E*−07 | 4.65*E*−01 |
| 0042273 | Ribosomal large subunit biogenesis | 6.15*E*−07 | 5.64*E*−02 |
| 0006413 | Translational initiation | 9.92*E*−07 | 1.12*E*−01 |
| 0007035 | Vacuolar acidification | 7.77*E*−05 | 3.29*E*−01 |
| 0000082 | G1/S transition of mitotic cell cycle | 1.43*E*−04 | 4.29*E*−01 |
| 0006606 | Protein import into nucleus | 1.59*E*−04 | 1.97*E*−02 |
| 0000086 | G2/M transition of mitotic cell cycle | 2.41*E*−04 | 6.91*E*−02 |
| 0016071 | mRNA metabolism | 2.73*E*−04 | 4.86*E*−02 |
| 0015992 | Proton transport | 5.22*E*−04 | 6.20*E*−01 |
| 0000054 | Ribosome export from nucleus | 5.80*E*−04 | 1.50*E*−01 |
| 0006270 | DNA replication initiation | 6.60*E*−04 | 6.83*E*−01 |
| 0006273 | Lagging strand elongation | 6.82*E*−04 | 4.86*E*−01 |
| 0042255 | Ribosome assembly | 9.97*E*−04 | 9.80*E*−02 |

was YCR057C (*PWP2*), which is a conserved 90S pre-ribosomal component essential for proper endonucleolytic cleavage of the 35S rRNA precursor, whose deletion leads to defects in cell cycle (Dosil and Bustelo, 2004; Shafaatian *et al*., 1996).

## 4 DISCUSSION

In this article, we proposed a novel edge-based scoring and searching approach to identify responsive sub-networks under particular conditions. First, every edge (interaction) was given an active score according to the gene expression information under the investigated conditions. Then, the overall sub-network score was calculated by all the interactions in the sub-network. The edge-based searching was implemented by the edge-based simulated annealing algorithm, which optimizes the connected edges, instead of the vertices as in the conventional vertex-based algorithm. Therefore, what we constructed was a genuine sub-network with specific active interactions, rather than merely a set of proteins connected by all the interactions in the original PPI network as what the vertex-based methods obtained. Furthermore, we demonstrated that the proposed method was able to discover the condition-relevant functional modules efficiently. It is worth noting that what we found were 'condition-responsive' or 'condition-relevant' sub-networks, including interactions likely to happen under current conditions. Because we only identified the interactions under particular conditions, it is impossible to identify the 'condition-specific' interactions, i.e. the interactions merely happen in the studied conditions but not under any other conditions.

It has been suggested that genes with similar expression profiles are more likely to encode interacting proteins (Ge *et al*., 2001), and as discussed in the Introduction section, it is a popular way to use the gene expression information to measure the 'activity' of interactions or a molecular network in response to a particular condition. However, it should be noted that the relationship between PPI and gene expression is complicated. Although several studies have found that mRNA and protein expression levels in yeast cells to be correlated to various degrees (Mijalski *et al*., 2005), the gene expression level does not necessarily represent the true protein abundance. Furthermore, expression data are of limited ability for identifying the most interesting 'switches' in PPI behavior, which are generally determined by other factors such as ligand binding and posttranslational modification. Therefore, integrating other data such as DNA–protein interaction, transcription factor-binding information (Lee *et al*., 2002; Matys *et al*., 2003) will be imperative to understanding how cellular networks coordinately adapt to changing environments (Barabasi and Oltvai, 2004; Tanay *et al*., 2004). Furthermore, other functional relationships such as metabolic and signaling pathways can be represented in the form of interaction networks, and our method can be applied to such networks to extract condition-responsive sub-networks.

Another problem that may degrade the proposed algorithm is the lack or incomplete coverage of interaction data. However, given the ever-increasing amount of interaction data, we expect that the approach described here will enhance the understanding how the responsive sub-networks, enriched with multiple functional modules, is assembled into an entire living system. Obviously, it is an important future task to find higher-level interactions among the multiple modules coordinately carrying out cellular functions responding to the same experimental condition.

## ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (grant nos. 30170515, 30370388 and 30370798).

*Conflict of Interest*: none declared.

## REFERENCES

Aalinkeel,R. *et al*. (2004) Gene expression of angiogenic factors correlates with metastatic potential of prostate cancer cells. *Cancer Res*., **64**, 5311–5321.

Agoulnik,I.U. *et al*. (2005) Role of SRC-1 in the promotion of prostate cancer cell growth and tumor progression. *Cancer Res*., **65**, 7959–7967.

Azare,J. *et al*. (2007) Constitutively activated Stat3 induces tumorigenesis and enhances cell motility of prostate epithelial cells through Integrin β6. *Mol. Cell Biol*., **27**, 4444–4453.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet*., **5**, 101–113.

Bogomolnaya,L.M. *et al*. (2004) A new enrichment approach identifies genes that alter cell cycle progression in Saccharomyces cerevisiae. *Curr. Genet*., **45**, 350–359.

Chen,J.Y. *et al*. (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac. Symp. Biocomput*., **11**, 367–378.

Chen,S.Y. *et al.* (2006) c-Jun enhancement of androgen receptor transactivation is associated with prostate cancer cell proliferation. *Oncogene*, **25**, 7212–7223.
Cohen,D.M. (2005) SRC family kinases in cell volume regulation. *Am. J. Physiol. Cell Physiol.*, **288**, C483–C493.
Comuzzi,B. *et al.* (2004) The androgen receptor co-activator CBP is up-regulated following androgen withdrawal and is highly expressed in advanced prostate cancer. *J. Pathol.*, **204**, 159–166.
Dez,C. and Tollervey,D. (2004) Ribosome synthesis meets the cell cycle. *Curr. Opin. Microbiol.*, **7**, 631–637.
Ding,C.H. (2003) Unsupervised feature selection via two-way ordering in gene expression analysis. *Bioinformatics*, **19**, 1259–1266.
Dosil,M. and Bustelo,X.R. (2004) Functional characterization of Pwp2, a WD family protein essential for the assembly of the 90 S pre-ribosomal particle. *J. Biol. Chem.*, **279**, 37385–37397.
Draghici,S. *et al.* (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**, 3775–3781.
Du,Y.C. and Stillman,B. (2002) Yph1p, an ORC-interacting protein: potential links between cell proliferation control, DNA replication, and ribosome biogenesis. *Cell*, **109**, 835–848.
Dudoit,S. and Fridlyand,J. (2003) Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, **19**, 1090–1099.
Edwards,J. and Bartlett,J.M. (2005) The androgen receptor and signal-transduction pathways in hormone-refractory prostate cancer. Part 2: androgen-receptor cofactors and bypass pathways. *BJU Int.*, **95**, 1327–1335.
Fornaro,M. *et al.* (2001) Integrins and prostate cancer metastases. *Cancer Metastasis Rev.*, **20**, 321–331.
Gandhi,T.K. *et al.* (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.*, **38**, 285–293.
Gasparian,A.V. *et al.* (2002) Selenium compounds inhibit I kappa B kinase (IKK) and nuclear factor-kappa B (NF-kappa B) in prostate cancer cells. *Mol. Cancer Ther.*, **1**, 1079–1087.
Ge,H. *et al.* (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet.*, **29**, 482–486.
Han,J.D. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.
Harris,M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
Hudes,G.R. (2002) Signaling inhibitors in the treatment of prostate cancer. *Invest New Drugs*, **20**, 159–172.
Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (Suppl. 1), S233–S240.
Jansen,R. *et al.* (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, **12**, 37–46.
Jeong,H. *et al.* (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
Kirkpatrick,S. *et al.* (1983) Optimization by Simulated Annealing. *Science*, **220**, 671.
Lapointe,J. *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 811–816.
Lee,T.I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.
Li,L.C. *et al.* (2003) PGDB: a curated and integrated database of genes related to the prostate. *Nucleic Acids Res.*, **31**, 291–293.
Loog,M. and Morgan,D.O. (2005) Cyclin specificity in the phosphorylation of cyclin-dependent kinase substrates. *Nature*, **434**, 104–108.
Matys,V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
Mendenhall,M.D. and Hodge,A.E. (1998) Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.*, **62**, 1191–1243.
Mijalski,T. *et al.* (2005) Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues. *Proc. Natl. Acad Sci. USA*, **102**, 8621–8626.
Neuhausen,S.L. *et al.* (2005) Prostate cancer risk and IRS1, IRS2, IGF1, and INS polymorphisms: strong association of IRS1 G972R variant and cancer risk. *Prostate*, **64**, 168–174.
Oeffinger,M. and Tollervey,D. (2003) Yeast Nop15p is an RNA-binding protein required for pre-rRNA processing and cytokinesis. *EMBO J.*, **22**, 6573–6583.
Peri,S. *et al.* (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.*, **32**, D497–D501.
Rahnenfuhrer,J. *et al.* (2004) Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat. Appl. Gen. Mol. Biol.*, **3**, Article 16.
Salwinski,L. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
Scott,M.S. *et al.* (2005) Identifying regulatory subnetworks for a set of genes. *Mol. Cell Proteomics*, **4**, 683–692.
Segal,E. *et al.* (2003) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, **19** (Suppl. 1), i264–i271.
Shafaatian,R. *et al.* (1996) PWP2, a member of the WD-repeat family of proteins, is an essential Saccharomyces cerevisiae gene involved in cell separation. *Mol. Gen. Genet.*, **252**, 101–114.
Sobel,S.G. (1997) Mini review: mitosis and the spindle pole body in Saccharomyces cerevisiae. *J. Exp. Zool.*, **277**, 120–138.
Sohler,F. *et al.* (2004) New methods for joint analysis of biological networks and expression data. *Bioinformatics*, **20**, 1517–1521.
Sorensen,K.D. *et al.* (2006) Identification of FYN kinase as a new tumor suppressor in prostate cancer. *AACR Meet. Abstr.*, **2006**, 611-a..
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.
Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.
Stewart,J.R. and O'Brian,C.A. (2005) Protein kinase C-{alpha} mediates epidermal growth factor receptor transactivation in human prostate cancer cells. *Mol. Cancer Ther.*, **4**, 726–732.
Tanay,A. *et al.* (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genome-wide data. *Proc. Natl Acad. Sci. USA*, **101**, 2981–2986.
Ubersax,J.A. *et al.* (2003) Targets of the cyclin-dependent kinase Cdk1. *Nature*, **425**, 859–864.
Uetz,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.
Wang,D. *et al.* (2006) Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules. *Bioinformatics*, **22**, 2883–2889.
Wittenberg,C. (2005) Cell cycle: cyclin guides the way. *Nature*, **434**, 34–35.
Xu,J. and Li,Y. (2006) Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, **22**, 2800–2805.
Xu,J.Z. *et al.* (2006) Peeling off the hidden genetic heterogeneities of cancers based on disease-relevant functional modules. *Mol. Med.*, **12**, 25–33.
Yu,L. *et al.* (2006) A survey of essential gene function in the yeast cell division cycle. *Mol. Biol. Cell*, **17**, 4736–4747.
Zien,A. *et al.* (2000) Analysis of gene expression data with pathway scores. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 407–417.