Mol. Cells *27*, 271-277, March 31, 2009
DOI/10.1007/s10059-009-0035-x

Molecules and Cells

Minireview

Identifying Responsive Functional Modules from Protein-Protein Interaction Network

Zikai Wu[1,2], Xingming Zhao[1], and Luonan Chen[1,*]

Proteins interact with each other within a cell, and those interactions give rise to the biological function and dynamical behavior of cellular systems. Generally, the protein interactions are temporal, spatial, or condition dependent in a specific cell, where only a small part of interactions usually take place under certain conditions. Recently, although a large amount of protein interaction data have been collected by high-throughput technologies, the interactions are recorded or summarized under various or different conditions and therefore cannot be directly used to identify signaling pathways or active networks, which are believed to work in specific cells under specific conditions. However, protein interactions activated under specific conditions may give hints to the biological process underlying corresponding phenotypes. In particular, responsive functional modules consist of protein interactions activated under specific conditions can provide insight into the mechanism underlying biological systems, e.g. protein interaction subnetworks found for certain diseases rather than normal conditions may help to discover potential biomarkers. From computational viewpoint, identifying responsive functional modules can be formulated as an optimization problem. Therefore, efficient computational methods for extracting responsive functional modules are strongly demanded due to the NP-hard nature of such a combinatorial problem. In this review, we first report recent advances in development of computational methods for extracting responsive functional modules or active pathways from protein interaction network and microarray data. Then from computational aspect, we discuss remaining obstacles and perspectives for this attractive and challenging topic in the area of systems biology.

INTRODUCTION

The classical molecular biology has dissected cells in a living organism into proteins, DNAs, metabolites and other small molecules. However, it is known that a complicated living organism cannot be fully understood by merely analyzing individual components instead of regarding the organism as a system or an interacting network. Therefore, one of major challenges in post-genomic biology is to capture the interactions among proteins, DNAs, metabolites and other small molecules, and understand how these interactions are organized to form an active network in a concert manner (Albert-Laszlo et al., 2004; Chen et al., 2009; Han et al., 2004; Watts and Atrogatz, 1998).

Recently, it has been found that a cell can be organized into modules in nature (Albert-Laszlo et al., 2004; Chen et al., 2009; Han et al., 2004; Watts and Atrogatz, 1998), where distinct sets of proteins and corresponding interactions constitute different building blocks underlying different biological processes. Based on these findings, a number of computational methods have been proposed to identify functional modules, which may be named as complexes, pathways, and so on (Adamcsek et al., 2006; Bader and Hogue, 2003; Cho et al., 2007; Chu and Ghahramani, 2006; Hirsh and Sharan, 2006; Hwang et al., 2006; King et al., 2004; Mete et al., 2008; Pereira-Leal et al., 2004; Qi et al., 2008; Scholtens et al., 2005; Sharan et al., 2005; Spirin and Mirny, 2003; Turanalp et al., 2008; Wang et al., 2006; Zhang et al., 2006; 2007). Despite different techniques adopted so far, most of the computational methods aim to identify functional modules by finding a dense connected subnetwork mainly based on network topology (Cho et al., 2007; Qi et al., 2008). On the other hand, a living cell is dynamic in nature, and a biological process may be activated at some time but turned off at another time. The biological processes activated under specific conditions can be expressed as specific responsive functional modules (RFM) in the protein network. For example, if one node of the protein network is disturbed, it will affect the downstream nodes in the same pathway, which can be expressed as a responsive functional module responding to the disturbance or the phenotype manifested. Hence, it is vital to identify the responsive functional modules or active pathways corresponding to specific phenotype within the genome-scale protein-protein interaction (PPI) network.

In recent years, with the rapid progress of high-throughput technologies, it becomes possible and traceable to capture the interactions among cellular components. For example, two-hybrid assay, affinity purification and co-immunoprecipitation technology are used to test physical interactions between two proteins, and ChIP-chip approach is also applied to detect physical interactions between proteins and DNAs. On the other hand, microarray data can also capture functional interaction

[1]Institute of Systems Biology, Shanghai University, Shanghai 200444, China, [2]School of Communication and Information Engineering, Shanghai University, China
*Correspondence: lnchen@staff.shu.edu.cn

Received January 23, 2009; accepted January 26, 2009; published online March 19, 2009

Keywords: optimization, protein-protein interaction network, responsive functional module, signaling pathway, systems biology

Table 1. Summary of heuristic approaches for responsive functional module detection. NA means not available.

References	Functional module	Data sources	Software
Ideker et al. (2002)	General module	PPI/Protein-DNA/microarray	Cytoscape
Sohler et al. (2004)	General module	Biological network/microarray	ToPNet
Scott et al. (2005)	General module	PPI/Protein-DNA/list of genes	NA
Nacu et al. (2007)	General module	PPI/microarray	GXNA
Liu et al. (2007)	General module	PPI/microarray	NA
Guo et al. (2007)	General module	PPI/microarray	NA
Ulitsky et al. (2008)	General module	PPI/microarray	NA
Steffen et al. (2002)	Signaling pathway	PPI/microarray	Netsearch
Liu et al. (2004)	Signaling pathway	PPI/microarray	NA
Scott et al. (2006)	Signaling pathway	PPI/microarray	NA
Arga et al. (2007)	Signaling pathway	PPI/microarray/GO	NA
Berek et al. (2007)	Signaling pathway	PPI/microarray/GO	Pathfinder

between two proteins, where corresponding genes' co-expression implies functional interaction between them (Guo et al., 2007; Han et al., 2004; Jansen et al., 2002; Rahnenfuhrer et al., 2004). Nowadays, high-throughput experiments have populated the public databases with large amount of protein-protein interaction data, Protein-DNA interaction data, and so on. Since protein interaction data are more abundant than Protein-DNA interaction data and more creditable than co-expression based functional interactions, it becomes the preferred source to reconstruct the cellular systems or protein networks. Although PPI network can provide insight into functional relationships among proteins, it is still a difficult task to identify RFMs by exploring only PPIs due to inconsistent conditions of those recorded interactions. On the other hand, gene expression data can be viewed as a snapshot of the dynamic behavior of cellular system at specific time points. Therefore, integrating PPI data with gene expression data can provide further insight into the dynamic behavior underlying the biological system, and help to reveal the functional modules activated under certain conditions, i.e. RFMs.

Generally, identifying RFMs from high throughput data can be formulated as an optimization problem. From the algorithmic viewpoint, the existing computational methods can be classified into two groups: heuristic approaches and exact approaches, which will be described next in details respectively. In this review, we first report the recent progress in development of computational methods used for extracting responsive functional modules from PPI network based on high-throughput data although it is by no means comprehensive due to the rapid evolvement of the field. Then from computational aspect, we discuss remaining obstacles and perspectives for this attractive and challenging topic in the area of systems biology.

Extracting responsive functional modules by heuristic approaches

From computational viewpoint, extracting RFMs can be formulated as an optimization problem with various type of objective functions defined for different purposes. With PPI network and other available information, many computational methods have been proposed to identify responsive function modules or pathways by utilizing heuristic algorithms for such an optimization problem. Table 1 lists several heuristic approaches that are widely used in practice. The details of these methods can be found in the referred references.

Recently, gene set analysis becomes an active research topic and many methods have been developed to reveal disturbance or phenotype related pathways (Backes et al., 2007; Bild and Febbo, 2005; Chen et al., 2008; Holden et al., 2008; Huang et al., 2006; Nettleton et al., 2008; Oron et al., 2008; Subramaniana et al., 2005). The rationale behind gene set analysis is that the biological process underlying the gene set would be the responsive one under specified phenotype if specified disturbance or phenotype related genes are enriched significantly or the cohesiveness level of transcription changed significantly in a priori gene set. In other words, the gene set obtained by gene set analysis is actually a functional module responding to some specific disturbance. Owing to the merits of PPI data, many computational methods have been proposed to extract the responsive functional module under disturbance or phenotype by mining PPI network (Alon et al., 1995; Arga et al., 2007; Bebek and Yang, 2007; Cabusora et al., 2005; Chu and Chen, 2008; Chuang et al., 2007; Dittrich et al.,2008; Guo et al., 2007; Ideker and Sharan, 2008; Ideker et al., 2002; Kann, 2007; Liu and Zhao, 2004; Liu et al., 2007; Murali and Rivera, 2008; Nacu et al., 2007; Qiu and Zhang, 2008; Qiu et al., 2009; Rajagopalan and Agarwal, 2005; Scott et al., 2005; 2006; Sohler et al., 2004; Steffen et al., 2002; Suderman and Michael, 2007; Ulitsky et al., 2008; Wang and Xia, 2008; Zhao et al., 2008a; 2008b; 2009).

Ideker et al. (2002) are among the first groups to investigate mechanism underlying the observed gene expression changes from the perspective of network. They proposed a new method to extract subnetworks from back-ground large-scale protein interaction network (in this network, protein-DNA interactions are also included). In their method, each gene or node is first weighted by a Z score measuring the gene's differential expression level, and then a aggregate score for subnetwork is defined accordingly. Finally, the simulated annealing algorithm is applied to search high score subnet-works from the background network. Later, several other methods were proposed (Chen et al., 2009; Chuang et al., 2007; Nacu et al., 2007; Rajagopalan and Agarwal, 2005; Scott et al., 2005; Sohler et al., 2004) by several groups. For example, an algorithm was developed to identify high score subnetworks expanding from seed nodes based on several score functions (Nacu et al., 2007), In Liu et al. (2007), Ideker et al's algorithm was combined with conventional gene set analysis to identify deregulated biological processes in Type 2 Diabetes. They assume that the biological process will be deemed as deregulated in the phenotype if the genes belonging to a high score subnetwork are enriched in it in many patient samples.

Guo et al. (2007) proposed a method to identify condition-responsive subnetworks from PPI network. The main difference from Ideker et al's work is that the relation between edges or interactions and disturbance is considered when scoring a subnetwork. In other words, all the interactions in background network are assumed to exist under disturbance in Ideker et al's work, while only protein-protein interactions with high co-expressions between corresponding genes are considered in Guo's work. It is evident that the latter assumption is more reasonable as many studies found that not all protein interactions occur at a specific tissue and at a specific time (Han et al., 2004; Jansen et al., 2002; Rahnenfuhrer et al., 2004). With the filtered PPI network, the simulated annealing algorithm is utilized to identify high scoring subnetworks. Most recently, Ulitsky et al. (2008) proposed an approach to detect minimal connected subnetworks, where distinct sets of genes are dysregulated in different patient samples. Finally, the goal is approached by integrating two greedy algorithms and Covering Using Shortest Paths algorithm with heuristic rules. The flexibility that allows for distinct affected gene sets in different patients just matches the diversity of cellular state under disease, and therefore this method fits better for analyzing multiple disease samples simultaneously.

Signal transduction is the primary means that cells receive extracellular stimuli and accordingly mediates sophisticated biological processes. Therefore, signaling pathway is one specific responsive functional module, by which cell responds to external stimuli. At present, molecular components of some signaling pathways have been identified by time-consuming gene knockout experiments and epistasis (Albert et al., 2007; Li et al., 2006), thereby raising the requirement for computational methods that discover molecular components and biochemical reactions involved in signaling pathways. Since signal transduction itself is a series of biochemical reactions achieved by a cascade of protein interactions, many new computational methods have been proposed to capture the details of signaling pathways by exploiting high-throughput genomic and proteomic data (Albert et al., 2007; Alon et al., 1995; Arga et al., 2007; Bebek and Yang, 2007; Li et al., 2006; Liu and Zhao, 2004; Scott et al., 2006; Steffen et al., 2002; Zhao et al., 2008a; 2008b; 2009).

Scott et al. (2006) developed a computational method, namely color coding (Alon et al., 1995; Chen et al., 2009), to reconstruct signaling pathways from yeast PPI network. In the color coding method, a number of possible pathways are firstly found with a score assigned to each candidate and the top scoring pathways are then assembled into a signaling network. In more detail, finding candidate pathways from a PPI network is formulated as a graph theory problem: given an undirected weighted graph, one tries to find some high score linear paths with fixed length that start and end at specified vertices. After assigning one color to each vertex randomly, color coding method works in a similar way as dynamic programming. In their work, Scott et al put additional constraints to dynamic programming algorithm for identifying more general sequential pathways that contain specific nodes.

Although PPI data are valuable sources for detecting signaling pathways, PPI data is notorious for false positives and false negatives. Furthermore, the available PPI data are far from complete. On the other hand, microarray data is much more abundant and easier to retrieve, thereby can complement the PPI data to some extent. By integrating PPI and microarray data, Steffen et al. (2002) developed an algorithm, namely Netsearch, to reconstruct signaling pathways. In Netsearch method, every possible linear path of a specified length starting at any membrane protein and ending at any DNA-binding protein (e.g. transcription factor) is ranked by the degree of correlations in the expression profiles of pathway members. Finally, top ranked pathways are aggregated into a signaling pathway (Chen et al., 2009). In addition, Liu and Zhao (2004) proposed a new score function for predicting the order of signaling pathway components by employing both protein interaction and microarray data. Since the incompleteness and false positives exist in PPI data, a new computational method was proposed for recovering signaling pathway by reconstructing PPI networks with functional information (Arga et al., 2007). The method is actually the same as the one proposed in (Steffen et al., 2002) except that a refined background network is reconstructed with PPIs occurring between a set of signaling related proteins, and that gene co-expression level of each adjacent protein pair is used to rank pathway.

Most recently, a new data mining method, namely Pathfinder (Bebek and Yang, 2007), was developed. In Pathfinder, it is assumed that the linear path segments in the PPI network that have similar characteristics with known signaling pathways are more possibly putative signaling pathways. The functional association rules between interacting members of known signaling pathways are first mined, and each interaction in PPI network is weighted by the level of co-expression of corresponding genes. Then, all the linear paths with length within a specified range are extracted, where each path starts from a membrane protein and ends at a transcription factor in the PPI network. Finally, pathways with average interaction weight higher than a given threshold are obtained if these pathways satisfy a certain number of functional association rules mined from known pathways.

Despite various technical differences, above mentioned heuristic methods follow the same framework. Firstly, a background interaction network is reconstructed by integrating PPI data with other high-throughput data. Secondly, the background network is transformed into a weighted network by weighting every gene (protein) and/or edge (interaction) with microarray data or other data sources. Thirdly, a heuristic algorithm is applied or developed to find high score subnetworks in the weighted background network. Finally, high score subnetworks are identified as putative responsive functional modules. The framework is depicted in Fig. 1.

Extracting responsive functional modules by exact approaches

Although different background networks are considered by the computational methods described in Table 1, they are actually heuristic methods by utilizing heuristic algorithms, e.g. simulated annealing, to identify the responsive functional modules. However, the heuristic methods cannot ensure that the solutions are optimal ones. To obtain high accurate solutions under certain conditions, some exact approaches based on mathematical programming and graph theory have been proposed. Table 2 lists the popular approaches used in literature.

Qiu et al. (2008; 2009) proposed a novel method to detect differentially expressed pathways from molecular networks. The difference of this work from Ideker's work is that a non-parametric statistical measure of differential expression level, namely signal to noise ratio t_i (SNR), is utilized to score a subnetwork and an exact searching strategy based on a mixed

Table 2. Summary of popular exact approaches for responsive functional module detection. NA means no software available

References	Functional module	Data sources	Software
Qiu et al. (2008; 2009)	General module	PPI/microarray	NA
Dittrich et al. (2008)	General module	PPI/microarray	Heinz
Wang et al. (2008)	General module	PPI/condition specific data	NA
Zhao et al. (2008a)	Signaling pathway	PPI	NA
Zhao et al. (2008b)	Signaling pathway	PPI/microarray	NA
Zhao et al. (2009)	Signaling pathway	PPI	NA

integer linear model is proposed to detect high score subnetworks. Specifically, the molecular interaction network was represented as an undirected graph $G = (V,E,W)$, where V represents the set of genes, E represents the set of interactions between nodes and W is the set of weights which are assigned by the SNR value t_i of each gene. The problem of finding maximum-score subnetwork with size R including the root node v_1 is modeled as a constrained maximum-weighted connected graph problem, which is solved by an integrating mixed integer linear programming (MILP) with breadth-first search strategy.

$$\max_{\{x_i,c_{ij}\}} \quad W = \frac{1}{R}\sum_{i=1}^{|V|} t_i x_i \tag{1.1}$$

$$\text{s.t.} \quad \sum_{j=1}^{|V|} c_{1j} = R - 1 \tag{1.2}$$

$$\sum_{j=1}^{|V|} c_{ji} - \sum_{j\neq 1}^{|V|} c_{ij} = x_i, \quad i = 2,\cdots,|V| \tag{1.3}$$

$$c_{ij} \le (R-1)x_i, \quad i,j = 1,\cdots,|V| \tag{1.4}$$

$$x_i \in \{0,1\}, \qquad i = 1,2,\cdots,|V| \tag{1.5}$$

where x_i is a binary variable representing if node v_i is selected ($x_i = 1$) or not ($x_i = 0$) in the subgraph, c_{ij} are dummy variables (real numbers) representing the flow between selected nodes. $|V|$ is the total number of nodes in the graph. The whole constraints ensure the connectivity of the selected nodes, which is a major advantage of the method. Ideally, running MILP in every node in the graph from $R = 1$ to $R = |V|$, all possible connected subgraphs with the highest scores of all sizes were extracted, thereby finding the largest one with a significantly high level.

Dittrich et al. (2008) developed a similar approach to identify functional modules in PPI network. The novelty of the method is that it can quickly extract optimal subnetworks. Specifically, a new score based on signal-noise decomposition is proposed to weighted subnetworks. Then, the problem of finding maximum weighted subnetworks is transformed into the well known prize-collecting Steiner tree problem (PCST), where a mathematical programming based algorithm is employed to solve the PCST problem.

The two methods described above consider only the information of nodes in the background network. However, it is necessary to consider the information of both nodes and edges while extracting condition specific subnetwork from large background network. Wang et al. proposed a continuous optimization model to take into account both nodes and edges (Wang and Xia, 2008). In their work, identifying a responsive functional module is to find a subnetwork with both high overall weight of its nodes and high overall weight of its edges. To accomplish this goal, the problem is formulated as an optimization model. The model is described as follows:

$$\begin{aligned} \max_{\{x_i\}} \quad & \sum_{i=1}^{|V|}\sum_{j=1}^{|V|} w_{ij}x_i x_j + \lambda\sum_{i=1}^{|V|} f_i x_i \\ \text{s.t.} \quad & x_1^{\beta} + x_2^{\beta} + \cdots + x_{|V|}^{\beta} = 1 \\ & x_i \ge 0, \qquad i = 1,2,\cdots,|V| \end{aligned} \tag{2.1}$$

where w_{ij} is the weight of edge E_{ij}, f_i is a non-negative weight to quantify the strength of association between node i and the condition, and $x_i \in [0,1]$ is a variable denoting the degree to which the i-th node belongs to the condition specific network. To make sure that the final subnetwork is not too big, a regularization constraint that limit the number of nodes selected is introduced. In the constraint, parameter β is introduced to adjust the strength of regularization applied to the variable xi, and is usually taken as 1 or 2. These two terms are balanced by a parameter λ. Finally, a gradient discent method is adopted to solve the model. The method can extract a subnetwork from a larger network quickly due to the formulation of continuous model rather than discrete combina-torial optimization model.

Zhao et al. (2008b) proposed a novel integer linear programming (ILP) method for uncovering signal transduction networks from PPI network by integrating PPI with microarray data. Different from the heuristic methods, Zhao et al. extract a signal transduction network (STN) directly from PPI network rather than identifying possible linear paths and assembling them into a STN. Given an undirected weighted PPI network, starting node (e.g. membrane protein) and ending node (e.g. transcription factor), STN to be found is regarded as a compact connected subnetwork with maximum weights. The objective function of the model (ILP model) as well as constraints is described as follows:

$$\min_{\{x_i,y_{ij}\}} \quad S = -\sum_{i=1}^{|V|}\sum_{j=1}^{|V|} w_{ij}y_{ij} + \lambda\sum_{i=1}^{|V|}\sum_{j=1}^{|V|} y_{ij} \tag{3.1}$$

$$\text{s.t} \quad y_{ij} \le x_i \tag{3.2}$$

$$y_{ij} \le x_j \tag{3.3}$$

$$\sum_{j=1}^{|V|} y_{ij} \ge 1, \quad \text{if } i \text{ is a starting or ending protein} \tag{3.4}$$

$$\sum_{j=1}^{|V|} y_{ij} \ge 2x_i, \text{if } i \text{ is not a starting or ending protein} \tag{3.5}$$

$$x_i = 1, \quad \text{if } i \text{ is a protein known in STN} \tag{3.6}$$

$$x_i \in \{0,1\}, \qquad i = 1,2,\cdots,|V| \tag{3.7}$$

$$y_{ij} \in \{0,1\}, \qquad i,j = 1,2,\cdots,|V| \tag{3.7}$$

where w_{ij} is the weight of edge $E(i,j)$. x_i is a binary variable to denote whether protein i is selected as a component of the STN, and y_{ij} is also a binary variable to indicate whether the biochemical reaction represented by $E(i, j)$ is involved in STN. In the objective function, the first term aims to find a STN with

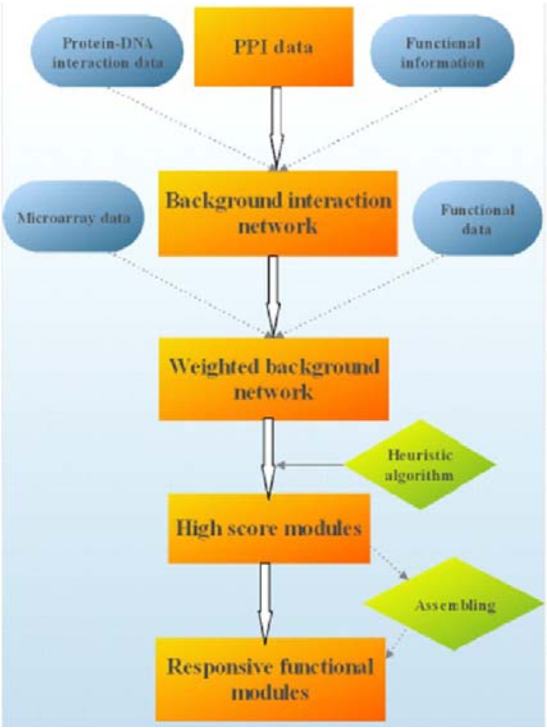
Protein-DNA interaction data
PPI data
Functional information
Background interaction network
Microarray data
Functional data
Weighted background network
Heuristic algorithm
High score modules
Assembling
Responsive functional modules

Fig. 1. The framework of heuristic approaches for extracting responsive functional modules. The dotted line denotes optional operation.

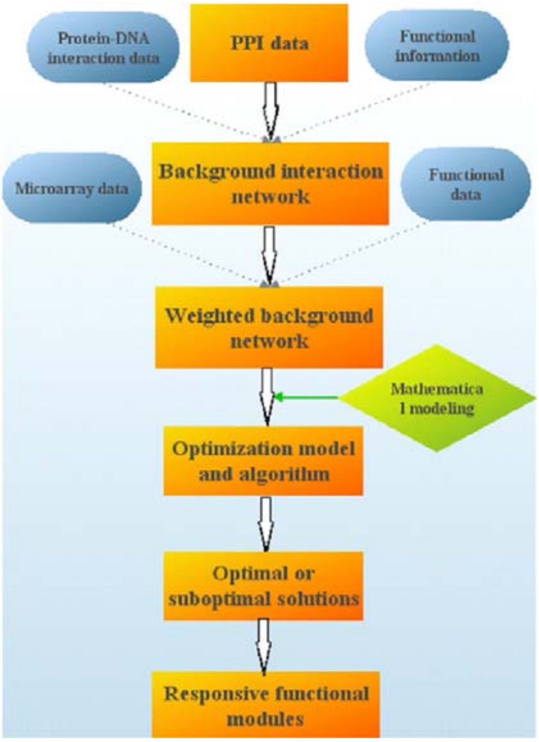
Protein-DNA interaction data
PPI data
Functional information
Microarray data
Background interaction network
Functional data
Weighted background network
Mathematical modeling
Optimization model and algorithm
Optimal or suboptimal solutions
Responsive functional modules

Fig. 2. The framework of exact approaches for extracting responsive functional modules. The dotted line denotes optional operation.

maximum weights while the second term is used to control the size of STN so that the obtained subnetwork is sparse. There exists a tradeoff between the two terms, which can be balanced by the parameter λ. $|V|$ is the total number of proteins in the PPI network. The constraint $\sum_{j=1}^{|V|} y_{ij} \geq 2x_i$ is to ensure that x_i has at least two linking edges once it is selected as a component of the STN, whereas the constraint $\sum_{j=1}^{|V|} y_{ij} \geq 1$ means that each starting protein or ending protein has at least one link to or from other proteins. These two constraints ensure that the components in the subnetwork are as connected as possible. The constraints $y_{ij} \leq x_i$ and $y_{ij} \leq x_j$ mean that if and only if proteins i and j are selected as the components of STN, the biochemical reaction denoted by the edge $E(i, j)$ should be considered. Equation $x_i = 1$ is the condition for any protein known involved in the STN, e.g. from the experiment results or literature. Finally, a relaxed linear programming (LP) algorithm is adopted to solve the optimization problem.

The ILP model works well in practice, but it is not easy to control the size of the identified subnetwork by manipulating parameter λ in some case. To cope with this problem, Zhao et al. proposed an improved integer linear programming model by introducing network flow to extract STNs from PPI network (Zhao et al., 2009), which is denoted as network flow model here. In the network flow model, it is guaranteed that a fixed number of components can be chosen while these components are ensured to be connected in the PPI network. Their idea behind the model is to find out a maximum weighted subnetwork of a specific size that accomplishes the signal transduction process with as few biochemical reactions as possible (i.e. parsimonious principle).

Figure 2 summarizes the work flow of the above mentioned exact approaches. It can be seen that the mission of identifying RFMs is an optimization problem of finding high score subnetworks in weighted background network once genes (proteins) and/or edges (interactions)' relevance to condition have been weighted. Therefore, the optimization problem is formulated as a different optimization model in a different exact approach and the corresponding optimization algorithm is applied or developed to find the optimal or suboptimal solutions. Ultimately, this optimality of the subnetwork is equated with subnetwork's relevance to condition under investigation in these exact approaches.

Since the problem of finding high score responsive functional modules from a large interaction network is NP-hard (Dittrich et al., 2008; Ideker et al., 2002) from the viewpoint of computational complexity, most computational methods for detecting responsive functional modules are heuristic approaches. Hence, the extracted functional modules by them are not necessarily the optimal ones and sometimes difficult to interpret the results. By contrast, exact approaches, such as mathematical programming based methods proposed in (Dittrich et al., 2008; Qiu et al., 2008; 2009; Wang and Xia, 2008; Zhao et al., 2008a; 2008b; 2009) generally identify maximum-score functional modules in reasonable time from real molecular networks for most of cases empirically. In addition, mathematical programming formulations can be interpreted as polyhedra in high-dimensional space (Dittrich et al., 2008). Theoretical analysis of these results often leads to new insights into understanding the original problem (Dittrich et al., 2008). However for a large background network, exact approaches may suffer from the dimensionality problem due to NP-hard nature of the formulation, whereas heuristic approaches are more efficient to derive an approximate solution.

CONCLUSIONS

Elucidating the mapping from a bio-molecular interaction network to a function is an active research field in systems biology in recent years (Chen et al., 2009). In this paper, we surveyed recent advances in developing computational methods to extract responsive functional modules from a PPI network. Actually, instead of RFMs or active molecular networks, the similar or same framework can be adopted to extract non-physical interacting networks, e.g. functional modules or functional subnetworks, i.e. constructing background functional network, weighting nodes and/or edges, scoring functional subnetworks, and searching high score functioanl subnetworks. Although many biological relevant functional modules have been extracted by these methods, there are still many obstacles that hamper their further application.

In all the surveyed methods, protein-protein interaction data are utilized to construct background network. It is known that the protein-protein interaction data are both noisy and incomplete. Some interactions actually do not exist under specific conditions, which makes the extracted putative connected subnetwork disconnected in fact. On the other hand, some truly biological relevant responsive functional modules are disconnected in the background network since some interactions have not been identified. As a result, these truly biological relevant RFMs may be missed by existing computational methods. One possible solution to the first problem is to filter the PPI data based on additional biological information as done by Bebek and Yang (2007) and Chu and Chen (2008). As for the second problem, integrating other information, e.g. Protein-DNA regulation and gene expression, with PPI data can alleviate the problem of the incompleteness of PPIs to some extent. With the further advance of high throughput biotechnologies, both the quality and amount of PPI data will be improved greatly, and thereby computational methods' performance is expected to be enhanced accordingly.

In many computational methods, it is almost a routine to overlay gene expression data onto the background PPI network to measure each gene (protein) or edge (interaction)'s relevance to conditions under investigation. However, gene expression is controlled by complex factors. Therefore, differential expression does not imply the dysregulation or relevance to disturbance necessarily. Furthermore, gene expression level cannot honestly reflect the true protein concentration. In other words, high gene co-expression does not necessarily mean a true interaction between their corresponding proteins. On the other hand, integrating gene expression data with other information, such as phenotype, can measure one gene's relevance to specific disturbance more accurately. With the progress of biotechniques that can measure protein abundance, such as mass spectrometry, the evaluation of one interaction (edge)'s relevance to disturbance is expected to be improved significantly and therefore the performance of computational methods for extracting responsive functional moodules will also be enhanced accordingly.

In addition, designing an appropriate score scheme for subnetworks of interest also plays a key role in developing computational methods. For example, a good score scheme should aggregate component's relevance to specific conditions as accurate as possible, and make a high score subnetwork more easier to find. Clearly, designing a good score scheme mainly requires biological insight on the problem of interest. On the other hand, with a given score scheme, a problem-specific algorithm can be developed to efficiently find optimal or suboptimal solutions stochastically or deterministically, where computation time is an important criterion due to NP-hard nature of the combinatorial problem. Therefore, provided that an appropriate score scheme is given based on the biological knowledge and insight, how to design an efficient algorithm to obtain optimal or suboptimal solutions within acceptable time becomes an important task from the computational viewpoint.

ACKNOWLEDGMENTS

Dr Zikai Wu was partly supported by the National Natural Science Foundation of China (NSFC) under Grant No. 60873129 and 10802043. Xingming Zhao was supported by the National High Technology Research and Development Program of China (2006AA02Z309), JSPS-NSFC collaboration project (107111-40116), and partly supported by Innovation Funding of Shanghai University (A.10-0112-08-407), and Luonan Chen is partly supported by Key Project of Shanghai Education Committee (B.10-0412-08-001). The authors would like to thank Dr. Yong Wang and Yuqing Qiu for helpful discussion.

REFERENCES

Adamcsek, B., Palla, G., Farkas, I., Derényi, I., and Vicsek, T. (2006). Cfinder: locating cliques and overlapping modules in biological networks. Bioinformatics *22*, 1021-1023.

Albert, R., DasGupta, B., Dondi, R., Kachalo, S., Sontag, E., Zelikovsky, A., and Westbrooks, K. (2007). A novel method for signal transduction network inference from indirect experimental evidence. J. Comput. Biol. *14*, 927-949.

Alon, N., Yuster, R., and Zwick, U. (1995). Color-coding. J. ACM. *42*, 844-856.

Arga, K., Önsan, Z., Kiidar, B., Ölgen, K., and Nielsen, J. (2007). Understanding signaling in yeast: insights from network analysis. Biotechnol. Bioeng. *97*, 1246-1258.

Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y., Müller, R., Meese, E., and Lenhof, H. (2007). GeneTrail-advanced gene set enrichment analysis. Nucleic Acids Res. *35*, W186-W192.

Bader, G., and Hogue, C. (2003). An automated method for finding molecular complexes in large protein interaction networks, BMC Bioinformatics *4*, 2.

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

Bebek, G., and Yang, J. (2007). Pathfinder: mining signal transduction pathway segments from protein-protein interaction networks. BMC Bioinformatics *8*, 335.

Bild, A., and Febbo, P. (2005). Application of a priori established gene sets to discover biologically important differential expression in microarray data. Proc. Natl. Acad. Sci. USA *102*, 15278-15279.

Cabusora, L., Sutton, E., Fulmer, A., and Forst, C. (2005). Differential network expression during drug and stress response. Biofinromatics *21*, 2898-2905.

Chen, X., Wang, L., Smith, J., and Zhang, P. (2008). Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. Bioinformatics *24*, 2474-2481.

Chen, L., Wang, R., and Zhang, X.S. (2009). Biomolecular networks: methods and applications in systems biology (New Jersey, USA: Wiley Interscience).

Cho, Y., Hwang, W., Ramanathan, M., and Zhang, A. (2007). Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinformatics *8*, 265.

Chu, W., and Ghahramani, Z. (2006). Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. Pacific Symposium on Biocomputing *11*, 231-242.

Chu, H., and Chen, B. (2008). Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. BMC Syst. Biol. *2*, 56.

Chuang, H., Lee, E., Liu, Y., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. Mol. Syst. Biol. *3*, 140.

Dittrich, M., Klau, G., Rosenwald, A., Dandekarand, T., and Muller,

T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics *24*, i223-i231.

Guo, Z., Li, Y., Gong, X., Yao, C., Ma, W., Wang, D., Li, Y., Zhu, J., Zhang, M., Yang, D., et al. (2007). Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. Bioinformatics *23*, 2121-2128.

Han, J., Bertin, N., Hao, T., Goldberg, D., Berriz, G., Zhang, L., Dupuy, D., Walhout, A., Cusick, M., Roth, F., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature *430*, 88-93.

Hirsh, E., and Sharan, R. (2006). Identification of conserved protein complexes based on a model of protein network evolution. Bioinformatics *23*, e170-e176.

Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics *24*, 2784-2785.

Huang, R., Wallqvist, A., and Covell, D. (2006). Targeting changes in cancer: assessing pathway stability by comparing pathway gene expression coherence levels in tumor and normal tissues. Mol. Cancer Ther. *5*, 2417-2427.

Hwang, W., Cho, Y., Zhang, A., and Ramanathan, M. (2006). A novel functional module detection algorithm for protein-protein interaction networks. Algorithms Mol. Biol. *1*, 24.

Ideker, T., and Sharan, R. (2008). Protein networks in disease. Genome Res. *18*, 644-652.

Ideker, T., Ozier, O., schwikowski, B., and Siegel, A. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics *18*, S233-S240.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. Genome Res. *12*, 37-46.

Kann, M. (2007). Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief. Bioinform. *8*, 333-346.

King, A., Pržulj, N., and Jurisica, I. (2004). Protein complex prediction via cost-based clustering. Bioinformatics *20*, 3013-3020.

Li, S., Assmann, S., and Albert, R. (2006). Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. PLoS Biol. *4*, e312.

Liu, Y., and Zhao, H. (2004). A computational approach for ordering signal transduction pathway components from genomics and proteomics data. BMC Bioinformatics *5*, 158.

Liu, M., Liberzon, A., Kong, S., Lai, W., Park, P., Kohane, I., and Kasif, S. (2007). Network-based analysis of affected biological processes in type 2 diabetes models. PLOS Genet. *3*, e96.

Mete, M., Tang, F., Xu, X., and Yuruk, N. (2008). A structural approach for finding functional modules from large biological networks. BMC Bioinformatics *9*, S19.

Murali, T., and Rivera, C. (2008). Network legos: buiding blocks of cellular wiring diagrams. J. Comput. Biol. *15*, 829-844.

Nacu, S., Critchley-Thorne, R., Lee, P., and Holmes, S. (2007). Gene expression network analysis and applications to immunology. Bioinformatics *23*, 850-858.

Nettleton, D., Recknor, J., and Reecy, J. (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. Bioinformatics *24*, 192-201.

Noisel, J., Sanguinetti, G., and Wright, P. (2008). Identifying differentially-expressed subnetworks with MMG. Bioinformatics *24*, 2792-2793.

Oron, A., Jiang, Z., and Gentleman, R. (2008). Gene set enrichment analysis using linear models and diagnostics. Bioinformatics *24*, 2586-2591.

Pereira-Leal, J., Enright, A., and Ouzounis, C. (2004). Detection of functional modules from protein interaction networks. Proteins *54*, 49-57.

Qi, Y., Balem, F., Faloutsos, C., Klein-Seetharaman, J., and Bar-Joseph, Z. (2008). Protein complex identification by supervised graph local clustering. Bioinformatics *24*, i250-i258.

Qiu, Y., and Zhang, S. (2008). Uncovering Differentially expressed Pathways with protein Interation and gene expression data. Lecture Notes in Operations Res. *9*, 74-82.

Qiu, Y., Zhang, S., Zhang, X-S., and Chen, L. (2009). Identifying differentially expressed pathways by high throughput data. IET Syst. Biol. (in press).

Rahnenfuhrer, J., Domingues, F., Maydt, J., and Lengauer, T. (2004). Calculating the statistical significance of changes in pathway activity from gene expression data. Stat. Appl. Gen. Mol. Biol. *3*, Article 16.

Rajagopalan, D., and Agarwal, P. (2005). Inferring pathways from gene lists using a literature-derived network of biological relationships. Bioinformatics *21*, 788-793.

Scholtens, D., Vidal, M., and Gentleman, R. (2005). Local modeling of global interactome networks. Bioinformatics *21*, 3548-3557.

Scott, M., Perkins, T., Bunnell, S., Pepin, F., Thomas, D., and Hallett, M. (2005). Identifying regulatory subnetworks for a set of genes. Mol. Cell. Proteomics *4*, 683-692.

Scott, J., Ideker, T., Karp, R., and Sharan, R. (2006). Efficient algorithms for detecting signaling pathways in protein interaction networks. J. Comput. Biol. *13*, 133-144.

Sharan, R., Ideker, T., Kelley, B.P., Shamir, R., and Karp, R.M. (2005). Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. J. Comput. Biol. *12*, 835-846.

Sohler, F., Hanisch, D., and Zimmer, R. (2004). New methods for joint analysis of biological networks and expression data. Bioinformatics *20*, 1517-1521.

Spirin, V., and Mirny, L. (2003). Protein complexes and functional modules in molecular networks. Proc. Natl Acad. Sci. USA *100*, 12123-12128.

Steffen, M., Petti, A., Aach, J., D'haeseleer, P., and Church, G. (2002). Automated modelling of signal transduction networks. BMC Bioinformatics *3*, 34.

Subramaniana, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545-15550.

Suderman, M., and Michael, H. (2007). Tools for visually exploring biological networks. Bioinformatics *23*, 2651-2659.

Turanalp, M., and Can, T. (2008). Discovering functional interaction patterns in protein-protein interaction networks. BMC Bioinformatics *9*, 276.

Ulitsky, I., Karp, M., and Shamir, R. (2008). Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. Lect. N. Bioinformat. (RECOMB2008) *4955*, 347-359.

Wang, Y., and Xia, Y. (2008). Condition specific subnetwork identification using an optimization model. Lecture Notes in Operations Res. *9*, 333-340.

Wang, R., Zhang, S., Zhang, X., and Chen, L. (2006). Identifying modules in complex networks by a graph-theoretical method and its application in protein interaction networks. Lect. N. Bioinformat. *4682*, 1090-1101.

Watts, D., and Atrogatz, S. (1998). Collective dynamics of 'small word' networks. Nature *393*, 440-442.

Zhang, S., Ning, X., and Zhang, X. (2006). Identification of functional modules in a PPI network by clique percolaion clusering. Comput. Biol. Chem. *30*, 445-451.

Zhang, S., Jin, G., Zhang, X., and Chen, L. (2007). Discovering functions and revealing mechanisms at molecular level from biological networks. Proteomics *7*, 2856-2869.

Zhao, X., Wang, R., Chen, L., and Aihara, K. (2008a). Automatic modeling of signal pathways from protein-protein interaction networks. In A., Brazma, S., Miyano, and T., Akutsu, eds., Proceedings of The 6th Asia Pacific Bioinformatics Conference, Vol. 6 of Serias on advances in bioinformatics and computational biology Imperial College Press, Singapore, 287-296.

Zhao, X., Wang, R., Chen, L., and Aihara, K. (2008b). Uncovering signal transduction networks from high-throughput data by integer linear programming. Nucleic Acids Res. *36*, e48.

Zhao, X., Wang, R., Chen, L., and Aihara, K. (2009). Automatic modeling of signaling pathways based on network flow model. J. Bioinformat. Computational Biol. (in press).