The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Condition specific subnetwork identification using an optimization model

Yong Wang[1,2] Yu Xia[1]

[1]Bioinformatics Program, Department of Chemistry, Boston University, Boston, MA 02215, USA
[2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China

Abstract Subnetworks can reveal the complex patterns of the whole-genome network by extracting the interactions that depend on temporal, spatial, or condition specific context. In this paper we present an optimization framework to identify condition specific subnetworks. This framework allows us to identify the most coherent subnetwork by integrating the information from both nodes and edges in the graph. Importantly we design an algorithm to solve the optimization problem efficiently. It is very fast and can extract subnetworks from large-scale network with about 10000 nodes. As a pilot study we apply our method to identify type 2 diabetes related subnetworks in the human protein-protein interaction network.

Keywords Condition specific subnetwork; Optimization model and algorithm; Diabetes

1 Introduction

The classical view of biology focuses on the functions of single biomolecules. Recently, a new and expanded view called network biology has emerged that emphasizes the interactions among biomolecules and the subsequent biomolecular networks [1]. The key idea of network biology is that the function of biomolecules can be understood by studying the interacting neighbors, and by examining the structure of the interaction network [2]. Usually graphs are used to represent these complex biological networks. On one hand, the global topological properties of a graph can reveal the wholegenome connectivity, robustness, modularity and hierarchical structure. On the other hand, the local patterns of interactions such as network motifs, complexes, pathways and functional modules, enable one to view the whole interactome as overlapping subnetworks, each associated with specific contexts or conditions.

Recently, several subnetwork identification methods have been developed to extract information from the global networks For example, network clustering algorithms are designed to identify protein complexes [3] pathways [4], or functional modules [5] from the protein-protein interaction network. Regulatory modules [6] or feed-forward/feedback motifs [7] are extracted from transcriptional regulatory network. Cross-species network alignment or comparison methods are used to reveal the evolutionarily conserved subnetworks [8]. Recent studies integrate protein-protein interactions and gene expression profiles to select subnetworks, which are then used as novel markers for prognosis of

metastasis formation [9] and type 2 diabetes [10]. Importantly, they showed that protein subnetwork markers outperform the predictors based on collections of non-interconnected genes for predicting breast cancer [11].

Though the concept of subnetwork is very important and extensively applied in different contexts, novel subnetwork identification methods that are flexible and efficient are still much needed. In this paper, we study the subnetwork identification problem from an optimization viewpoint. First, to facilitate date integration, we use a weighted graph to represent the protein network, where edge and node weights encode network and condition specific information. Second, we present an optimization framework to identify the most coherent subnetwork by simultaneously choosing a subset of condition specific nodes and as many interconnections among these nodes as possible.

2 Optimization model for subnetwork identification

Many biological networks can be represented as an undirected graph $G = (V, E)$ The n nodes of the graph G are biological molecules $V_1, V_2, \cdots, V_n$, and the set of edges $E = \{e_i, i = 1, 2, \cdots, m\}$ are the connectivity relationships among these nodes. Depending on the weights associated with edges, the graph G can be binary (nodes are either connected or not connected, for example in the protein-protein interaction network) or weighted (weights represent connectivity strength, for example in the protein functional linkage network). Without loss of generality, we use a symmetric weight matrix W to quantify the connectivity strength (for example, W can be the edge confidence scores for biomolecular interaction or functional linkage networks). Here we require all weights to be nonnegative $(W_{ij} \geq 0, i, j = 1, 2, \cdots, n)$. W_{ij}=0 if nodes i and j are not connected. In

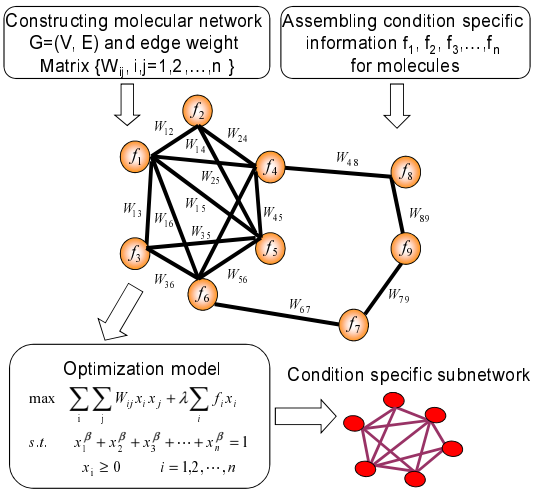

Figure 1: The scheme of our subnetwork identification method. Node and edge weighted graph is used to encode the biomolecular interaction (the weights on the edges) and condition specific information (the weights on the nodes). Optimization model is designed to extract the subnetwork which is both densely connected and condition specific.

addition to the biological network, there usually exist additional condition specific data that are useful for identify subnetworks. Usually they are phenotype data, gene expression data, functional annotation, or other biological context data. We add these information to the graph G by assuming that every node has a condition specific non-negative weight $(f_1, f_2, f_3, \cdots, f_n)$ quantifying the strength of association between the node and the specific condition under consideration. For example, the node weight can measure how much the gene is differentially expressed under a particular condition, or how strongly the gene is related to a particular disease.

In Figure 1, we show that the node and edge weighted graph representation can encode both biomolecular interactions and condition specific information. The main task is to identify the condition specific subnetwork by exploiting the structure of the graph and integrating weights associated with nodes and edges. Our basic assumptions are: 1) Nodes in the subnetwork should be condition specific. 2) These nodes should densely connect to one another within the subnetwork. 3) The selected subnetwork should have a proper size and cannot be too big or too small.

Based on the above assumptions, we need to simultaneously include as many condition specific nodes as possible, and maximize the interconnectivity of these nodes. Thus the optimization model for identifying condition specific subnetworks from graph G can be formulated as follows,

$$\begin{array}{ll} \max & \sum_{\mathrm{i}}\sum_{\mathrm{j}} W_{ij}x_i x_j + \lambda \sum_i f_i x_i \\ s.t. & x_1^\beta + x_2^\beta + x_3^\beta + \cdots + x_n^\beta = 1 \\ & x_{\mathrm{i}} \geq 0 \qquad i = 1, 2, \cdots, n \end{array} \tag{1}$$

Where the n-dimensional non-negative vector $x = (x_1, x_2, \cdots, x_n)$, determined by solving our optimization model, represent the degree to which each node belongs to the condition specific subnetwork. The variable x_i can be interpreted as whether the i-th node in graph G is included in the condition specific subnetwork. The first term in the objective function measures the interconnectivity within the subnetwork, while the second term measures the degree of association between the subnetwork nodes and the specific condition. Finally, we introduce a positive parameter λ to balance and integrate the above two terms This model, when unconstrained, has a trivial solution where all nodes from the original network are included in the condition specific subnetwork. To make sure that the final subnetwork is not too big, we introduce a regularization constraint that limit the number of nodes selected.

Parameter β is introduced in model (1) to adjust the strength of regularization applied to the variable $x = (x_1, x_2, \cdots, x_n)$. When $\beta = 2$ this is a trust region problem which optimizes a quadratic function subject to a ball constraint (L2-type constraint). It is very attractive in many cases since the optimization of a quadratic function over a sphere is polynomially solvable in contrast to general nonconvex programming [12] but tends to select all the nodes in the network to the final subnetwork. When $\beta = 1$, this L1-type constraint will lead to a sparse solution, i.e., many of the entries in the final optimal solution x will be zeros [13, 14]. Usually we use $\beta = 1$ in model (1) in order to extract small-sized subnetworks from a very large network

To better understand the above model, we consider two extreme cases for $\beta = 1$. If we only consider the second term in the objective function (node weights) model (1) is

simplified as the production (profit maximization) problem which is a linear programming problem. To the other extreme, if we only consider the first term of the objective function (edge weights), the problem is called standard quadratic optimization problem (QP) by finding (global) maximum of a quadratic form over the standard simplex. Several important problems can be cast into a standard QP in a straightforward way. As an example, a new continuous reformulation of the maximum weight clique problem in undirected graphs can be presented in this formulation [13]. Our model (1) lies somewhere between these two extreme cases by both considering the nodes weights and edge weights.

To roughly estimate the computational complexity of model (1), we can relate it to the well-known clique problem. If we focus on the first term of model (1), restrict the weight matrix W to be the adjacency matrix, and restrict the variable x to only take binary values (0 or a positive constant), model (1) can be used to find the maximum clique in an unweighted graph. Consideration of edge weights generalizes the concept of cliques to weighted graphs. Both the maximum cardinality and the maximum weight clique problems are NP-hard. These problems are extensively studied and many practically efficient heuristic algorithms are developed using combinatorial optimization techniques, such as sequential greedy heuristics, local search heuristics, simulated annealing methods, neural networks, genetic algorithms, and tabu search.

Biomolecular networks are often large in scale. For example in yeast the protein-protein interaction network is estimated to have about 6,000 nodes and 50,000 interactions. It is impossible to solve such a large combinatorial optimization problem exactly in reasonable time. To address this issue, we formulate the subnetwork identification problem as a continuous (nonconvex) optimization problem, as described in model (1). The approximation of the discrete combinatorial problem by a continuous optimization problem is based on the theorem due to Motzkin and Straus [15] which relates maximal cliques of an unweighted undirected graph to the optimization of a quadratic function.

The Lagrange function of optimization model (1) is:

$$L = -\sum_{\mathrm{i}}\sum_{\mathrm{j}} W_{ij}x_i x_j - \lambda\sum_i f_i x_i + \alpha(\quad x_1^\beta + x_2^\beta + x_3^\beta + \cdots + x_n^\beta - 1) - \sum_i \mu_i x_i$$

Then the KKT condition is:

$$\begin{array}{ll}
\frac{\partial L}{\partial x_i} = 0 \Rightarrow \mu_i = -2(WX)_i - \lambda f_i + \alpha\beta x_i^{\beta-1} & i = 1,2,\cdots,n \\
\mu_i x_i = 0 & i = 1,2,\cdots,n \\
x_{\mathrm{i}} \geq 0, \qquad \mu_i \geq 0 & i = 1,2,\cdots,n \\
x_1^\beta + x_2^\beta + x_3^\beta + \cdots + x_n^\beta = 1 &
\end{array}$$

The Lagrange factor α can be easily solved as

$$\alpha = (2X^T WX + \lambda\sum_i f_i x_i) \Big/ \beta$$

Then we can use the following iterative algorithm to quickly find a local minimum from a predetermined initial solution:

$$x_i^{t+1} = (x_i^t \frac{2(WX)_i + \lambda f_i)}{\alpha\beta})^{\frac{1}{\beta}} = (x_i^t \frac{2(WX)_i + \lambda f_i)}{2X^T WX + \lambda\sum_i f_i x_i})^{\frac{1}{\beta}}$$

It can be proven using a strategy similar to that in [15] that the algorithm is convergent. Furthermore the convergent solution satisfies the constraints and the KKT condition. Finally the non-zero entries in solution x (determined in practice as entries that are greater than a cutoff) define the final subnetwork (Motzkin-Straus theorem [13]).

3 Pilot Study on type 2 diabetes related subnetwork

Type 2 diabetes (T2D) mellitus is a complex disease with profound impact on health and longevity [10]. It is estimated to affect more than 150 million people worldwide by the World Health Organization statistics. The symptom of T2D is that the body is unable to respond appropriately to insulin produced by the pancreas. T2D is defined by elevations in plasma glucose levels (hyperglycemia). At the same time, it encompasses a variety of metabolic abnormalities, including reduced responsiveness to insulin (insulin resistance) in key insulintargeted tissues such as muscle, adipose tissue, liver, kidney and brain; abnormal accumulation of lipids in non-adipose tissue, and abnormal pancreatic beta-cell function leading to insufficient insulin secretion [16].

In this pilot study T2D related subnetwork is reconstructed by the integration of protein-protein interaction network and T2D candidate gene information. The basic assumption is that each protein in the protein-protein interaction network is labeled with a confidence score that measures its degree of association with the specific phenotype of T2D. From these information we can find T2D related subnetwork which can in turn be used as novel biomarkers for diagnosis [11].

Our general procedure for identifying condition specific subnetworks consists of three steps: (1) Collect a set of genes associated with a specific condition, disease or function. (2) Assemble the cellular protein-protein interaction network. (3) Apply our optimization method to extract the condition specific subnetworks.

3.1 Collection of type 2 diabetes candidate genes

Currently many methods have been developed to identify T2D candidate genes by integrating data from phenotype, sequence, expression and annotation [16]. In [16] the authors summarize seven recent computational methods to identify T2D candidate genes and give a unified score to integrate different approaches In total there are 2503 genes related to T2D and each gene is assigned a confidence score which is defined as the number of methods that select this gene to be T2D candidate gene (range from 1 to 7, the higher the number, the more confident for its association with T2D) [16].

3.2 Assembly of human protein-protein interaction network

The protein-protein interaction data in human are downloaded from BioGRID (version 2.0.41)[17]. In total there are 7,903 proteins and 44,422 interactions. This interaction network is very sparse and the percentage of protein pairs that interact is only 0.14%. Here, we make the network denser by extending the definition of interaction to include not only protein pairs that directly interact, but also protein pairs that indirectly interact through a common neighbor. We assign a weight to every interacting pair that measures the strength of their interaction. In this way, we get a weighted protein-protein interaction network with 724,144 edges (23% of all protein pairs, a 16-fold increase in network size). Compared with other methods to make the network denser, for example the shortest path and diffusion kernel methods, our strategy is simple, robust, and flexible.

3.3 Extracting type 2 diabetes related subnetwork

We apply our subnetwork identification method to integrate the human protein-protein interaction network and T2D candidate gene list. In our computation, we choose the parameter$\beta = 1$ to perform the L1 regularization in our model. Parameter λ is chosen to be 0.01, which makes the sum of node weights and sum of edge weights roughly equal. Then we use model (1) and the approximate algorithm previously described to identify disease related subnetworks. After one locally optimal solution is obtained, the subnetwork corresponding to non-zero entries in the solution vector is extracted; these nodes are eliminated from the network, and the whole procedure is then iterated, i.e., we solve for another locally optimal solution and its corresponding subnetwork based on the new network. In total we select four locally optimal subnetworks as shown in Figure 2. Our method is able to pick out the relatively dense substructure in the protein-protein interaction network and most of the chosen nodes have high-confidence association with T2D (Figure 2). However we do find some non-T2D candidate genes in the selected subnetwork. This fact suggests that disease related subnetwork revealed by our method is more than just selecting candidate gene sets without considering network information.

3.4 Assessing the effectiveness of our method

We validate the extracted subnetworks related to T2D mainly by their GO function annotation information We found these four selected subnetworks in Figure 2 are closely related to insulin-degradation, signal transduction, and metabolism functions. In the subnetwork (a) of Figure 2, the protein IDE is an insulin-degrading enzyme and has such functions as signal transduction, cell-cell signaling, insulysin activity, and metalloendopeptidase activity. It closely interacts with 9 other proteins. Among them MAPK3 and PACSIN1 possess extracellular signal-regulated kinase and kinase activity respectively. In the subnetwork (b) of Figure 2, C1QR1 is a C1q receptor and IL18R1 is an IL1 receptor related protein. These membrane-bound receptors are related to blood coagulation and antimicrobial humoral response. In the subnetwork (c) of Figure 2, the protein PRSS25 is an HtrA like serine protease and has such functions as induction of apoptosis by intracellular signals, regulation of multi-cellular organism growth, and forebrain development. It interacts with 4 other proteins and shares the cell organization and biogenesis function with GEMIN5 and ELN. In the subnetwork (d) of Figure 2, the protein PSEN1 performs both signal transduction and metabolism function. It interacts with three proteins with unknown function and works together with protein degradation protein CASP10, pest, pathogen or parasite responsive protein IFI27, and metabolism related protein ZBTB16.

4 Discussion and conclusion

In our formulation, we use the undirected graph to represent the biological network. Our optimization model can be easily extended to directed graphs. For example the transcriptional regulatory network is a directed graph in which transcription factors regulate the expression of target genes. Our method can be easily extended to identify condition specific subnetworks in these directed networks.

As a pilot example, we apply our method to identify type 2 diabetes related subnetworks. Our optimization model is very general and has many potential applications due to the importance of subnetworks. Possible applications include: 1. Biological context

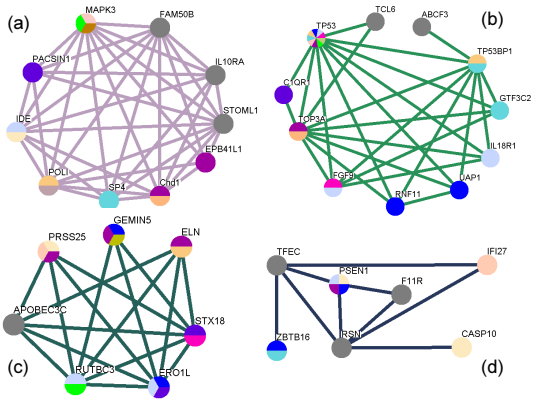

Figure 2: Four selected type 2 diabetes related subnetworks by our method. The node is labeled by its gene name and colored by its GO function (unknown function is colored by grey).

or disease related subnetwork identification by extensively integrate heterogeneous data sources (protein-protein interaction, transcriptional regulatory, metabolic and functional linkage network, gene expression data, mass spectrum data, linkage disequilibrium/SNP data, GO function, and evolution). 2. Condition specific bicluster structure identification (for example subnetworks in protein-small molecule interaction network or transcriptional regulatory network). 3. Network alignment (identification of subnetworks in the global alignment graph which is formed by combination of two networks [8]). 4. Protein local structure alignment (a protein is viewed as an amino acid interaction network and the evolutionary or chemical information of the amino acids are considered, then further alignment of two proteins is achieved by aligning two amino acids networks).

In conclusion, we proposed an optimization model to identify subnetworks by integrating biological network and condition specific information. As a pilot study, we apply our method to identify type 2 diabetes related subnetworks. There are two challenges in human disease subnetwork identification. First, the present protein-protein interaction network in human is noisy and far from complete. Second, our basic assumption is that subnetworks are better biomarkers than single proteins, which needs further experimental and clinical verification especially for complex diseases such as T2D. Further research directions include validation of the effectiveness of subnetwork biomarkers, and improvement of the subnetwork identification algorithm.

Acknowledges

YW is supported by the Grant No. 2006CB503905 from the Ministry of Science and Technology, China and the Grant No. 10801131 from the National Natural Science Foundation of China. YX is supported by a Research Starter Grant in Informatics from the PhRMA Foundation. The authors would like to thank Prof. Xiang-Sun Zhang for helpful suggestions and the revision for the manuscript.

References

[1] Eisenberg, D., Marcotte, E.M., Xenarios, I., Yeates, T.O.: Protein function in the post-genomic era. Nature **405** (2000) 823-826

[2] Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nature Reviews Genetics **5** (2004) 101-113

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

[4] Zhao, X.M., Wang, R.S., Chen, L., Aihara, K.: Uncovering signal transduction networks from high-throughput data by integer linear programming. Nucleic Acids Research **36** (2008) e48

[5] Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A.: Detection of functional modules from protein interaction networks. Proteins **54** (2001) 49 - 57

[6] Segal, E., Shapira, M., Regev, A., Peer, D., Botstein, D., Koller, D., Friedman, N.: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet **34** (2003) 166-176

[7] Alon, U.: An Introduction to Systems Biology: Design Principles of Biological Circuits. Chapman & Hall/CRC (2007)

[8] Li, Z., Zhang, S., Wang, Y., Zhang, X.S., Chen, L.: Alignment of molecular networks by integer quadratic programming. Bioinformatics **23** (2007) 1631

[9] Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., Ideker, T.: Network-based classification of breast cancer metastasis. Molecular Systems Biology **3** (2007)

[10] Liu, M., Liberzon, A., Kong, S.W., Lai, W.R., Park, P.J., Kerr, K.: Network-based analysis of affected biological processes in type 2 diabetes models. PLoS Genet **3** (2007) e96

[11] Auffray, C.: Protein subnetwork markers improve prediction of cancer outcome. Molecular Systems Biology **3** (2007)

[12] Ye, Y.: A new complexity result on minimization of a quadratic function with a sphere constraint. Princeton Series In Computer Science (1992) 19-31

[13] Motzkin, T.S., Straus, E.G.: Maxima for graphs and a new proof of a theorem of Turan. Canad. J. Math **17**(1965) 533-540

[14] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B **58** (1996) 267-288

[15] Ding, C., He, X., Xiong, H., Peng, H.: Transitive closure and metric inequality of weighted graphs: detecting protein interaction modules using cliques. International Journal of Data Mining and Bioinformatics **1** (2006) 162-177

[16] Tiffin, N., Adie, E., Turner, F., Brunner, H.G., van Driel, M.A., Oti, M., Lopez-Bigas, N., Ouzounis, C., Perez-Iratxeta, C., Andrade-Navarro, M.A.: Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Research **34** (2006) 3067

[17] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucl. Acids Res. **34** (2006) D535-539