The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008
Copyright © 2008 ORSC & APORC, pp. 341–349

Revealing Disease Related Interactions by Correlation Analysis

Zi-Kai Wu[1,2] Zhi-Yong Zhang[1,2] Lv-Wen Zhang[1,3]
Katsuhisa Horimoto[4]

[1]Institute of Systems Biology, Shanghai University, Shanghai 200444
[2]School of Communication and Information Engineering, Shanghai University, Shanghai 200444
[3]School of Computer Engineering and Science, Shanghai University, Shanghai 200444
[4]Computational Biology Research Center, National Institute of Advanced
Industrial Science and Technology 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan

Abstract The computational identification of disease related lesions is still a key open problem in biomedicine and systems biology. Dysregulated interactions may be an important reason that causes disease. In this paper, we aim to identify dysregulated interactions so as to elucidate the mechanism of disease in a systematic manner. Specially, we present a method to detect which protein-protein interactions or genetic interactions are downregulated or upregulated due to disease process. The proposed method was applied to a human molecular interaction network and a prostate cancer microarray dataset to reveal dysregulated interactions. The enrichment analysis of cancerous genes and disease related GO terms in identified dysregulated interactions shows that the identified dysregulated interactions are disease related, which verifies the effectiveness of our method.

Keywords Network; Disease; Interaction; Correlation

1 Introduction

Life is a complex phenomenon, which cannot be clearly understood by merely studying individual components of cells. It is the interactions of those components or networks that ultimately hold responsibility of living organisms' forms and functions. Due to the recent rapid progress on biomedical science, the fundamental mechanisms on many diseases have be revealed at molecular level. For example, it has be elucidated that many cancers originate from some mutations on certain genes caused by chance or experimental factor because these mutations trigger downstream effect to the cellular system, i.e. on genes, proteins, partial pathway or entire pathway [1]. From the viewpoint of network biology, a disease can be viewed as a perturbation to the cellular system or biomolecular interaction network. In other words, the cellular system under disease state is a disturbed system which is rewired from the original undisturbed system (or control state) accordingly. As disease is considered to perturb the cellular system from the aspect of node and edge (connectivity), computational method of identifying disease related lesions can be grouped into two classes naturally, i.e. node-centric method and edge-centric method.

At present, computational identification methods are mainly node-centric. Take cancer research as an example. Until now, a number of methods have been proposed to identify cancer related genes. At earlier stage, most of these methods were based on differential expression analysis. In other words, the aberrantly expressed genes are identified as cancer related lesions. Although partical success has made in identifying cancer related genes, these methods are unable either to infer any details on how a protein's behavior has changed or to reveal what specific mechanisms lead to the pathologic transition [2].

To overcome this drawback, some gene-centric identification methods have been developed to embed themself in the context of cellular network. These methods utilize the known disease relatedness of other nodes in the cellular network to infer some node's disease relatedness. The rationale is that if some neighbors (direct or indirect neighbors) of a gene are disease related, then the gene can also be inferred to be disease related with certain confidence [3]. With such a scheme, Kartik M Mani et al. proposed a novel identification method [2]. Their analysis method works in two steps. That is, this method first identifies dysregulated interactions (interactions showing either a gain of correlation or a loss of correlation pattern) in the phenotype of interest, and then ranks genes according to the statistical significance of dysregulated interaction enrichment among the interactions in which they directly participate [2]. This method's rationale is that if a node or gene's relation with most of their neighbors are changed under the disease state, then it can be inferred with high confidence that the gene itself is arch-criminal and disease related.

Some other gene-centric identification methods aim at the entire pathway or a prior defined gene set[4,5,6,7]. Pathway-based methods use a metric to measure the cohesiveness level of the members of the pathway and represent the tightness of relation between its members. Their rationale is that if the cohesiveness level is descended or elevated under disease state, the pathway can be viewed as a disrupted or newly constructed subsystem under disease state [5]. Gene set-based methods first use some metric to measure the differential expression level of each gene and then a ranked list of differentially expressed gene is obtained. Enrichment analysis of differentially expressed gene in the prior defined gene set is conducted to find which gene set's overall differential expression level is statistically significant [6, 7].

At present, there are also some edge-centric computational identification methods, for which the key is how to define the edges between nodes in the network and how to capture the differential behavior of the edge. Essentially, the definition of edge depends on the data at hand. High-throughput technologies are now producing vast amounts of biological data representing the availability of specific molecular species in a cellular population[2]. These include, among many others, gene expression and genotypic profiles [8], DNA-binding profiles[9], genomic sequences, and protein abundance from mass spectrometry [10]. At the same time, another high-throughput experiments have populated the public databases with thousands of protein-protein interaction (PPI) data and genetic interaction data[11].

Some researchers use the gene co-expression to define the edge between genes [12,13 ,14,15,16]: if two genes's mRNA expression levels are highly correlated under certain condition, then it can say that there is a functional association between two genes, in other words, there exists an edge between two genes. Jung-Kyoon Choi et al.[12] constructed a normal and disease coexpression network respectively based on 10 cancer microarray datasets and 10 their normal counterparts, and then identified the differential coexpression in the network. There are also some other methods based on differential co-expression analysis that was proposed to identify disease related lesions[13,14,15,16].

Other research works use the physical PPI and genetic interaction to identify disease related edges. One weakness of the high-throughput PPI data and genetic interaction data is that it contains no information about the conditions under which the interactions may take place[17]. Under the hypothesis that higher expression correlation of the genes implies genuine interactions of the proteins under the investigated conditions, it is a popular way to use the gene expression information to measure the 'activity 'of an interaction in response to the investigated condition. Zheng Guo et al.[17] scored the edge in PPI network based on the correlation coefficient of two genes's expression levels and the deferential expression of two genes, and then used simulated annealing algorithm to find a statistically significant responsive subnetwork.

On the other hand, protein-protein interaction have recently been recognized as challenging but attractive targets for small chemical drugs[18]. Furthermore, recent research works suggest that PPI inhibition could lead treatments for some human disease[18-23]. Motivated by both the potential pharmaceutic and therapeutic applications of disease related interactions and sparseness of computational methods for identifying disease related PPI or genetic interactions, we propose a new method to identify dysrgulated interactions by exploiting the mechanism of diseases in this paper. Specially, we present a method to detect which protein-protein interactions or genetic interactions are downregulated or upregulated during disease process.

The remainder paper is organized as follows. Firstly, we describe the details of our method as well as the data set we used. Secondly, the results are presented through numerical tests on prostate cancer case. Finally, the features for the new method of identifying disease related interaction are discussed, and a brief conclusion and directions of further research works are presented in the last section.

2 Methods and materials

2.1 Dataset and data processing

The protein-protein interaction and genetic interaction data was first derived from the BIOGRID database(2008, 2.0.36 version). Then the self-interactions and reduplicate interactions were removed from the dataset. Finally, we have 23791 interactions in the interaction data set, which constitute a protein interaction network.

The prostate microarray data set [24] consists of about 7641 genes measured in 71 prostate tumors as well as 41 normal prostate specimens. In the microarray dataset, if there are multiple probes that correspond to the same gene, we choose the one that contains the least amount of missing values. Then, we only retain genes with missing data smaller than one third of the total sample size. Finally, we convert all values $<=$ 10 to 10, and then perform a base 2 log transform. The prostate cancer related genes were obtained from Prostate Gene Database (PGDB)[25].

2.2 Estimation of pairwise gene co-expression

In this paper, the Percentage Bend Correlation [26] with $\beta = 0.1$ is applied to obtain a robust correlation estimate. Percentage Bend Correlation is first adopted to detect outliers in expression values of each gene so as to reduce the effects of those outliers in the correlation calculation[15]. Since the Percentage Bend Correlation may have some bias due to sample size, Fisher's z-transform [27] is also performed to reduce sample size effect, which can be formulated as

$$Z = \frac{\sqrt{n-3}}{2} \times \log \sqrt{\frac{1+r}{1-r}} \tag{1}$$

where r and n denote correlation estimate and sample size respectively, while Z corresponds to the Fisher's Z scores. Z score divided by its theoretical standard deviation theoretically has an asymptotically standard normal distribution. However, Min Xu et al. observed that the distributions of the z-score are still different from dataset to dataset [15]. Hence, we further normalize z-scores to enforce the standard normal distribution. After that, standardized correlations r' are obtained by inverting the z-score with a fixed n of 30 as Min Xu did.

2.3 Active interactions under certain condition

We give different definition of active interaction with respect to physical protein-protein interaction and genetic interaction. Suppose a physical protein-protein interaction connects gene A and gene B in cellular interaction network. We define the interaction to be active under normal state if the expression correlation of gene A and gene B in normal data set is higher than some threshold (in this paper, the threshold is set to be 0.20). Otherwise, the physical interaction between A and B are defined as inactive. For genetic interaction, we define it to be active under normal state if the absolute value of its two genes's expression correlation is higher than some threshold. Otherwise, the genetic interaction between A and B are defined as inactive. Similarly, we can define how an interaction is active or inactive under disease state.

2.4 Downregulated and upregulated interactions under disease state

We define an interaction to be upregulated if it is inactive in normal state but active under disease state. We define an interaction to be downregulated if it is active in normal state but inactive under disease state.

2.5 Enrichment analysis

The GO term enrichment analysis is done by the hypergeometric test on genes involved in downregulated interactions and upregulated interactions respectively through submitting them to DAVID online webserver(http://david.abcc.ncifcrf.gov /home.jsp). The prostate cancer and cancer related gene enrichment analysis are also done by the hypergeometric test.

Finally, the whole procedure of the method is summarized as Figure 1.

3 Results and discussion

Under the different thresholds, there are different numbers of interactions being active under normal state or disease state. In this paper, we present the result obtained when setting threshold being 0.20.

Under the threshold of 0.20, there are 1289 interactions that are active under normal state, while there are 1310 interactions that are active under disease state. Accordingly,

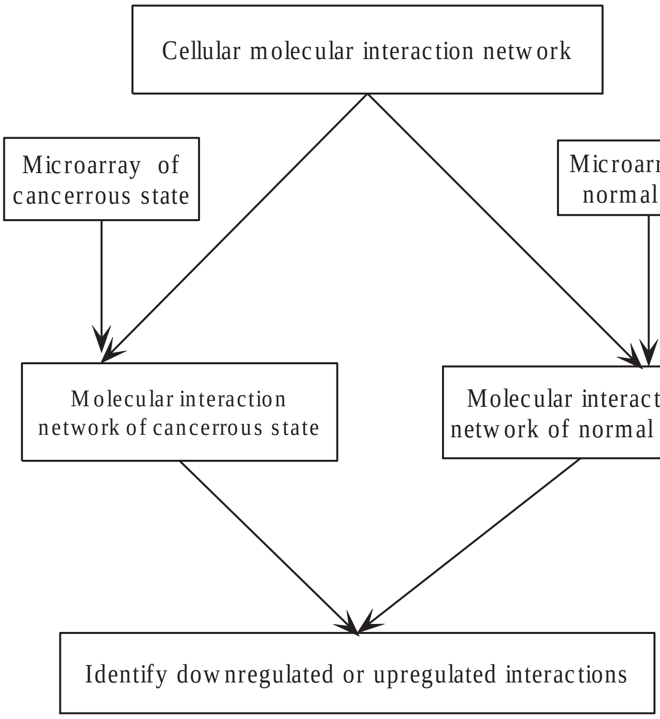
Cellular molecular interaction network
Microarray of cancerrous state
Microarray of normal state
Molecular interaction network of cancerrous state
Molecular interaction network of normal state
Identify downregulated or upregulated interactions

Figure 1: Flowchat of the proposed method

there are 213 interactions that are upregulated and 228 interactions that are downregulated. To evaluate the biological relevance of this identified dysregulated interactions, we perform some enrichment analysis. Firstly, the identified dysregulated interactions involve many genes. If these genes are cancer related, then we can infer that these interactions are also cancer related to some extent. There are 327 genes involved in upregulated interactions, of which 17 genes are known cancer related. There are 337 genes involved in downregulated interactions, of which 18 genes are known cancer related. Furthermore, there are 8042 genes involved in the interaction network. The known 118 cancer related genes that are included in these 8042 genes are used as background. We performed enrichment analysis on genes involved in downregulated and upregulated interactions respectively. The p-value of enrichment analysis is $4.9685e-006$ and $1.7217e-006$ respectively. The small p value shows that the enrichment of cancer related genes on the identified dysregulated interactions is statistically significant, and the identified dysregulated interactions are biological relevant and cancer related.

To further verify its biological relevance and cancer relatedness, we also performed the enrichment analysis of GO terms on the identified dysregulated interactions. There are many GO terms that are enriched. In this paper, we only present GO terms belonging to biological process category for the sake of simplicity. Some representative GO terms are listed in Tables 1 and 2 respectively. Enriched GO terms on downregulated inter-

Table 1: Representative enriched GO terms on downregulated interactions

GO term	P-value
Regulation of transcription	1.34E-10
Cell differentiation	2.77E-10
Programmed cell death	1.11E-10
Apoptotic program	0.0017
Cell proliferation	3.04E-7
Cell death	2.005E-10
Intracellular receptor-mediated signaling pathway	1.707E-6
RNA biosynthetic process	1.45E-10

actions include regulation of transcription, cell differentiation, programmed cell death, apoptotic program, cell proliferation, cell death, intracellular receptor-mediated signaling pathway, RNA biosynthetic process, which are all well known cancer related GO terms. Enriched GO terms on upregulated interactions include intracellular signaling cascade, negative regulation of metabolic process, regulation of transcription, cell differentiation, programmed cell death, apoptotic program, cell proliferation, cell death, intracellular receptor-mediated signaling pathway, and RNA biosynthetic process. It can be seen that most of these enriched terms were identified with small p-value. In summary, the enrichment of cancer related GO terms further verifies the biological relevance and cancer relatedness of our identified dysregulated interactions.

However, enrichment of cancer related genes and GO terms are just indirect evidence for the cancer relatedness of the identified dysregulated interactions. Finding direct evidence supporting cancer relatedness of dysregulated interactions is a challenging but important work.

Note that the method proposed in this paper is similar to the method presented in [15]. Next, we outline the main differences between the proposed method and the existing methods below.

(1). We use Percentage Bend Correlation to measure correlation, while mutual information was applied in [15].

(2). Method in [15] needs large background population to measure activity of interactions, while only counterpart samples of tissue samples are needed in our method.

(3). Difference between correlation of two genes in background population and tissue samples are used to define gain or loss of interactions in the existing methods. On the other hand, in our method, each interaction is classified as active or inactive under some condition, and thereby the downregulated or upregulated interactions are defined.

(4). The most important difference is that the goal of the research in [15] is to find disease related genes or perturbed target. However, our work aims to directly at dysregulated interactions. In other words, our goal is to exploit the impact of dysregulated

Table 2: Representative enriched GO terms on upregulated interactions

GO term	P-value
Intracellular signaling cascade	1.89E-13
Negative regulation of metabolic process	3.91E-7
Positive regulation of transcription	8.84E-10
Cell differentiation	3.30E-9
Programmed cell death	1.71E-8
Apoptotic program	2.85E-4
Regulation of cell proliferation	5.11E-6
Cell death	7.58E-8
Intracellular receptor-mediated signaling pathway	0.005
RNA biosynthetic process	0.0035

interactions on cellular function and their relation to disease.

4 Conclusion and future work

The computational identification of disease related lesions is still a key open problem in biomedicine and systems biology. In this paper, we proposed a new method to exploit the mechanism of disease by identifying dysregulated interactions. Specially, we present a method to detect which protein-protein interactions or genetic interactions are downregulated or upregulated. Experiment on a prostate cancer case shows that the identified dysregulated interactions are disease related, which confirms the effectiveness of our method.

However, our method indirectly verifies disease relatedness of the dysregulated interactions. This is still far away from our ultimate goal that elucidates the role that the dysregulated interactions play in disease. To reach this goal, some further work should be made in the future:

(1). Find direct evidence that can demonstrate cancer relatedness of dysregulated interactions.

(2). Identify the relation between downregulated interactions and upregulated interactions. For instance, we want to know if or not there exists some switch-like behavior from this study. We also want to know which cellular function or process is disturbed and which is newly emerged with the deletion of some old interactions and the addition of inclusion of new interactions.

(3). Integrate the methodology of pathway detection with our method. Now the computational identification of protein-protein target is mainly based on structural properties. We can exploit those techniques to provide a primary candidate list of protein-protein targets.

Acknowledgment

The authors would like to thank Dr Yong Wang, Xingming Zhao and Ruisheng Wang, Prof.Zengrong Liu and Prof. Luonan Chen for helpful discussions and suggestions. The authors would like to thank Dr Min Xu (Program in Molecular and Computational Biology, University of Southern California, Los Angeles, CA, USA) for providing the microarray data sets.

References

[1] Bert Volgelstein, Kenneth W Kinzler: Cancer genes and the pathways they control. NATURE MEDICINE 10(8) (2004) 789–799.

[2] Katik M Mani, Celine Lefebvre, Kai wang, Wei Keat Lim, Katia Basso, Riccardo Dalla-favera, Andrea Califano: A systems biology approach to prediction of oncogenes and molecular perturbation tragets in B-cell lymphomas. Molecular Systems Biology 4:169 (2008) doi:10.1038/msb.2008.2.

[3] Ramon aragues, Chris Sander, Baldo Oliva: Predicting cancer involvement of genes from hetrogeneous data. BMC Bioinformatics 9:172 (2008) doi:10.1186/1471-2105-9-172.

[4] Igor Ulitsky, Richard M Karp, Ron Shamir: Detecting Disease-Specific Dysregulated Pathways Via Analysis of Clinical Expression Profiles. Lecture Notes in Computer Science(RECOMB2008) (2008) 347-359.

[5] Ruili Huang, Anders Wallqvist, David G Covell: Targeting changes in cancer: assessing pathway stability by comparing pathway gene expression coherence levels in tumor and normal tissues. Molecular Cancer Therapeutics, 5(9) (2006) 2417–2427.

[6] Aravind Subramaniana, Pablo Tamayoa, Vamsi K Moothaa, Sayan Mukherjeed, Benjamin L Eberta, Michael A Gillettea, Amanda Paulovichg, Scott L Pomeroyh, Todd R Goluba, Eric S Landera, Jill P Mesirova: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy Science(USA), 102(43) (2005) 15545–15550.

[7] Trey Ideker, Owen Ozier, Benno Schwikowski, Andrew F Siegel: Discovering regulatory and signaling circuits in molecular interaction networks. Bioinformatics, 18(Suppl.1) (2002) S233–S240.

[8] M schena, D Shalon, R W Davis, P O Brown: Quatitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270 (2005) 467–470.

[9] B Ren, F Robert, J J Wyrick, O Aparicio, E G Jennings, I Simon, J Zeitlinger, J Schreiber, N Hannett, E Kanin, T L Volkert, C J wilson, S P Bell, R A Yong: Genome-wide location and function of DNA binding proteins. Science, 290 (2000) 2306–2309.

[10] O D Perez, G P Nolan: Simutaneous measurement of multiple active kinase states using polychromatic flow cytometry. Nature Biotechnology, 20 (2002) 155–162.

[11] P Uetz: A comprehensive analysis of protein-protein interactions in Sacharomyces cerevisiae. Nature, 403 (2000) 623–627.

[12] Jung Kyoon Choi, Ungsik Yu, Ook Joon Yoo, Sangsoo Kim: Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics, 21(24) (2005) 4348-4355.

[13] Kerby Shedden, Jeremy Taylor: Differential Correlation Detects Complex Associations Between Gene Expression and Clinical Outcomes in Lung Adenocarcinomas. Methods of Microarray Data Analysis, Springer-Verlag, Heidelberg (2005) 121–131.

[14] Yinglei Lai, Baolin Wu, Liang Chen, Hongyu Zhao: A statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics, 20(17) (2004) 3146–3155.

[15] Min Xu, Ming-Chih J Kao, Juan Nunez-Iglesias, Joseph R Nevins, Mike West, Xianghong Jasmine Zhou: An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. BMC Genomics, 9(Suppl 1):S12 (2008) doi:10.1186/1471-2164-9-S1-S12.

[16] Ker Chau Li: Genome-wide coexpression dynamics: Theory and application. Proceedings of the National Academy of Sciences(USA), 99(26) (2002) 16875–16880.

[17] Zheng Guo, Yongjin Li, Xue Gong, Chen Yao, Wencai Ma, Dong Wang, Yanhui Li, Jing Zhu, Min Zhang, Da Yang, Jing Wang: . Bioinformatics, 23(16) (2007) 2121–2128.

[18] M R Arkin, J A Wells: Samll-molecular inhibtors of protein-protein interactions: progressing towards the dream. Nature Reviews Drug Discovery, 3 (2004) 301–317.

[19] P L Toogood: Inhibtion of protein-protein association by small molecules: approaches and progress . Journal of Medicinal Chemistry, 45 (2002) 1543–1558.

[20] A I Archakov, V M Govorun, A V Dubanov, Y D Ivanov, A V Veselovsky, P Lewis, P Jassen: Protein-protein interactions as a target for drugs in proteomics. Proteomics, 3 (2003) 380–391.

[21] L Pagliaro, J Felding, K Audouze, S J Nielsen, R B Terry, K J Christian, S Butcher: Emerging classes of protein-protein interactions inhitors and new tools for their development. Current Opinion in Chemical Biology, 8 (2004) 442–449.

[22] S Fletcher, A D Hamilton: Targeting protein-protein interactions by rational design: mimicry of protein surfaces. Journal of the Royal Society Interface, 3 (2006) 215–233.

[23] Nobuyoshi Sugaya, Kazuyoshi Ikeda, Toshiyuki Tashiro, Shiru Takeda, Jun Otomo, Yoshiko Ishida, Akiko Shiratori, Atushi Toyoda, Hideki Noguchi, Tadayuki Takeda, Satoru Kuhara, Yoshiyuki Sakaki, Takao Iwayanagi: An integrative in silico approach for discovering candidates for drug-targetable protein-protein interactions in interactome data. BMC Pharmacology, 7:10 (2007) doi:10.1186/1471-2210-7-10.

[24] Jacques Lapointe, Chunde Li, John P Higgins, Matt van de Rijna, Eric Bair, Kelli Mont-Gomery, Michelle Ferrari, Lars Egevad, Walter Rayford, Ulf Bergerheim, Peter Ekman, Angelo M DeMarzo, Robert Tibshirani, David Botstein, Patrick O Brown, James D Brooks, Jonathan R Pollacka: Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proceedings of the Academy Science(USA), 101(3) (2004) 811–816.

[25] Longcheng Li, Hong Zhao, Hiroaki Shiina, Christopher J Kane, Rajvir Dahiya: PGDB: a curated and integrated database of genes related to the prostate. Nucleic Acids Research, 31 (2003) 291–293.

[26] R R Wilcox: Introduction to robust estimation and hypothesis testing, Academic Press, San Diego (1997).

[27] T W Anderson: An introduction to multivariate statistical analysis. Wiley-Interscience, Hoboken N.J. (2003).