Network-Based Identification of Smoking-Associated Gene Signature for Lung Cancer

Ying-Wooi Wan, Changchang Xiao, and Nancy Lan Guo
Mary Babb Randolph Cancer Center, West Virginia University, Morgantown, WV 26506-9300
Email: {ywan2, cxiao}@mix.wvu.edu, lguo@hsc.wvu.edu

Abstract

This study presents a novel computational approach to identifying a smoking-associated gene signature. The methodology contains the following steps: 1) identifying genes significantly associated with lung cancer survival, 2) selecting genes which are differentially expressed in smoker versus non-smoker groups from the survival genes, 3) from these candidate genes, constructing gene co-expression networks based on prediction logic for smokers and non-smokers, 4) identifying smoking-mediated differential components, i.e., the unique gene co-expression patterns specific to each group, and 5) from the differential components, identifying genes directly co-expressed with major lung cancer hallmarks. The identified 7-gene signature could separate lung cancer patients into two risk groups with distinct post-operative survival (log-rank $P < 0.05$, Kaplan-Meier analysis) in four independent cohorts (n=427). It also has implications in the diagnosis of lung cancer (accuracy = 74%) in a cohort of smokers (n=164). Computationally derived co-expression patterns were validated with Pathway Studio and STRING 8.

1. Introduction

Lung cancer remains the leading cause of cancer deaths for both men and women in the United States [1]. Non-small cell lung cancer (NSCLC) is the most common subtype. Studies have demonstrated that smoking contributes to about 90% of all lung cancer cases and it appears to be a strong risk factor in the development of lung cancer [5-7]. However, smoking is not an established prognostic factor in lung cancer as its effect in lung cancer progression remains unclear. In this study, we sought to identify a smoking-associated gene signature with implications in lung cancer diagnosis and prognosis using genome-wide transcriptional profiles from lung cancer patients.

Most studies in molecular biomarker discovery rank genes based on their association with the clinical outcome. The top-ranked genes are then selected as signature genes [3, 8, 9]. However, these approaches do not account for the interactions among genes. It's known that genes and proteins do not function in isolation. Instead, genes function through a series of interactions with one another and disease is one possible result of aberration in these interactions. Furthermore, recent studies suggest that molecular network analysis could be used to improve disease classification [10-12], and identify disease genes [13], novel therapeutic targets [14, 15], and disease related sub-networks [16]. Thus, by incorporating the study of gene associations with disease outcome and co-expression networks analysis, it could lead to discovery of biomarkers for precise disease prognosis.

Boolean networks can provide important biological insights into regulation functions [17-20]. The Boolean implication networks presented by Sahoo et al. [18] used scatter plots of expression between two genes to induce the implication relations. We developed an induction algorithm based on prediction logic [21] to derive implication relations. In our previous study, implication networks were employed to model disease-mediated genome-wide co-expression networks for the identification of a prognostic gene signature [22]. In this study, implication networks were used to infer the relevance to signaling pathways in a set of selected genes associated with smoking and lung cancer survival.

We hypothesized that an analysis of genes associated with smoking and major lung cancer signaling pathways will lead to the identification of a gene signature that provides a more accurate diagnosis and prognosis of lung cancer. The following steps were carried out to test the hypothesis: 1) Genes that were significantly associated with lung cancer survival were identified from genome-wide expression profiles using the training set (n=256). 2) Genes with differential expression in smokers versus non-smokers

were then selected for further analysis. 3) The implication network algorithm was employed to construct smoking mediated gene co-expression networks. 4) From the differential components that are unique to the smoker or non-smoker group, genes that had common co-expression with *EGF, EGFR, MET, KRAS, E2F3,* and *E2F5* were pinpointed. The identified 7-gene signature was then validated in three independent cohorts (*n*=427) for prognostic prediction.

2. Materials and Methods

2.1. Implication induction algorithm for pair-wise coexpression network construction

An implication network is a directed graph with variables as nodes, and adjacent nodes are connected with arch representing implications. The first induction algorithm for implication network was proposed by Liu et al. [23, 24] based on binomial distribution, which is suitable for binary datasets. An alternative network induction algorithm was proposed by Guo et al. [21] based on prediction logic [25], which is applicable for more general applications, including multinomial datasets and multi-classification problems. Prediction logic reveals the implication relationships among variables in a dataset and evaluates propositions in formal logic by integrating formal logic theory and statistics. The most important aspect of prediction logic is the conceptual value of prediction analysis in constructing and evaluating useful statements, particularly in complex multinomial problems with moderate sample sizes. This feature is vital for clinical applications, in which many clinical parameters are multinomial and the patient sample size is small.

We used prediction logic based on formal logic rules relating two dichotomous variables to induce the implication network. The six most important implication rules relating two dichotomous variables are shown in Fig. 1, where each table is a contingency table and the shaded cells represent the errors for the corresponding implication rule. For example, $A \wedge \neg B$ is the error cell for the implication rule $A \Rightarrow B$, $N_{A \wedge \neg B}$ represents the number of error occurrences. A modified *U*-Optimality method [25] (Fig. 2) was used to derive the implication relation between each pair of variables in the dataset.

In the implication induction algorithm (Fig. 2), U_p is the scope of the implication rule, representing the portion of the data covered by the implication relation, and ∇_p is the precision of the implication rule, representing the prediction success of the corresponding implication relation. An implication rule

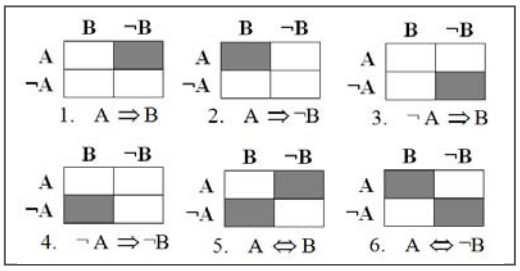

Figure 1. Six important implication rules relating two dichotomous variables.

has high precision when the number of error occurrences is a small portion of the data covered by the implication rule. The minimum scope and precision required by the implication rule are indicated respectively by U_{min} and $\nabla_{\min}$, which must be positive for a valid implication relation. The induction algorithm derives an implication rule if it has the maximum scope, U_p and it satisfies the constraint that its scope, U_p and precision, ∇_p are greater than the required minimum values, U_{min} and $\nabla_{\min}$, respectively. To simplify the computations of the maximization problem, the ∇_{ij} value of every error cell must be greater than that of the non-error cells for the corresponding implication rule [21].

For a single error cell, where N_{ij} is the number of error occurrences, scope, U_p and precision, ∇_p are defined as:

$$U_p = U_{ij} = \frac{N_{i.} * N_{.j}}{N^2},\ \nabla_p = \nabla_{ij} = 1 - \frac{N_{ij}}{N * U_p}.$$

For multiple error cells, they are defined as:

$$U_p = \sum_i \sum_j \omega_{ij} * U_{ij},\ \nabla_p = \sum_i \sum_j \left(\frac{\omega_{ij} * U_{ij}}{U_p} \right) \nabla_{ij}$$

where $\omega_{ij} = 1$ for error cells; otherwise, $\omega_{ij} = 0$.

This implication induction algorithm is general for discrete datasets. With the expansion of the contingency table M_{ij} (Fig. 2), implication rules can be induced for multinomial datasets, where error cells are those with top precision (∇_{ij} values) and satisfying all the constraints. The proposition can then be induced according to the error set.

The complexity of the induction algorithm is $O(Nv^2)$, where N is the sample size and v is the number of variables in the dataset (i.e. nodes in the implication networks) [21]. The difference between this algorithm and that of Hildebrand et al. [25] is that minimum requirements for deriving an implication rule were set for both scope (U_p) and precision (∇_p), instead of for precision alone.

The Implication Induction Algorithm
Begin

Set a significant level ∇_{min} and a minimal U_{min}

For $node_i$, $i \in [0, v_{max} - 1]$ and $node_j$, $j \in [i+1, v_{max}]$
(Note: v_{max} is the total number of nodes)

For all empirical case samples N

Compute a contingency table as in Figure 1

$$M_{ij} = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix}$$

For each relation type k out of the six cases, **find** the solution

$$\text{Subject to} \quad \begin{array}{c} Max\ U_p \\ Max\ U_p \geq U_{min} \\ \nabla_p \geq \nabla_{min} \end{array}$$

$$\nabla_{\text{error cells}} > \nabla_{\text{non-error cells}}$$

If the solution exists, **then return** a type k relation

End

Figure 2. Implication induction algorithm for building co-expression networks.

2.2. Microarray profiles and patient samples

Four sets of published microarray gene expression profiles were used in this study. The first set contains 442 lung adenocarcinoma patient samples obtained from a multi-center microarray study of lung cancer published by Shedden et al. [2]. The second set contains 130 adenocarcinoma and squamous cell lung cancer samples published by Raponi et al. [3]. The third set contains 111 non-small cell lung carcinoma samples published by Bild et al. [4]. The fourth set contains samples of airway epithelial cells from 164 current and former smokers published by Spira et al. [5]. Data used in the analysis was quantile-normalized and $\log_2$ transformed with dChip [26].

3. Results and Discussions

3.1. Identification of a smoking-associated gene signature

In this study, the UM and HLM cohorts from Shedden et al. [2] formed the training set (n=256), whereas MSK and DFCI formed the test set (n=186). Genes with missing values in at least half of the samples were removed, which left 19,866 genes for the analysis.

Survival genes were first selected from the whole genome. A total of 2,310 genes were significantly associated with overall survival ($P < 0.05$, univariate Cox modeling) in the training data. Second, from this set of 2,310 genes, 217 genes showed significant differential expression ($P < 0.05$, t-tests) in smokers versus non-smokers in the training data. These 217 survival and smoking-associated genes as well as six major signaling proteins, including *EGF*, *EGFR*, *MET*, *KRAS*, *E2F3*, and *E2F5,* were included in the network analysis. These signaling pathways are included in human non-small cell lung cancer disease mechanisms delineated by the KEGG Pathway Database (http://www.genome.jp/kegg/pathway/hsa/hsa05223.html). These six hallmarks were not significantly associated with survival nor differentially expressed in smokers.

To construct implication networks, expression profiles in each patient were partitioned into binary values using the mean expression profile of each gene as the cutoff. If the expression of a gene in a patient sample was greater than the mean in the cohort, this gene was denoted as *up-regulated* in this tumor sample; otherwise, it was denoted as *down-regulated* in the tumor sample. Patient samples in the training set were separated into two groups: smokers (patients who smoked in the past or who are currently smoking) and non-smokers (patients who never smoked). For each patient group, co-expression network among the 223 genes was constructed using the implication induction algorithm. Between each pair of the 223 genes, possible significant ($P < 0.05$; z-tests) co-expression relations (interactions) were derived in the smoker group and the non-smoker group, constituting smoking-mediated gene co-expression networks for lung cancer. By comparing the implication rules between each pair of nodes in the two networks, differential network components were identified. These differential components are interactions that were present in the smoker group but missing in the non-smoker group, or conversely, those present in the non-smoker group but absent in the smoker group.

From the differential components associated with the smoker group and the non-smoker group, genes having direct interactions with the six lung cancer hallmarks were identified. As a result, six genes were identified from the smoker group and one gene was identified from the non-smoker group. This constituted the smoking-associated 7-gene signature for lung cancer prognosis.

3.2. Prognostic evaluation of the signature

We sought to study if the gene signature identified could provide accurate prognostic prediction of survival for lung cancer patients. The six hallmarks were not fitted in the model as they were not significantly associated with survival. On the training cohort, the original continuous expression profiles of the seven probes were fitted into a Cox proportional hazard model as covariates. A survival risk score was generated for each patient in the training set. To

identify the best patient stratification scheme, various cutoff values of the risk scores were evaluated on the training set. The cutoff value that gave the shortest distance to the point of perfect prediction, i.e., point [0,1] in the 3-year ROC curve (Fig. 3A), produced the best patient stratification in the training set (Fig. 3B). Therefore, the training model and cutoff value were applied to the test set (Fig. 3C). In both training and test sets, this classification scheme generated significant patient stratifications (log-rank P<0.007, Kaplan-Meier analysis).

To evaluate the statistical significance of the signature identified from the proposed network analysis, a set of seven genes from the 217 survival and smoking-associated genes were randomly selected and constructed as a classifier using the same approach with the Cox proportional hazard model. Results showed that the identified signature gave significantly (P < 0.04) better prognosis compared with 1,000 random gene sets.

3.3. Smoking association and smoking cessation

To evaluate the association of the identified 7-gene signature with smoking, we evaluated the performance of the prognostic signature on smokers in the studied cohorts. Results showed that the signature gave accurate prognostic prediction in smokers in the test cohort (log-rank P < 0.01, Kaplan-Meier analysis) (Fig. 3D) but not in non-smokers (log-rank P < 0.12, Kaplan-Meier analysis, results not shown). In addition, gene expression-defined high- and low-risk groups showed significant association with smoking (P < 0.02, Chi-square tests) and smoking cessation (P < 0.00001, Chi-square tests) (Table 1). Specifically, smokers were significantly associated with high-risk group compared with non-smokers, and current smokers showed a stronger association with the high-risk group compared with former smokers.

Table 1. Associations between smoking status and gene expression-defined prognostic risk groups.

	Low-risk	High-risk	Chi-square Test
Smoker	143	157	**Smoking association** $\chi^2 = 5.76$ ($P = 0.02$)
Non-smoker	33	16	
Current Smoker	3	29	**Smoking cessation** $\chi^2 = 19.37$ (P = 1.08e-5)
Former Smoker	140	128	

3.4. Prognostic validation on other histology subtypes of NSCLC

The prognostic performance of the 7-gene signature was further evaluated on Raponi [3] and Bild [4] cohorts including squamous cell lung cancer. Due to small sample size, patient samples in the studied cohort were randomly partitioned into separate training and test sets. Then, a prognostic classifier was constructed

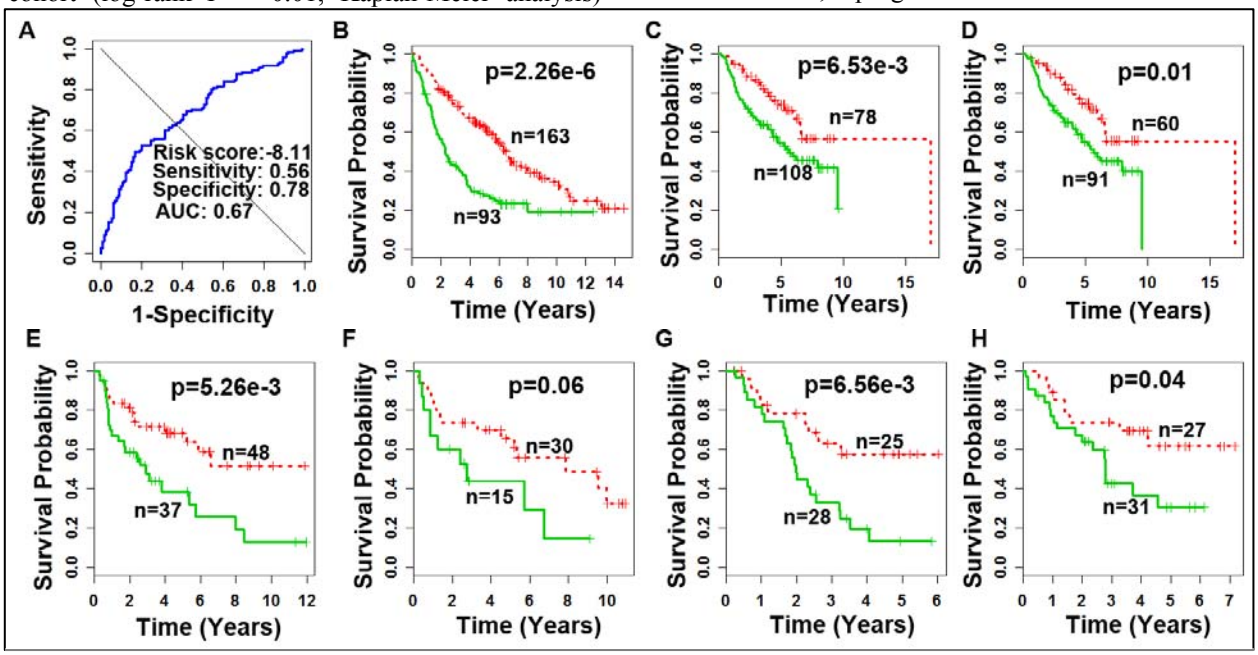

Figure 3. Prognostic prediction of patient survival by the smoking-associated gene signature. On the cohorts from Shedden et al. [2], the risk score giving the best prediction on the 3-year ROC curve was identified as the cutoff for patient stratification (A). This cutoff value generated significant patient stratification on the training set (B), test set (C), and smokers of test set (D) in Kaplan-Meier analyses. Significant patient stratifications were also obtained in the training and test sets on cohorts from Raponi et al. [3] (E, F) and Bild et al. [4] (G, H). Log-rank tests were used to assess the significance of the difference between survival probabilities in two prognostic groups.

on training set using the Cox proportional hazard model and validated on the test set without re-estimation of parameters. In both training and test sets, the 7-gene signature stratified patients into two distinct survival groups (Fig. 3E-3H).

3.5. Early detection of lung cancer

We further evaluated whether the 7-gene signature could be used for the diagnosis of lung cancer in smokers. The smoking cohort from Spira et al. [5] was separated into a training set (n=77) and two independent test sets (n=52 and n=35). With the nearest neighbor algorithm implemented in WEKA [27], the classifier could accurately identify lung cancer patients from normal patients with an overall accuracy greater than 73% in both test sets (Table 2). Furthermore, the 7-gene signature's performance was significantly ($P < 0.002$) better than that of random seven genes using the same classifier in 1,000 tests, on the same training and test sets.

Table 2. Prediction of lung cancer risk in smokers.

	Sensitivity (lung cancer)	Specificity (normal)	Overall Accuracy
Training (10-fold CV)	74% (26/35)	57% (24/42)	65% (50/77)
Test 1	72% (18/25)	74% (20/27)	73% (38/52)
Test 2	72% (13/18)	76% (13/17)	74% (26/35)

3.6. Confirmation of network topology

The co-expression network topology was evaluated. To increase the reproducibility, common differential network components that were present in both training and test sets were retrieved. These co-expression relations represent the smoking-mediated gene co-expressions in lung cancer patients. There were 17 common interactions specifically associated with smokers (Fig. 4A) and one interaction specifically associated with non-smokers (Fig. 4B). Statistical significance of the 18 interactions commonly found in both training and test sets was evaluated as $P < 0.18$ in 1000 permutation tests based on a metric, S. The metric S represents the proportion of the number of common interactions found in both training and test sets over the number of interactions found in the training set. Null distribution of the metric (S) was generated by permuting the class labels in the test set.

In order to confirm the biological relevance of the derived co-expression relations, literature-reported interactions related to these genes were retrieved by inputting these genes into bioinformatics tools including Pathway Studio (Fig. 4D) and other curated signal pathway databases. Three interactions specific to smokers that were derived from the implication network have been validated in experiments (Fig. 4A).

4. Conclusions and future work

This study identified a smoking-associated 7-gene signature that co-expressed with major lung cancer signaling pathways. The identified 7-gene signature could potentially be used for prognostic categorization and screening of lung cancer risk in smokers. The gene expression signature showed strong association with smoking and smoking cessation.

The results indicate that the implication network methodology based on prediction logic could identify biologically relevant co-expression patterns. The implication networks successfully revealed biological interactions reported in the literature. Currently, we are carrying out experiments to validate smoking mediated gene expression and the perturbation of signaling pathway mechanisms.

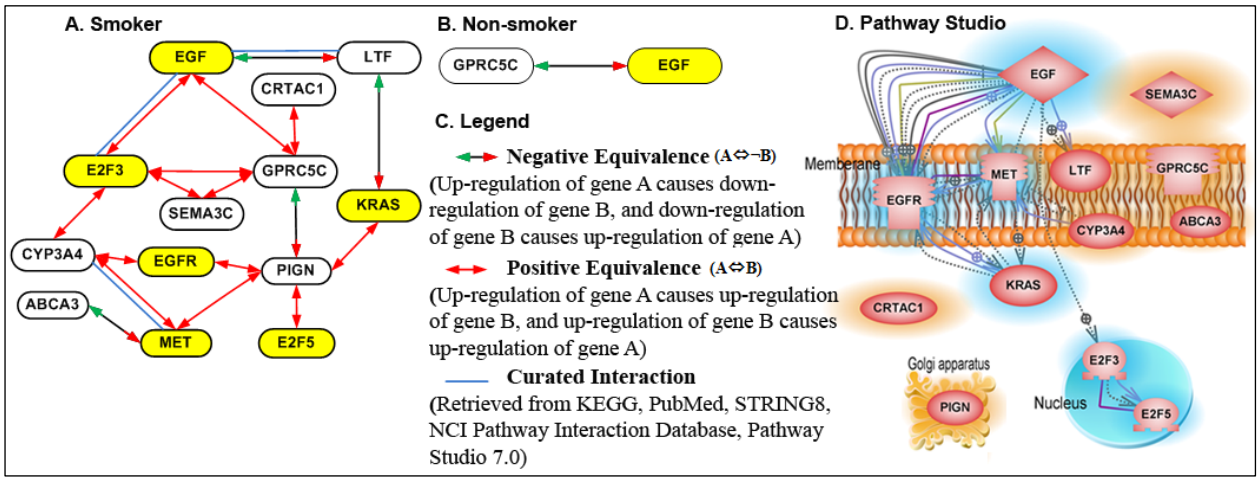

Figure 4. Interactions among the smoking-associated signature genes and lung cancer hallmarks. Gene co-expression patterns specific to smokers (A) and non-smokers (B) derived by the implication network algorithm ($P < 0.05$) commonly present in both training and test sets. The biological interpretation of the implication relations are described in (C). Interactions reported in literature were also retrieved from Pathway Studio (D).

5. Acknowledgement

We gratefully thank Dr. James Denvir and Rebecca Raese (West Virginia University) for helpful suggestions. This project is supported by NIH R01LM009500 and NCRR P20RR16440 (Guo). Software license and training for Pathway Studio is supported by NIH/NCRR P2016477.

Open Software Access: GeNet (R and C packages) is provided:
http://www.hsc.wvu.edu/mbrcc/fs/GuoLab/products.asp

6. References

[1] A. Jemal, R. Siegel, E. Ward et al., "Cancer statistics, 2009," *CA Cancer J.Clin.,* vol. 59, no. 4. pp.225-249, July, 2009.

[2] K. Shedden, J. M. Taylor, S. A. Enkemann et al., "Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study," *Nat.Med,* vol. 14, no. 8. pp.822-827, Aug., 2008.

[3] M. Raponi, Y. Zhang, J. Yu et al., "Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung," *Cancer Res.,* vol. 66, no. 15. pp.7466-7472, Aug., 2006.

[4] A. H. Bild, G. Yao, J. T. Chang et al., "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature,* vol. 439, no. 7074. pp.353-357, Jan., 2006.

[5] A. Spira, J. E. Beane, V. Shah et al., "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer," *Nat Med.,* vol. 13, no. 3. pp.361-366, Mar., 2007.

[6] P. P. Massion, Y. Zou, H. Chen et al., "Smoking-related genomic signatures in non-small cell lung cancer," *Am.J.Respir.Crit Care Med.,* vol. 178, no. 11. pp.1164-1172, Dec., 2008.

[7] M. Woenckhaus, L. Klein-Hitpass, U. Grepmeier et al., "Smoking and cancer-related gene expression in bronchial epithelium and non-small-cell lung cancers," *J.Pathol.,* vol. 210, no. 2. pp.192-204, Oct., 2006.

[8] D. G. Beer, S. L. Kardia, C. C. Huang et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nat.Med.,* vol. 8, no. 8. pp.816-824, Aug., 2002.

[9] H. Y. Chen, S. L. Yu, C. H. Chen et al., "A five-gene signature and clinical outcome in non-small-cell lung cancer," *N.Engl.J.Med.,* vol. 356, no. 1. pp.11-20, Jan., 2007.

[10] H. Y. Chuang, E. Lee, Y. T. Liu et al., "Network-based classification of breast cancer metastasis," *Mol.Syst.Biol.,* vol. 3. pp.140, 2007.

[11] F. J. Muller, L. C. Laurent, D. Kostka et al., "Regulatory networks define phenotypic classes of human stem cell lines," *Nature,* vol. 455, no. 7211. pp.401-405, Sept., 2008.

[12] I. W. Taylor, R. Linding, D. Warde-Farley et al., "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nat Biotechnol.,* vol. 27, no. 2. pp.199-204, Feb., 2009.

[13] V. Emilsson, G. Thorleifsson, B. Zhang et al., "Genetics of gene expression and its effect on disease," *Nature,* vol. 452, no. 7186. pp.423-428, Mar., 2008.

[14] P. Csermely, V. Agoston, and S. Pongor, "The efficiency of multi-target drugs: the network approach might help drug design," *Trends Pharmacol.Sci.,* vol. 26, no. 4. pp.178-182, Apr., 2005.

[15] M. A. Yildirim, K. I. Goh, M. E. Cusick et al., "Drug-target network," *Nat.Biotechnol.,* vol. 25, no. 10. pp.1119-1126, Oct., 2007.

[16] S. E. Calvano, W. Xiao, D. R. Richards et al., "A network-based analysis of systemic inflammation in humans," *Nature,* vol. 437, no. 7061. pp.1032-1037, Oct., 2005.

[17] R. Jansen, H. Yu, D. Greenbaum et al., "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science,* vol. 302, no. 5644. pp.449-453, Oct., 2003.

[18] D. Sahoo, D. L. Dill, A. J. Gentles et al., "Boolean implication networks derived from large scale, whole genome microarray datasets," *Genome Biol.,* vol. 9, no. 10. pp.R157, 2008.

[19] K. Sachs, O. Perez, D. Pe'er et al., "Causal protein-signaling networks derived from multiparameter single-cell data," *Science,* vol. 308, no. 5721. pp.523-529, Apr., 2005.

[20] R. Milo, S. Shen-Orr, S. Itzkovitz et al., "Network motifs: simple building blocks of complex networks," *Science,* vol. 298, no. 5594. pp.824-827, Oct., 2002.

[21] L. Guo, B. Cukic, and H. Singh, "Predicting Fault Prone Modules by the Dempster-Shafer Belief Networks," *18th IEEE International Conference on Automated Software Engineering (ASE'03)*. pp.249-252, 2003.

[22] Y. W. Wan, S. Bose, J. Denvir et al., "A Novel Network Model for Molecular Prognosis," *Proc.ACM International Conference on Bioinformatics and Computational Biology*, 2010.

[23] J. Liu and M. C. Desmarais, "A Method of Learning Implication Networks from Empirical Data: Algorithm and Monte-Carlo Simulation-Based Validation," *IEEE Transactions on Knowledge and Data Engineering,* vol. 9, no. 6. pp.990-1004, 1997.

[24] J. Liu, D. Maluf, and M. C. Desmarais, "A New Uncertainty Measure for Belief Networks with Applications to Optimal Evidential Inferencing," *IEEE Transactions on Knowledge and Data Engineering,* vol. 13, no. 3. pp.416-425, 2001.

[25] D. K. Hildebrand, J. D. Laing, and H. Rosenthal, *Prediction Analysis of Cross Classifications*: John Wiley & Sons, 1977.

[26] C. Li, "Automating dChip: toward reproducible sharing of microarray data analysis," *BMC.Bioinformatics.,* vol. 9. pp.231, 2008.

[27] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition)*: Morgan Kaufmann, 2005.