

A Novel Network Model for Molecular Prognosis

Ying-Wooi Wan, Swetha Bose, James Denvir, and Nancy Lan Guo*

Mary Babb Randolph Cancer Center, West Virginia University, Morgantown, WV 26506

ABSTRACT

Network-based genome-wide association studies (NWA) utilize the molecular interactions between genes and functional pathways in biomarker identification. This study presents a novel network-based methodology for identifying prognostic gene signatures to predict cancer recurrence. The methodology contains the following steps: 1) Constructing genome-wide coexpression networks for different disease states (metastatic vs. non-metastatic). Prediction logic is used to induct valid implication relations between each pair of gene expression profiles in terms of formal logic rules. 2) Identifying differential components associated with specific disease states from the genome-wide coexpression networks. 3) Dissecting network modules that are tightly connected with major disease signal hallmarks from the disease specific differential components. 4) Identifying most significant genes/probes associated with clinical outcome from the pathway connected network modules. Using this methodology, a 14-gene prognostic signature was identified for accurate patient stratification in early stage lung cancer.

Keywords: Implication networks, gene co-expression networks, molecular prognosis, personalized therapy

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

***Corresponding:** lguo@hsc.wvu.edu

1. INTRODUCTION

The accurate assessment of disease progression in individual patients is a critical prerequisite in personalized medicine. With the completion of the Human Genome Project, the emphasis of genome-wide association studies has shifted from cataloging the “parts list” of signature genes and proteins to elucidating the networks of interactions that take place among them [1]. Increasing evidence has suggested that molecular network analysis could be used to improve disease classification [2] and identify novel therapeutic targets [3]. Nevertheless, major challenges have been the development of methods for efficiently constructing genome-scale coexpression networks and the identification of a particular set of markers, from among the enormous number of potential markers, that has the highest predictive ability for disease outcome [4]. This study tests the hypothesis that the combined analysis of disease-mediated genome-wide coexpression networks, hallmark signal pathways, and clinical approaches leads to more informed clinical decision-making. This study will focus on the molecular diagnosis and

prognosis of lung cancer relapse and metastasis.

Lung cancer is the leading cause of cancer-related deaths in industrialized countries. Non-small cell lung cancer (NSCLC) accounts for about 80% of lung cancer cases. Currently, surgery is the major treatment option for patients with stage I NSCLC. However, 35–50% of stage I NSCLC patients will relapse within 5 years [5]. It remains an unsolved challenge for physicians to reliably identify patients at high risk for recurrence as candidates for chemotherapy. A few studies have described transcriptional profiling for lung cancer prognosis [6–8]. Nevertheless, there is no clinically applied gene test for this deadly disease.

In current genome-wide association studies, genes are ranked according to their association with the clinical outcome, and the top-ranked genes are included in the classifier. It has been noted that individual biomarkers showing strong association with disease outcome are not necessarily good classifiers [9]. Genes and proteins do not function in isolation, but rather interact with one another to form modular machines [10]. Molecular network analysis has led to promising applications in identifying new disease genes [11] and disease-related subnetworks [12], and classifying diseases [2].

Boolean networks can provide important biological insights into regulation functions [13]. Nevertheless, as the number of global states is exponential in the number of entities and the analysis relies on an exhaustive enumeration of all possible trajectories, this method is computationally expensive and only practical for small networks [14]. A recent formalism, causal Bayesian belief networks, have been utilized to model cellular networks [15]. Nevertheless, the number of possible networks is exponential in the number of nodes under consideration, which makes it impossible to evaluate all possible networks. Furthermore, it is not always possible to determine the causal relationships between nodes, i.e., the direction of the edges, owing to a property known as Markov equivalence [16]. More importantly, the acyclic Bayesian network structure was unable to model feedback loops, which are essential in signal pathways [17] and genetic networks [18–20]. To overcome this limitation, a more complex scheme, dynamic Bayesian networks, was explored for modeling temporal microarray data [21,22].

As an alternative to Bayesian networks, an implication network model employs a *partial order knowledge structure* (POKS) for structural learning and uses the Bayesian theory for inference propagation [23,24]. When using Dempster-Shafer theory for belief updating, this implication network methodology is termed a Dempster-Shafer belief network [25,26]. An implication network is a general methodology for reasoning under uncertainty. POKSs are closed under union and intersection of implication relations, and have the formal properties of directed acyclic graphs. The constraints on the partial order can be entirely represented by AND/OR graphs [23,27]. When the constraints on the partial order are relaxed, the implication networks can represent cyclic relations among the nodes. In this condition, the implication network structure is a directed graph with nodes connected by implication (causal) rules, which can contain cycles such as feedback loops.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM-BCB 2010, Nigara Falls, NY, USA.

Copyright (c) 2010 ACM ISBN 978-1-4503-0192-3... \$10.00

Motivated to model complex molecular patterns for assessing disease progression, we employed the implication network formalism for efficiently constructing disease-mediated genome-wide coexpression networks for the identification of prognostic gene signatures.

2. ALGORITHM

The implication network induction algorithm proposed by Liu et al. [25,26] is based on *binomial distribution*, which is suitable for binary datasets. We developed a network induction algorithm based on *prediction logic* [28,29], which can be used in general applications, including multinomial datasets and multi-classification problems. Prediction logic reveals the implication (causal) relationships among variables in a dataset and evaluates propositions in formal logic. It integrates formal logic theory and statistics to build a convenient predictive structure for a dataset. 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 patient sample size is usually small.

We used prediction logic based on formal logic rules relating two dichotomous variables to induce the implication network structure. A modified *U-Optimality* method [29] (Fig. 1) was used to derive the implication relation between each pair of attributes in a data set. In the implication induction algorithm (Fig. 1), 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 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 ∇_{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 ∇_{min} . To simplify the computation of the maximization problem, the ∇_p value of every error cell must be greater than that of the non-error cell for the corresponding implication rule [28,29].

For a single error cell, where N_{ij} is the number of error occurrences, we have:

$$U_p = U_{ij} = \frac{N_i * N_j}{N^2}, \quad \nabla_p = \nabla_{ij} = 1 - \frac{N_{ij}}{N * U_p}$$

For multiple error cells,

$$U_p = \sum_i \sum_j \omega_{ij} * U_{ij}, \quad \nabla_p = \sum_i \sum_j \left(\frac{\omega_{ij} U_{ij}}{U_p} \right) \nabla_{ij} \quad (\omega_{ij} = 1 \text{ for error cells; otherwise, } \omega_{ij} = 0)$$

The difference between our implication and that of Hildebrand et al. [29] is that we set minimum requirements for both scope (U_p) and precision (∇_p), instead of precision alone. Furthermore, each implication rule has an associated weight function that represents the conditional probability of the implied event.

The Implication Induction Algorithm

Begin

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

For $node_i, i \in [0, n_{max} - 1]$ and $node_j, j \in [i + 1, n_{max}]$
(Note: n_{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 to

$$\text{Subject to } \begin{aligned} & \text{Max } U_p \\ & \text{Max } U_p > U_{min} \\ & \nabla_p \geq \nabla_{min} \end{aligned}$$

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

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

End

Fig. 1. Implication induction algorithm.

3. IMPLEMENTATION AND RESULTS

In this study, an implication induction algorithm (Fig. 1) was used to construct pair-wise genome-scale coexpression networks for predicting recurrence in lung adenocarcinomas. In a published (dChip normalized) dataset [8], UM and HLM cohorts formed the training set ($n = 256$), whereas MSK ($n = 104$) and DFCI ($n = 82$) constituted two independent validation sets.

Genes with missing measurements in more than half of the samples were removed from analysis. Furthermore, for genes measured with multiple probes, the average expression of the duplicates was used to represent the expression profile of a unique gene for the network analysis (with 12,566 unique genes). To construct implication networks, the mean expression of each gene in a patient cohort was used as a cutoff to partition the expression profiles. 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. In the training set, patients who died within 5 years were labeled as poor-prognosis ($n = 125$), and those who survived 5 years after surgery were labeled as good-prognosis ($n = 104$). Censored cases (those with follow-up of less than 5 years) were removed from the analysis ($n = 27$). For each patient group in the training set, a genome-scale coexpression network was constructed using the implication induction algorithm. Between each pair of genes, possible significant ($P < 0.05$; z -tests) coexpression relations were derived in each patient group, constituting disease-mediated gene coexpression networks. By comparing the connectivity patterns (implication relations) of each pair of genes between the two networks, disease-specific differential network components were identified. These differential components contain the coexpression relations that were either present in the poor-prognosis group but missing in the good-prognosis group, or conversely, those present in the good-prognosis group but missing in the poor-prognosis group. In this analysis, more than 67 million interactions were derived in the good-prognosis group and more than 69 million interactions were derived in the poor-prognosis group. Of these interactions, more than 38 million were common to both disease states, more than 29 million were unique to the good-prognosis group, and more than 31 million were unique to the poor-

prognosis group. The computation was completed in 40 min by an Intel® Core™2 Duo processor with a 2.83-GHz CPU, 4 GB of memory (RAM) allocated, and 455 GB of hard disk space.

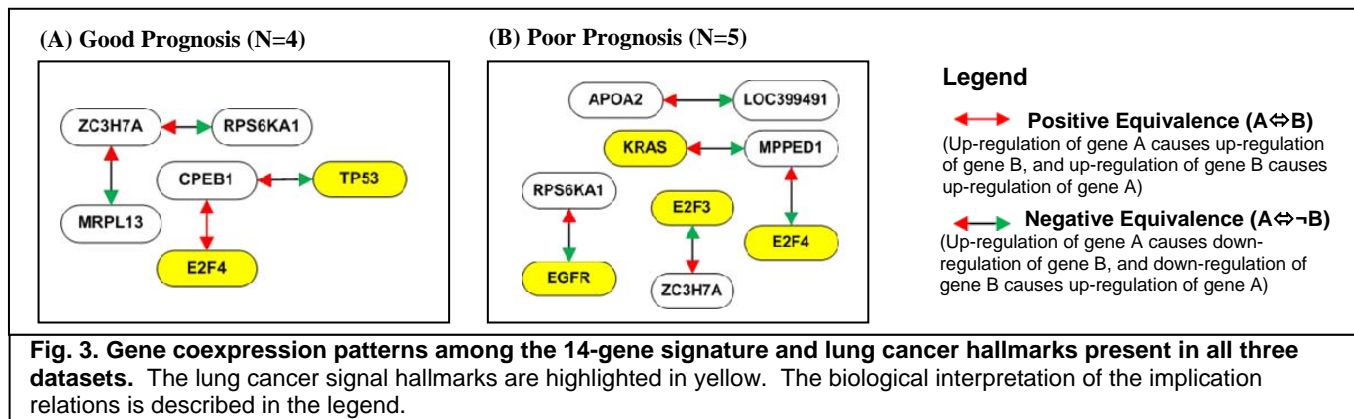
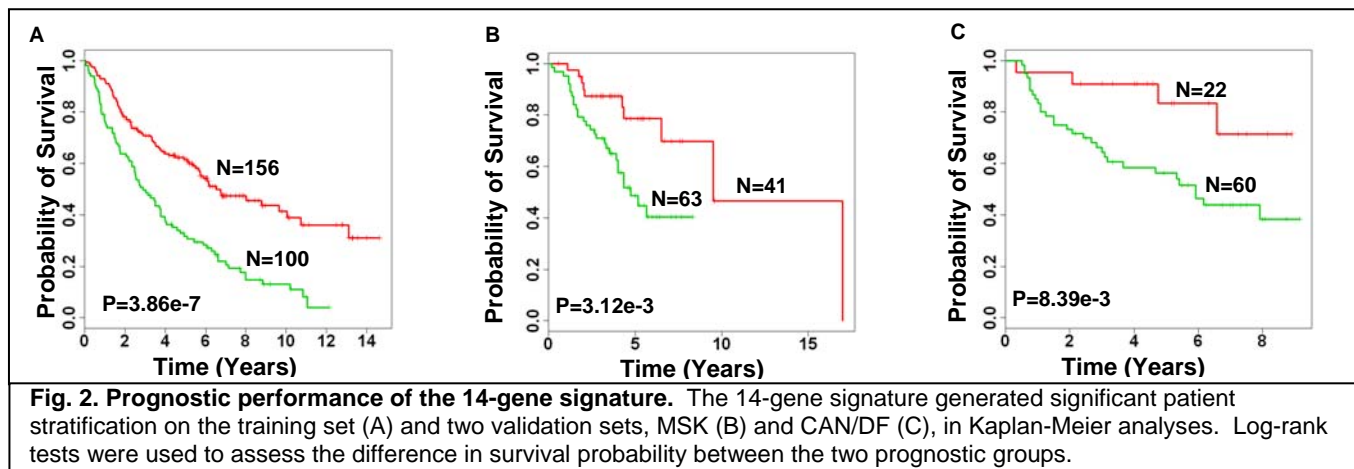
Next, genes displaying direct co-regulation with major NSCLC signal proteins were identified from the disease-specific network modules. Genes of a significant ($P < 0.05$) coexpression relation with *TP53*, *KRAS*, *EGF*, *EGFR*, *E2F3*, and *E2F4* were pinpointed from the differential components associated with each patient group. As a result, 63 genes were identified from the poor-prognosis group, 48 genes from the good-prognosis group, and 9 genes common in both groups, yielding a set of 102 genes.

We sought to evaluate whether the genes identified from the proposed network analysis could generate accurate prognostic prediction. From the training set of the original continuous microarray data, 19 probes were significantly associated with overall survival ($P < 0.05$, univariate Cox modeling), from which the top 14 genes ranked by RELIEF [30] were identified as the most accurate prognostic gene signature. By fitting a multivariate Cox proportional hazard model with the 14 genes as covariates, a survival risk score was generated for each patient. A risk score of -11.79 was identified as a cutoff value for patient stratification in the training set (Fig. 2A). This training model and cutoff value were applied to the two validation sets (Fig. 2B and 2C). In all three patient cohorts, this scheme stratified patients into prognostic groups with distinct overall survival (log-rank $P < 0.008$, Kaplan-Meier analyses).

The coexpression patterns of these 14 signature genes and six NSCLC hallmarks derived from the differential components in the training set were compared with those derived in the two validation sets. The common gene coexpression patterns presented in all three datasets are shown in Fig. 3, indicating the reproducibility of the gene/protein interactions derived from transcriptional profiles. Among all three patient cohorts, there are 4 common gene coexpression relations specifically associated with good-prognosis (Fig. 3A) and 5 common coexpression relations specifically associated with poor-prognosis (Fig. 3B). The coexpression relations among these genes are elucidated by the implication network structure. The coexpression networks in Fig. 3 are significant at $P < 0.24$ as evaluated in 1000 permutation tests. Specifically, a metric (S) was computed to represent the proportion of the number of common coexpression relations among three datasets over the number of coexpression relations found in the training set. The null distribution was generated by permuting the class labels in two validation sets.

4. CONCLUSIONS

This study demonstrates that the implication network methodology based on prediction logic is suitable for constructing genome-scale coexpression networks for analyzing perturbed gene expression patterns in different disease states. The disease-mediated differential network components may contain important information for the discovery of biomarkers and pathways for



targeted therapy and prognostic prediction. The implication network methodology provides a convenient and more predictive structure of gene regulation than the networks constructed based on correlation coefficients.

ACKNOWLEDGMENTS

We thank Dr. David Beer (University of Michigan) and Dr. Trey Ideker (University of California in San Diego) for thoughtful discussions. We appreciate the comments from anonymous reviewers. This project is supported by NIH R01LM009500 (PI: Guo) and NCRP P20RR16440 and Supplement (PD: Guo).

REFERENCES

- [1] Ideker T, Sharan R. 2008. Protein networks in disease. *Genome Res* 18: 644-652.
- [2] Chuang HY, Lee E, Liu YT, Lee D, Ideker T. 2007. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3: 140.
- [3] Csermely P, Agoston V, Pongor S. 2005. The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol Sci* 26: 178-182.
- [4] Sotiriou C, Piccart MJ. 2007. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat Rev Cancer* 7: 545-553.
- [5] Hoffman PC, Mauer AM, Vokes EE. 2000. Lung cancer. *Lancet* 355: 479-485.
- [6] Chen HY, Yu SL, Chen CH, Chang GC, Chen CY et al. 2007. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 356: 11-20.
- [7] Potti A, Mukherjee S, Petersen R, Dressman HK, Bild A et al. 2006. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N Engl J Med* 355: 570-580.
- [8] Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ et al. 2008. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 14: 822-827.
- [9] Emir B, Wieand S, Su JQ, Cha S. 1998. Analysis of repeated markers used to predict progression of cancer. *Stat Med* 17: 2563-2578.
- [10] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. *Nature* 402: C47-C52.
- [11] Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F et al. 2008. Genetics of gene expression and its effect on disease. *Nature* 452: 423-428.
- [12] Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV et al. 2005. A network-based analysis of systemic inflammation in humans. *Nature* 437: 1032-1037.
- [13] Albert R, Othmer HG. 2003. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J Theor Biol* 223: 1-18.
- [14] Karlebach G, Shamir R. 2008. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol* 9: 770-780.
- [15] Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* 303: 799-805.
- [16] Zhu J, Zhang B, Smith EN, Drees B, Brem RB et al. 2008. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet* 40: 854-861.
- [17] Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308: 523-529.
- [18] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D et al. 2002. Network motifs: simple building blocks of complex networks. *Science* 298: 824-827.
- [19] Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S et al. 2004. Superfamilies of evolved and designed networks. *Science* 303: 1538-1542.
- [20] Wuchty S, Oltvai ZN, Barabasi AL. 2003. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet* 35: 176-179.
- [21] Kim SY, Imoto S, Miyano S. 2003. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief Bioinform* 4: 228-235.
- [22] Pe'er D, Regev A, Elidan G, Friedman N. 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17 Suppl 1: S215-S224.
- [23] Desmarais MC, Maluf A, Liu J. 1996. User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-Adapted Interaction* 5: 283-315.
- [24] Desmarais MC, Meshkinfam P, Gagnon M. 2006. Learned Student Models with Item to Item Knowledge Structures. *User Modeling and User-Adapted Interaction* 16: 403-434.
- [25] Liu J, Desmarais MC. 1997. A Method of Learning Implication Networks from Empirical Data: Algorithm and Monte-Carlo Simulation-Based Validation. *IEEE Transactions on Knowledge and Data Engineering* 9: 990-1004.
- [26] Liu J, Maluf D, Desmarais MC. 2001. A New Uncertainty Measure for Belief Networks with Applications to Optimal Evidential Inferencing. *IEEE Transactions on Knowledge and Data Engineering* 13: 416-425.
- [27] Falmagne JC, Doignon JP, Koppen M, Villano M, Johannesen L. 1990. Introduction to knowledge spaces: how to build, test and search them. *Psychological Review* 97: 201-224.
- [28] Guo L, Cukic B, Singh H. 2003. Predicting Fault Prone Modules by the Dempster-Shafer Belief Networks. 18th IEEE International Conference on Automated Software Engineering (ASE'03) 249-252.
- [29] Hildebrand, D. K., Laing, J. D., and Rosenthal, H. 1977. Prediction Analysis of Cross Classifications. John Wiley & Sons.
- [30] Witten, I. H. and Frank, E. 2005. Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition). Morgan Kaufmann.