Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Network and Pathway Analysis of Cancer Susceptibility (A)

Network-Based Identification of Biomarkers Coexpressed with Multiple Pathways

Nancy Lan Guo and Ying-Wooi Wan

Mary Babb Randolph Cancer Center/School of Public Health, West Virginia University, Morgantown, WV, USA.

ABSTRACT: Unraveling complex molecular interactions and networks and incorporating clinical information in modeling will present a paradigm shift in molecular medicine. Embedding biological relevance via modeling molecular networks and pathways has become increasingly important for biomarker identification in cancer susceptibility and metastasis studies. Here, we give a comprehensive overview of computational methods used for biomarker identification, and provide a performance comparison of several network models used in studies of cancer susceptibility, disease progression, and prognostication. Specifically, we evaluated implication networks, Boolean networks, Bayesian networks, and Pearson's correlation networks in constructing gene coexpression networks for identifying lung cancer diagnostic and prognostic biomarkers. The results show that implication networks, implemented in *Genet* package, identified sets of biomarkers that generated an accurate prediction of lung cancer risk and metastases; meanwhile, implication networks revealed more biologically relevant molecular interactions than Boolean networks, Bayesian networks, and Pearson's correlation networks when evaluated with MSigDB database.

KEYWORDS: implication networks, coexpression with signaling pathways, lung cancer biomarkers

SUPPLEMENT: Network and Pathway Analysis of Cancer Susceptibility (A)

CITATION: Guo and Wan. Network-Based Identification of Biomarkers Coexpressed with Multiple Pathways. *Cancer Informatics* 2014:13(S5) 37–47 doi: 10.4137/CIN.S14054.

RECEIVED: April 23, 2014. **RESUBMITTED:** June 25, 2014. **ACCEPTED FOR PUBLICATION:** June 29, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Review

FUNDING: This study was funded by NIH R01/R56LM009500 (PI: Guo) and NCRR P20RR16440 and Supplement (PD: Guo).

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: lguo@hsc.wvu.edu

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction

The accurate assessment of disease susceptibility, progression, and treatment response in individual patients is a critical prerequisite for personalized therapy. High-throughput genome-scale profiling technologies have the potential to allow such molecular diagnostics. To date, there have been few gene expression-based tests applied in clinics for disease intervention. This fact puts a premium on developing innovative methodologies to embed biological relevance into biomarker identification.

With the completion of the Human Genome Project, the emphasis of genome-wide studies has shifted from cataloging a "parts list" of signature genes and proteins to elucidating the networks of interactions that occur among them.[1,2] Molecular network analyses have been used to improve disease classification[3–6] and identify novel therapeutic targets.[7–11] Nevertheless, major challenges include the development of methods for efficiently constructing genome-scale interaction networks[12] and the identification, from among the enormous number of genes, of a particular set of markers with the highest capacity for molecular diagnostics, prognostics, and prediction of treatment response.[13,14]

Here, we will give a comprehensive overview of computational methods used for biomarker identification, including rank-based feature selection methods and major network methodologies used in systems biology. Furthermore, we provide a performance comparison of several network models used in studies of cancer susceptibility, disease progression, and prognostication. Specifically, implication networks, as implemented in the *Genet* package, were used in conjunction with other rank-based feature selection algorithms to identify lung cancer diagnostic and prognostic biomarkers. The molecular interactions among the identified biomarkers were revealed with implication networks, Boolean networks, Bayesian networks, and Pearson's correlation networks. Each was then evaluated with five collections of gene sets and biological pathways from the MSigDB[1].

Rank-Based Methods for Biomarker Identification

The emerging use of biomarkers may enable physicians to make treatment decisions based on the specific characteristics of individual patients and their tumors, instead of population statistics.[15] 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. To identify the most powerful biomarkers in individualized prognostication, state-of-the-art feature selection methods[16–18] should be widely applied.

Attribute selection techniques can be categorized as those that rank *individual* attributes (filters) or those that rank *subsets* of attributes. Commonly used individual feature filtering methods include Cox models,[19] ANOVA, Bhattacharyya distance, divergence-based methods,[20] gain ratio, information gain, relief,[21,22] linear discriminant analysis,[23] and random forests.[24–26] Algorithms that evaluate subsets of features include correlation-based feature selection, consistency-based subset evaluation, wrapper,[21,22] self-organizing maps (SOM),[27] independent component analysis,[28–30] partial least squares,[31] principal component analysis (PCA),[32–34] kernel PCA,[35,36] sliced inverse regression,[37] and logistic regression.[38] Exhaustive search, branch-and-bound search, sequential search (forward or backward), floating search, "plus *l*-take away *r*" selection,[39] Tabu search,[40] ant colony optimization,[41,42] genetic algorithms,[43,44] simulated annealing,[45–47] and stochastic hill climbing[48] can be used as search strategies in feature selection. Only the first two search methods guarantee the optimal subset; the rest generate suboptimal results. However, the worst-case complexity of the first two search methods is exponential, and therefore, these two methods are not feasible for a large dataset. Some feature selection algorithms such as *significant analysis of microarray (SAM)*[49] and the *multivariate permutation test (MPT)* are designed specifically for gene filtering.[50] As the number of variables is much greater than the sample size in high-throughput applications, feature pre-selection using the *t*- or *F*-test[51] and nonparametric Wilcoxon statistics[52,53] are used in processing raw high-throughput data.

Regularized Linear Models

Regularized linear models can also be used to identify biomarkers. Linear models are used to study the effects of multiple factors on the response variable or used to construct a prediction model. In microarray studies, linear models such as ANOVA or ordinary least square (OLS) linear regression models were used to analyze gene expression changes or to construct classification models.[54,55] In the general context, an OLS linear regression model predicts the response of variable y (formulated in Figure 1A), which estimates the set of coefficients β by minimizing the residual squared error.

In genomic studies, where the curse of dimensionality phenomenon with the large p (number of predictors) small n (number of samples) is common, linear models are fitted along with certain penalty terms called regularized linear models. Two common regularized linear models used in genomic studies are *lasso* (least absolute shrinkage and selection operator)[56] and elastic net.[57] *Lasso* imposes an L1-norm penalty (Fig. 1B) to the model to enforce shrinkage and to avoid the over-fitting problem in the large p small n situation commonly present in genome studies. However, *lasso* performs poorly in data with high colinearity[58] and selects only one out of a group of genes sharing the same biological process. In order to enable the selection of genes belonging to the same biological process or pathway, elastic net was proposed.[57] This is basically an extension of *lasso* through combining the L2-norm along with the L1-norm penalty (Fig. 1C). The combination of both L1-norm and L2-norm penalties aims to allow both shrinkage and grouping of gene variables. However, the grouping feature of elastic net would lead to the selection of highly redundant genes and therefore the incapability of pinpointing a small subset of predictive genes.

With the abundant resources and increasing knowledge of biological regulatory networks, protein–protein interactions (PPI), signaling pathways, and known relationships among genes could be incorporated into the regression model. The network could be represented by a graph, and the graph's corresponding Laplacian matrix could then be applied as a penalty in the regression models (Fig. 1D). By having the graph Laplacian matrix as the penalty term, the smoothness of the coefficients is applied over the topography of the graph instead of solely to the correlations among the genes. In other words, the a priori knowledge of the functional relations among genes is embedded into the model through the network (graph) and could reveal a set of genes that are more biologically relevant instead of a set of correlated genes (which could be redundant). The network-constraint regularized model has been proposed to identify biomarkers associated with patient survival time,[59] and a network-constraint logistic model was used to identify biomarkers for tumor subtype[60] with cancer genomic data. These network-regularized regression models outperform *lasso* and elastic net with simulation data in both studies.[59,60] In a cancer susceptibility study of glioblastoma and tumor subtype analysis with breast cancer profiles of The Cancer Genome Atlas (TCGA) consortium, these two network-constraint regularized regression models identified biomarkers confirmed in published literature.[59,60]

A

$$\widehat{\beta} = \arg\min_{\beta} [(y - X\beta)^T (y - X\beta)]$$

B

$$\widehat{\beta}_{lasso} = \arg\min_{\beta} [(y - X\beta)^T (y - X\beta) + \underset{\textit{Shrinkage}}{\lambda|\beta|_1}]$$

C

$$\widehat{\beta}_{e.net} = \arg\min_{\beta} [(y - X\beta)^T (y - X\beta) + \underset{\textit{Shrinkage}}{(1-\alpha)|\beta|_1} + \underset{\textit{Grouping}}{\alpha|\beta|_2}]$$

D

$$\widehat{\beta}_{net} = \arg\min_{\beta} [(y - X\beta)^T (y - X\beta) + \underset{\textit{Shrinkage}}{\lambda_1|\beta|_1} + \underset{\substack{\textit{Grouping based}\\ \textit{on graph structure}}}{\lambda_2 \beta^T L \beta}]$$

Definition for each object in the above formula:

Suppose there are *n* observations and *p* predictors

Response vector : $y = (y_1, ..., y_n)^T$

Input matrix : $X_{nxp} = (x_1|, ..., |x_p)$

Coefficient vector : $\beta = (\beta_1, ..., \beta_p)^T$

L1-norm of $\beta : |\beta|_1 = \sum_{i=1}^{p} |\beta_i|$

L2-norm of $\beta : |\beta|_2 = (\sum_{i=1}^{p} \beta^2)^{1/2} = (\beta^T \beta)^{1/2}$

Given a network *G* = (*V*, *E*) where *V* is the set of nodes for p predictors, *E* is the adjacency matrix

Laplacian matrix : $L = I - D^{-\frac{1}{2}} E D^{-\frac{1}{2}}$, where degree matrix : $D = \mathrm{diag}(E \cdot 1_p)$

Figure 1. Coefficient estimation for regularized linear models. Equations to estimate the coefficient vectors in (**A**) OLS linear regression model, (**B**) *lasso*, (**C**) elastic net, and (**D**) network-constrained regularization models.

General Methodologies for Modeling Molecular Networks

It has been noted that individual biomarkers showing strong association with disease outcome are not necessarily good classifiers.[61–63] Because genes and proteins do not function in isolation, but rather interact with one another to form modular machines,[64] understanding the interaction networks is critical to unraveling the molecular basis of disease. Molecular network analysis has led to promising applications in identifying new disease genes[65–70] and disease-related subnetworks,[71–80] mapping cause-and-effect genetic perturbations,[81–84] and classifying diseases.[3–6,85,86] The various computational models that have been developed for molecular network analysis can be roughly categorized into three classes[12]: logical models to demonstrate the state of entities (genes/proteins) at any time as a discrete level[87–90]; continuous models to represent real-valued network processes[91–96] and activities[97–102]; and single-molecule models[103–105] to simulate small regulatory networks and mechanisms.[106–110]

Logical models. In the category of logical models, Boolean networks[87] were recently used to analyze the relationship between regulation functions and network stability in a yeast transcriptional network[111] and the dynamics of cell-cycle regulation.[112] The structure of Boolean networks can be learned from gene expression profiles.[113–115] Boolean networks can provide important biological insights into regulation functions and the existence and nature of *steady states* (ie, polarity gene expression)[116] and network *robustness*. 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.[12] Because of insufficient experimental data or incomplete understanding of a system, several candidate regulatory functions may be possible for an entity. To express uncertainty in regulatory logic, the probabilistic Boolean network (PBN) was developed[117] and used to model a 15-gene subnetwork inferred from human glioma expression data.[118] The synchronous dynamics of a Boolean network can be captured by a Petri net,[119] which is a nondeterministic model widely used for detecting active pathways and state cycles[120] and for analyzing large metabolic pathways[121–124] and regulatory networks.[125] Another model, module networks, infers the regulation logic of gene modules as a decision tree, given gene expression data.[126] The Boolean implication networks presented by Sahoo et al.[127,128] used scatter plots of the expression between two genes to derive the implication relations in the whole genome. To date, Boolean implication networks have not been applied in biomarker discovery.

Markov networks are another family of logical models used to infer the inter-relationships among genes. Markov network, also known as Markov random field, is a statistical framework to analyze and visualize conditional relationships between sets of random variables. The structure of the conditional relationships could be exhaustively explored because of the Markov properties.[129] In the graphical form, vertices represent random variables and the edges between vertices denote the conditional dependencies between the variables.

For example, variables *A* and *B* are connected if *A* is predictive of *B*, independent of all other variables. Markov networks are efficient in representing the distributions over a very high dimension of variables. Therefore, Markov networks could be used to infer the underlying structure of relationships among the genes in cancer patients. These methods are advantageous when only the genomic data are available and the clinical covariates are not available or not predictive of the disease. A commonly used Markov network in high-throughput genomic data is Graphical Gaussian Model (GGM). GGM has been applied to infer the relationships among sets of random variables with continuous values. In bioinformatics, GGM has been applied to study the patterns of relationships and associations between a large-scale of genes based on DNA microarray gene expression data.[130,131] Because of the large dimensionality of genomic data, the original GGM has a few challenges when applied to high-throughput data. The first challenge is that a large number of observations are required in order to obtain reliable estimates of the conditional dependencies between variables[132]; whereas genomic data have tens of thousands of genes involved but only with a few hundred observations. The second challenge lies in the model selection. As the number of models grows super-exponentially with the number of genes, only a small subset of models can be tested.[132] The third challenge is that a dense network with a large number of edges connecting numerous genes involved makes interpretability unfeasible.[133] It has been known that biological networks are not fully connected. Instead, a biological network is sparse and free scale. To overcome these challenges, variations and extensions of GGM were proposed. Among these include the modified GGM approach that first infers small subnetworks of three genes (tri-graph) and then combines the subnetworks into the proposed complete network.[131] This modified GGM was applied to elucidate the regulatory network of two isoprenoid biosynthesis pathways in *Arabidopsis thaliana*.[131] Another example is the application of a regularization procedure to estimate a sparse precision matrix in the setting of GGM.[134] A novel threshold gradient descent (TGD) regularization is applied for imposing penalization estimation of the GGM and thus accounts for the curse of dimensionality issue in high-throughput genomic data.

With the advancement of technology in recent years, sequencing technology is gradually replacing DNA microarray to measure genome-wide gene expression profiles as RNA-sequencing technologies yield less technological variation than microarrays.[135] Different from the typical log-ratio expression values from microarray data that follow approximately a Guassian distribution, RNA-seq data measurements are in read counts of how many times a transcript has been mapped to the specific genomic location. These read counts are non-negative integer values, which follow approximately a Poisson distribution.[135–137] Therefore, Poisson graphical models should then be used for analyzing next-generation sequencing data instead of Gaussian graphical models. Various methods proposed to model the underlying structure of multivariate count data of Poisson distribution suffer from deficiencies, including infeasibility of applying the contingency table-based approach when the number of variables is extremely high,[129] and limitations of modeling only the marginal distributions of independent variables[138] or modeling only negative dependencies.[139] Recently, an approach was proposed to overcome these deficiencies in modeling regulatory networks from sequencing count data.[140] Specifically, the proposed log-linear Poisson graphical model estimates the model parameters locally via neighborhood selection by fitting L1-norm penalized data to a log-linear model and provides high computational efficiency with the employment of a fast parallel algorithm. The proposed log-linear Poisson graphical model was applied on breast cancer microRNA data and revealed known regulator modules of breast cancer. It also discovered novel microRNA clusters and hubs that provide further insights into regulatory mechanisms of breast cancer.[140]

Since the last decade, TCGA consortium has been profiling various genomic data from hundreds of patients of various cancer types to facilitate the understanding of molecular mechanisms underlying these deadly diseases. A few logical models have been proposed to utilize this abundant resource. One example is the Multi-Dendrix method, which is a linear programming algorithm to learn a set of driver pathway modules with both high mutual exclusivity and coverage of patients from somatic mutation data.[141] Applications of Multi-Dendrix to glioblastoma and breast cancer from TCGA consortium identified mutation genes overlapping with known oncogenic pathways, including *PI*(3)*K* in glioblastoma and *p*53 and *GATA*3 in breast cancer. Another example is a method known as pathway recognition algorithm using data integration on a genomic model (PARADIGM).[142] PARADIGM employs a probabilistic graphical model based on factor graphs to infer network modules perturbed in cancer patients through integration of various genomic data. The strength of PARADIGM over other methods is that it integrates various genomic data, ranging from gene expression, copy number data, methylation data, and even known interactions from known signaling pathways.

Bayesian belief networks. A recent formalism, Bayesian belief networks, is recognized as one of the most promising methodologies for prediction under uncertainty.[48,143] Bayesian networks express complex causal relations within the model and predict events based on partial or uncertain data computed by joint probability distributions and conditionals.[144–147] Bayesian networks have been utilized to aid clinical decision-making[148] and to model cellular networks,[149] including genome-wide gene interactions,[150] protein interactions,[151–153] and causal influences in cellular signaling networks.[154] In modeling signal pathway interactions, Bayesian networks not only automatically elucidated most of the traditionally reported signaling relationships but also predicted novel inter-pathway network causalities, which were verified experimentally.[154]

A Bayesian belief network (BBN) is a directed acyclic graph that represents probabilistic relationships among uncertain variables. The graph is made of nodes and arcs where the nodes represent uncertain variables and the arcs the causal/relevance relationships between the variables. Each node is associated with a node probability table (NPT). The NPT captures the conditional probabilities of a node given the value of its parent nodes. For nodes without parents, the NPTs are simply the marginal probabilities or prior distributions. There are several ways to determine the probabilities for the NPTs. We can accommodate both subjective probabilities elicited from domain experts and probabilities based on objective data. Each uncertain variable represents an event or a proposition.

The acyclic structure of Bayesian networks clearly represents the primary cause in the directed graph, which is appealing in predictions. Nevertheless, the number of possible networks is exponential in the number of nodes under consideration, which makes it impossible to evaluate all possible networks. Thus, heuristic searches are used to construct Bayesian networks. Furthermore, it is not always possible to determine the causal relationships between nodes, ie, the direction of the edges, owing to a property known as Markov equivalence.[155,156] More importantly, the acyclic Bayesian network structure was unable to model feedback loops, which are essential in signaling pathways[154] and genetic networks.[157–159] To overcome this limitation, a more complex scheme, dynamic Bayesian networks, was explored for modeling temporal microarray data.[160,161] As an expansion of Bayesian networks, a probabilistic version of the MetaReg model,[162] represented as a factor graph,[163,164] was developed[165] to facilitate changes in the network structure (refinement) and inclusion of additional entities (expansion).[166]

Implication networks. 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.[167,168] When using Dempster–Shafer theory for belief updating, this implication network methodology is termed as a Dempster–Shafer belief network.[169,170] An implication network is a general methodology for reasoning under uncertainty, as are other alternative formalisms such as neural networks,[171,172] dependency networks,[173] Gaussian networks,[174] Mycin's certainty factors,[175] Prospector's inference nets,[176,177] and fuzzy sets.[167] 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.[167,178] 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.

Recently, implication networks have been used to model concurrent coexpression with major disease signaling hallmarks for lung cancer prognostic biomarker identification.[179,180] In these studies, genome-wide coexpression networks specifically associated with different prognostic groups were constructed using implication networks. Candidate genes coexpressed with six or seven major lung cancer signaling hallmarks were identified from these disease-associated genome-wide coexpression networks. These candidate genes were further selected to form prognostic gene signatures using rank-based methods including Cox model, Relief and random forests.[180] The selected biomarker sets form biologically relevant networks when evaluated with curated databases of PPI, chromosome locations, signaling pathways, cis-regulatory motifs/transcription factor binding sites, cancer related gene sets, and gene ontology. This network-based approach identified extensive prognostic gene signatures that outperformed existing ones that were identified using traditional rank-based methods. These results demonstrate that rather than using traditional methods to merely evaluate statistical association with disease outcome, embedding biological relevance into network modeling of the human genome could identify clinically important disease biomarkers.

Approach and Implementation of Implication Networks: Genet

The implication networks can be inducted automatically and dynamically from a dataset by using prediction logic. The structure of the implication network does not represent causal relationship as in the Bayesian network. Instead, it represents implication relationship among the nodes, such as A = >B. Unlike the Bayesian networks that need the complete knowledge of the real-world in order to build the correct causal model once and for all, the implication networks can be constructed dynamically and efficiently based on available data. Therefore, the implication network construction is more flexible than that of the Bayesian networks. The inducted implication network is a directed graph. Each node represents an individual variable or hypothesis. Each arc in the graph signifies the existence of a direct implication (eg, influence) rule between two adjacent nodes. The value of one variable is dependent on the values of all variables that influence it. When evidence from distinct sources is observed for certain nodes, it is combined by the Dempster–Shafer scheme.[181]

Genet[2] is an implementation of the novel implication networks based on prediction logic to construct disease-mediated genome-wide coexpression networks, permitting cyclic relations. To model crosstalk with signaling pathways, Genet allows users to input major disease signaling hallmark proteins for identifying candidate genes that are concurrently coexpressed with these signaling pathways. To identify the final biomarker set, Genet could conveniently link to state-of-the-art feature selection methods, including univariate Cox model, random forests, and the Relief algorithm.

The overall process of identifying gene signatures using Genet is as follows: (1) constructing genome-wide coexpression networks using prediction logic ($P < 0.05$, z-tests) for each disease state or patient group. (2) By comparing the logic relations connecting each pair of genes between the disease-associated coexpression networks, differential network components were obtained, constituting the disease-specific coexpression networks. (3) Selecting candidate genes displaying a direct significant ($P < 0.05$, z-tests) coexpression relation with the specified signaling hallmark genes from the disease-specific coexpression networks. (4) Identifying a final biomarker set (gene signature) from the candidate genes by using *Relief*, random forests, or Cox model.

Since the implication network induction algorithm takes dichotomous variables, the gene expression profiles need to be discretized into binary values; whereas the final step of gene selection and disease classification is performed with the original microarray data. The collection of signaling hallmark genes should be selected according to disease relevance. For example, multiple signaling proteins from the KEGG human non-small cell lung cancer signaling pathways[3] were selected as disease hallmarks to identify coexpressed prognostic gene biomarkers.[180]

Genet is implemented with a combination of C and R, where the C-executable is run through the R interface. This implementation was used as we employed extensive dynamic memory relocation in C to keep track of the derived genome-scale coexpression relations, which makes the package highly efficient in computation time and memory use. The package runs on Windows OS (Windows Vista or higher) with a minimum of 4 GB of RAM. It requires only 40 minutes for an analysis with 20K genes in 256 patient samples. We have linked Genet with Cox model and random forests implemented in R. JavaScript was written to invoke *Relief* implemented in WEKA (22)[1].

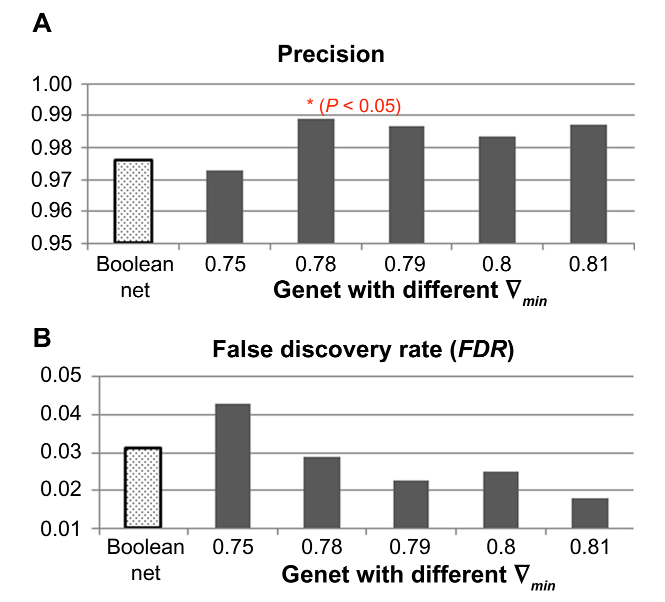

Figure 2. Precision (**A**) and FDR (**B**) of the disease-specific coexpression networks derived with Boolean implication networks and Genet. Genome-wide coexpression networks were constructed for good prognosis and poor prognosis patient groups, respectively, in the training cohort from Shedden et al.[185] The disease-specific networks derived with both models were compared in terms of precision and FDR. An asterisk (*) above the bar indicates that the precision is significantly ($P < 0.05$) higher than the null precisions in 1,000 permutations.

Comparison of Network Models in Cancer Signature Identification

Genet was employed in a few genome-wide coexpression network studies to identify prognostic gene signatures for lung cancer.[179,180] The proposed methodology identified a total of 21 gene signatures[180] that outperformed previously reported ones identified using traditional feature selection methods on the same datasets.[182] Genet was also applied to model smoking-mediated coexpression networks on a selected set of genes associated with lung cancer survival and smoking history. A seven-gene[183] and a six-gene smoking-associated signature[184] were identified for accurate diagnosis and prognosis of lung cancer in smokers.

Next, the biological relevance of the coexpression networks derived with Genet was compared with other network models. Based on five collections of gene sets and pathways from the MSigDB, a coexpression relation was considered a true positive (TP) if the pair of genes satisfy any of the following: (1) present on the same chromosome or cytogenetic band; (2) in the same curated or canonical pathway; (3) share cis-regulator motif, binding motif, or transcription factor binding site; (4) annotated by the same GO term; or (5) within the same computational gene sets mined from cancer-oriented microarray data. The coexpression relation was considered a false positive (FP) if the gene pair does not satisfy all five conditions listed above. If at least one gene in the pair is not annotated, a coexpression relation was labeled as nondiscriminatory (ND). Coexpression relations labeled as ND were excluded in the evaluation as they were not confirmed. Once all relations were labeled, precision (TP/[TP + FP]) and q-value (FP/[TP + FP]) of the disease-mediated coexpression networks were computed. Null distributions of precisions and q-values were generated in 1,000 random permutations of the class labels in the test cohort. From the null statistics, the statistical significance (P) of the precision is indicated by the chance of getting higher precision from the null distribution. The false discovery rate (FDR) of the disease-mediated coexpression networks is the average of the q-value from the null distribution.

Comparison with Boolean networks. On the lung cancer patient cohorts from the Director's Challenge Study,[185] coexpression networks derived with Genet and Boolean implication networks were compared. Results showed that coexpression relations derived from Boolean implication networks did not include many of the major lung cancer hallmarks,

[1] http://wvucancer.org/guoLab/Products

which made it unfeasible to select marker genes coexpressed with multiple signaling pathways.

The large number of relations derived from the implication networks had been a source of concern for false discovery on the derived coexpression relations. This limitation could be overcome by tuning the minimum precision (∇min) parameter in the induction algorithm employed in Genet. In contrast, the Boolean implication network does not provide further information on tuning the parameters. This makes Genet more flexible than the Boolean implication networks. While comparing the networks derived from two methods, ∇min was tuned to be within 0.75 and 0.81 so that the coexpression networks derived from Genet are at a size comparable to those derived from the Boolean implication networks. Results showed that the precisions for the networks derived from both methods were greater than 95%. However, only precision of the implication networks with ∇min = 0.78 was statistically significant ($P < 0.04$). The precision of the implication networks with ∇min = 0.75 was borderline significant ($P < 0.06$) and that of the Boolean implication networks was not significant ($P < 0.21$) (Fig. 2A). On the other hand, the FDR of the derived networks was all less than 5% (Fig. 2B).

Comparison with Bayesian networks. In comparison with the Bayesian networks (Bayesnet) modeled with TETRAD IV[5] for the 21 signatures identified, the disease-specific coexpression networks derived using Genet and Bayesnet have comparably high precisions and low FDR (FDR < 0.1) on the training cohort from the Director's Challenge Study. However, in the more robust approach that is based on the coexpression relations commonly present in the networks derived on the training cohort and the two test cohorts, for all 21 signatures, there was no relation commonly found in the disease-specific coexpression networks derived in all three cohorts using Bayesnet. On the other hand, the relations derived from the training cohort using Genet could be successfully reproduced in both test cohorts with significantly high precision (precision = 1 for 18 signatures; Fig. 3A) and low FDR (FDR < 0.1; Fig. 3B).

Comparison with Pearson's correlation networks. In comparing the smoking-mediated coexpression networks of the signatures and hallmark genes with those derived from the Pearson's correlation coexpression networks, the precisions and FDR are comparable.[184] However, as discussed in the Introduction, the relations represented in Pearson's correlation coexpression network do not describe the directions of the associations.

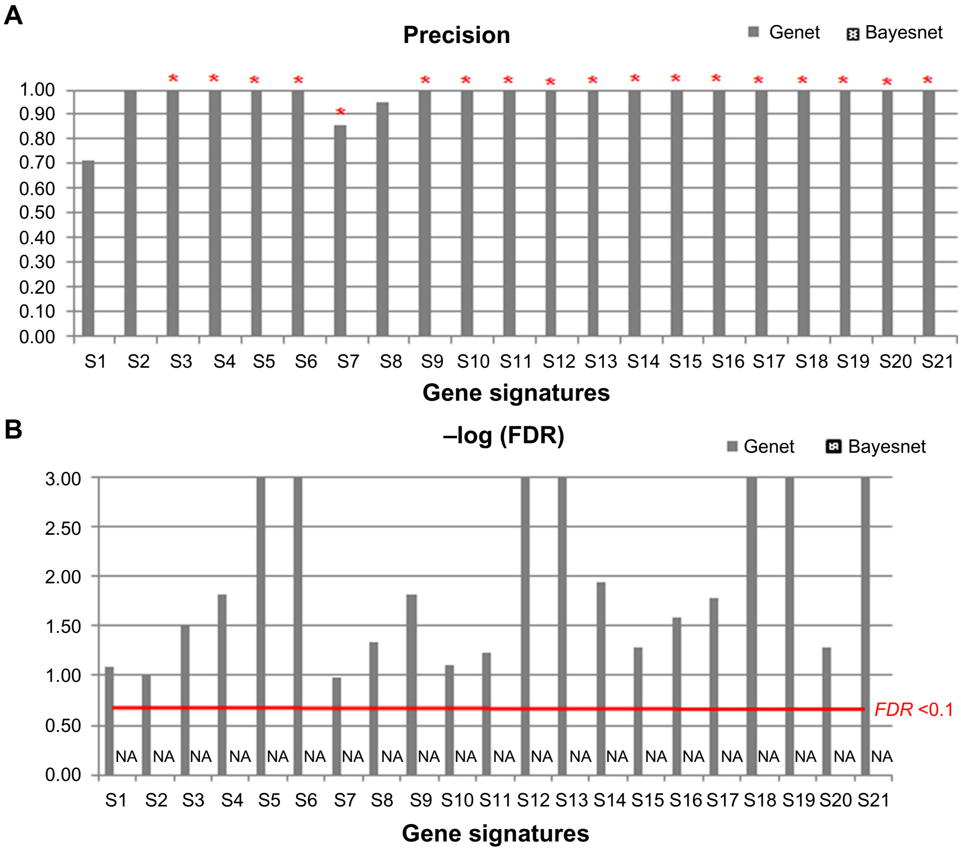

Figure 3. Comparison of the disease-specific coexpression networks derived with Genet and Bayesian networks. Comparisons of the disease-specific coexpression relations validated in two test cohorts in terms of precision (**A**) and FDR (**B**) for the 21 identified prognostic lung cancer gene signatures. For the Bayesian networks, the precision is zero for all 21 gene signatures in (**A**) and the FDR is NA in (**B**) because no coexpression relation was validated by both test cohorts. The asterisk (*) above the bar indicates that the precision is significantly ($P < 0.05$) greater than null precisions in 1,000 permutations.

In contrast, coexpression networks derived with Genet describe both the directions of the regulation between pair of genes.

In the few studies with Genet, mean expression of each gene was used as the cutoff to discretize the gene expression into binary values, which would include all patient samples for the network induction. Instead of using the mean expression values, more stringent statistics, such as mean +/– standard deviation, could be used to partition gene expression into discrete values. However, it would lead to the removal of patient samples that do not meet the predefined threshold.

Conclusions

Unraveling complex molecular interactions and networks and incorporating clinical information in the modeling will present a paradigm shift in molecular medicine. Embedding biological relevance via modeling molecular networks and pathways has become increasingly important for biomarker identification in cancer susceptibility and metastasis studies. As guidance, a few commonly used methods in biomarker identification are summarized in Table 1. In summary, the rank-based methods and regularized models are used when a response variable, ie, clinical outcome, is available; whereas network models would not require any outcome or response variable to be fitted in the model. These methods could be used for different kinds of high-throughput data, including mRNA/miRNA expression from microarrays, mutation from Single Nucleotide Polymorphism (SNP) arrays, and read counts from next-generation sequencing data. Our studies show that a combination of network models and rank-based feature selection methods could identify gene signatures with accurate diagnostic and prognostic performance, and reveal biologically relevant molecular networks. In this review, multiple network-based models were evaluated in several case studies, with implication networks outperforming Bayesian belief networks, Boolean networks, and Pearson's correlation coexpression networks.

Acknowledgments

We thank Rebecca Raese for editing this paper.

Author Contributions

Conceived and designed the experiments: NLG. Analyzed the data: NLG, YW. Wrote the first draft of the manuscript: NLG. Contributing to the writing of the manuscript: YW. Agree with manuscript results and conclusions: NLG, YW. Jointly developed the structure and arguments for the paper: NLG, YW. Made critical revisions and approved final version: NLG, YW. Both authors reviewed and approved of the final manuscript.

REFERENCES

1. Ideker T, Sharan R. Protein networks in disease. *Genome Res*. 2008;18(4):644–52.
2. Han JD. Understanding biological functions through molecular networks. *Cell Res*. 2008;18(2):224–37.

Table 1. Summary of commonly used methods for biomarker identification.

METHOD	INPUT HIGH-THROUGHPUT DATA*	RESPONSE VARIABLE
Rank-based methods		
SAM[49]	Expression	Comparison of two conditions
Random forests[24–26]	Expression	Gaussian
Cox model[19]	Expression, Mutation	Survival time and Event
F-statistics[51]	Mutation	Comparison of two conditions
Regularized models		
Lasso[56]	Expression, Mutation, Read counts	Gaussian
Elastic net[57]	Expression, Mutation, Read counts	Gaussian
Network-constraint regularized model[59]	Expression, Mutation, Read counts, Gene-interaction Network	Gaussian
Network-constraint logistic model[60]	Expression, Mutation, Read counts, Gene-interaction Network	Binary
Network based methods		
Graphical Gaussian Models[130,131]	Expression	Not required
Poisson graphical models[135–137]	Read counts	Not required
Multi-Dentrix[141]	Mutation	Not required
PARADIGM[142]	Integration of Expression, Mutation, and Biological function and interaction data	Not required
Bayesian belief networks[149]	Expression	Not required
Implication networks[179,180]	Expression, Mutation	Not required

Notes: *Expression, mRNA/miRNA expression profiled with microarray. Mutation, mutation profiled with SNP array or next-generation sequencing. Read counts, mRNA/miRNA expression profiled with next-generation sequencing.

3. Muller FJ, Laurent LC, Kostka D, et al. Regulatory networks define phenotypic classes of human stem cell lines. *Nature*. 2008;455(7211):401–5.
4. Slavov N, Dawson KA. Correlation signature of the macroscopic states of the gene regulatory network in cancer. *Proc Natl Acad Sci U S A*. 2009; 106(11):4079–84.
5. Taylor IW, Linding R, Warde-Farley D, et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol*. 2009;27(2):199–204.
6. Segal E, Friedman N, Kaminski N, Regev A, Koller D. From signatures to models: understanding cancer using microarrays. *Nat Genet*. 2005;37(suppl):S38–45.
7. Kitano H. A robustness-based approach to systems-oriented drug design. *Nat Rev Drug Discov*. 2007;6(3):202–10.
8 Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. *Nat Biotechnol*. 2007;25(10):1119–26.
9. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*. 2008;4(11):682–90.
10. Altieri DC. Survivin, cancer networks and pathway-directed drug discovery. *Nat Rev Cancer*. 2008;8(1):61–70.
11. Araujo RP, Liotta LA, Petricoin EF. Proteins, drug targets and the mechanisms they control: the simple truth about complex networks. *Nat Rev Drug Discov*. 2007;6(11):871–80.
12. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol*. 2008;9(10):770–80.
13. Sotiriou C, Piccart MJ. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat Rev Cancer*. 2007;7(7):545–53.
14. Quackenbush J. Microarray analysis and tumor classification. *N Engl J Med*. 2006;354(23):2463–72.
15. Dalton WS, Friend SH. Cancer biomarkers – an invitation to the table. *Science*. 2006;312(5777):1165–8.
16. Gold C, Holub A, Sollich P. Bayesian approach to feature selection and parameter tuning for support vector machine classifiers. *Neural Netw*. 2005;18(5–6):693–701.
17. Huang TM, Kecman V. Gene extraction for cancer diagnosis by support vector machines-An improvement. *Artif Intell Med*. 2005;35(1–2):185–94.
18. Xiong M, Li W, Zhao J, Jin L, Boerwinkle E. Feature (gene) selection in gene expression-based tumor classification. *Mol Genet Metab*. 2001;73(3):239–47.
19. Cox D. Regression models and life-tables (with discussion). *J R Stat Soc B Methodol*. 1972;34:187–220.
20. Theodoridis S, Koutroumbas K. *Pattern Recognition*. 3rd ed. San Diego: Academic Press; 2006.
21. Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans Knowl Data Eng*. 2003;15(3):1437–47.
22. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann; 2005.
23. Kim SJ, Magnani A, Boyd SP. Robust fisher discriminant analysis. In: Weiss Y, Schoch G, Platt J eds. *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press; 2006:659–66.
24. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
25. Jiang H, Deng Y, Chen HS, et al. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*. 2004;5:81.
26. Diaz-Uriarte R, Alvarez DA. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*. 2006;7:3.
27. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*. 1999;96(6):2907–12.
28. Klampfl S, Legenstein R, Maass W. Information bottleneck optimization and independent component extraction with spiking neurons. In: Scholkopf B, Platt J, Hoffman PC eds. *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press; 2007:713–20.
29. Lindgren JT, Hyvärinen A. Emergence of conjunctive visual features by quadratic independent component analysis. In: Scholkopf B, Platt J, Hoffman PC eds. *Advances in Neural Information Processing Systems* 19. Cambridge, MA: MIT Press; 2007:897–904.
30. Theis F. Towards a general independent subspace analysis. In: Scholkopf B, Platt J, Hoffman PC eds. *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press; 2007:1361–8.
31. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*. 2002;18(1):39–50.
32. Braun M, Buhmann J, Müller K. Denoising and Dimension Reduction in Feature Space. In: Scholkopf B, Platt J, Hoffman T, editors. *Advances in Neural Information Processing Systems 19*. MIT Press; 2007:185–92.
33. Moghaddam B, Weiss Y, Avidan S. Spectral bounds for sparse PCA: exact and greedy algorithms. In: Weiss Y, Schoch G, Platt J eds. *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press; 2005:915–22.
34. Opper M. An approximate inference approach for the PCA reconstruction error. In: Weiss Y, Schoch G, Platt J eds. *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press; 2005:1035–42.
35. Schraudolph NN, Günter S, Viswanadhan VN. Fast Iterative Kernel PCA. In: Schoch G, Platt J, Hoffman T, editors. *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press; 2007:1225–32.
36. Zwald L, Blanchard G. On the convergence of eigenspaces in kernel principal components analysis. In: Weiss Y, Schoch G, Platt J eds. *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press; 2005: 1649–56.
37. Li K. Sliced inverse regression for dimension reduction. *J Am Stat Assoc*. 1991;86:316–42.
38. Li W, Sun F, Grosse I. Extreme value distribution based gene selection criteria for discriminant microarray data analysis using logistic regression. *J Comput Biol*. 2006;11(2–3):215–26.
39. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. *IEEE Trans Pattern Anal Mach Intell*. 2000;22(1):4–37.
40. Good AC. Novel DOCK clique driven 3D similarity database search tools for molecule shape matching and beyond: adding flexibility to the search for ligand kin. *J Mol Graph Model*. 2007;26:656–66.
41. Karpenko O, Shi J, Dai Y. Prediction of MHC class II binders using the ant colony search strategy. *Artif Intell Med*. 2005;35(1–2):147–56.
42. Pratt SC, Sumpter DJ. A tunable algorithm for collective decision-making. *Proc Natl Acad Sci U S A*. 2006;103(43):15906–10.
43. Baluja S. Genetic algorithms and explicit search statistics. In: Mozer MC, Jordan M, Petsche T eds. *Advances in Neural Information Processing Systems 9*. Cambridge, MA: MIT Press; 1997:319–25.
44. Juels A, Wattenberg M. Stochastic hill climbing as a baseline method for evaluating generic algorithms. In: Touretzky DS, Mozer MC, Hasselmo ME eds. *Advances in Neural Information Processing Systems 8*. Cambridge, MA: MIT Press; 1996:430–6.
45. Falk CT. Preliminary ordering of multiple linked loci using pairwise linkage data. *Genet Epidemiol*. 1992;9(5):367–75.
46. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. *Science*. 1983;220(4598):671–80.
47. Rodrigo G, Carrera J, Jaramillo A. Genetdes: automatic design of transcriptional networks. *Bioinformatics*. 2007;23(14):1857–8.
48. Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. 2nd ed. Englewood Cliffs: Prentice Hall; 2003.
49. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*. 2001;98(9): 5116–21.
50. Raza M, Gondal I, Green D, Coppel RL. Feature Selection and Classification of Gene Expression Profile in Hereditary Breast Cancer. *Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*. 2004. pp 315–20.
51. Dudoit S, Shafer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci*. 2003;18(1):71–103.
52. Dettling M, Buhlmann P. Boosting for tumor classification with gene expression data. *Bioinformatics*. 2003;19(9):1061–9.
53. Dudoit S, Yang Y, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sin*. 2002;12:111–39.
54. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *J Comput Biol*. 2000;7(6):819–37.
55. Wolfinger RD, Gibson G, Wolfinger ED, et al. Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol*. 2001;8(6):625–37.
56. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc B Methodol*. 1996;58(1):267–88.
57. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B Stat Methodol*. 2005;67:301–20.
58. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat*. 2009;37(4):1733–51.
59. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*. 2008;24(9):1175–82.
60. Zhang W, Wan YW, Allen GI, Pang K, Anderson ML, Liu Z. Molecular pathway identification using biological network-regularized logistic models. *BMC Genomics*. 2013;14(suppl 8):S7.
61. Baker SG, Kramer BS, Srivastava S. Markers for early detection of cancer: statistical guidelines for nested case-control studies. *BMC Med Res Methodol*. 2002;2:4.
62. Emir B, Wieand S, Su JQ, Cha S. Analysis of repeated markers used to predict progression of cancer. *Stat Med*. 1998;17(22):2563–78.
63. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159(9):882–90.
64. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature*. 1999;402(6761 suppl):C47–52.
65. Franke L, van BH, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*. 2006;78(6):1011–25.

66. Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. *J Med Genet*. 2006;43(8):691–8.
67. Mani KM, Lefebvre C, Wang K, et al. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol*. 2008;4:169.
68. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci U S A*. 2008;105(11):4323–8.
69. Ortutay C, Vihinen M. Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res*. 2009;37(2):622–8.
70. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. *Mol Syst Biol*. 2008;4:189.
71. Calvano SE, Xiao W, Richards DR, et al. A network-based analysis of systemic inflammation in humans. *Nature*. 2005;437(7061):1032–7.
72. Ghazalpour A, Doss S, Zhang B, et al. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet*. 2006;2(8):e130.
73. Goehler H, Lalowski M, Stelzl U, et al. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell*. 2004;15(6):853–65.
74. Lim J, Hao T, Shaw C, et al. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*. 2006;125(4):801–14.
75. Pujana MA, Han JD, Starita LM, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet*. 2007;39(11):1338–49.
76. Sanchez I, Mahlke C, Yuan J. Pivotal role of oligomerization in expanded polyglutamine neurodegenerative disorders. *Nature*. 2003;421(6921):373–9.
77. Limviphuvadh V, Tanaka S, Goto S, Ueda K, Kanehisa M. The commonality of protein interaction networks determined in neurodegenerative disorders (NDDs). *Bioinformatics*. 2007;23(16):2129–38.
78. Li CY, Mao X, Wei L. Genes and (common) pathways underlying drug addiction. *PLoS Comput Biol*. 2008;4(1):e2.
79. Guo Z, Li Y, Gong X, et al. Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*. 2007;23(16):2121–8.
80. He L, He X, Lim LP, et al. A microRNA component of the p53 tumour suppressor network. *Nature*. 2007;447(7148):1130–4.
81. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. *Nat Genet*. 2007;39(10):1202–7.
82. Goring HH, Curran JE, Johnson MP, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet*. 2007;39(10):1208–16.
83. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. *Nat Genet*. 2007;39(10):1217–24.
84. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat Rev Genet*. 2009;10(3):184–94.
85. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol Syst Biol*. 2007;3:140.
86. Efroni S, Schaefer CF, Buetow KH. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One*. 2007;2(5):e425.
87. Glass L, Kauffman SA. The logical analysis of continuous, non-linear biochemical control networks. *J Theor Biol*. 1973;39(1):103–29.
88. Thomas R. Boolean formalization of genetic control circuits. *J Theor Biol*. 1973;42(3):563–85.
89. Davidson EH, Rast JP, Oliveri P, et al. A genomic regulatory network for development. *Science*. 2002;295(5560):1669–78.
90. Smith J, Theodoris C, Davidson EH. A gene regulatory network subcircuit drives a dynamic pattern of gene expression. *Science*. 2007;318(5851):794–7.
91. Kingsmore SF. Multiplexed protein measurement: technologies and applications of protein and antibody arrays. *Nat Rev Drug Discov*. 2006;5(4):310–21.
92. Yeung MK, Tegner J, Collins JJ. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A*. 2002;99(9):6163–8.
93. Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z. Reconstructing dynamic regulatory maps. *Mol Syst Biol*. 2007;3:74.
94. Ciliberti S, Martin OC, Wagner A. Innovation and robustness in complex regulatory gene networks. *Proc Natl Acad Sci U S A*. 2007;104(34):13591–6.
95. Shi Y, Klustein M, Simon I, Mitchell T, Bar-Joseph Z. Continuous hidden process model for time series expression experiments. *Bioinformatics*. 2007;23(13):i459–67.
96. Cao Y, Liang J. Optimal enumeration of state space of finitely buffered stochastic molecular networks and exact computation of steady state landscape probability. *BMC Syst Biol*. 2008;2:30.
97. Klipp E, Nordlander B, Kruger R, Gennemark P, Hohmann S. Integrative model of the response of yeast to osmotic shock. *Nat Biotechnol*. 2005;23(8):975–82.
98. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*. 2004;429(6987):92–6.
99. Herrgard MJ, Lee BS, Portnoy V, Palsson BO. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res*. 2006;16(5):627–35.
100. Shlomi T, Eisenberg Y, Sharan R, Ruppin E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol*. 2007;3:101.
101. Chechik G, Oh E, Rando O, Weissman J, Regev A, Koller D. Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nat Biotechnol*. 2008;26(11):1251–9.
102. Dunlop MJ, Cox RS III, Levine JH, Murray RM, Elowitz MB. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nat Genet*. 2008;40(12):1493–8.
103. Gibson M, Bruck J. Efficient exact stochastic simulation of chemical systems with many species and many channels. *J Phys Chem*. 1999;104:1876–89.
104. Gillespie D. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comp Phys*. 1976;22:403–34.
105. Gillespie D. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*. 1977;81:2340–61.
106. Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics*. 1998;149(4):1633–48.
107. Gonze D, Goldbeter A. Circadian rhythms and molecular noise. *Chaos*. 2006;16(2):026110.
108. Niemitalo O, Neubauer A, Liebal U, Myllyharju J, Juffer AH, Neubauer P. Modelling of translation of human protein disulfide isomerase in *Escherichia coli*-A case study of gene optimisation. *J Biotechnol*. 2005;120(1):11–24.
109. Schultz D, Ben JE, Onuchic JN, Wolynes PG. Molecular level stochastic model for competence cycles in Bacillus subtilis. *Proc Natl Acad Sci U S A*. 2007;104(45):17582–7.
110. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer DV. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell*. 2005;122(2):169–82.
111. Kauffman S, Peterson C, Samuelsson B, Troein C. Random Boolean network models and the yeast transcriptional network. *Proc Natl Acad Sci U S A*. 2003;100(25):14796–9.
112. Li F, Long T, Lu Y, Ouyang Q, Tang C. The yeast cell-cycle network is robustly designed. *Proc Natl Acad Sci U S A*. 2004;101(14):4781–6.
113. Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac Symp Biocomput*. 1999:17–28. PMID 10380182.
114. Lähdesmäki H, Shmulevich I, Yli-Harja O. On learning gene regulatory networks under the Boolean network model. *Mach Learn*. 2003;52:147–67.
115. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*. 1998;9(12):3273–97.
116. Albert R, Othmer HG. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J Theor Biol*. 2003;223(1):1–18.
117. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002;18(2):261–74.
118. Shmulevich I, Gluhovsky I, Hashimoto RF, Dougherty ER, Zhang W. Steady-state analysis of genetic regulatory networks modelled by probabilistic boolean networks. *Comp Funct Genomics*. 2003;4(6):601–8.
119. Steggles LJ, Banks R, Shaw O, Wipat A. Qualitatively modelling and analysing genetic regulatory networks: a Petri net approach. *Bioinformatics*. 2007;23(3):336–43.
120. Peterson J. *Petri Net Theory and the Modeling of Systems*. New Jersey: Prentice Hall PTR; 1981.
121. Koch I, Schueler M, Heiner M. STEPP – search tool for exploration of Petri net paths: a new tool for Petri net-based path analysis in biochemical networks. *In Silico Biol*. 2005;5(2):129–37.
122. Kuffner R, Zimmer R, Lengauer T. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*. 2000;16(9):825–36.
123. Reddy VN, Liebman MN, Mavrovouniotis ML. Qualitative analysis of biochemical reaction systems. *Comput Biol Med*. 1996;26(1):9–24.
124. Simao E, Remy E, Thieffry D, Chaouiya C. Qualitative modelling of regulated metabolic pathways: application to the tryptophan biosynthesis in *E. Coli*. *Bioinformatics*. 2005;21(suppl 2):ii190–6.
125. Chaouiya C, Remy E, Ruet P, Thieffry D. Proceedings of the 25th International Conference on Applications and Theory of Petri Nets. Berlin: Springer; 2004.
126. Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*. 2003;34(2):166–76.
127. Sahoo D, Dill DL, Gentles AJ, Tibshirani R, Plevritis SK. Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol*. 2008;9(10):R157.

128. Sahoo D, Seita J, Bhattacharya D, et al. MiDReG: a method of mining developmentally regulated genes using Boolean implications. *Proc Natl Acad Sci U S A*. 2010;107(13):5732–7.
129. Lauritzen SL. *Graphical Models*. Oxford: Clarendon Press; 1996.
130. Dobra A, Jones B, Hans C, Nevis J, West M. Sparse graphical models for exploring gene expression data. *J Multivar Anal*. 2004;90:196–212.
131. Wille A, Zimmermann P, Vranova E, et al. Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol*. 2004;5(11):R92.
132. Wang J, Nygaard V, Smith-Sorensen B, Hovig E, Myklebost O. MArray: analysing single, replicated or reversed microarray experiments. *Bioinformatics*. 2002;18(8):1139–40.
133. Waddell PJ, Kishino H. Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Inform Ser Workshop Genome Inform*. 2000;11:129–40.
134. Li H, Gui J. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*. 2006;7(2):302–17.
135. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509–17.
136. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010;11:94.
137. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*. 2012;13(3):523–38.
138. Karlis D. An EM algorithm for multivariate Poisson distribution and related models. *J Appl Stat*. 2003;30(1):63–77.
139. Yang E, Ravikumar P, Allen G, Liu Z. Graphic models via generalized linear models. Advances in Neural Information Processing Systems 25 (NIPS 2012), Oral Presentation; 2012.
140. Allen GI, Liu Z. A Log-Linear Graphical Model for Inferring Genetic Networks from High-Throughput Sequencing Data. *IEEE International Conference on Bioinformatics and Biomedicine*. 2012. Philadelphia, USA.
141. Leiserson MD, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol*. 2013;9(5):e1003054.
142. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010;26(12):i237–45.
143. Fenton NE, Neil M. A critique of software defect prediction models. *IEEE Trans Software Eng*. 1999;25(5):675–89.
144. Heckerman D. *Probabilistic Similarity Networks*. Cambridge, MA: The MIT Press; 1991.
145. Henrion M. Propagating uncertainty in Bayesian networks by probabilistic logic sampling. *Proceedings of Uncertainty in Artificial Intelligence*, 4, 1988;149–63. Minneapolis, MN, USA. Published By: AUAI Press, Corvallis, Oregon.
146. Lauritzen S, Spiegelhalter D. Local computations with probabilities on graphical structures and their applications to expert systems. *J R Stat Soc B*. 1988;50(157):224.
147. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann; 1988.
148. Rodin AS, Boerwinkle E. Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma apoE levels). *Bioinformatics*. 2005;21(15):3273–8.
149. Friedman N. Inferring cellular networks using probabilistic graphical models. *Science*. 2004;303(5659):799–805.
150. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol*. 2000;7(3–4):601–20.
151. Jansen R, Yu H, Greenbaum D, et al. Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*. 2003;302(5644):449–53.
152. Yeang CH, Ideker T, Jaakkola T. Physical network models. *J Comput Biol*. 2004;11(2–3):243–62.
153. Yeang CH, Vingron M. A joint model of regulatory and metabolic networks. *BMC Bioinformatics*. 2006;7:332.
154. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*. 2005;308(5721):523–9.
155. Zhu J, Wiener MC, Zhang C, et al. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol*. 2007;3(4):e69.
156. Zhu J, Zhang B, Smith EN, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet*. 2008;40(7):854–61.
157. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science*. 2002;298(5594):824–7.
158. Milo R, Itzkovitz S, Kashtan N, et al. Superfamilies of evolved and designed networks. *Science*. 2004;303(5663):1538–42.
159. Wuchty S, Oltvai ZN, Barabasi AL. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet*. 2003;35(2):176–9.
160. Ong IM, Glasner JD, Page D. Modelling regulatory pathways in *E. coli* from time series expression profiles. *Bioinformatics*. 2002;18(suppl 1):S241–8.
161. Pe'er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*. 2001;17(suppl 1):S215–24.
162. Gat-Viks I, Tanay A, Shamir R. Modeling and analysis of heterogeneous regulation in biological networks. *J Comput Biol*. 2004;11(6):1034–49.
163. Kschischang FR, Frey BJ, Loeliger HA. Factor graphs and the sum-product algorithm. *IEEE Trans Info Theory*. 2001;47:498–519.
164. MacKay DJC. Introduction to Monte Carlo Methods in Learning in Graphical Models. New York: Kluwer Academic Press; 1998.
165. Gat-Viks I, Tanay A, Raijman D, Shamir R. A probabilistic methodology for integrating knowledge and experiments on biological networks. *J Comput Biol*. 2006;13(2):165–81.
166. Gat-Viks I, Shamir R. Refinement and expansion of signaling pathways: the osmotic response network in yeast. *Genome Res*. 2007;17(3):358–67.
167. Desmarais MC, Maluf A, Liu J. User-expertise modeling with empirically derived probabilistic implication networks. *User Model User Adapted Interact*. 1996;5(3–4):283–315.
168. Desmarais MC, Meshkinfam P, Gagnon M. Learned Student Models with Item to Item Knowledge Structures. *User Model User Adapted Interact*. 2006;16(5):403–34.
169. Liu J, Desmarais MCA. Method of learning implication networks from empirical data: algorithm and Monte-Carlo simulation-based validation. *IEEE Trans Knowl Data Eng*. 1997;9(6):990–1004.
170. Liu J, Maluf D, Desmarais MC. A new uncertainty measure for belief networks with applications to optimal evidential inferencing. *IEEE Trans Knowl Data Eng*. 2001;13(3):416–25.
171. Honavar V, Uhr L. Artificial Intelligence and Neural Networks: Steps Toward Principled Integration. New York, NY: Academic Press; 1994.
172. O'Neill MC, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. *BMC Bioinformatics*. 2003;4:13.
173. Heckerman D, Chickering DM, Meek C, Rounthwaite R, Kadie C. Dependency networks for inference, collaborative filtering, and data visualization. *J Mach Learn Res*. 2001;1:49–75.
174. Kundu S, Sorensen DC, Phillips GN Jr. Automatic domain decomposition of proteins by a Gaussian network model. *Proteins*. 2004;57(4):725–33.
175. Cruz GP, Beliakov G. On the interpretation of certainty factors in expert systems. *Artif Intell Med*. 1996;8(1):1–14.
176. Campbell AN, Hollister VF, Duda RO, Hart PE. Recognition of a hidden mineral deposit by an artificial intelligence program. *Science*. 1982;217(4563):927–9.
177. Finlay AY, Sinclair J, Alty JL. Expert system diagnosis of ichthyosis. *Clin Exp Dermatol*. 1987;12(3):239–40.
178. Falmagne JC, Doignon JP, Koppen M, Villano M, Johannesen L. Introduction to knowledge spaces: how to build, test and search them. *Psychol Rev*. 1990;97(2):201–24.
179. Guo NL, Wan YW, Bose S, Denvir J, Kashon ML, Andrew ME. A novel network model identified a 13-gene lung cancer prognostic signature. *Int J Comput Biol Drug Des*. 2011;4(1):19–39.
180. Wan YW, Beer DG, Guo NL. Signaling pathway-based identification of extensive prognostic gene signatures for lung adenocarcinoma. *Lung Cancer*. 2012;76(1):98–105.
181. Shafer G. *A Mathematical Theory of Evidence*. Princeton: Princeton University Press; 1976.
182. Wan YW, Beer DG, Guo NL. Signaling pathway-based identification of extensive prognostic gene signatures for lung adenocarcinoma. *Lung Cancer*. 2011;76(1):98–105.
183. Wan YW, Raese RA, Fortney JE, et al. A smoking-associated 7-gene signature for lung cancer diagnosis and prognosis. *Int J Oncol*. 2012;41(4):1387–96.
184. Guo NL, Wan YW. Pathway-based identification of a smoking associated 6-gene signature predictive of lung cancer risk and survival. *Artif Intell Med*. 2012;55(2):97–105.
185. Shedden K, Taylor JM, Enkemann SA, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med*. 2008;14(8):822–7.