[Frontiers in Bioscience 13, 263-275, January 1, 2008]

Inferring regulatory networks

Huai Li [1], **Jianhua Xuan**[2], **Yue Wang**[2], **Ming Zhan**[1]

[1]*Bioinformatics Unit, Branch of Research Resources, National Institute on Aging, NIH, Baltimore, MD, USA,* [2]*Department of Electrical, Computer, and Biomedical Engineering, Virginia Polytechnic Institute and State University, Arlington, VA, USA*

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Computational approaches for identifying gene modules
3.1. Advanced statistical approaches
3.2. Matrix decomposition approaches
4. Computational approaches for inferring gene connectivity
4.1. ODE-based models
4.2. Bayesian networks
4.3. Coexpression networks
4.4. Probabilistic boolean networks
4.5. Inference from multiple sources of data
5. Network analysis in silico
5.1. Steady state analysis by markov chain simulation
5.2. Intervention analysis by markov chain model
6. Closing remarks
7. Acknowledgments
8. References

1. ABSTRACT

The discovery of regulatory networks is an important aspect in the post genomic research. The process requires integrated efforts of experimental and computational strategies by employing the systems biology approach. This review summarizes some of the major themes in computational inference of regulatory networks based on gene expression and other data sources, including transcriptional module identification, network topology inference, and network analysis. Popular solutions to each of these problems and their relative merits are discussed.

2. INTRODUCTION

The most important and widespread mechanism used by cells to regulate molecular functions or biological process is the coordinate transcriptional and post-transcriptional network of the interacting genes or their products. To understand how physiological and pathological phenotypes arise from gene regulatory networks is a major challenge in post genomic research and requires computational systems biology approaches (1-3). Systems biology is an emergent field that aims at system-level understanding of biological systems (1, 2). Genes in regulatory networks are often connected through interlocking positive and negative loops. An intuitive understanding of the structure and dynamics of the network is difficult to obtain. Using systems' approaches, such as mathematical modeling and *in silico* simulation study, the structure of regulatory networks can be described precisely and their dynamic behavior can be predicted in a systematic way (4, 5). The systems study on the regulatory network can directly benefit the identification of biomarker for drug discovery and the development of effective preventive and therapeutic intervention in disease or aging (6-8).

Various high-throughput technologies like microarrays (9) and factor-binding profiling (10) provide researchers with valuable resources for elucidating how genes interact with each other and how a cell's regulatory networks control vast batteries of genes simultaneously. Recently, many computational algorithms (6, 11-41) have been proposed to unravel regulatory networks from gene

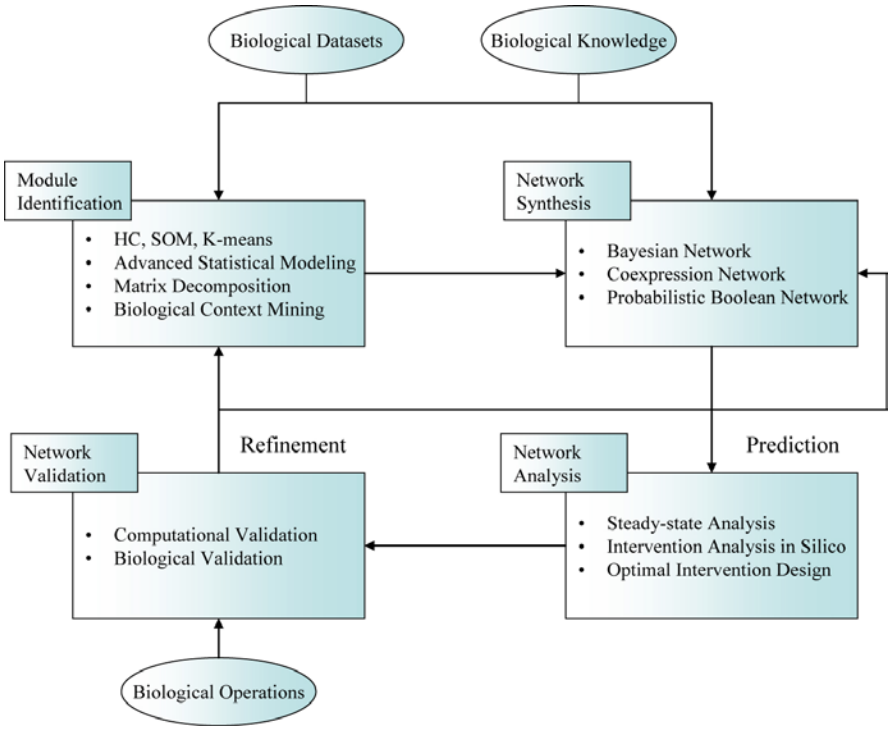

Figure 1. Flow diagram for inferring regulatory networks.

expression or other "omics" data. The *in silico* generated hypothetical models are further tested against biological experiments or published reports. Even though, two major challenges remain in inferring regulatory networks from large "omics" data sets: the statistical limitations posed by these data sets (e.g. few samples but larger number of genes in most microarray data sets) and the high computational complexity due to large and complex structures of regulatory networks. Analysis by integrating multiple data sources (e.g. DNA sequences, protein-protein interactions, protein structural information, and protein-DNA binding data) shows promise to overcome the statistical limitations existed in an individual data set (37, 42). On the other hand, there is growing evidence that suggests a multi-scale and hierarchical modular architecture in biological networks (30, 43-46). This is consistent with the fact that many of these networks exhibit a scale-free topology (30, 47). In the context of genetic networks, this implies that genes form small clusters or modules within each of which the constituent genes have close interactions; some of these clusters form larger 'meta-clusters' that themselves exhibit interactions and this process may continue on several different scales. Therefore, regulatory networks may be broken down to sub-network with small number of genes, and each sub-network can be separately modeled (46). This will decrease the computational complexity.

In this review, we summarize some of the major themes in inferring regulatory networks, including gene module identification, network topology inference, and network dynamics analysis. Figure 1 illustrates an overall flow diagram for inferring regulatory networks. Important aspects of the network inference are discussed in this paper.

3. COMPUTATIONAL APPROACHES FOR IDENTIFYING GENE MODULES

Genes with coordinate activities for certain biological functions often have tightly regulated interactions and form contextual modules. It is important to identify such regulatory network structures for understanding the biological events associated with different experimental conditions and identifying gene expression signatures.

3.1. Advanced Statistical Approaches

Clustering of genes and clustering of experiments are unsupervised modeling approaches that are in common use for identifying the co-regulated type of local patterns (48-50). Clustering methods consider only correlative or linear relationships between genes, so that often fail to capture the contextual modularity that might result from highly nonlinear interactions among genes. Clustering methods also partition genes into mutually exclusive clusters, but in reality a gene may be parts of several different biological processes. More biologically meaningful modules can be uncovered by employing more sophisticated algorithms recently developed, using multiple sources of data, or by integrating the algorithms with prior biological knowledge (13, 15, 21, 34, 51, 52).

Segal *et al* proposed a class of probabilistic graphical models, module networks, for inferring regulatory modules from gene expression data (13). In this

framework, a regulatory module is a set of genes that are regulated in concert by a shared regulation program. The regulation program specifies the behavior of the genes in the module as a function of the express level of regulators. Clearly, this approach relies on the assumption that the expression levels of regulated genes depend on the expression levels of regulators. The method was demonstrated for its ability to generate detailed testable hypotheses relating to both regulatory modules and their control programs. The experimental results supported their computationally generated hypotheses, suggesting regulatory roles for previously uncharacterized proteins.

Bar-Joseph *et al* described an algorithm that uses genomic expression and transcription factor binding data to discover transcriptional modules (15). The algorithm performs an efficient exhaustive search over all possible combinations of transcription factors implied by the protein-DNA interaction data. Once a set of genes bound by a common set of transcription factors is found, the algorithm proceeds to find a smaller subset of genes that are co-expressed. The algorithm then seeks to add additional genes to the module that are similarly expressed and considered bound by the same set of transcription factors. They applied their algorithm to 106 yeast transcription factors profiled in rich medium conditions and yeast expression data from over 500 experiments. The results indicated that the algorithm can assign group of genes to regulators more accurately by integrating the binding information.

Zhou *et al* introduced an approach, termed second-order expression analysis, for the identification of transcriptional modules (34). They defined the first-order expression analysis as the extraction of expression patterns from one microarray data set. They then proposed the second-order expression analysis as a study of the correlated occurrences of those expression patterns across multiple data sets measured under different types of conditions. Using yeast as a model system, they demonstrated that the second-order analysis can identify genes of the same function yet without coexpression patterns. The approach could also reveal network relationships among different transcriptional modules.

Wang *et al* (52-54) developed an algorithm called 'visual and statistical data analyzer' (VISDA) for gene cluster discovery and visualization. VISDA uses a hierarchical normal mixture model to approximate the overall distribution of gene expression data. Based on the model, genes can be partitioned into clusters and sub-clusters hierarchically. VISDA also incorporate human interaction into the clustering process that makes it unique in comparison with other methods. VISDA has recently been adopted as one of the core data analysis components by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) on its new cancer biomedical informatics grid (caBIG^{TM}) initiative (55).

3.2. Matrix Decomposition Approaches

Matrix decomposition methods have been recently introduced for uncovering transcriptional modules from microarray data. These methods treat microarray data as a mixture of unknown signals that may correspond to specific biological sources. These methods do not assume that genes with similar expression profiles share the same pathway or similar functions. The methods can also partition genes to mutually inclusive modules to reflect the fact that genes may have multiple functions or are active in multiple biological processes. A variety of matrix decomposition methods have been proposed for microarray data analysis, including singular value decomposition (56, 57), independent components analysis (58-60), non-negative matrix factorizations (61-65), network component analysis (28), and probabilistic sparse matrix factorization (66).

Alter *et al* described the use of singular value decomposition (SVD) in transforming gene expression data from a genes × arrays space to a reduced diagonalized "eigengenes" × "eigenarrays" space (56). The eigengenes and eigenarrays are orthonormal superpositions of the genes and arrays. Sorting the data according to the correlations of the genes (and arrays) with eigengenes (and eigenarrays) gives a global picture of the dynamics of gene expression. With yeast cell-cycle data sets, they showed that the SVD method can classify individual genes and arrays into groups of similar regulation and function, or similar cellular state and biological phenotype, respectively.

Independent components analysis (ICA) is a statistical method for revealing hidden factors that underlie sets of random variables or signals. Lee *et al* applied ICA to project microarray data into statistically independent components that correspond to putative biological processes, and to cluster genes according to over- or under-expression in each component (58). The results showed that ICA outperforms other clustering methods, such as principal component analysis, *k*-means clustering, in constructing functionally coherent clusters on microarray datasets from yeast, C. elegans and human. Similarly, Frigyesi *et al* applied ICA to two tumor data sets and one time series experiment (60). They used an iterated ICA algorithm to estimate independent components and proposed a scheme to identify those genes that have significant loadings in each component. The results demonstrated that ICA can identify gene clusters with high biological relevance compared with results based on correlated expression.

Non-negative matrix factorization (NMF) is a recently developed machine learning technique, capable of finding smaller and more localized patterns as well as global patterns. The approach can be particularly useful in identifying biological subsystems (i.e., sets of genes that function in concert in a relatively tightly regulated manner) (61, 62). Kim *et al* (61) applied the NMF approach to a large data set consisting of 300 genome-wide expression measurements of yeast to identify the cellular subsystems. The results showed that local features detected by NMF were mapped well to functional cellular subsystems. Most recently, Wang *et al* developed a algorithm, least squares non-negative matrix factorization (LS-NMF), for

integrating uncertainty measurements of gene expression data into NMF updating rules (65). The LS-NMF algorithm maintains the advantages of the original NMF algorithm but exceeds NMF significantly in terms of identifying functionally related genes as determined from annotations in the MIPS database.

Dueck *et al* proposed a probabilistic sparse matrix factorization (PSMF) model and variational Bayesian learning scheme to cluster microarray data (66). PSMF is a probabilistic extension of sparse matrix factorization which can account for uncertainties due to the different level of noise in the data. The PSMF approach model the gene expression as linear weighted combinations of a small number of predominant transcriptional regulators. The PSMF method is appropriate for modeling gene expression data, in which while many genes are involved in gene regulation, only a small number of regulators (e.g. transcription factors) have predominant impact to the expression of the most genes. The results demonstrated that PSMF can better recover functionally relevant clusters from expression data than standard clustering techniques, including hierarchical clustering, k-means clustering, and self-organizing maps**.**

Most recently, Li and Zhan presented a new method, ModulePro, for transcriptional module discovery from microarray data (67). The new method is based on two-stage decomposition of microarray data, firstly by nonlinear independent component analysis and then by probabilistic sparse matrix decomposition. ModulePro offers several advantages: a) it takes into account the nonlinear structure existed in the data; b) the approach does not need the assumption that genes with similar functions or of the same pathway share similar expression profiles; and c) the method can assign one gene into different modules. In comparison with other methods such as hierarchical clustering, k-means, self-organizing maps, and probabilistic sparse matrix decomposition approach, ModulePro leads to a significant improvement in uncovering biologically relevant transcriptional modules.

4. COMPUTATIONAL APPROACHES FOR INFERRING GENE CONNECTIVITY

Inferring gene connectivity involves the selection of a network model and the inference of topology and functions of the network from data. There have been considerable efforts to build models for mimicking gene regulatory networks, covering from fine-scale continuous modeling to coarse-scale discrete modeling. By treating concentrations of gene products as time-dependent variables, three kinds of computational models are proposed so far: a) continuous-time and continuous-variable models (e.g. differential equations); b) discrete-time and continuous-variable models (e.g. Bayesian networks); and c) discrete-time and discrete-variable models (e.g. Boolean networks). Learning the connectivity and relationship between genes in a network model has been studied recently by various signal processing (68), pattern recognition (27, 69, 70), and Bayesian approach (21, 71, 72). Many of these studies have focused on discrete-time networks. Although there have been some successes on modeling continuous-time networks (33, 73), currently available biological observations often lack sufficient richness to identify the parameters of these complex structures.

4.1. ODE-based Models

Ordinary differential equations (ODEs) have been widely used to model the dynamics of genetic regulatory systems (74-85). More specifically, if $x_i(t)$ denotes the state of the ith vertex of the system at time t (e.g. the concentration of the particular proteins, mRNAs, or small molecules associated with that vertex), then its evolution in time, is described by a system of ODEs:

$$\frac{dx_i(t)}{dt} + ax_i(t) = f_i(x_j(t), j \in N_i), i = 1,2,\cdots,l. \quad (1)$$

where f_i is a nonlinear real-valued function of the states $x_j(t), j \in N_i$, of the vertices $j \in N_i$ that interact with the ith vertex. The ODEs in Eq. 1 are also known as kinetic equations. Due to the nonlinearity of f_i, analytical solution of Eq. 1 is not possible. General-purpose numerical ODE solvers, such as Runge-Kutta method (86), are usually applied to solve these ODEs. Differential equations can describe the dynamic regulatory behavior of cellular systems more quantitatively but may require high resolution time series data for the inference of its model parameters as well as more quantitative and detailed information for the parameters, which are not easy to acquire (87).

4.2. Bayesian Networks

A Bayesian network is a representation of a joint probability distribution as a directed acyclic graph (DAG) (16, 21). The vertices of a DAG correspond to random variables [X_1 ..., X_N] and the edges correspond to parent-child dependencies among variables. The random variables may be either discrete or continuous-valued. In the context of gene regulatory networks, X_i may represent the expression level of gene *i*. The joint probability distribution can thus be written in the simple product form:

$$P[X_1, X_2, \cdots, X_N] = \prod_{i=1}^{N} P[X_i \mid \mathrm{Pa}(X_i)] \quad (2)$$

Bayesian networks have a number of features which make them attractive candidates for modeling gene expression data, such as their ability to handle noisy or missing data, to handle hidden variables such as protein levels which may have an effect on mRNA steady state levels, to describe locally interacting processes and the possibility of making causal inferences from the derived models. Based on Bayesian networks, Friedman and colleagues (21) proposed a heuristic algorithm and produced networks which appeared biologically plausible for the yeast cell cycling array data. Bayesian networks have the disadvantage of excluding dynamical aspects of gene regulation. To some extent, this can be overcome through generalizations like dynamical Bayesian networks, which allow feedback relations between genes in a network. Murphy and Mian (41) proposed the use of a

dynamical Bayesian network to model time series gene expression data. Lately, many other Bayesian network models have been proposed for analyzing gene expression data. Most published work to date has considered either static Bayesian networks with fully observed data (29) or static Bayesian networks which model quantized data but incorporate some hidden variables (35, 88).

An understanding of causal relationship in a network is crucial in determining the impact of interventions at the genetic level and performing counterfactual reasoning that leads to finding 'causes'. In general, dependence relations in Bayesian networks do not give unique causal inferences. There are multiple graphs that yield the same joint distribution. Measurements of gene expression, in the absence of interventions, are insufficient to uniquely determine the underlying causal mechanisms. Recently a few studies provided methods for uniquely inferring causal mechanisms for certain cases of Bayesian networks based on perturbation data (22, 35, 89). Even though, most researches on reverse engineering of gene regulatory networks by either Boolean or differential equation-based models do not take the 'causal' aspect of gene connections into consideration (21, 27, 31, 33, 41, 69-72, 90, 91). How to learn causal relationships between genes? In wet-labs, this can be done by knocking out all possible subsets of genes of a given set and studying the impact on the other genes in the set. This is not often feasible when the number of genes in the set is more than a handful. An alternative approach is to use time series gene expression data. Unfortunately such data can only be obtained for cells of particular organisms such as yeast. For human tissues, high-throughput gene expression data is only available for the steady-state. Therefore, how to infer causal relationships between genes from steady-state data is an open question for researchers of this field.

4.3. Coexpression Networks

The study of gene coexpression allows exploration of transcriptional responses that involve coordinated expression of genes encoding proteins which work in concert in the cell. With recent interests in biological networks, the study of gene coexpression has emerged as a novel holistic approach for microarray data analysis (92-95). Most coexpression studies have been based on Pearson's correlation coefficient (19, 92, 93, 96) and mutual information measurement (11, 19, 38, 97). Butte *et al* developed a methodology, termed relevance network, that computes comprehensive pair-wise mutual information (MI) for all gene pairs in a microarray dataset (19). By picking only associations at or above the threshold of MI, they constructed several relevance networks from a public microarray data set and explained the biological significance of each relevance network. A recent paper by Basso *et al.* (38) described a statistical algorithm, ARACNE, for more accurately inferring pair-wise interactions among genes and their protein products. ARACNE first identifies statistically significant gene-gene coregulation by mutual information, and then eliminates indirect relationships. Relationships included in the final reconstructed network have a high probability of representing either direct regulatory interactions or interactions mediated by post-transcriptional modifiers that are undetectable from gene-expression profiles. ARACNE was used to recover a network from gene expression profiles of human B-cell populations. The results suggested that the B-cell regulatory network has both a scale-free and hierarchical architecture, implying the presence of a few 'hubs' that are highly connected and preferentially connected to one another.

The linear-model-based correlation coefficient provides a good first approximation of coexpression, but is also associated with certain pitfalls. When the relationship between log-expression levels of two genes is nonlinear, the degree of coexpression would be underestimated (24). Since the correlation coefficient is a symmetrical measurement, it can not provide evidence of directional relationship in which one gene is upstream of another (16). Similarly, mutual information is also not suitable for modeling directional relationships. The coefficient of determination (CoD), on the other hand, is capable of uncovering nonlinear relationship of coexpression and measuring the directionality, thus it is particularly useful in prediction analysis of gene expression, determination of connectivity in regulatory pathways, and network inference (6, 12, 23, 32, 98). The CoD is a measure for the relative improvement in prediction accuracy owing to the presence of the observed variables, i.e., how much better the combination of given genes (predictors) predicts the behavior of the target gene in comparison to the absence of predictors. It is mathematically defined as $\theta_{opt} = (\varepsilon_0 - \varepsilon_{opt})/\varepsilon_0$, where ε_{opt} is the error for the optimal predictors and ε_0 is the prediction error in the absence of predictors. Since $\varepsilon_{opt} \le \varepsilon_0$, we have $0 \le \theta_{opt} \le 1$.

Recently, Li and Zhan proposed an algorithm, CoExPro, which provides a more biologically meaningful and comprehensive model for gene coexpression, functional relationship, and network structure (99). This algorithm is based on B-spline approximation followed by CoD estimation. The new algorithm is capable of uncovering both linear and nonlinear relationships of coexpression and measuring the directionality. Thus it is particularly useful in prediction analysis of gene expression, determination of connectivity in regulatory pathways, and network inference. The computation by this algorithm requires no quantization of microarray data, thus avoiding significant loss or mis-presentation of biological information, which would otherwise occur in the conventional application of CoD (23, 98). The algorithm was used in modeling the coexpression patterns and exploring biological information from microarray data of several cancers and their normal tissue counterparts. The algorithm allowed correct identification of coexpressed ligand-receptor pairs specific to cancerous tissues and shed light on the understanding of cancer development.

4.4. Probabilistic Boolean Networks

Boolean network models, originally introduced by Kauffman (100, 101), can provide useful insights in

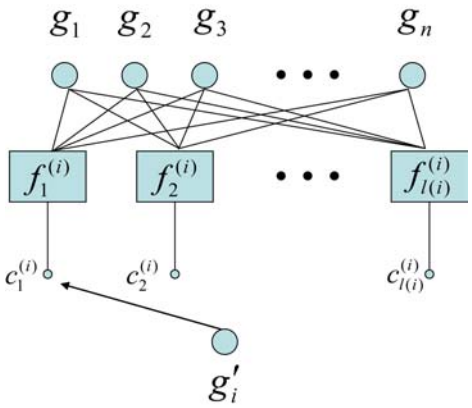

Figure 2. A basic building block of a PBN.

network dynamics at the coarse level. For modeling large-scale genetic regulatory systems, Boolean networks may represent the only practical alternative (5). Recently Boolean networks have been extended to Probabilistic Boolean networks to cope with the randomness inherent in biological systems (32). Microarray data exhibit uncertainty on several levels. First, there is biological uncertainty in the form of intrinsic and extrinsic noise. Second, there is experimental noise due to the complex measurement process, ranging from hybridization conditions to microarray image processing techniques. Third, there may be interacting latent variables, such as proteins, various environmental conditions, or other genes that we simply do not measure, which are further sources of variability in the measurements. To address the uncertainty, Shmulevich *et al* introduced probabilistic Boolean networks (PBNs) by associating several predictors with each target gene (32). If target gene g_i' has $l(i)$ associated predictor functions, $f_1^{(i)}, f_2^{(i)}, \ldots, f_{l(i)}^{(i)}$ then at each point in time t one of these functions is selected to form the transition rule for g_i' at time t+1. Clearly, if $l(i)$ =1 for all $i = 1,2,\ldots,n$, the PBN simply reduces to a standard Boolean network. The basic building block of a PBN is shown in Figure 2. The wiring diagram for the entire PBN consists of n such building blocks. Conceptually, the probabilistic predictor of each target gene can be thought of as a random switch, where at each time point in the network, the function $f_k^{(i)}$ is chosen with probability $c_k^{(i)}$ to predict gene g_i'. One way to assign these probabilities is to employ the CoD, normalized such that $\sum_{k=1}^{l(i)} c_k^{(i)} = 1$. That is, $c_k^{(i)} = \theta_k^{(i)} / \sum_{j=1}^{l(i)} \theta_j^{(i)}$, where $\theta_k^{(i)}$ is the CoD for the target gene g_i' relative to the genes used as inputs to predictor $f_k^{(i)}$.

Within the context of PBNs, Hashimoto *et al* have developed a method to grow a network started from a smaller number of genes of interest, or seed genes (23). The proposed algorithm is flexible and permits various designer choices regarding how to proceed such as the measure of connection strength between genes, search protocol, and stopping conditions. As an example, one can assign the CoD (32) as the strength measuring function. Identifying the seed genes of interest is a critical step in this algorithm. The seed genes are usually selected with the aid of prior biological knowledge. They applied the algorithm to a melanoma data set and constructed a network that consists of only 30 genes.

While good at abstracting uncertainty in biological system, the PBN model fails in describing the context specific determinism of regulatory systems. Context can be defined as a certain condition under which a limited number of genes are tightly regulated by each other for a specific cellular mechanism or a specific task. This specific task can be a different developmental stage, or tissue specific function, resulting in a specific cell-type. The change of this context will result in the change in the set of genes that are highly interactive, and probably their connectivity and relationships. Different biological contexts can also correlate with different diseases or might be a reason why a certain group of patients respond to a therapy while others do not. Kim and Li developed a context-sensitive Boolean network (cBN) model to describe the behavior of cellular systems (102). A cBN can be considered as a constrained PBN, where the constraint is the way to assign the probability for the model. The rule inference is based on the assumption that the inferred rules and the observations are consistent within a (given) context. Figure 3 shows an example of cBNs that contain two contexts and fifteen genes.

4.5. Inference from Multiple Sources of Data

Most early researches on automatic learning of transcriptional regulatory networks employ only gene expression data. Recent simulation studies suggest that regulatory networks learned from gene expression data alone can be considerably obscured by the recovery of spurious interactions when the number of observations is small (103). Integrating findings from multiple data sources (e.g. DNA sequences, gene and protein expression profiles, protein-protein interactions, protein structural information, and protein-DNA binding data) can overcome this drawback (42). Two major yet related approaches have been developed in joint learning transcriptional regulation from multiple data sources. In one approach, various types of data are used to identify sets of genes that interact together in the cell, or are co-regulated in modules (13, 15). In the other approach, various types of data are used to supplement gene expression data in learning regulatory networks (51, 104).

Bernard and Hartemink presented a method for jointly learning dynamic models of transcriptional regulatory networks from gene expression data and transcription factor binding data, based on dynamic Bayesian network inference algorithms (104). Results obtained from analyzing yeast cell cycle data demonstrate that the recovery of dynamic regulatory networks from multiple types of data by this joint learning algorithm is more accurate than that from each data type alone.

Imoto *et al* proposed a statistical method for estimating a gene network based on Bayesian networks

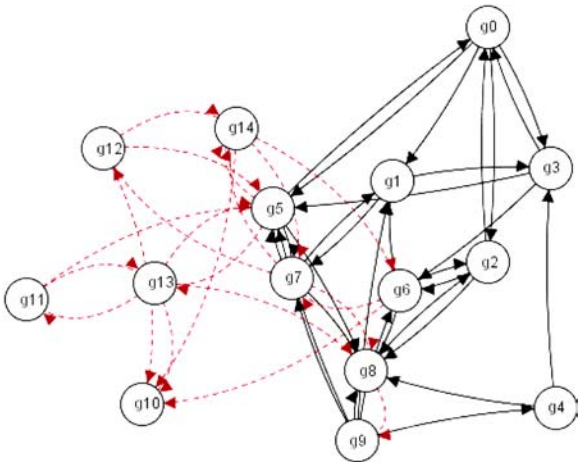

Figure 3. An example of cBN with two contexts.

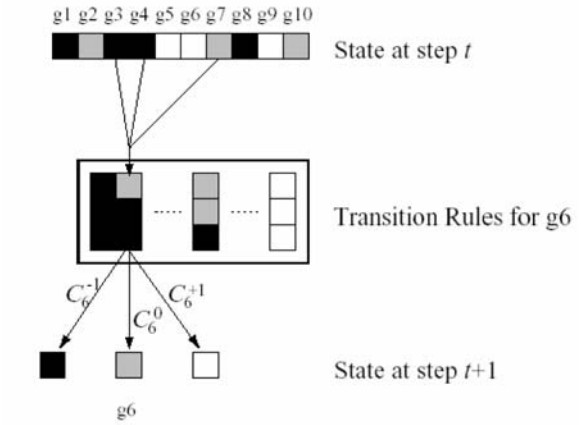

Figure 4. The structure of the Markov chain model.

from microarray gene expression data together with biological knowledge including protein-protein interactions, protein-DNA interactions, transcriptional factor binding information, existing literature and so on (51). An advantage of the method is that the balance between microarray information and biological knowledge is optimized automatically by the proposed criterion. Monte Carlo simulations showed the effectiveness of the proposed method in extracting more information from microarray data and estimating the gene network more accurately.

Yeang *et al* developed a framework for inferring transcriptional regulation (105). The models they developed, called physical network models, are annotated molecular interaction graphs. The attributes in the model correspond to verifiable properties of the underlying biological system such as the existence of protein–protein and protein–DNA interactions, the directionality of signal transduction in protein–protein interactions, signs of the immediate effects of these interactions, etc. Possible configurations of these variables are constrained by the available data sources. Application of this algorithm on datasets related to the pheromone response pathway in yeast demonstrated that the derived model was consistent with previous knowledge of the pathway.

5. NETWORK ANALYSIS *IN SILICO*

In silico network analysis involves studying the long run behavior of the system (steady-state analysis), observing the effects caused by perturbation in the network structure (perturbation analysis), and predicting what changes in the network structure or functions should be imposed to achieve desired effects (intervention analysis). *In silico* simulation has been particularly important in network analysis since network activity is constrained by the various complex forms of interactions (4, 5). Various algorithms have been employed in examining dynamic behaviors of biological networks *in silico*, including the Markov chain (6, 12) and probabilistic Boolean network (7, 8). Here, we describe steady-state analysis and intervention analysis to show how a network analysis is formulated.

5.1. Steady State Analysis by Markov Chain Simulation

Mathematical modeling that allows estimation of steady state behavior in biological systems is useful for examining two ubiquitous forms of biological system behavior. The first is homeostasis, the ability of cells to maintain their ongoing processes within a narrow range compatible with the survival. The second is a switch-like functionality, which allows cells to rapidly transit limited process segments between metastable states. Kim *et al* proposed a finite-state Markov chain model and explored whether the mode can capture the biological behavior above described (12). The proposed model contains n nodes, each of which represents one of the n genes selected. Each gene has a ternary value, which is assigned from over-expressed (1), equivalently-expressed (0), and under-expressed (-1). The state space of the Markov chain has 3^n states. For capturing the dynamics of the network, they consider a "wiring rule" such that the expression state of each gene at step $t + 1$ is predicted by the expression levels of the other genes at step t in the same network. For each target gene, a set of three predictor genes is chosen with the highest CoD value. Instead of using many possible Boolean functions that are independent of the state of the system, as in the PBN model, they use the state of three predictor genes at step t and the corresponding conditional probabilities, which are estimated from observed data, to derive the state of the target gene at step $t + 1$. Eq. (3) describes the definition of the Markov chain between a state at step t and the state at step $t + 1$.

$$S^{(t)} =: (g_1^{(t)} g_2^{(t)} \cdots g_n^{(t)}) \longrightarrow S^{(t+1)} =: (g_1^{(t+1)} g_2^{(t+1)} \cdots g_n^{(t+1)}) \quad (3)$$

The transition rule is depicted in Figure 4. In the simulation, gene perturbation is added to guarantee the chain converge to be a steady-state distribution (7). Considering gene perturbation, the transition probability is generalized as follows:

$$\Pr\{S^{(t)} \to S^{(t+1)}\} = \left(\prod_{l=1}^{n} C_l^{g_l^{(t+1)}} \right) \times (1-p)^n + p^{n_0} (1-p)^{n-n_0} p_0^{n_0} \times 1_{[S^{(t)} \neq S^{(t+1)}]} \quad (4)$$

where p is the perturbation probability for each gene, $n_0 = \sum_{l=1}^{n} 1_{[g_l^{(t)} \neq g_l^{(t+1)}]}$ is the number of genes to be perturbed, $p_0 = 1/(q\text{-}1)$, and q is the level of gene expression.

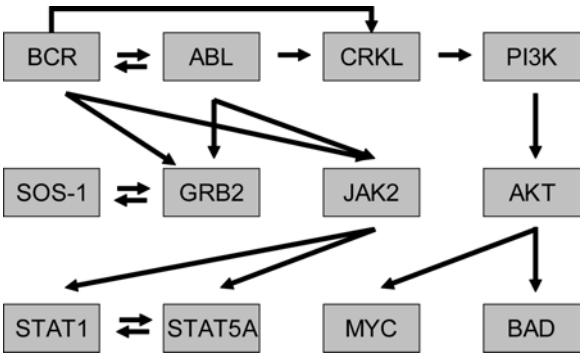

Figure 5. The topology of leukemia-related BCR-ABL pathway. The arrows represent the directions of the causal relationships among genes. BCR are ABL are linked to the cytoplasm as a part of a large signaling complex with a variety of cellular substrates, related to the development of leukemia. The drug Gleevec is a selective BCR-ABL inhibitor in this pathway.

In a steady state analysis using Markov chain simulation on the gene expression profiles of 31 melanoma cell lines, 50 genes capable of both predicting other genes as well as being predicted by other genes with high CoDs were chosen out of all genes (12). From the 50 genes, 10 genes were further selected based on their roles in classifying malignant melanoma and known biological functions (106). The results indicated that the steady state distributions produced approximate the initial observations. Only a limited number of states possessed significant probability of occurrence. These behaviors are congruent with biological behaviors, as cells appear to occupy only a negligible portion of the state space available to them. The transition rules generated for the model produces localized stability. The study suggests that, in the limited context, Markov chain simulation emulates well the dynamic behavior of a small regulatory network. By systematically examining the characteristics of the rules and interconnections that lead to stabilization and switch-like transitions, we may gain a better understanding of biological regulation.

5.2. Intervention Analysis by Markov Chain Model

Intervention analysis can not only open up a window on the biological behavior of an organism and disease progression, but also translate into accurate diagnosis, target identification, drug development, and treatment. Shmulevich *et al* (7) used a PBN model to study gene perturbation and intervention, and developed several computational tools based on first-passage times in Markov chains. Pal *et al* treated intervention with external control variables in a context-sensitive PBN (8). They applied the control theory to find optimal strategies for manipulating external control variables that affect the transition probabilities of states in the network. However, few studies have so far demonstrated a systematic understanding of the dynamic behavior of a regulatory network in response to each internal gene intervention or external perturbation in a one-to-one relationship.

Recently, Li and Zhan developed a algorithm, PathwayPro, to mimic the behavior of a biological pathway through a series of interventions made *in silico* upon each gene or gene combination (6). The inputs to the algorithm are experiment-specific regulatory pathways and gene expression data. The outputs are the estimated probabilities of a network transit across different cellular conditions under each transcriptional intervention. The algorithm can provide answers to two questions. First, whether or how much a gene or external perturbation contributes to the transition behavior of a regulatory pathway in instances such as disease development or recovery, aging process, and cell differentiation. Second, in what specific ways is this contribution manifested. The first-passage times allow capturing the goals of intervention by a quick transition to (or avoiding) certain states or by maximizing the probability of reaching certain states before a certain time. They are thus used to decide which genes are the best candidates for intervention. The first passage time from state x to state y can be defined as the probability $F_k(x,y)$ that, starting in state x, the first time the network will reach a given state y will be at step k. It is easy to see that for k=1, $F_1(x,y)=A(x,y)$, which is just the transition probability from x to y. For $k \geq 2$, $F_k(x,y)$ satisfies (107)

$$F_k(x,y)=\sum_{z\in\{-1,0,1\}^n-\{y\}} A(x,z)F_{k-1}(z,y) \tag{5}$$

In Eq. 5, each element $A(x,y)$ of the transition matrix A can be computed using Eq. 4. For a fixed K, a $3^n \times K$ matrix **F** can be created in which each column contains the probability $F_k(x,y)$ from all possible starting states x to a given target state y at k steps. One can then use $H_K(x,y)=\sum_{k=1}^{K} F_k(x,y)$ as a measurement index. In PathwayPro, the intervention information matrix **H** is constructed by fixing $K=3$. In this matrix, each row $H_3(x,:)$ represents the probability that the network, from a starting state x, will visit all desired ending states before step $K=3$. Each column $H_3(:,y)$ represents the probability that the network, starting in all possible intervened states, will visit state y before step K=3. To simulate a simple stimulus, the expression level of one gene, two genes, or three genes is mathematically changed each time while the rest of the genes are kept unchanged for a starting state x. For a ternary expression, $C_n^3 \times 3^3$ intervened states are generated for intervening one, two, and three genes, including the original state x.

PathwayPro was used for analyzing the leukemia-related BCR-ABL pathway (6). The analysis profiled the dynamic behavior of the pathway in response to leukemia development and identified possible disease and drug targets. Figure 5 shows the network topology of the ABL-BCR pathway. *In silico* simulation was conducted by transcriptional intervention on each gene (referred to as single-gene intervention), each combination of two genes (double-gene intervention), and each combination of three

Table 1. Probabilities of network transition by serial interventions on genes in the ABL-BCR pathway of human

Gene	Transcriptional Intervention	Transition Probability
(A) Transition from normal to CML states by single-gene intervention[a]		
BCR	0 => -1 => 1	0.00639
(B) Transition from CML to normal states by single-gene intervention[b]		
ABL1	1 => 0 => -1	0.000299
(C) Transition from the normal to CML states by double-gene intervention[c]		
BCR ABL1	0 -1 => 1 1 => 1 1	0.0109
BCR BAD	0 1 => -1 0 => 1 0	0.00639
BCR MYC	0 -1 => -1 0 => 1 0	0.00639
BCR BAD	0 1 => -1 -1 => 1 0	0.00639
BCR MYC	0 -1 => -1 1 => 1 0	0.00639
BCR STAT5A	0 1 => -1 -1 => 1 1	0.00639
BCR STAT5A	0 1 => -1 0 => 1 1	0.00639
BCR STAT1	0 0 => -1 1 => 1 0	0.00639
BCR STAT1	0 0 => -1 -1 => 1 0	0.00639
BCR CRKL	0 -1 => -1 1 => 1 0	0.00539
BCR CRKL	0 -1 => -1 0 => 1 0	0.00399
BCR PIK3CG	0 -1 => -1 0 => 1 -1	0.00384
BCR JAK2	0 0 => -1 1 => 1 0	0.00224
BCR AKT1	0 0 => -1 -1 => 1 0	0.00107
(D) Transition from the CML to normal states by double-gene intervention[d]		
ABL1 AKT1	1 0 => 0 1 => -1 0	0.00185
ABL1 AKT1	1 0 => 0 -1 => -1 0	0.00179
BCR ABL1	1 1 => 0 -1 => 0 -1	0.00111
(E) Transition from normal to CML states by triple-gene intervention[e]		
BCR ABL1 BAD	0 -1 1 => 1 1 0 => 1 1 0	0.0109
BCR ABL1 MYC	0 -1 -1 => 1 1 0 => 1 1 0	0.0109
BCR ABL1 BAD	0 -1 1 => 1 1 -1 => 1 1 0	0.0109
BCR ABL1 MYC	0 -1 -1 => 1 1 1 => 1 1 0	0.0109
BCR ABL1 STAT5A	0 -1 1 => 1 1 0 => 1 1 1	0.0109
BCR ABL1 STAT5A	0 -1 1 => 1 1 -1 => 1 1 1	0.0109
BCR ABL1 STAT1	0 -1 0 => 1 1 -1 => 1 1 0	0.0109
BCR ABL1 STAT1	0 -1 0 => 1 1 1 => 1 1 0	0.0109
(F) Transition from CML to normal states by triple-gene intervention[f]		
BCR ABL1 AKT1	1 1 0 => 0 -1 1 => 0 -1 0	0.00684
BCR ABL1 AKT1	1 1 0 => 0 -1 -1 => 0 -1 0	0.00662
ABL1 CRKL AKT1	1 0 0 => 0 -1 1 => -1 -1 0	0.00297
ABL1 CRKL AKT1	1 0 0 => 0 -1 -1 => -1 -1 0	0.00288
BCR ABL1 AKT1	1 1 0 => -1 -1 1 => 0 -1 0	0.00274
BCR ABL1 AKT1	1 1 0 => -1 -1 -1 => 0 -1 0	0.00265
ABL1 CRKL AKT1	1 0 0 => 0 1 1 => -1 -1 0	0.00250
ABL1 CRKL AKT1	1 0 0 => 0 1 -1 => -1 -1 0	0.00242

The gene expression profile of each state is presented as: initial state (e.g. normal state) => state after intervened => end state (e.g. disease state). Transcriptional intervention is presented as: initial state (e.g. normal state) => state after intervened => end state (e.g. disease state). In each state, expression levels of each gene are presented by ternary values. [a]Probability cutoff 1E-4; [b]Probability cutoff 1E-4; [c]Probability cutoff 1E-3; [d]Probability cutoff 1E-3; [e]Probability cutoff 1E-2; [f]Probability cutoff 2E-3.

genes (triple-gene intervention). In each intervention, the observed expression of a gene was altered to the opposite direction or remained unchanged. The transition probabilities of the BCR-ABL pathway were measured between the normal condition and leukemia state under a series of transcriptional interventions. The probability of the network transitioning from normal to leukemia states reveals disease susceptibility of genes involved. The higher the probability is, the more likely a gene or gene combination under a certain intervention is responsible for the development of the disease. On the other hand, the probability of the transition from leukemia to normal states is a measure of the potential usefulness of a drug or therapeutic intervention. The analysis showed that more genes and gene combinations had high probabilities to contribute to regulatory network transitions from normal to leukemia states than from leukemia to normal states (Table 1). The result suggests that the chance is higher for a human to develop leukemia than to recover from the disease. The importance of BCR and ABL to the network transition was further illustrated by the single-gene intervention, where both BCR and ABL were associated with the highest transition probability (Table 1). Moreover, BCR and ABL showed high frequencies in all of their partnerships with other genes in the double or triple interventions positive for network transition. As illustrated in Figure 5, BCR and ABL were on the top by the frequency of partnership with other genes in the normal to leukemia transition, while BCR and ABL, along with AKT and CRKL, were on the top in the leukemia to normal transition in the triple-gene invention. These results suggest that BCR and ABL are the most contributive genes to the network transition between the normal condition and the leukemia state, and therefore the most susceptible for the development of the CML leukemia as well as the recovery from the disease to a normal condition. The two genes can

thus serve as good drug targets for the treatment of CML leukemia. This result, reached independently by computational analysis, is in agreement with the conclusion of previous laboratory-based studies. It has been shown that CML is associated in most cases with the fusion of the genes ABL and BCR, and the activation of BCR-ABL represses apoptosis and allows transformed cells to divide, resulting in the development of CML (108-110). The drug Gleevec is a selective BCR-ABL inhibitor, effective in the treatment of CML (111). PathwayPro not only correctly identified the drug targets, but further indicated that BAD and MYC played critical roles in leukemia development while AKT appeared important in the leukemia recovery to normal. The results provide new insights into our understanding of the leukemia disease.

6. CLOSING REMARKS

Systems biology is aimed at elucidating how genes interact to each other to perform specific biological processes or functions, and how disease or cellular phenotypes arise from networks of genes and their products. Multidisciplinary efforts have been made for modeling and inferring regulatory networks from microarray or other data sources. These studies facilitate our understanding of cellular systems. The generated hypotheses can be further tested via independent biological experiments. The studies can eventually open up a window for *in silico* development of systematic approaches for effective preventive and therapeutic intervention in disease.

7. ACKNOWLEDGMENTS

The authors would like to thank Dr. William Baumann and Dr. Zhiping Gu for their comments on this review, the Intramural Research Program, National Institute on Aging, NIH and the National Institutes of Health (under Grants CA109872, EB000830) for generous support.

8. REFERENCES

1. Kitano, H.: Computational systems biology. *Nature*, 420, 206-10 (2002)
2. Ideker, T., T. Galitski & L. Hood: A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2, 343-72 (2001)
3. New Pathways to Discovery [http://nihroadmap.nih.gov/newpathways/]
4. de Jong, H.: Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *Journal of Computational Biology*, 9, 67--103 (2002)
5. Smolen, P., D. A. Baxter & J. H. Byrne: Modeling transcriptional control in gene networks--methods, recent results, and future directions. *Bull Math Biol*, 62, 247-92 (2000)
6. Li, H. & M. Zhan: Systematic intervention of transcription for identifying network response to disease and cellular phenotypes. *Bioinformatics*, 22, 96-102 (2006)
7. Shmulevich, I., E. R. Dougherty & W. Zhang: Gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics*, 18, 1319-31 (2002)
8. Pal, R., A. Datta, M. L. Bittner & E. R. Dougherty: Intervention in context-sensitive probabilistic Boolean networks. *Bioinformatics*, 21, 1211-8 (2005)
9. Schulze, A. & J. Downward: Navigating Gene Expression Using Microarrays - A Technology Review. *Nature Cell Biology*, 3, E190-E195 (2002)
10. Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell & R. A. Young: Genome-wide location and function of DNA binding proteins. *Science*, 290, 2306-9 (2000)
11. Margolin, A. A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera & A. Califano: ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1, S1-7 (2006)
12. Kim, S., H. Li, E. R. Dougherty, N. Chao, Y. Chen, M. L. Bittner & E. B. Suh: Can Markov chain models mimic biological regulation? *J. Biological Systems*, 10, 337-357 (2002)
13. Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller & N. Friedman: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34, 166-76 (2003)
14. Akutsu, T., S. Miyano & S. Kuhara: Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16, 727-34 (2000)
15. Bar-Joseph, Z., G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young & D. K. Gifford: Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, 21, 1337-42 (2003)
16. Imoto, S., T. Goto & S. Miyano: Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pac Symp Biocomput*175-86 (2002)
17. Zhang, Z. & M. Gerstein: Reconstructing genetic networks in yeast. *Nat Biotechnol*, 21, 1295-7 (2003)
18. Bolouri, H. & E. H. Davidson: Modeling transcriptional regulatory networks. *Bioessays*, 1118-29 (2002).
19. Butte, A. J. & I. S. Kohane: Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*418-29 (2000)
20. D'Haeseleer, P., S. Liang & R. Somogyi: Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16, 707-26 (2000)
21. Friedman, N., M. Linial, I. Nachman & D. Pe'er: Using Bayesian networks to analyze expression data. *J Comput Biol*, 7, 601-20 (2000)
22. Guelzim, N., S. Bottani, P. Bourgine & F. Kepes: Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31, 60-3 (2002)
23. Hashimoto, R., S. Kim, I. Shmulevich, W. Zhang, M. L. Bittner & E. R. Dougherty: Growing genetic regulatory networks from seed genes. *Bioinformatics*, 20, 1241-1247 (2004)
24. Herrgard, M. J., M. W. Covert & B. O. Palsson: Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res.*, 13, 2423-34. (2003)

25. Jenssen, T. K., A. Laegreid, J. Komorowski & E. Hovig: A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28, 21-8 (2001)
26. Kauffman, S., C. Peterson, B. Samuelsson & C. Troein: Random Boolean network models and the yeast transcriptional network. *Proc Natl Acad Sci U S A*, 100, 14796-9 (2003)
27. Liang, S., S. Fuhrman & R. Somogyi: Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*18-29 (1998)
28. Liao, J. C., R. Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti & V. P. Roychowdhury: Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A*, 100, 15522-7 (2003)
29. Pe'er, D., A. Regev, G. Elidan & N. Friedman: Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17 Suppl 1, S215-24 (2001)
30. Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai & A. L. Barabasi: Hierarchical organization of modularity in metabolic networks. *Science*, 297, 1551-5 (2002)
31. Sachs, K., D. Gifford, T. Jaakkola, P. Sorger & D. A. Lauffenburger: Bayesian network approach to cell signaling pathway modeling. *Sci STKE*, 2002, PE38 (2002)
32. Shmulevich, I., E. R. Dougherty, S. Kim & W. Zhang: Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18, 261-74 (2002)
33. Tegner, J., M. K. Yeung, J. Hasty & J. J. Collins: Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc Natl Acad Sci U S A*, 100, 5944-9 (2003)
34. Zhou, X. J., M. C. Kao, H. Huang, A. Wong, J. Nunez-Iglesias, M. Primig, O. M. Aparicio, C. E. Finch, T. E. Morgan & W. H. Wong: Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol*, 23, 238-43 (2005)
35. Yoo, C. & G. F. Cooper: Discovery of gene-regulation pathways using local causal search. *Proc AMIA Symp*914-8 (2002)
36. Kolchanov, N. A., M. P. Ponomarenko, A. S. Frolov, E. A. Ananko, F. A. Kolpakov, E. V. Ignatieva, O. A. Podkolodnaya, T. N. Goryachkovskaya, I. L. Stepanenko, T. I. Merkulova, V. V. Babenko, Y. V. Ponomarenko, A. V. Kochetov, N. L. Podkolodny, D. V. Vorobiev, S. V. Lavryushev, D. A. Grigorovich, Y. V. Kondrakhin, L. Milanesi, E. Wingender, V. Solovyev & G. C. Overton: Integrated databases and computer systems for studying eukaryotic gene expression. *Bioinformatics*, 15, 669-86 (1999)
37. Hartemink, A. J. & E. Segal: Joint learning from multiple types of genomic data. *Pac Symp Biocomput*445-6 (2005)
38. Basso, K., A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera & A. Califano: Reverse engineering of regulatory networks in human B cells. *Nat Genet*, 37, 382-90 (2005)
39. Luscombe, N. M., M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann & M. Gerstein: Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431, 308-12 (2004)
40. Bonneau, R., D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga & V. Thorsson: The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol*, 7, R36 (2006)
41. Murphy, K. & S. Mian: Modelling gene expression data using dynamic Bayesian networks. In: Division of Computer Science, University of California, Berkeley (1999)
42. Li, J., X. Li, H. Su, H. Chen & D. W. Galbraith: A framework of integrating gene relations from heterogeneous data sources: an experiment on Arabidopsis thaliana. *Bioinformatics*, 22, 2037-43 (2006)
43. Hartwell, L. H., J. J. Hopfield, S. Leibler & A. W. Murray: From molecular to modular cell biology. *Nature*, 402, C47-52 (1999)
44. Resendis-Antonio, O., J. A. Freyre-Gonzalez, R. Menchaca-Mendez, R. M. Gutierrez-Rios, A. Martinez-Antonio, C. Avila-Sanchez & J. Collado-Vides: Modular analysis of the transcriptional regulatory network of E. coli. *Trends Genet*, 21, 16-20 (2005)
45. Ideker, T., V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold & L. Hood: Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292, 929-34 (2001)
46. Alon, U.: Biological networks: the tinkerer as an engineer. *Science*, 301, 1866-7 (2003)
47. Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai & A. L. Barabasi: The large-scale organization of metabolic networks. *Nature*, 407, 651-4 (2000)
48. Eisen, M. B., P. T. Spellman, P. O. Brown & D. Botstein: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95, 14863-8 (1998)
49. Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho & G. M. Church: Systematic determination of genetic network architecture. *Nat Genet*, 22, 281-5 (1999)
50. Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander & T. R. Golub: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96, 2907-12 (1999)
51. Imoto, S., T. Higuchi, T. Goto, K. Tashiro, S. Kuhara & S. Miyano: Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *J Bioinform Comput Biol*, 2, 77-98 (2004)
52. Wang, Z., Y. Wang, J. Lu, S. Y. Kung, J. Zhang, R. Lee, J. Xuan, J. Khan & R. Clarke: Discriminatory Mining of Gene Expression Microarray Data. *Journal of VLSI Signal Processing*, 35, 255-72 (2003)
53. Wang, Y., L. Luo, M. T. Freedman & S. Y. Kung: Probabilistic principal component subspaces: a hierarchical finite mixture model for data visualization. *IEEE Trans. Neural Networks*, 11, 625-36 (2000)
54. Bakay, M., Z. Wang, G. Melcon, L. Schiltz, J. Xuan, P. Zhao, V. Sartorelli, J. Seo, E. Pegoraro, C. Angelini, B. Shneiderman, D. Escolar, Y. W. Chen, S. T. Winokur, L. M. Pachman, C. Fan, R. Mandler, Y. Nevo, E. Gordon, Y. Zhu, Y. Dong, Y. Wang & E. P. Hoffman: Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration. *Brain*, 129, 996-1013 (2006)

55. caBIG - Welcome to the caBIG™ Web site [http://caBIG.nci.nih.gov]
56. Alter, O., P. O. Brown & D. Botstein: Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, 97, 10101-6 (2000)
57. Holter, N. S., A. Maritan, M. Cieplak, N. V. Fedoroff & J. R. Banavar: Dynamic modeling of gene expression data. *Proc Natl Acad Sci U S A*, 98, 1693-8 (2001)
58. Lee, S. I. & S. Batzoglou: Application of independent component analysis to microarrays. *Genome Biol*, 4, R76 (2003)
59. Liebermeister, W.: Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18, 51-60 (2002)
60. Frigyesi, A., S. Veerla, D. Lindgren & M. Hoglund: Independent component analysis reveals new and biologically significant structures in microarray data. *BMC Bioinformatics*, 7, 290 (2006)
61. Kim, P. M. & B. Tidor: Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res*, 13, 1706-18 (2003)
62. Lee, D. D. & H. S. Seung: Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788-91 (1999)
63. Brunet, J. P., P. Tamayo, T. R. Golub & J. P. Mesirov: Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101, 4164-9 (2004)
64. Gao, Y. & G. Church: Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, 21, 3970-5 (2005)
65. Wang, G., A. V. Kossenkov & M. F. Ochs: LS-NMF: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics*, 7, 175 (2006)
66. Dueck, D., Q. D. Morris & B. J. Frey: Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics*, 21 Suppl 1, i144-i151 (2005)
67. Li, H., Y. Sun & M. Zhan: The discovery of transcriptional modules by a two-stage matrix decomposition approach. *Bioinformatics*, 23, 473-9 (2007)
68. Kim, S., E. R. Dougherty, M. L. Bittner, Y. Chen, K. L. Sivakumar, P. S. Meltzer & J. M. Trent: A General Nonlinear Framework for the Analysis of Gene Interaction via Expression Array. *Journal of Biomedical Optics*, 5, 411-424 (2000)
69. Lukashin, A. V., M. E. Lukashev & R. Fuchs: Topology of gene expression networks as revealed by data mining and modeling. *Bioinformatics*, 19, 1909-16 (2003)
70. Nakahara, H., S. Nishimura, M. Inoue, G. Hori & S. Amari: Gene interaction in DNA microarray data is decomposed by information geometric measure. *Bioinformatics*, 19, 1124-31 (2003)
71. Husmeier, D.: Reverse engineering of genetic networks with Bayesian networks. *Biochem Soc Trans*, 31, 1516-8 (2003)
72. Perrin, B.-E., L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet & d'Alché-Buc: Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, 19, ii138-ii148 (2003)
73. Gardner, T. S., D. di Bernardo, D. Lorenz & J. J. Collins: Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301, 102-5 (2003)
74. Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford & R. A. Young: Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298, 799-804 (2002)
75. von Dassow, G., E. Meir, E. M. Munro & G. M. Odell: The segment polarity network is a robust developmental module. *Nature*, 188-92 (2000).
76. von Dassow, G. & G. M. Odell: Design and constraints of the Drosophila segment polarity module: robust spatial patterning emerges from intertwined cell state switches. *J Exp Zool*, 179-215 (2002).
77. Savageau, M. A.: Design principles for elementary gene circuits: Elements, methods, and examples. *Chaos*, 142-159 (2001).
78. Davidson, E. H.: Genomic Regulatory Systems: Development and Evolution. Academic Press, San Diego, CA (2001)
79. Elowitz, M. B. & S. Leibler: A synthetic oscillatory network of transcriptional regulators. *Nature*, 335-8 (2000).
80. Gardner, T. S., C. R. Cantor & J. J. Collins: Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403, 339-42 (2000)
81. Yildirim, N. & M. C. Mackey: Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data. *Biophys J*, 84, 2841-51 (2003)
82. Akutsu, T., S. Miyano & S. Kuhara: Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J Comput Biol*, 7, 331-43 (2000)
83. Ronen, M., R. Rosenberg, B. I. Shraiman & U. Alon: Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A*, 99, 10555-60 (2002)
84. Hatzimanikatis, V. & K. H. Lee: Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metab Eng*, 1, 275-81 (1999)
85. Chen, T., H. L. He & G. M. Church: Modeling gene expression with differential equations. *Pac Symp Biocomput*29-40 (1999)
86. Press, W. H., B. P. Flannery, S. A. Teukolsky & W. T. Vetterling: Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge, England (1993)
87. Goutsias, J. & S. Kim: A Nonlinear Discrete Dynamical Model for Transcriptional Regulation: Construction and Properties. *Biophys. J.*, 86, 1922-1945 (2004)
88. Hartemink, A. J., D. K. Gifford, T. S. Jaakkola & R. A. Young: Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks. *Proceedings of the Pacific Symposium on Biocomputing'01*, R. B. Altman, K. Lauderdale, A. K. Dunker, L. Hunter&T. E. Klein, Eds., 422--433 (2001)

89. Xing, B. & M. J. van der Laan: A causal inference approach for constructing transcriptional regulatory networks. *Bioinformatics*, 21, 4007-13 (2005)
90. Kim, S., E. R. Dougherty, M. L. Bittner, Y. Chen, K. L. Sivakumar, P. S. Meltzer & J. M. Trent: Multivariate measurement of gene-expression relationships. *Genomics*, 67, 201--209 (2000)
91. Mateos, A., J. Dopazo, R. Jansen, Y. Tu, M. Gerstein & G. Stolovitzky: Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res*, 12, 1703-15 (2002)
92. Stuart, J. M., E. Segal, D. Koller & S. K. Kim: A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302, 249-55 (2003)
93. Lee, H. K., A. K. Hsu, J. Sajdak, J. Qin & P. Pavlidis: Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14, 1085-94 (2004)
94. van Noort, V., B. Snel & M. A. Huynen: The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep*, 5, 280-4 (2004)
95. Carter, S. L., C. M. Brechbuhler, M. Griffin & A. T. Bond: Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20, 2242-50 (2004)
96. Graeber, T. G. & D. Eisenberg: Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat Genet.*, 29, 295-300 (2001)
97. Zhou, X., X. Wang & E. R. Dougherty: Construction of Genomic Networks Using Mutual-Information Clustering and Reversible-Jump Markov-Chain Monte-Carlo Predictor Design. *Signal Processing*, 83, 745-61 (2003)
98. Dougherty, E. R., S. Kim & Y. Chen: Coefficient of determination in nonlinear signal processing. *Signal Processing*, 80, 2219--2235 (2000)
99. Li, H., Y. Sun & M. Zhan: Analysis of gene coexpression by B-spline based CoD estimation. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007, 1-10 (2007)
100. Kauffman, S. A.: Requirements for Evolvability in Complex Systems: Orderly Dynamics and Frozen Components. *Physica D*, 42, 135-152 (1990)
101. Kauffman, S. A.: The Origins of Order, Self-Organization and Selection in Evolution. New York (1993)
102. Li, H., J. Whitmore, E. Suh, M. Bittner & S. Kim: Learning context-sensitive Boolean network from steady-state observations and its analysis. *Research in Computational Molecular Biology (RECOMB 2004)*, (2004).
103. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19, 2271-82 (2003)
104. Bernard, A. & A. J. Hartemink: Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac Symp Biocomput*459-70 (2005)
105. Yeang, C. H., T. Ideker & T. Jaakkola: Physical network models. *J Comput Biol*, 11, 243-62 (2004)
106. Bittner, M., P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts & V. Sondak: Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406, 536-40 (2000)
107. Cinlar, E.: Introduction to Stochastic Processes. Prentice Hall, New Jersey (1975)
108. Lugo, T. G., A. M. Pendergast, A. J. Muller & O. N. Witte: Tyrosine kinse activity and transformation potency of bcr-abl oncogene products. *Science*, 247, 1079–1082 (1990)
109. Raitano, A. B., Y. E. Whang & C. L. Sawyers: Signal transduction by wild-type and leukemogenic Abl proteins. *Biochim Biophys Acta*, 1333, 201–216 (1997)
110. Zou, X. & K. Calame: Signaling pathways activated by oncogenic forms of Abl tyrosine kinase. *J Biol Chem.*, 274, 18141-18144 (1999)
111. Druker, B. J., C. L. Sawyers & H. Kantarjian: Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *N Engl J Med.*, 344, 1038-1042 (2001)

Abbreviations: CoD: coefficient of determination; SVD: singular value decomposition; ICA: independent components analysis; NMF: non-negative matrix factorizations; PBN: probabilistic Boolean network; GO: gene ontology; MIPS: Munich Information Center for Protein Sequences; PSMF: probabilistic sparse matrix factorization; VISDA: visual and statistical data analyzer; ODE: ordinary differential equation; CBN: context-sensitive Boolean network

Key words: Systems Biology, Regulatory Network, Coexpression, Pathway Dynamics, Modeling And Inference, Transcriptional Intervention, Review

Send correspondence to: Dr. Huai Li, Bioinformatics Unit, Branch of Research Resources, National Institute on Aging, NIH, 333 Cassell Drive, Baltimore, MD, USA, Tel: 410-558-8535, Fax: 410-558- 8674, E-mail: huaili@mail.nih.gov

http://www.bioscience.org/current/vol13.htm