



For reprint orders, please contact:
reprints@future-medicine.co.uk

Genomic approaches in dissecting complex biological pathways

Ning Sun¹ &
Hongyu Zhao^{1,2}

[†]Author for correspondence
¹Department of Epidemiology
and Public Health, Yale
University School of
Medicine, New Haven,
CT 06520, USA
²Department of Genetics, Yale
University School of
Medicine, New Haven,
CT 06520, USA
E-mail: hongyu.zhao@
yale.edu

Advances in genomic research have provided many types of large-scale data that contain rich information on various biological pathways. Intensive efforts have been made to qualitatively or quantitatively model biological pathways using these genomic data. Some general network properties, such as the scale-free property and network motifs, have been discussed and various network models have been applied to reconstruct pathways. However, there is a lack of systematic integration of prior knowledge and different genomic data in these analyses. In this review, we discuss pathway reconstruction under the consideration of the complexity embedded in the biological system, and the global and local properties of biological pathways. We review major methodologies, including clustering methods, scale-free networks models, Bayesian networks models, Boolean networks models, systems of differential equations, and data integration methods. We focus on the difficulty of each methodology in modeling biological pathways, and emphasize that different models capture different aspects of biological pathways or genomic data. The 'noisy' large-scale genomic data require the mathematical models and computational methods to be both robust and identifiable. In addition, we believe that ideal models should have the capability of incorporating various data types and these models need to be assessed through rigorous comparisons with empirical data.

Introduction

The identification and validation of drug targets relies critically on knowledge of the biochemical pathways in which potential target molecules function within cells. For this reason, the study of biochemical pathways is one major focus of drug discovery research, and findings from such research are essential for drug developments by biopharmaceutical and genomic companies. Despite intensive studies, these biological pathways are far from being completely understood. In recent years, advances in high-throughput biotechnologies have led to the accumulation of large amounts of genomic data of various types. These data have been analyzed in the context of known pathways to both test the utility of genomic data and to gain a better and more comprehensive understanding of these pathways (e.g., see [1-4]). However, these analyses are mostly descriptive without sophisticated modeling, prediction, and testing components to systematically integrate various types of experimental data as well as known biological knowledge. This is in part due to the fact that it is not a trivial task to extract useful information from the large amounts of diverse genomic data to model and understand complex biological systems. In fact, most existing mathematical and

computational approaches are not yet well suited for this important task. Therefore, although the question of how to optimally use genomic data to analyze biological pathways remains open, we believe that a systematic review on the previous efforts is important to understand the advantages and limitations of each method so as to stimulate further progress in this area.

We note that several review articles in this area have appeared in the literature. In their review, van Someren *et al.* [5] recorded a time line of various analyses or models on using large-scale gene expression data to reconstruct genetic regulatory networks. They discussed various models (static versus dynamic, deterministic versus stochastic, linear versus nonlinear), data types (continuous versus discrete), and estimation objectives (pairwise or triplewise versus combinatorial relationships among the genes). They also reviewed numerical algorithms to estimate model parameters and described the similarity or particularity of each method. De Jong [6] reviewed methodologies for pathway reconstruction and focused on the mathematical formalisms of possible models, including directed graphs, Bayesian networks; Boolean networks and their generalizations; ordinary and partial differential equations; qualitative differential

Keywords: Bayesian network, biological pathway, Boolean network, clustering, data integration, differential equation, genomics, proteomics

future
Medicine

equations; stochastic equations; and rule-based formalisms. De Jong discussed how these models had been employed in pathway analyses. Although these two reviews provided us an overview of the computational methodologies for biological pathway reconstructions, neither described the properties of known biological pathways with regard to their complexity, discussed the reasons for the limited applications of these mathematical models, or emphasized the importance of data and knowledge integration in model developments.

Given the vast literature on pathway reconstruction, it is neither possible nor our intention to cover all the methods that have been proposed to date to model biological pathways. Instead, we aim to discuss the models used in understanding the nature of known biological pathways, the limitations of previous modeling efforts, and the future directions of computational pathway reconstructions. This review is organized into the following sections:

- the complexity of biological pathways
- various types of genomic data
- different modeling approaches
- future directions

The section on modeling approaches is further divided into subsections, including clustering methods, scale-free networks, Bayesian networks (standard and dynamic), Boolean networks (standard and probabilistic), system of the differential equations, and data integration methods. We cover the discussions on the methodologies in the order of their potential to reveal the underlying mechanisms of biological pathways instead of following a historical time line of their applications. The overarching goal of this review is to identify unsolved difficulties encountered in the current efforts and to discuss future directions in reconstructing biological pathways.

The complexity of biological pathways

A biological pathway can be viewed as a network of chemical reactions or physical interactions, but it also serves as a route for mass transfer/transport or signal transduction to accomplish certain cellular functions. Some broad classes of pathways include metabolic pathways, transcriptional regulatory pathways, and signal transduction pathways. A biological pathway is often very complex, which includes a large number of components, the intricacy of the interfaces between them, different degrees of nesting, various types of data structures, and

channeling of chemical reactions, as well as the dynamic assembly, translocation and degradation of biological processes. It is apparently impossible to identify all the properties of a biological pathway in one step. For different types of pathways, different modeling and analysis strategies may be required. Here, we focus on how to obtain a graphic representation that captures the core relationships of proteins in a biological pathway. Even for this simplified task, we still need to battle with great complexity posed by a biological pathway.

The graphic presentation of a pathway includes a number of nodes or vertices representing metabolites/proteins/genes, and the directed or undirected edges between any pair of nodes measuring the dependence among the nodes. For example, in a signal transduction pathway, the nodes usually represent proteins or inorganic chemicals, and the edges represent the physical interactions (e.g., adsorption) or chemical reactions (e.g., phosphorylation/phosphorylase). In pathway modeling, these detailed interactions or reactions for the edges are generally classified as 'activation/inhibition' or simply annotated as 'interaction'. The measurement on 'interaction' can be in various forms such as binary data using 1 to indicate interaction and 0 to indicate the absence of interaction, or probabilistic values to quantify the probability of having certain association. On examining a single node in the graph, we find that it links with multiple neighboring nodes through the edges with input (manifesting the incoming edges of that node) or the edges with output (indicating the outgoing edges of that node). The fraction of the realized edges among all possible edges defines the connectivity of a pathway. In genomic studies, we need to identify the nodes (metabolites, proteins, genes) of a pathway as well as the edges (the dependences among these nodes) from the data for all or a large amount of the annotated or predicted genes of a genome. Szathmáry *et al.* [7] demonstrated two aspects of genomic complexity: one is the number of genes (nodes) and the other is the connectivity of genetic regulatory networks. Weng *et al.* [8] investigated the complexity in biological signaling systems. They illustrated that complexity may arise from the large number of components (nodes in graph), many with isoforms that have partially overlapping functions; from the connections among components (edges); and from the spatial relationship between components (nodes or edges). Many types of genomic data (e.g., gene expression

data) are the result of the combinatorial effects from many cellular pathways. Therefore, the process of constructing a biological pathway using such data is to identify a specific pathway from a complicated web of highly connected biological pathways. These functionally connected pathways can be genetic regulatory networks, metabolic pathways, and signal transduction cascades.

The first task in pathway reconstruction is to define a specific pathway for analysis. The complicated structure of biological pathways requires the definition of a pathway with a certain degree of abstraction. In principle, all the pathways are functionally related to support cell activities. However, it is commonly assumed that cellular functionality of pathways can be partitioned into a collection of modules, where each module is a discrete entity of several elementary components and performs an identifiable task, separable from the functions of other modules [9-14]. Ravasz *et al.* [15] studied the metabolic pathways of *Escherichia coli* and identified the hierarchical organization of the modularity of the metabolic pathways. The hierarchical structure of functional modules in pathways directly leads to different degrees of nesting in those pathways. Gagner *et al.* [16] applied hierarchical analysis of dependency in *E. coli* metabolic pathways to define individual pathways. However, the interactions between pathways [8] and the existence of protein isoforms lead to overlapping among functional modules. This complexity makes it difficult to mathematically define a pathway separate from other pathways, and expert knowledge may be needed to better define pathways of interest.

In addition to defining specific pathways for analysis, we also need to understand the topological properties of biological pathways. Many of the complex biological pathways share global statistical features: the modularity of the network topology (highly clustered connectivity), and the 'small-world' property that is characterized by short paths between any two nodes. Recent research showed that the metabolic pathways have a scale-free topology, where the probability that a substrate can react with k other substrates decays as a power law $P(k) \sim k^{-\gamma}$ with $\gamma \cong 2.2$ [17,18]. Wagner [19] reported that the protein-protein interaction networks are also scale-free and follow the power law. Maslov and Sneppen [20] quantified the correlations between the connectivity of interacting nodes (proteins in protein-protein interaction network or genes in genetic regulatory network). They observed a

power law relationship between the connectivity of the interacting pair of nodes for both the protein-protein interaction and genetic regulatory networks. They found that in both interaction and genetic regulatory networks, links (edges) between highly connected nodes are systematically suppressed, whereas those between highly connected and lowly connected pairs of nodes are favored.

Besides the investigations on the global properties of biological pathways, scientists have also studied the basic units of biological pathways. Milo *et al.* [21] identified the 'network motifs' from the directed networks. They applied their methods on the genetic regulatory networks of yeast and *E. coli* and found that the network motifs are present in both organisms: a three-node motif termed 'feedforward' loop and a four-node motif termed 'bi-fan'. These network motifs appear numerous times in the real network, and they are much more frequent than what is expected in a randomized network. Lee *et al.* [22] identified six genetic regulatory network motifs and presented an automated process of using the network motifs to assemble the transcriptional regulatory network structure. Papin *et al.* [23] used the extreme pathways (a unique minimal set of vectors) to completely characterize the steady-state capabilities of genome-scale metabolic networks. Based on the extreme pathway matrix and the stoichiometric matrix of the network, the reaction participation and the extreme pathway lengths can be computed. Those attribute values serve to elucidate systematic biological features.

The dynamic characteristic of biological pathways also contributes to the complexity of biological pathways. Although we currently ignore the detailed kinetics of biochemical reactions or transport mechanisms, we cannot avoid the dynamic feature of biological processes. Bhalla and Iyengar [24] indicated that the properties of signaling cascades, such as the feedback loop, likely support the 'learned behavior' of the biological system through intracellular biochemical reactions. Therefore, an external signal that activates the cellular responses may lead to a series of transient states of the biological pathways. It is important to understand which or what transient states are captured by the available genomic data. Therefore, the observed genomic data are not only the end products of the combinatorial effects from all the pathways but also are time dependent. This transient effect should be taken into account in pathway analyses.

Various types of genomic data

Recent technological advances allow us to collect many different types of data at a genome-wide scale, including DNA sequences, gene and protein expression measurements, protein–protein interactions, protein structural information, protein–DNA binding data, protein localizations, and chromatin structures. Previous knowledge on biological systems, such as pathway information, and gene functions are available from many databases (e.g., Swiss-Prot, Gene Ontology [GO], and Kyoto Encyclopedia of Genes and Genomes [KEGG]). Although all genomic data elucidate the large-scale modular organization of the cell, the information embedded in each data type reveals different aspects of cell activities, for example, different types of the biological pathways:

- Gene expression data more directly reflect the outcomes of transcription regulation and metabolic activities under a certain condition.
- Protein–protein interaction data and protein structure data are important to study signal transduction pathways or the formation of protein complexes.
- Protein–DNA binding data are critical to investigate genetic regulatory networks.
- Protein concentrations as a property of gene products can be associated with any data types in the analyses.

In addition, protein and DNA sequences, and accumulated knowledge on biological pathways and gene functions are also essential to build pathways.

Note that the genomic data are collected under various experimental designs. Large-scale gene expression data can be designed to measure the difference among organs/tissues, the responses to different experimental conditions, or the progression of certain biological processes over time. The utilization of those data in pathway analyses deserves additional attention so as to extract the most relevant information on pathways.

It is common, however, that experimental genomic data sets often contain errors due to imperfections in the applied technologies. The noise of genomic data plus the transient aspects of biological pathways add more uncertainty in the analysis and understanding of biological pathways.

Modeling approaches

Because of the significance of pathway reconstruction in understanding biological systems,

especially its application in drug discovery, scientists from multiple disciplines have made great efforts to develop computational methods to reconstruct pathways using various genomic data. The reconstructed pathways should possess similar properties to real biological pathways at both the global (scale-free and modularity) and local (network motifs) scales. In this section, we first review clustering methods as they have been widely applied in gene expression and pathway analyses. Following the discussion of clustering methods, we cover topics in the order of resolution and determinism of the network models: scale-free networks, Bayesian networks, Boolean networks, and system of differential equations. Specific issues, such as the uncertainty involved in the Boolean networks and the dynamic property of the Bayesian networks, are also discussed under each network subsection. Finally, we discuss methods based on integrating various types of data to infer biological pathways.

Clustering methods

Cluster analysis is one standard statistical pattern recognition method. Its goal is to group individuals in a population to discover the structure in the data. In genomic studies, clustering methods are often used to identify gene groups that are highly co-expressed over a large number of experiments (e.g., Eisen *et al.* [25]). It is generally accepted that genes in the same expression cluster tend to share similar biological functions (e.g., Wen *et al.* [26]). The direct applications of gene expression clustering (D'haeseleer *et al.* [27]) are:

- to study the upstream sequences of the genes in the same expression cluster to identify shared sequence patterns that are likely to be regulatory motifs
- to infer functions between the annotated and predicted genes in the same cluster
- to distinguish cell or tissue types using the genes defining the clusters as the molecular signature

In pathway analyses, the first application will potentially provide the nodes (the genes sharing similar regulation) in a regulatory network. The second application needs to be combined with other genomic information, such as functional annotation and known pathways, to assist pathway analysis. In any case, it is apparent that clustering methods alone do not lead to an understanding of the intricate relationships among all the genes in a pathway.

Ideker *et al.* [3] integrated genomic and proteomic analyses to investigate a metabolic network in yeast. They found that genes linked by physical interactions in the network tend to have more strongly correlated expression profiles than genes chosen at random. However, such correlation is weak (e.g., Jansen *et al.* [28]). The network interactions that likely transmit a change in expression from one gene (or protein) to another may be more easily identified by the group of highly co-expressed genes. Therefore, genes in one cluster that are highly co-regulated may not be in the same pathway. Nevertheless, the rich information in gene expression clusters can be integrated with other genomic data to discover or refine biological pathways (e.g., [2,3]).

Despite the limitation of using gene expression clusters to infer biological pathways, the clustering methods may extract valuable information from the genomic data. In practice, essentially all clustering methods have been applied to gene expression data to identify gene and experimental clusters, including:

- hierarchical methods [25,26,29,30]
- the sum-of-squares methods, such as the K-means method [31,32], the fuzzy K-means method [33], and the self-organized maps [34,35]
- the multivariate mixture models [36]
- the mixed-effects models [37]

Among these methods, the most frequently cited work is by Eisen *et al.* [25] who applied a standard agglomerative hierarchical clustering algorithm to yeast gene expression data. Their software, Cluster and Treeview, has many clustering and visualization functions. In many clustering algorithms a dissimilarity matrix is first calculated from the pairwise distances between genes, where the distance can be defined in many ways, for example, the Euclidean distance or $(1 - r)$, where r is Pearson's correlation coefficient. Because of the differences in distance measures and clustering algorithms, different clusters may result from the same genomic data. In fact, results from various clustering methods have been used to better annotate genes [38]. Another application of the clustering method is to cluster the metabolites in metabolic pathways based on their connectivity. The details are discussed in the next subsection. As for the robustness of the clustering results using the gene expression data, Zhang and Zhao [39] performed sensitivity analyses on hierarchical clustering algorithms to two large-scale data sets to identify clusters that are more reliable than others. Similar ideas have

been proposed by other groups (Kerr and Churchill [40], and McShane *et al.* [41]).

Clustering methods can be improved to better identify co-expressed gene groups as co-expression is defined in the context of the set of experiments being performed. If the objective is to investigate gene responses under a subset of conditions, the clustering results may be quite different from those using all the experimental data. As there are many different ways to partition the overall set of experiments, it is not a trivial matter to piece together clustering results under different sets of conditions to extract information in the identified gene clusters to infer gene functions, and to infer biological pathways.

Scale-free networks

Investigation on the topology of various biological pathways revealed that genetic regulatory networks, metabolic pathways and protein-protein interaction networks (closely related to signal transduction cascades) all have the scale-free property.

A network (G) can be represented by a graph with a set of nodes (V) and the links or edges (E) that connect the nodes ($G = (V, E)$). The number of nodes connected to a given node indicates the degree of the connectivity of that node. In a scale-free network, the frequency of a node with k degree of connectivity follows a power law, and such a relationship is kept throughout various scales of the network. Although most work on the scale-free network to date is concerned with the global property of biological pathways, some research groups have gone beyond examining the overall topology to study the genesis of such topological structures.

Ravasz *et al.* [15] showed that the hierarchical modularity is intrinsically embedded in the scale-free network of metabolites. They demonstrated that the clusters of metabolites on the basis of their connectivity in the network potentially encode their biochemical similarity. This method was then used to dissect metabolic pathways. Again, the uncertainty encountered in the clustering analysis is also present in this type of analyses.

Rzhetsky and Gomez [42] suggested a stochastic model to generate a genetic regulatory network on the basis of the annotated DNA motifs and protein-binding domains. They obtained a unique list of genes and connected these genes according to their DNA motifs and the binding domains in their gene products – proteins. They assumed that each gene or gene product contains

one upstream 'domain' and one downstream 'domain'. The dependence between any pair of genes was represented by the edges pointing from the upstream 'domain' of one gene or gene product to the downstream 'domain' of the other gene or gene product. They assumed that the initial network underwent an evolutionary process, where the duplication of domains and the innovation of the edges were considered to realize more possible DNA-protein bindings. The resulting network shares the statistical scale-free properties with real genetic regulatory networks. This scale-free property allows the estimates of model parameters obtained from the minimization of the model error between a subnet of the model network and the limited known genetic regulatory network, to be utilized in the prediction of the entire network. Thus, Rzhetsky and Gomez's network model is capable of both estimating the model parameters (the domain duplication constant and the edge innovation rates) and predicting unknown DNA-protein binding based on known genetic regulatory networks. This model, however, mainly captures the scale-free structure of the genetic regulatory network. It is easy to see that the completeness of the subnet of the known genetic regulatory network will significantly affect the estimation of model parameters. The assumption on the constant evolutionary parameters for different genes or gene products also limits the model's ability to accurately identify important genes in genetic regulatory networks. In addition, there was no discussion on integrating other available genomic data, such as protein-protein interaction data, to provide more information in the inference of genetic regulatory networks.

To date, work on the scale-free network has mainly emphasized the connectivity distribution in biological pathways. The application of hierarchical clustering methods on the network connectivity can reveal the modularity of the biological network to a certain degree, and these properties have been evaluated at the global network level. However, global analysis does not lend itself to discover local information on each specific biological pathway, which is the central focus for biologists.

Bayesian networks

As discussed above, all biological pathways have two topological properties: the modularity and scale-free properties. The topological modularity may indicate certain functional modularity of the networks. Some graph-based networks

models, such as the Bayesian and Boolean networks, were utilized to directly capture the functional modularity of complex biological systems. The local information of the network plays the key role in building both network models. The Boolean network is a deterministic model, whereas the Bayesian network is an explicit probabilistic network model. Both Boolean and Bayesian network models delineate more local properties than the scale-free network model, hence they are more relevant in biological inference than the scale-free network model. In this subsection, we focus on the Bayesian network model.

In Bayesian networks, the nodes are treated as one set of n random variables $\mathbf{X} = \{X_1, \dots, X_n\}$. The configurations of the variable set ($\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ and x_i is the value of the random variable X_i at the given system state) define the domain of the Bayesian network. The variables are connected with directed arcs, which point from the parent variables to the child variables. The set of variables \mathbf{X} and the arcs linking the variables consist of the structure of the Bayesian network S . The child variables are conditionally independent given the parent variables. Hence, the structure S encodes the conditional independence assertions about the variables in \mathbf{X} . For each variable in \mathbf{X} , there is a local probability distribution $p(x_i|pa_i)$, to quantify the conditional probability of the variable value given its parent variables. In summary, a Bayesian network (G) can be represented using a directed graph with certain structure S and local probability distribution P , that is $G = (S, P)$, where $S = (\mathbf{X}, \hat{\mathbf{E}})$ and $\hat{\mathbf{E}}$ represents the directed links or arcs; and $P = \{p(x_i|pa_i), i = 1, \dots, n\}$. An excellent review on the Bayesian network was written by Heckerman [43].

For a given Bayesian network ($G = (S, P)$), the joint probability distribution for a certain configuration (\mathbf{x}) of \mathbf{X} describes the probability of observing \mathbf{x} from the given Bayesian network. This joint probability distribution for \mathbf{X} can be simplified by the conditional independence among \mathbf{X} as:

$$p(\mathbf{x}|G) = \prod_{i=1}^n p(x_i|pa_i)$$

From the above definition, the Bayesian network has the following features:

- The uncertainty is explicitly considered in the Bayesian network.
- The embedded conditional independence in the structure S allows a partition of the large

computation on the entire network into small portions of the network.

- The probability of observing \mathbf{x} given the Bayesian network (G) can be computed, therefore the Bayesian network model can be used for prediction based on the previous observations.

The probabilities encoded by the Bayesian networks may be interpreted as Bayesian probability if the prior knowledge is incorporated in the estimation, or the physical probability if the probabilities are estimated only from data. In the modeling of biological pathways, there usually is some prior information available. A Bayesian approach coupling with the Bayesian networks model for the biological pathways allows us to combine prior information and the observed data, and provides a principled approach to learning relationships among variables and local probability distributions. Hence, we focus on the Bayesian approach in the following paragraphs.

In learning biological pathways using the Bayesian networks model, there are two general cases:

- learning the local probabilities P from a known structure S of the proposed Bayesian networks model
- learning both the local probabilities P and the structure S of the proposed Bayesian network

For the case with a given structure S , the posterior distribution of the unknown local probabilities, denoted as θ , is used to evaluate the learned local probabilities:

$$p(\theta|D, S) = \prod_{i=1}^n \prod_{j=1}^{q_i} p(\theta_{ij}|D, S)$$

In this formula, each variable X_i has q_i sets of possible configurations of its parent variables, θ_{ij} denotes the local probabilities corresponding to each set of configuration of its parent variables and is the posterior distribution of θ_{ij} with respect to r possible configurations of variable X_i ($\{x_{i1}, x_{i2}, \dots, x_{ir}\}$). The above equation implies an assumption of mutual independence among parameters θ_{ij} for $j=1, \dots, q_i$. Although this assumption is quite strong, similar assumptions have been adopted in the Bayesian network modeling to reduce the complexity of the system.

When both the structure (S^h ; where h stands for 'hypothesis') and the local probabilities (P) are unknown, learning the structure is usually the goal and the local probabilities can be treated as the nuisance parameters. Given the observations,

the posterior of the network structure is used to evaluate the learned Bayesian network:

$$p(S^h|D) \propto p(S^h)p(D|S^h)$$

In this formula, the network prior $p(S^h)$ can be specified. The conditional likelihood $p(D|S^h)$ may have the closed form for the complete and discrete data, but the Monte Carlo methods or the Gaussian approximation may need to be employed for incomplete data. The effect of the nuisance parameters θ is omitted through the integration for the marginal likelihood. The integration can be carried out either through the Monte Carlo methods or the Gaussian approximation. In the Monte Carlo methods, the marginal likelihood is calculated as:

$$p(D|S^h) = \frac{p(\theta|S^h)p(D|\theta, S^h)}{p(\theta|D, S^h)},$$

where the prior of the unknown parameters θ ($p(\theta|S^h)$) can be specified, the likelihood of a given Bayesian network ($p(D|\theta, S^h)$) can be assessed directly from the Bayesian inference, and the denominator term ($p(\theta|D, S^h)$) can be computed using Gibbs sampling. In the Gaussian approximation, the likelihood function can be evaluated in a closed form:

$$p(D|S^h) = \int p(D|\theta, S^h)p(\theta|S^h)d\theta$$

In this formula, $p(D|\theta, S^h)p(\theta|S^h)$ is assumed to be a multivariate normal distribution.

When we use the Bayesian network to model a biological pathway, we can either assess the behavior of a given Bayesian network or learn the probabilities (P) and/or the structure (S) of the network. To assess the behavior of the network, all the configurations of the random variables can be described through a set of joint probabilities of the network. The set of the variable configurations with the highest joint probability indicates the most likely outcome or 'behavior' of the network. This is a forward problem to simulate the behavior of the network. To identify a biological pathway, however, is to learn the local probabilities and/or the structure of the Bayesian network from the given observation D . This becomes an inverse problem. The posterior probabilities for the unknowns are used to evaluate the learned Bayesian network with respect to the observation D . When both the local probabilities and the structure have to be identified, the posterior probability of the unknown structure needs to be evaluated using the above Bayesian approach with respect to all possible network realizations and the variable configurations. The model with the highest Bayesian score

(e.g., the posterior probability for a hypothetical structure S of the network) is chosen to be the learning results. However, an exhaustive search over all realizations of the structure (which is to an exponential order of n) combining all configurations of the variables is an NP (non-deterministic polynomial time)-hard problem. Therefore, researchers have proposed heuristic search algorithms, including greedy search, greedy search with restarts, best-fit search, and Monte Carlo methods. The application of Bayesian network involves model selection or selective model averaging. Both methods will guard the model network from overfitting of data.

Friedman *et al.* [44] demonstrated the use of the Bayesian network to recover gene interactions through the analysis of yeast microarray data. Hartemink *et al.* [45,46] used the Bayesian network to infer biological pathways through the incorporation of latent variables and applied the selective model averaging method to score each Bayesian network. They evaluated a galactose system with three variables to distinguish two biologically meaningful hypotheses. However, there is no feedback in the Bayesian network framework and the common existence of cyclic biological pathways also challenges the application of the Bayesian network. Smith *et al.* [47] simulated the complex biological system using their NETWORK-INFERENCE algorithm. Their algorithm is based on the dynamic Bayesian network. The dynamic process of the Bayesian network follows the Markov chain. The transition probabilities are time independent. They model the cyclic paths as the different states of the Bayesian network at the adjacent two time-steps t and $t+\Delta t$. Ong *et al.* [48] also applied the dynamic Bayesian network to explore time-course expression data.

To a certain degree, some understanding on the biological pathways has been gained using the Bayesian network model. In addition to the computation difficulty in identifying a Bayesian network and the difficulty caused by the presence of incomplete data, one critical limitation of this approach is that many Bayesian network realizations can have equivalent joint distributions and this simple fact directly limits the application of the Bayesian network model for the purpose of the biological causal inference. As stated by Heckerman 'Given the causal Markov condition, we can infer causal relationships from the conditional independence and the conditional dependence relationships that we learn from the data' [43]. In the case for studying

biological pathways, the causal Markov condition is unclear. The lack of causal inference from the conditional dependence/independence of the Bayesian network indicates that the biological pathway modeled by the Bayesian network may be unidentifiable. In addition, all observations are assumed to stem from the same distribution, which clearly cannot model the dynamics of biological systems as well as responses to environmental perturbations. Although the dynamic Bayesian network may partially address these problems, the computational and theoretical implications of extension to more general models require further investigation. It has been reported in the literature [45] that the Bayesian network methodology was able to correctly identify the true biological model from two competing hypotheses, however, it became clear that this particular analysis was driven by 2 outlying observations from a total of 55 observations (H Zhao and B Wu, unpublished results). The Bayesian networks also failed to detect the galactose pathway from genomics data reported in Ideker *et al.* [3]. Furthermore, when a dynamic Bayesian network was applied to time-course data in *Drosophila* [49], it failed to identify the correct transcriptional regulatory network among three genes showing expression patterns clearly consistent with known biology (H Zhao and B Wu, unpublished results). A closer inspection of the cause of dynamic Bayesian network failure shows that the stationarity assumption underlying this approach may be too strong and inappropriate. Our experience with Bayesian networks and dynamic Bayesian networks suggests that a considerable amount of work needs to be done to improve current methods before meaningful results can be reliably extracted from genomic data.

Despite these issues, it is possible that association information among genes may be abstracted from data using the Bayesian networks model. However, it remains questionable whether the Bayesian networks model is the best modeling approach to extract associations among the variables. To be more specific, whether the directed graph itself is necessary in identifying the associations among genes? Bayesian statistical models, without invoking network structures, can also combine prior information and data, avoid overfitting the data, and detect the dependences/independences among the variables. In addition, it avoids the large complexity coupled with the network structure. Therefore, using the Bayesian

network for gene association studies does not appear to be an attractive approach.

Boolean networks

In contrast to the Bayesian network, a Boolean network is a deterministic network model. It reveals the logic relationship between one node and its surrounding nodes. The state of each node is determined by a series of k other nodes. The parameter k is called the in-degree of the node. These k neighboring nodes, following certain rules or logics (the Boolean functions), control the binary state of the node (On or Off state, recorded as 1 or 0, respectively). In the standard Boolean network, there is one set of n nodes and one set of n Boolean functions. The graph G for the Boolean network is presented by $G = (V, f)$. The V ($V = \{X_i; i = 1, \dots, n\}$) is the list of nodes. The f ($f = \{f_i; i = 1, \dots, n\}$) is the vector of Boolean functions (f). In a given Boolean network, the known Boolean functions (f) are used to update the states of all n nodes at time $t+1$ from their previous states at time t . The node states of the network define the state of the system. The state of the system evolves across time. Thus, the Boolean network is used to realize the dynamic behavior of the network. The long-term dynamics of a Boolean network leads the network to re-enter one of the previous state patterns of the system. This periodic state cycle of the Boolean network is called an attractor. The set of system states consists of an attractor of the system, called a confluent. For each attractor, the final cycle can be reached starting from any state in its confluent. Multiple attractors may exist for one Boolean network. No system state presents in more than one attractor. Therefore, each attractor of the network is behaviorally isolated from the others.

Akutsu *et al.* [50] discussed four typical problems involved in the Boolean network: consistency, counting, enumeration, and identification. Given a set of Boolean functions and the initial states of the nodes (the input of the network), the output of the Boolean network is determined. The consistency between the Boolean network and the observed output of the real system can be evaluated. This is the consistency problem. When there exist more than one set of Boolean functions supporting the consistency of their Boolean network realizations to the real system, the count of these Boolean networks forms the counting problem. The enumeration problem is to output all the possible Boolean networks that are consistent with the real system.

When the configuration of the Boolean network is unknown, the set of Boolean functions (f) needs to be identified from the Input/Output pairs of observations on the states of all nodes. The encountered identification problem is to determine whether there exists only one unique set of Boolean functions to consistently describe Input/Output pairs of the real system. Akutsu and colleagues [50] proved that a number of $O(2^{2k}(2k+\alpha) \log_n)$ of uniformly and randomly generated Input/Output pairs is sufficient for the identification of the unique Boolean network. However, for the time-course data, the Output at time t serves as the Input at time $t+1$. If the time-course data follows the dynamics of the Boolean network, the time-course data will tend to reach an attractor state after a long run. Therefore, Akutsu *et al.* [50] argued that such data are basically from one pair of Input/Output data and not enough to identify the unique Boolean network. Hence, a large amount of data from independent perturbation experiments is necessary to identify large-scale networks.

The Boolean network was originally introduced to model the biological system more than 30 years ago [51-53]. In this model, the genes are assigned as the nodes of the network. The states of the genes are binary (0 for OFF and 1 for ON) representing the presence and absence of a gene expression. The edge indicates the activation or inhibition of one gene to another. The edge information is summarized in the Boolean function associated with each node, which represents a combined effect from its surrounding nodes following the Boolean rules (AND, OR, Exclusive OR, and Not). Liang *et al.* [54] developed the REVerse Engineering ALgorithm (REVEAL) to build the network according to the states of all nodes. They used mutual information to determine the rules for the node with k degrees of connectivity. The search was repeated with the increasing value of k . Yuh *et al.* [1] investigated the *cis*-regulatory logic in the upstream sequences of a sea urchin gene: endo 16. The seven modules (G to A) were modeled as the nodes. Their quantitative model successfully uncovered the logic interactions among these modules. However, real biological pathways contain large uncertainty, the inherent determinism of the standard Boolean network tends to overfit biological data. This becomes one salient limitation of the Boolean network in describing real networks.

Akutsu *et al.* [55] presented a Boolean network with noise. They assumed the Boolean function

is held with a certain probability in the noisy Boolean network. Shmulevich *et al.* [56,57] applied the Probabilistic Boolean Network (PBN) to reconstruct the genetic regulatory network using gene expression data. In PBN, $G = (V, F)$, V is similarly defined as in the standard Boolean network $F = \{F_i; i = 1, \dots, n\}$. Due to the imprecision in experimental data, they allowed a random selection between several Boolean functions (F_i) instead of choosing one single Boolean function (f_i in the standard Boolean network) for each node. The total number of the realizations of the network N equals to $\prod_{i=1}^n l_i$. In this formula, l_i is the number of Boolean functions in F_i for node i . All the network realizations can be represented in a \mathbf{K} matrix with rows corresponding to network realization and columns corresponding to the nodes, and the element indicating the identity of the Boolean function used for the specific node in the particular network realization. Based on the truth table of all the Boolean functions, the \mathbf{K} matrix, and the network probability, one can obtain the state transition matrix \mathbf{A} . The simulation on the prediction of all the networks in \mathbf{K} matrix can be performed. The output of the system can be estimated from the input states of the Boolean network and transition matrix \mathbf{A} . The coefficient of determination (COD) [58] can be used to measure the degree to which the transcriptional levels of an observed gene set can be used to improve the prediction of the transcriptional level of a target gene relative to the best possible prediction in the absence of observations. The COD value determines the weights of selection of these Boolean functions. Finally, the probabilities of different network structures can be calculated based on the selection probabilities of the Boolean functions in that network realization. In their model, Shmulevich *et al.* also discussed using simulations to reveal the dynamics of the PBN and the identification of the significant genes in the selected PBN.

When the data have no 'noise', the standard Boolean network can successfully capture the functional modularity of the network. The Boolean functions inherit the logical determinism and represent a clear causal inference. However, this determinism is a serious limitation in pathway reconstruction when the 'noisy' genomic data are utilized. The PBN model averages the networks models by assigning the Boolean functions with different selection probabilities. This approach reduces the overfitting of the 'noisy' data so that it relaxes the

rigidity from the determinism of the Bayesian network. The causal inference is statistically embedded in the selection probabilities of the Boolean functions for each node. However, the introduction of the selection probabilities of the Boolean functions for each of n network nodes leads to a significant increase in computational complexity. More importantly, the identifiability of the PBN has not yet been theoretically addressed.

It seems that we encounter a dilemma here in that the implicit or explicit introduction of the probability into the network structure will avoid the overfitting of the noisy data so that the network model can be learned from these noisy data, but at the same time, it may limit us from making causal inference from the annotation of the edges, arcs or links of the network graph. Both the dynamic Bayesian network and the Boolean network are designed to describe the dynamic process. In the dynamic Bayesian network, a Markov chain process is assumed for the transition of the network along the time coordinate. In the Boolean network, the Boolean functions serve as the transition rules of the network along the time coordinate. Both transition functions are time independent, which may not be true to describe the dynamics of a biological pathway because it is controlled by the kinetics of the biochemical reactions/interactions or mechanism of transport among the chemical compounds. Although the Boolean network has stronger causal inference than the Bayesian network, both network models are less deterministic than the system of differential equations describing biological pathways, especially in terms of capturing the dynamic behaviors of the system.

System of differential equations

Although our goal is to identify the causal graph of biological pathways, the dynamic effect of the real network on the observed genomic data cannot be totally ignored. Whether the imposed dynamic behavior in the network models, such as the Boolean and dynamic Bayesian networks, is consistent with the reality may be revealed from studies on the mechanisms of the pathway dynamics. Since the associations or causal inferences among the network variables or nodes represent biochemical reactions, physical interactions, and/or transport of the chemical compounds, each edge, arc or link can be further annotated with a quantitative formula instead of the annotation on the relationships. Normally,

the quantitative formulae display the frame of the differential equations. All the edges, arcs or links of the network are then depicted as a system of coupled differential equations.

The system of the differential equations can be derived from the general properties of the system, such as the mass continuity, or the more accurate kinetics of the biochemical reactions. The differential system could be ordinary differential equations (ODEs) or partial differential equations (PDEs), linear or nonlinear.

In genomic research, the system of differential equations has been utilized to capture the dynamics embedded in the time-course data. The ODEs are mainly used to describe time-dependent kinetics, whereas the PDEs also include the spatial effects so that the latter system can also reveal transport mechanisms like molecular diffusion. Although De Jong [6] discussed the PDE as one possible model for reconstructing biological pathways, there is no report on the successful application of PDE on reconstructing genetic networks using genomic data. The ODE systems have several applications, and they can be classified as linear or nonlinear ODE systems. One general formation of a linear system was given by Gardner *et al.* [59]:

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x} + \mathbf{u}$$

In this formula, the states of the nodes in the network have continuous values and are stored in the vector \mathbf{x} . The vector \mathbf{u} represents the external effects on the system (like 'sink' or 'source' terms). The matrix \mathbf{A} is the network model. With suitable perturbations, the positive or negative effects among the nodes can be identified. Several other studies also adopted similar linear differential equation models [27,60]. However, the kinetics of the biological reactions is limited to first-order reactions in the linear ODE model, and real systems may be much more complex.

Savageau [61-65] proposed the multivariate power-law functions (the \mathcal{S} -system) to approximate the kinetic rate laws of the reactions. The \mathcal{S} -system is a typical nonlinear ODE system. In an \mathcal{S} -system, the n variables consist of both the dependent and independent variables, and the kinetics for the i^{th} variable (x_i) is represented as:

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^n x_j^{g_{ij}} - \beta_i \prod_{j=1}^n x_j^{h_{ij}}$$

In this formula, α_i and β_i are the kinetics rate constants, and g_{ij} and h_{ij} are the kinetic orders. Voit *et al.* [66] have developed such a system to interpret the glycolysis gene expression pattern

of heat-shocked yeast. Kikuchi *et al.* [67] studied the \mathcal{S} -system of Hlavacek and Savageau [68] and applied the genetic algorithm to estimate the model parameters through the optimization steps. Their goal was to extend the \mathcal{S} -system formalism to a relatively larger scale of the pathways. The graph realization of the genetic network can be retrieved from the stoichiometric matrix underlying the \mathcal{S} -system. For a network with n variables, there are $O(2n+2n^2)$ parameters in the model. The identification problem of such a system is rather difficult. How to avoid the local minimums in searching the parameter space and how to overcome the stiffness of the differential system are also major difficulties in the applications of the ODE model.

Because of the above difficulties in applying the \mathcal{S} -system, it has been mainly used to simulate the behaviors or refine the structure of the well-understood metabolic pathways (e.g., see [69-72]). The computational outcome of the system can be compared to the time-course experimental observations. Since there is a lack of direct measurements on metabolites or pathway fluxes in genomic studies, gene expression data from perturbed or time-course experiments were used in the above studies. However, gene expression data mainly reflect the end products of gene regulation and metabolic activities, and there exists a big gap between gene expression and protein expression, and furthermore, the activity of the enzyme proteins, which control the kinetics of the reactions. In fact, our recent studies on the Calvin cycle of *Arabidopsis* showed that the gene expression data contain some sensitivity information on the regulation potential of enzyme genes in the specific pathway [73]. However, the sensitivity information was not explicitly present in the gene expression data, neither did it directly reflect the structure of the metabolic pathway.

Owing to the unknown pathway structure and higher degree of complexity than graph mapping in the differential equation system, the large-scale genomic data containing the most relevant information on the biological pathways are required. Gene expression data alone are not sufficient to reconstruct such a complex system. Even if the large-scale enzyme activities or pathway fluxes were measured, the large number of kinetics parameters that are required to be estimated coupled with possible incomplete data would make the application of the system of the differential equations to large-scale genomic studies unrealistic. However, this method could be the ultimate approach to

discover and describe the dynamics of the pathways with a smaller scale study after enough information is collected.

Data integration methods

Reconstruction of biological pathways is a process of reverse engineering. Identifiability is one main issue to evaluate the applicability of mathematical models and computational methods. The methods discussed so far have the following order of determinism: the scale-free network model can be used to capture the global properties of the networks, which can then be followed by the indeterministic Bayesian network model; the logically deterministic Boolean network model; and, finally, to the quantitatively deterministic differential equation systems in order to uncover the dynamics of the pathways. Note that these different network models are suitable for different kinds of genomic data, and the extracted information reveals different aspects of biological pathways. Therefore, data integration is an essential part of biological pathway reconstruction, and many published studies have employed this approach in order to better understand biological pathways.

Roberts *et al.* [2] combined their expert knowledge and gene expression pattern to reveal the signaling and circuitry of multiple mitogen-activated protein kinase (MAPK) pathways. Ideker *et al.* [3] more explicitly demonstrated an integrated approach of building, testing and refining galactose utilization (GAL) by adopting a genetic model for GAL based on prior knowledge, and qualitatively integrating the information from the global networks of protein–protein interaction, protein–DNA binding, and other known physical interactions with gene expression data of various biological perturbations. Davidson *et al.* [4] derived a genomic regulatory network for development from large-scale perturbation analyses, in combination with computational methodologies, *cis*-regulatory analysis, and molecular embryology. It is obvious that such analyses required a great deal of expert opinions, genomic data analyses, and model realizations. A data integration model should inherit the basic idea of these inductive and heuristic analyses, and properly symbolize knowledge-based information and automate the induction procedures. With a model, one can also perform predictions and simulations to explore the global as well as local properties of the pathways. Through this approach, we can extract information from known pathways to develop a

deduction framework for different types of networks, thus extending our knowledge of these.

Here we discuss the efforts toward building an integrated model, including those in symbolizing the descriptive expert knowledge on biological pathways and those on modeling pathway structures, to allow the incorporation of prior information as well as the knowledge taken from various types of genomic data.

Valuable expert knowledge on a given biological pathway is usually made available through large numbers of publications, and there have been recent efforts to develop a computational system to extract information from the literature. For example, the Natural Language Processing (NLP) technique has been developed to extract such information [74]. In NLP, a given string is tokenized, the frequency or conditional frequency distributions of particular tokens are evaluated, and a list of related tokens are then provided through certain methods (e.g., the count method, the N method, or the maximum method). The information is stored in a database with a schema built under certain ontology. The ontology defines the concepts representing domain-specific entities and the relationships among the concepts. The ontology allows the complexity and the richness of domains stored within a hierarchical tree. Such information is useful to map biological pathways. Among the many types of biological ontology, several target bioinformatics studies include: the Molecular Biology Ontology (MBO) [75], GO, and the TAMBIS Ontology (TaO). MBO aims to contain concepts and relationships that are required to describe biological objects, experimental procedures and computational aspects of molecular biology. GO narrows the scope to capture information about the role of gene products within an organism. TaO targets bioinformatics tasks and resources. Since we seek to reconstruct biological pathways, it is mainly at the cellular level. Therefore, GO with rich information on genes may be appropriate for this goal. EcoCyc [76], however, has developed its own ontology to structure the domain entities (nodes of the pathways) and the domain relationships (the potential edges of the pathways) under the scope of its specific biological questions. Through the structure of the EcoCyc ontology, complicated biological relations can be reduced to simple relationships, like activation and inhibition. The simplified relationships can be represented by edges without orientations. An initial model for a pathway may be visualized as a two-dimensional map.

Many other databases have been developed to store biological pathways, including KEGG, Swiss-Prot, Tremble, and Pathway Microarray Processor for Arabidopsis (PathMAPA). The node and edge information of the pathways are retrievable by querying the database, and such extracted information can be utilized for computational modeling. Some databases also support visualization of predicted pathways using various models. For example, Genetic Networks Analyser (GNA) [77] applies the Boolean network model to learn genetic regulatory networks from the gene expression data. CelleratorTM [78] uses an ODE system to describe single and multi-cellular signal transduction networks. eXPatGen [79] is an online gene expression pattern generator that allows the users to simulate gene expression data and systematically simulate data with observed data to evaluate different models.

However, the more important step is the identification of model structures, which allows the automatic incorporation of existing knowledge and observed data. Since model identification is an NP-hard problem, heuristic search following biologically meaningful model selection rules is necessary. The biological information can be introduced as the prior or thresholds in model selections, while the model should also have the learning ability to gather new information from genomic data. Gomez *et al.* [80] presented a Bayesian approach to predicting protein–protein interactions. The scale-free network is used to describe the general topology of the overall network. For a given network, the set of network nodes can be decomposed into multiple bins in which the topological property (in- and out-degree) of all nodes is identical. In each bin, the probability of a node having the given topology is computed by a multinomial distribution. The power law of the scale-free network is used to estimate these probabilities. The represented scale-free property of the network is directly associated with the network nodes. For the same given network, the probability of having an edge between any pairs of proteins is evaluated by the protein–protein attraction probability. To obtain the protein–protein attraction probability, they represented proteins as a collection of domains or motifs, and transferred the domain–domain interactions to evaluate the protein–protein attraction probability. Finally, the given network is scored from both the probability of having the scale-free topological property and the probability of having individual edges based on

protein–protein interactions. The author also suggested the following Bayesian model to automate the entire searching procedure over the network realization space:

$$P(\text{network}_i | \text{data}) = \frac{P(\text{network}_i | \text{data})P(\text{network}_i)}{\sum_{\text{all networks}} P(\text{network}_i | \text{data})P(\text{network}_i)}$$

This model structure combining powerful search algorithms has strong learning ability. We believe there is a great potential in this type of model in reconstructing biological pathways.

Other than statistical learning models, Boolean networks, Bayesian networks and decision trees all contain strong machine learning capacity. In order to decide which model to use, it is first necessary to decide what properties of the network we would like to model and how available data can be incorporated into the network. For instance, gaining knowledge of the network structure is no longer crucial if we are mainly concerned with identifying pairwise associations among the genes.

As for formal statistical analysis to integrate different data types, most research has focused on inferring protein–protein interaction networks through various genomic data. For example, gene expression data and DNA sequence data can be utilized to better predict protein–protein interactions [81], and the inference can be further strengthened by the integration of mutation data, protein localization data, and ontology data [82]. Gene expression data and DNA–protein interaction data have been combined to infer transcription regulatory networks [83].

Discussion on future studies

The above discussion demonstrates that many efforts have been made to overcome the complexity of biological pathways and to reconstruct pathways from genomic data. However, there is a lack of connection between the properties of known biological pathways and the properties of mathematical and computational models. The modularity of biological pathways has been studied with respect to network connectivity, which may refer to the functionality of modular pathways. The application of this property may lead to the reduction of the complexity of biological pathways by decomposing the network into modules so that we can focus on modeling individual modules.

There also lacks a connection between mathematical models and available genomic data. Much research has been conducted to model genetic regulatory networks using gene

Highlights

- Most biological pathways share the global properties of the topology: the scale-free network and the modularity. They also have the tendency of having certain network motifs/modules.
- The clustering method may capture the functional relationship among genes but require other information to realize a network structure.
- The scale-free network models are mainly based on the global properties of the biological pathway.
- Bayesian network models delineate the association among the genes, robust to the noisy genomic data but lack clear biological causal inference.
- Boolean network models are logically deterministic. They tend to overfit the noisy genomic data so their application is limited.
- The system of differential equations may uncover the dynamics of the biological pathways. However, its application may be limited to well-studied pathways involving a limited number of genes.
- Data integration methods represent one of the most promising directions for pathway reconstruction. Two major efforts had been made: one is to symbolize the descriptive biological knowledge, and the other is to the joint analysis of various data types.
- Future work needs to emphasize the connection between mathematical models and available genomic data, quantitative integration between known pathway properties and various data types, and automatic incorporation of descriptive biological knowledge into mathematical models.

expression data alone and some have achieved great success, for example, Segal *et al.* [84]. Given the large noise in gene expression data, the deterministic Boolean network model will fail for its lack of robustness, and the Bayesian network model will fail for its inherent identifiability problem. In addition, recent studies show that gene expression data provide weaker signals than protein–DNA binding data when dissecting the gene regulation pathways.

The third general deficiency in current modeling approaches is the lack of connection between known descriptive knowledge and mathematical

models. An efficient heuristic search algorithm is important in extracting the pathway from the observation. The symbolized or digitalized known information from the literature may limit the search domain of the mathematical models greatly to reduce the computation burden.

Although we should develop mathematical models to target certain biological pathways and utilize data with the strongest signals, related genomic data can also be used to refine the model. There lacks a model structure with the ability of incorporating both the known information and the knowledge learned from other genomic data.

The noisy genomic data leave us with a dilemma in regards to choosing between a deterministic model and a flexible probabilistic model. We think a qualitative or a narrow domain of the real biological pathway is the main goal at the current stage. A probabilistic model may be favored to sketch the pathway structure, while a more deterministic model may provide detailed information on the smaller-scale pathways.

No matter what models we apply to reconstruct the unknown biological pathways, the pathways should preserve the global properties of the biological pathways. The simulation tools could play important roles in deducting latent variables in the pathways, and eventual success will come from integrated analysis and modeling of various types of prior knowledge and data types.

Acknowledgments

This work was supported in part by the NSF grant DMS-0241160 and NIH grant R01 GM59507.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Yuh C, Bolouri H, Davidson EH: Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 279, 1896-1902 (1998).
2. Roberts CJ, Nelson B, Marton MJ *et al.*: Signaling and circuitry of multiple MAPK pathways revealed by a matrix of Global gene expression profiles. *Science* 287, 873-880 (2000).
3. Ideker T, Thorsson V, Ranish JA *et al.*: Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929-934 (2001).
4. Davidson EH, Rast JP, Oliveri P *et al.*: A genomic regulatory network for development. *Science* 295, 1669-1678 (2002).
5. Van Someren EP, Wessels LFA, Backer E, Reinders MJT: Genetic network modeling. *Pharmacogenomics* 3, 1-19 (2002).
6. De Jong H: Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9, 67-103 (2002).
7. Szathmáry E, Jordán F, Pál C: Can genes explain biological complexity? *Science* 292, 1315-1316 (2001).
8. Weng G, Bhalla US, Lyengar R: Complexity in biological signaling systems. *Science* 284, 92-96 (1999).
9. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: From molecular to modular cell Biology. *Nature* 402(Suppl. 6761), C47-C52 (1999).
10. Lauffenburger DA: Cell signaling pathways as control modules: complexity for simplicity? *Proc. Natl. Acad. Sci. USA* 97, 5031-5033 (2000).
11. Rao CV, Arkin AP: Control motifs for intracellular regulatory networks. *Ann. Rev. Biomed. Eng.* 3, 391-419 (2000).
12. Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR: Dynamic

- modeling of gene expression data. *Proc. Natl. Acad. Sci. USA* 98, 1693-1698 (2001).
13. Hasty J, McMillen D, Isaacs F, Collins JJ: Computational studies of gene regulatory networks: *in numero* molecular biology. *Nat. Rev. Genet.* 2, 268-279 (2001).
 14. Shen-Orr S, Milo R, Mangan S, Alon U: Network motifs in the transcriptional regulatory network of *Escherichia coli*. *Nat. Genet.* 31, 64-68 (2002).
 15. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL: Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551-1555 (2002).
 - **The authors systematically investigated the scale-free property and the modularity of the metabolic networks and identified the hierarchical organization of modularity.**
 16. Gagneur J, Jackson DB, Casari G: Hierarchical analysis of dependency in metabolic networks. *Bioinformatics* 19, 1027-1034 (2003).
 17. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL: The large-scale organization of metabolic networks. *Nature* 407, 651-654 (2000).
 18. Wagner A, Fell D: The small world inside large metabolic networks. *Proc. R. Soc. Lond., B, Biol. Sci.* 268, 1803-1810 (2001).
 19. Wagner A: The yeast protein interaction network evolves rapidly and contains few duplicate genes. *Mol. Biol. Evol.* 18, 1283-1292 (2001).
 20. Maslov S, Sneppen K: Specificity and stability in topology of protein networks. *Science* 296, 910-913 (2002).
 21. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: Network motifs: simple building blocks of complex networks. *Science* 298, 824-827 (2002).
 22. Lee TI, Rinaldi NJ, Robert F *et al.*: Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799-804 (2002).
 - **The authors analyzed several network motifs and built transcriptional regulatory circuitry using the statistically selected genes from both gene expression data and DNA-protein binding data.**
 23. Papin JA, Price ND, Palsson BO: Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome Res* 12(12), 1889-1900 (2002).
 24. Bhalla US, Iyengar R: Emergent properties of networks of biological signaling pathways. *Science* 283, 381-387 (1999).
 25. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863-14868 (1998).
 - **The authors clustered genes according to their gene expression patterns.**
 26. Wen X, Fuhrman S, Michaels GS *et al.*: Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* 95, 334-339 (1998).
 27. D'haeseleer P, Liang S, Somogyi R: Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707-726 (2000).
 28. Jansen R, Greenbaum D, Gerstein M: Relating whole-genome expression data with protein-protein interactions. *Genome Res* 12, 37-46 (2002).
 29. 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* 9, 3273-3297 (1998).
 30. Alon U, Barkai N, Notterman DA *et al.*: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probe by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96, 6745-6750 (1999).
 31. Ben-Dor A, Shamir R, Yakhini Z: Clustering gene expression patterns. *J. Comput. Biol.* 6, 281-297 (1999).
 32. Zhu J, Zhang MQ: Cluster, function and promoter: analysis of yeast expression array. *Pac. Symp. Biocomput.* 5, 476-487 (2000).
 33. Mjolsness E, Castano R, Gray A: Multi-parent clustering algorithms for large-scale gene expression analysis. *Technical Report JPL-ICTR-99-5 Jet Propulsion Laboratory* (1999) (Section 367).
 34. 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. USA* 96, 2907 (1999).
 35. Marvroudi S, Papadimitriou S, Bezerianos A: Gene expression data analysis with a dynamically extended self-organized map that exploits class information. *Bioinformatics* 18, 1446-1453 (2002).
 36. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, 977-987 (2001).
 37. Luan Y, Li H: Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* 19, 474-482 (2003).
 38. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31, 255-265 (2002).
 39. Zhang K, Zhao H: Assessing reliability of gene clusters from gene expression data. *Funct. Integr. Genomics* 1, 156-173 (2000).
 40. Kerr MK, Churchill GA: Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA* 98, 8961-8965 (2001).
 41. McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R: Methods of assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* 18, 1462-1469 (2002).
 42. Rzhetsky A, Gomez SM: Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics* 17, 988-996 (2001).
 - **The authors applied the scale-free networks model to predict the protein-protein interaction network.**
 43. Heckerman D: A tutorial on learning with Bayesian networks. *Technical Report MSR-TR-95-06* Microsoft Research (1996).
 44. Friedman N, Linial M, Nachman I, Pe'er D: Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601-620 (2000).
 45. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA: Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* 6, 422-433 (2001).
 46. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA: Combining location and expression data for principled discovery of genetic regulatory networks. *Pac. Symp. Biocomput.* 7, 434-449 (2002).
 47. Smith VA, Jarvis ED, Hartemink AJ: Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics* 18, S216-S224 (2002).
 48. Ong IM, Glasner JD, Page D: Modeling regulatory pathways in *E. coli* from time series expression profiles. *Bioinformatics* 18, S241-S248 (2002).
 - **The author systematically described the properties of the Bayesian networks.**
 49. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, White K: Gene expression during the life cycle of *Drosophila melanogaster*. *Science* 297, 2270-2275 (2002).

50. Akutsu T, Miyano S, Kuhara S: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Proc. 9th Ann. CAN-SIAM SODA*, 695-702 (1998).
- **The authors investigated the identification problem in the Boolean network model.**
51. Kauffman SA: Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* 22, 437-467 (1969).
52. Kauffman SA: *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press, New York (1993).
53. Glass K, Kauffman SA: The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.* 39, 103-129 (1973).
54. Liang S, Fuhrman S, Somogyi R: REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* 3, 18-29 (1998).
55. Akutsu T, Miyano S, Kuhara S: Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics* 16, 727-734 (2000).
56. Shmulevich I, Dougherty ER, Kim S, Zhang W: Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* 18, 261-274 (2002).
- **The authors applied the probabilistic Boolean networks in reconstructing genetic regulatory networks.**
57. Shmulevich I, Dougherty ER, Zhang W: Gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics* 18, 1319-1331 (2002).
58. Dougherty ER, Kim S, Chen Y: Coefficient of determination in nonlinear signal processing. *Signal Process.* 80, 2219-2235 (2000).
59. Gardner TS, di Bernardo D, Lorenz D, Collins JJ: Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301, 102-105 (2003).
- **The authors applied the linear ODE system to infer genetic networks.**
60. Tengér J, Yeung MKS, Hasty J, Collins JJ: Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. USA* 100, 5944-5949 (2003).
61. Savageau MA: Biochemical system analysis, I. Some mathematical properties of the rate law for the component enzymatic reactions. *J. Theor. Biol.* 25, 365-369 (1969).
62. Savageau MA: Biochemical system analysis, II. The steady-state solutions for an n-pool system using a power-law approximation. *J. Theor. Biol.* 25, 370-379 (1969).
63. Savageau MA: Biochemical system analysis, III. Dynamic solutions using a power-law approximation. *J. Theor. Biol.* 26, 215-226 (1970).
64. Savageau MA: The behavior of intact biochemical control systems. *Curr. Top. Cell. Regul.* 6, 63-129 (1972).
65. Savageau MA: *Biochemical Systems Analysis. A Study of Function and Design in Molecular Biology*. Addison-Wesley, Reading, MA, USA (1976).
66. Voit EO, Radivoyevitch T: Biochemical systems analysis of genome-wide expression data. *Bioinformatics* 16, 1023-1037 (2000).
- **The authors applied the S-system (a nonlinear ODE system) to interpret the glycolysis gene expression patterns of heat-shocked yeast.**
67. Kikuchi S, Tominaga D, Arita M, Takahashi K, Tomita M: Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics* 19, 643-650 (2003).
68. Hlavacek WS, Savageau MA: Rules for coupled expression of regulator and effector genes in inducible circuits. *J. Mol. Biol.* 255, 121-139 (1996).
69. Sauro HM: SCAMP: a general-purpose simulator and metabolic control analysis program. *Comput. Appl. Biosci.* 9, 441-450 (1993).
70. Mendes P: GEPASI: a software package for modeling the dynamics, steady states and control of biochemical and other systems. *Comput. Appl. Biosci.* 9, 563-571 (1993).
71. Mendes P: Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* 22, 361-363 (1997).
72. Holzhutter HG, Colosimo A: SIMFIT: a microcomputer software toolkit for modelistic studies in biochemistry. *Comput. Appl. Biosci.* 6, 23-28 (1990).
73. Sun N, Ma L, Pan D, Zhao H, Deng XW: Evaluation of regulatory potential of Calvin cycle pathway steps based on large-scale gene expression profiling data. *Plant Mol. Biol.* (2004) (In Press).
74. Rzhetsky A, Koike T, Kalachikov S *et al.*: A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics* 16, 1120-1128 (2000).
- **The authors used the NLP to collect the knowledge from the literatures. This is the step of symbolizing the descriptive knowledge.**
75. Schulze-Kremer S: Ontologies for molecular biology. *Pac. Symp. Biocomput.* 3, 693-704 (1998).
76. Karp PD: Metabolic databases. *Trends Biochem. Sci.* 23, 114-116 (1998).
- **The author emphasized the idea of symbolizing the descriptive biological knowledge.**
77. De Jong H, Geiselman J, Hernandez C, Page M: Genetic network analyzer: qualitative simulation of genetic regulatory networks. *Bioinformatics* 19, 336-344 (2003).
78. Shapiro BE, Mjolsness ED: Developmental simulation with cellerator. *Proceedings of the 2nd International Conference on Systems Biology (ICSB)*, Pasadena, CA, USA (2001).
79. Michaud DJ, Marsh AG, Dhurjati PS: eXPatGen: generating dynamic expression patterns for the systematic evaluation of analytical methods. *Bioinformatics* 19, 1140-1146 (2003).
80. Gomez SM, Lo S, Rzhetsky A: Probability prediction of unknown metabolic and signal-transduction networks. *Genetics* 159, 1291-1298 (2001).
- **The authors presented a Bayesian approach to incorporating known protein domains and scale-free property of the protein-protein interaction networks into the modeling.**
81. Deane CM, Salwinski L, Xenarios I, Eisenberg D: Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* 1, 349-356 (2002).
82. Jansen R, Yu H, Greenbaum D *et al.*: A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302, 449-453 (2003).
- **The authors built the 'golden standard' for both the positive and the negative protein-protein interactions to estimate the likelihood of pairs of proteins presenting in the same complex within different data types. Then they separated the data types into two groups according to the dependence among the different data: one group contains four protein-protein interactions data, and the other group contains gene expression data, MIPS (Munich Information Center for Protein Sequences) function annotation, GO biological process, and essentiality data. The authors then joined the likelihood from various data types through Bayes' rule for both groups.**
83. Bar-Joseph Z, Gerber GK, Lee TI *et al.*: Computational discovery of gene modules

and regulatory networks. *Nat. Biotechnol.* 21(11), 1337-1342 (2003).

- **The authors combined the DNA–protein binding data and the gene expression data to identify the regulatory modules.**
84. Segal E, Shapira M, Regev A *et al.*: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166-176 (2003).