BIOCHIP JOURNAL, Vol. 3, No. 3, 190-202, September 2009

Mini Review

# Statistical Approaches to Genome-wide Biological Networks

**Jin Hwan Do**[1,2], **Satoru Miyano**[2]
**& Dong-Kug Choi**[3]

[1]Bio-Food and Drug Research Center, Konkuk University, Chungju 380-701, Korea
[2]Human Genome Center, Institute of Medical Science, University of Tokyo, Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[3]Department of Biotechnology, Konkuk University, Chungju 380-701, Korea
Correspondence and requests for materials should be addressed to D.-K. Choi (choidk@kku.ac.kr)

Accepted 4 August 2009

## Abstract

**The experiments based on high-throughput technology have been producing massive genomic data including protein-protein interactions, genome-wide mRNA expression and whole genome sequences, which allows the reconstruction of genome-wide biological networks representing relationships or interactions between genes or proteins. The network approach to biology is becoming the main framework to understand biological systems consisting of numerous dynamic networks of biochemical reactions and signaling interactions between cellular components. This is mainly due to efficient representation of a large amount of biological information. Many statistical models have been built and applied to construction of genome-wide biological networks from various type of high-throughput data. In this study, we survey statistical approaches to construction of four main biological networks with their pros and cons: gene regulatory networks, protein-protein interaction networks, metabolic networks and signal transduction networks. In addition, we also investigate the methods describing dynamic behavior of gene regulatory networks and signal transduction networks.**

**Keywords:** Bayesian networks, Boolean networks, State space approach, Flux balance analysis, Petri nets

## Introduction

The massive amount of 'omics' data including genomics, proteomics and metabolomics have offered opportunities to construct various biological networks such as, respectively, gene regulatory networks, protein-protein interaction (PPI) networks and metabolic networks. Although current technology does not represent the whole cellular system as a network because of its redundancy and complexity, a master global reaction network could still be formulated to represent the complete repertoire of possible biochemical reaction systems within the cell[1-3]. Genome-wide biological networks can be built based on interaction information between biological elements from genomic data produced by high-throughput technology. Thus, interaction analysis has emerged as an increasingly popular framework for exploring the complex system of relationships that characterize the functional organization of cellular environments[4].

The large-scale interactions are estimated directly by experiments or indirectly inferred by statistical models with experimental data. Here, we will focus on the statistical approaches to the construction of genome-wide biological networks from various types of high-throughput data. As the representation of a biological process by networks enables a systematic characterization of its structural properties such as the underlying design principles via the analysis of network topology, network has been becoming a popular framework in genome-wide biological research. Prior to construction of biological networks, it is useful to know the mathematical concept on network.

## Network

Network can be mathematically modeled as graphs consisting of nodes and edges representing elements and connections, respectively. A graph can be of two types: a directed graph with arcs or arrows, and an undirected graph with edges. To analyze the graph, many features are measured including degree, distance and clustering coefficient. The degree of a node is determined by the number of edges connected to that node and the distance between two nodes is defined as the shortest path length between these two nodes. The clustering coefficient ($\gamma_v$) describes the connectivity of the neighbors of a given node, *i.e.*, the existing edges in proximity of that node divided by all possible connections among the neighbors (Figure 1). The neighbors of a given node in network with strong local clustering are more likely to be connected to one another than would be expected through chance alone. It is crucial to measure which nodes in a network are more influential than others. For example, economic

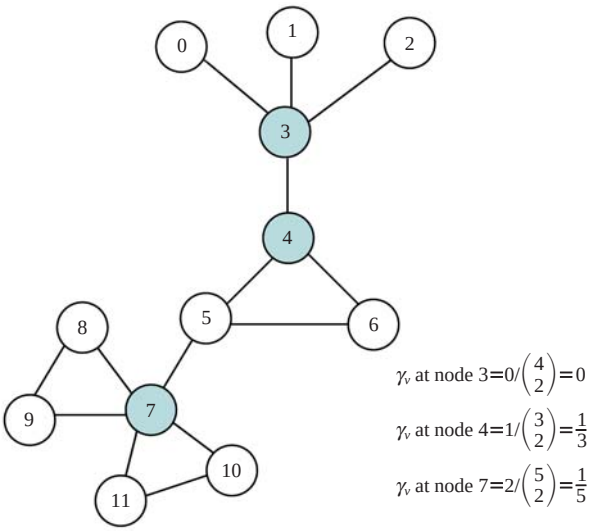



**Figure 1.** Calculation of clustering coefficient at three selected nodes (blue node).

considerations can dictate that one important target protein is selected to carry out an experiment from a protein interaction network suggesting multiple potential targets.

The importance of each node in the graph is often measured by assigning a numerical value to each node in the network and ranking of it; this is called centrality. There are many different concepts for computing the centrality[5]. Table 1 shows the values of various centrality indices corresponding to each node of an undirected graph shown in Figure 2. Node 6 has the highest centrality in distance-based centralities including eccentricity and closeness as well as shortest-path-based centrality, while node 13 has the highest degree centrality and feed-back based centralities including Katz status, eigenvector and PageRank. The use of centrality may lead to more rational approaches in experimental design, but it should be considered within an exploratory process under the relevant biological question.

**Table 1.** Various centrality indices corresponding to each node of Figure 1. The highest centrality index in each method is marked in gray. The centrality index was calculated using CentiBiN program (http://centibin.ipk-gatersleben.de/).

| Node | Type of centrality* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Degree $C_{deg}(v)$ | Eccentricity $C_{ecc}$ | Closeness $C_{clo}(v)$ | Shortest path betweenness $C_{spb}(v)$ | Katz status $C_{Katz}$ ($\alpha$=0.2) | Eigenvector $C_{eiv}$ ($\lambda$=$\lambda_{max}$) | PageRank $C_{pr}$ ($d$=0.3) |
| 0 | 1 | 0.125 | 0.010 | 0 | 1.473 | 0.009 | 0.032 |
| 1 | 1 | 0.125 | 0.010 | 0 | 1.473 | 0.009 | 0.032 |
| 2 | 1 | 0.125 | 0.010 | 0 | 1.473 | 0.009 | 0.032 |
| 3 | 4 | 0.143 | 0.013 | 51 | 2.363 | 0.026 | 0.096 |
| 4 | 3 | 0.167 | 0.015 | 60 | 2.397 | 0.050 | 0.060 |
| 5 | 1 | 0.2 | 0.014 | 0 | 1.448 | 0.042 | 0.030 |
| 6 | 3 | 0.25 | 0.018 | 98 | 2.238 | 0.123 | 0.064 |
| 7 | 3 | 0.2 | 0.017 | 84 | 2.376 | 0.074 | 0.059 |
| 8 | 3 | 0.2 | 0.018 | 90.5 | 2.365 | 0.248 | 0.061 |
| 9 | 2 | 0.167 | 0.017 | 38.5 | 2.293 | 0.305 | 0.044 |
| 10 | 2 | 0.167 | 0.017 | 38.5 | 2.293 | 0.305 | 0.044 |
| 11 | 1 | 0.125 | 0.012 | 0 | 1.820 | 0.221 | 0.030 |
| 12 | 1 | 0.143 | 0.011 | 0 | 1.449 | 0.016 | 0.030 |
| 13 | 8 | 0.143 | 0.016 | 93.5 | 4.102 | 0.655 | 0.172 |
| 14 | 1 | 0.125 | 0.012 | 0 | 1.820 | 0.221 | 0.030 |
| 15 | 1 | 0.125 | 0.012 | 0 | 1.820 | 0.221 | 0.030 |
| 16 | 1 | 0.125 | 0.012 | 0 | 1.820 | 0.221 | 0.030 |
| 17 | 1 | 0.125 | 0.012 | 0 | 1.820 | 0.221 | 0.030 |
| 18 | 1 | 0.125 | 0.012 | 0 | 1.820 | 0.221 | 0.030 |
| 19 | 3 | 0.167 | 0.014 | 18 | 2.244 | 0.047 | 0.064 |

*Each centrality is defined as follows. Let G=(V,E) be an undirected graph. The deg ($v$) denotes the degree of the node $v$ in an undirected graph; dist ($s$,$t$) denotes the length of a shortest path between the nodes $s$ and $t$; $\delta_{st}$ denotes the number of shortest paths from $s$ to $t$ and $\delta_{st}(v)$ the number of shortest path from $s$ to $t$ that use the node $v$; $A$ represents the adjacent matrix of the graph G. For the more detailed description please see the paper of Junker *et al.*[5].

$C_{\deg}(v)=\deg(v)$, $C_{ecc}=\frac{1}{\max\{dist(s,t):t\in V\}}$, $C_{clo}(v)=\frac{1}{\sum_{t\in V} dist(s,t)}$, $C_{clo}(v)=\frac{1}{\sum_{t\in V} dist(s,t)}$, $C_{spb}(v)=\sum_{s\in\vec{V}\wedge s\neq v}\sum_{t\in\vec{V}\wedge t\neq v}\delta_{st}(v)$, $C_{katz}(v)=\sum_{k=1}^{\infty}\alpha^k(A^T)^k\vec{1}$,

$\lambda C_{eiv}=AC_{eiv}$, $C_{pr}=dPC_{pr}+(1-d)\vec{1}$

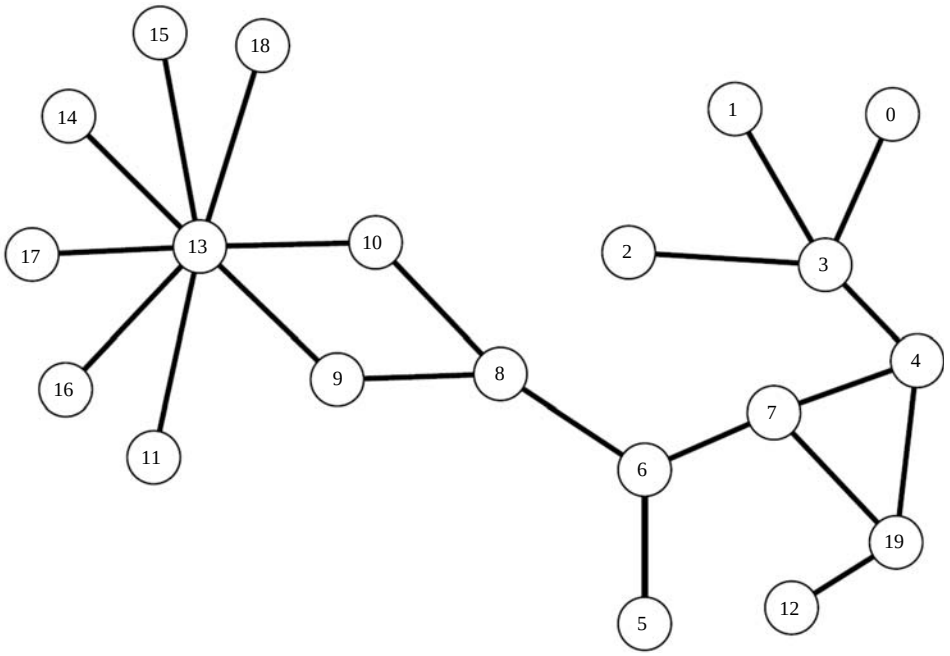



**Figure 2.** An undirected graph with twenty nodes.

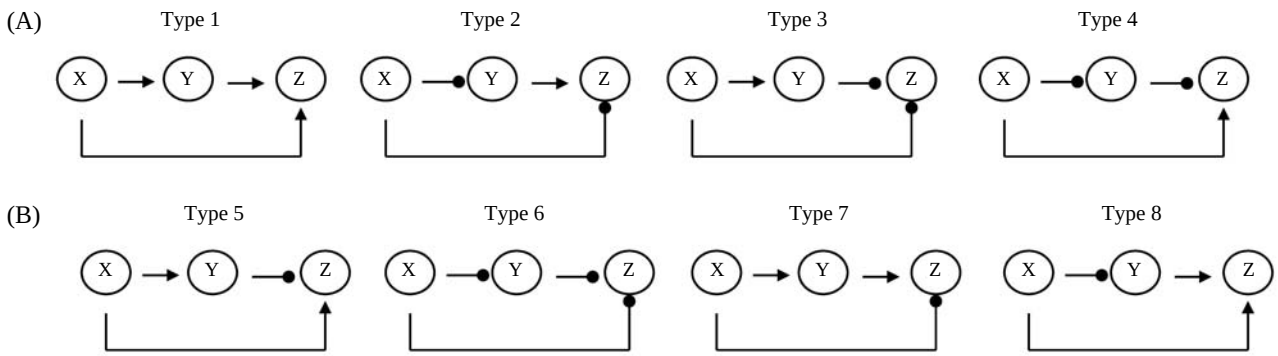



**Figure 3.** The eight types of FFLs consisting of three nodes according to signs of three edges. The symbols of → and ⊸ represent activation and repression, respectively.

Networks can also be characterized by network motifs, which are patterns of interconnections occurring in the networks at numbers that are significantly higher than those in randomized networks. The randomized networks of a given network are usually generated by arbitrarily rewiring the connections of the network locally, keeping the degree distribution at a global level. One of the most significant motifs in both *Escherichia coli* and yeast transcription networks is the feed-forward loop (FFL), a three-gene-pattern composed of two input transcription factors, one of which regulates the other, which both jointly regulate a target gene[6,7]. There are eight types of FFLs according to signs of three edges (Figure 3). These can be further classified into two groups: coherent FFL and incoherent FFL. The indirect path in coherent FFLs has the same overall effect on the target as the direct path, while the effect of the indirect path is opposite to that of direct path. The most abundant FFL type in transcriptional networks is type 1 coherent FFL[8].

## Gene Regulatory Networks

Biological processes such as morphogenesis, cell proliferation, differentiation, apoptosis and homeostasis are basically related to gene regulation mediated by transcription factors that can recognize and bind specific DNA sequence elements in the regulatory regions of genes. The gene regulatory network is usu-

ally represented by a relationship between transcription factors and their targets. The network representation unveils the global organization of transcriptional regulation such as its modular and hierarchical structure[9-12] and the fact that, on average, every target gene is controlled by two transcription factors[13,14]. The modeling of gene regulatory network from experimental data is a challenging task because the problem is combinatorial mission finding the right combination of regulators and available data can be an infrequent and inaccurate process[15]. The gene expression data from DNA microarrays have been utilized for inference of transcriptional regulatory networks using models such as Bayesian networks, Boolean networks, state space approach and differential equations. Among these models, Bayesian networks and Boolean networks have been frequently used for the inference of gene regulatory networks as large-scale qualitative modeling frameworks.

Bayesian networks can be used to represent conditional dependencies and independencies among variables corresponding to gene expression measurements. Given a set of genes and their expression patterns, a Bayesian approach finds the network that explains the observed patterns with the maximum of probabilities (Figure 4). Thus, a genetic regulatory system can be modeled by a directed acyclic graph (DAG) G=($V$, $E$), with $V$ representing a set of nodes and $E$ representing a set of edges in Bayesian networks. The nodes represent genes and correspond to random variables $x_i$ describing the expression level of gene $i$. The conditional distributions $p(x_i|\text{parents}(x_i))$ specify a joint probability $p(x)$ as:

$$p(X)=\prod_{i=1}^{n} p(x_i|\text{parents}(x_i)),$$

where parents ($x_i$) indicate the variables corresponding to the direct regulators of gene $i$ in G and $n$ represents the number of nodes (genes). There are three essential parts for learning a Bayesian network: model selection (choice of a DAG as candidate), parameter fitting (searching for best conditional probabilities for each node given a graph and experimental data) and fitness rating (scoring each candidate model).

The model selection is the most critical step because the number of all possible DAGs on $N$ nodes (genes) grows super-exponentially as the node number $N$ increases[15]. Figure 5 shows the Bayesian learning process for small data set including expression values of six genes (A-F) in seven experiments. The number of DAGs in this system is around $3.78 \cdot 10^6$. This shows that heuristics are required to efficiently learn a Bayesian network. The structure of DAG in Bayesian networks limits feedback loops. Dynamic Bayesian networks (DBNs), on the contrary, are able to feedback regulatory mechanisms by separating input nodes from output nodes (Figure 4(B)). However, DBNs suffer from various computational challenges and necessitate time-course data, which in some domains are not feasibly attainable in an applicable form[16]. Sachs and Itani[17] proposed a method for representing cyclic structures using Generalized Bayesian networks (GBNs) which enable structure learning in a cyclic domain, relying an perturbations which break the cyclic structure.

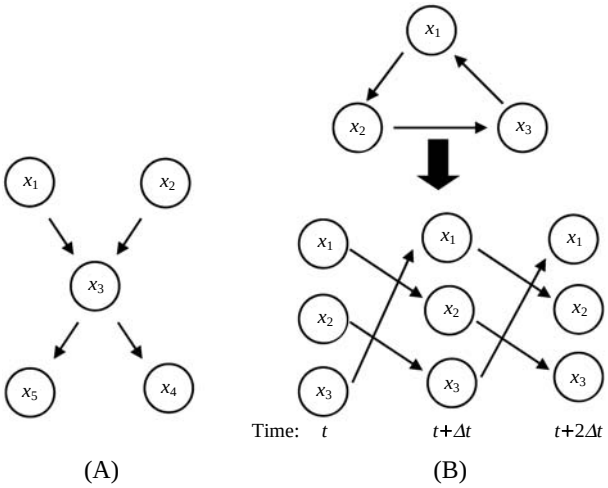


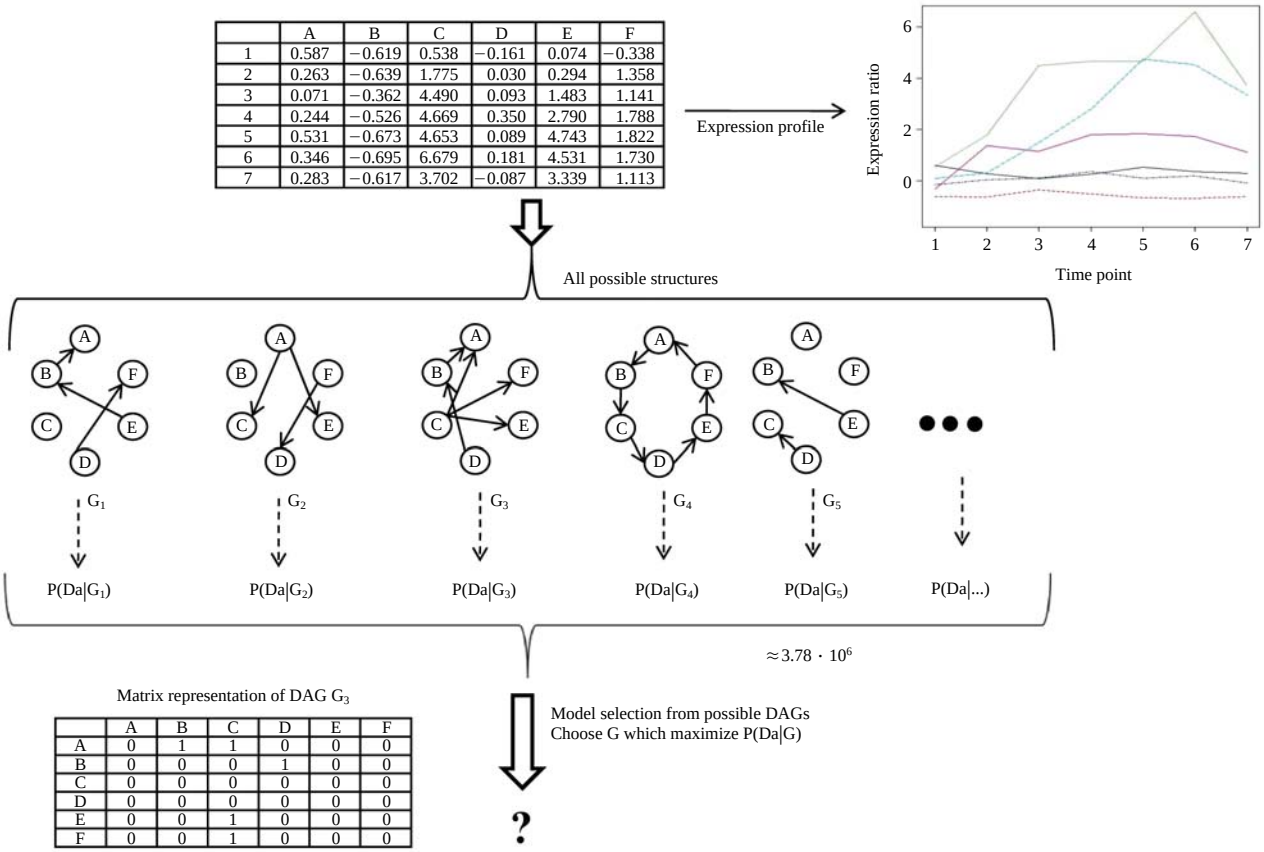

**Figure 4.** Graphical representation of Bayesian networks. (A) A static Bayesian network, (B) A dynamic Bayesian network including a cyclic regulation.

Another extensively investigated qualitative model for gene regulatory system is the Boolean network model, originally introduced by Kauffman[18]. A Boolean network can be also defined by a graph $G=(V, F)$, that is, a set of nodes (genes) $V=\{x_1, x_2,..., x_n\}$ (where $x_i \in \{0,1\}$) and a set of Boolean functions $F=\{f_1, f_2,..., f_n\}$, which represents the transitional relationships between time points. As Boolean networks utilize a Boolean variable $x_i \in \{0,1\}$ that define the state of a gene $i$ expressed by a network node as active (on, 1), inactive (off, 0), the continuous gene expression values should be changed into binary data prior to estimation of a Boolean network. Discrete assignment can be performed by clustering and thresholding using support vector regression[19]. The gene status at time point $t+1$ is determined by the values of some other genes at a previous time point $t$ using a Boolean functionas $f_i$ as:

$$x_i(t+1)=f(x_{j1(i)}(t),x_{j2(i)}(t),...,x_{jk(i)}(t))$$

where $j_{k(i)}$ indicates the mapping between gene networks at different time points.

The rules of regulatory interactions between genes are obtained through the Boolean function $F$; that is,




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | 0.587 | −0.619 | 0.538 | −0.161 | 0.074 | −0.338 |
| 2 | 0.263 | −0.639 | 1.775 | 0.030 | 0.294 | 1.358 |
| 3 | 0.071 | −0.362 | 4.490 | 0.093 | 1.483 | 1.141 |
| 4 | 0.244 | −0.526 | 4.669 | 0.350 | 2.790 | 1.788 |
| 5 | 0.531 | −0.673 | 4.653 | 0.089 | 4.743 | 1.822 |
| 6 | 0.346 | −0.695 | 6.679 | 0.181 | 4.531 | 1.730 |
| 7 | 0.283 | −0.617 | 3.702 | −0.087 | 3.339 | 1.113 |

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | 1 | 1 | 0 | 0 | 0 |
| B | 0 | 0 | 0 | 1 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 0 | 0 | 1 | 0 | 0 | 0 |
| F | 0 | 0 | 1 | 0 | 0 | 0 |

**Figure 5.** The Bayesian learning process for small data set including expression values of six genes (A-F). Da and G represent data and DAG, respectively.

the target gene is predicted by other genes through a Boolean function. In case gene expression cannot be described adequately by only two states, probabilistic Boolean networks can be used. Probabilistic Boolean networks (PBNs) combine more than one possible transition Boolean functions, so that each one can be randomly selected to update the target gene based on the selection probability, which is proportional to the coefficient of determination (COD) of each Boolean function[20]. PBN generalizes the standard rule-based interactions of Boolean networks into the stochastic settings. PBN has certain equivalences to DBN. Lähdesmäki *et al.*[21] demonstrated that PBNs and a certain subclass of DBNs can represent the same joint probability distribution over their common variables. The PBN has been used to construct networks in the context of several cancer studies, including glioma[22], melanoma[23] and leukemia[24].

We surveyed two main approaches including Bayesian and Boolean networks for qualitative models in gene regulatory networks. Besides these two approaches, the state-space approach has been also used for modeling gene regulatory network[25]. The state-space model is one of the most powerful methods to modeling a dynamic system and has been widely employed for engineering control systems[26]. Hirose *et al.*[27] proposed a module-based gene network estimation using state space model as below

$$x_n = Fx_{n-1} + v_n, n \in N,$$

$$y_n = Hx_n + w_n, n \in N_{obs},$$

where $F \in R^{k \times k}$ is the state transition matrix, $H \in R^{p \times k}$ is the observation matrix, $v_n$-$N_k(0_k, Q)$ and $w_n$-$N_p(0_p, R)$ are the system noise and the observation noise, respectively, and $y_n \in R^p$ and $N$ represent a series of vectors containing observed expression levels of $p$ genes at the $n$th time points and the total number of time point, respectively. This approach can be nicely applied to the case of short time course microarray data with several replicated data in each time point. The dimension of system variable $x_n$ is usually determined by Bayesian Information Criterion (BIC) or Probabilistic Principal Component Analysis (PPCA) from ob-

servation data. All models described above do not consider detailed information such as kinetics given by the system but capture the essential information about network such as interactions between elements.

Quantitative modeling of gene regulatory networks can be simulated by differential equations. That is, differential equations can describe gene expression changes as a function of the expression of other genes and environmental factors. Ordinary differential equations (ODEs) widely utilized to model dynamical systems in science and engineering may be applied to the modeling of the gene regulatory system:

$$\frac{dx}{dt}=f(x,p,u,t),$$

where $x(t)$ is the gene expression vector at time $t$, $f$ is the function that describes the rate of change of $x$ under the model parameter set $p$ and external perturbation $u$. For identification of large-scale gene regulatory networks from time-series microarray data, the differential equation is usually approximated by difference equation with linearization of function $f$ as:

$$\frac{x_i[t+\Delta t]-x_i[t]}{\Delta t}=\sum_{j=1}^{N} w_{i,j} \cdot x_j[t]+b_i \cdot u,\ i=1,...,N,$$

where $w_{i,j}$ represents the interaction matrix, $b_i$ indicates the effect of the perturbation effect $u$ on gene $i$ and $x_i[t]$ represents the expression value of gene $i$ at time $t$. Information of gene interaction is obtained by weight matrix $w_{i,j}$ which should be reduced to a sparse matrix.

The regularized least squares regression such as LASSO (Least Absolute Shrinkage and Selection Operator) can solve the linear equation system with constraint of sparseness of the weight matrix (interaction matrix). Chen *et al.*[28] proposed the stochastic differential equation (SDE) to reflect the stochasticity in gene expression. The SDE is widely used for modeling irregular motion, variability or uncertainty due to time series. From time $t$ to $t+\Delta t$, the dynamic transcription and degradation process can be modeled as:

$$\frac{x[t+\Delta t]-x[t]}{x[t]}=(g_t-\lambda)\Delta t+\sigma\Delta W_t,$$

where $x(t)$ represents the expression value of target gene at time $t$, $g_t$ is the transcription rate, $\lambda$ is the degradation rate and $\sigma\Delta W_t$ is the noise or random error captured by normal distribution $N(0,\sigma^2\Delta t)$. When $\Delta t \rightarrow 0$, SDE can be represented as:

$$\frac{dx_t}{dt}=(g_t-\lambda)dt+\sigma dW_t,$$

where $W_t$ is the standard Brownian motion. The SDE is useful especially when the network local connections such as the strict neighborhood of one target are interested[29].

## Protein-protein Interaction Networks

Protein-protein interactions (PPIs) are playing important roles in many cellular processes such as transcription, splicing, translation, cell cycle control, secretion and the assembly of enzymatic complexes. With availability of large-scale and high-throughput screening technology such as the yeast two-hybrid (Y2H) system, the protein network is constructed by combining pairwise interactions among all proteins considered. The Y2H approach has superior speed and robustness, but there are some drawbacks. For example, it can only consider interactions that occur within the nucleus of the yeast cell where the active transcription factor is reconstituted. Thus, proteins localized into other cellular compartments cannot interact, even if they have real interactions. In addition, the system cannot catch the dependence of interaction on post-transitional modifications such as phophorylation, acetylation or glycosylation.

This limitation of Y2H can be overcome by the affinity purification coupled to mass spectrometry (AP-MS). That is, AP-MS approach can identify interactions that occur in the native cellular environment. Interactions depending on post-transitional modifications of one or more components of the complex can be identified by the AP-MS approach. Thus, AP-MS can detect higher order interactions, while the Y2H system determines binary interactions. Protein interactions can be modeled by graphs similar to gene regulatory networks. The proteins correspond to the nodes of the graph, and edges between protein pairs indicate an interaction. Unlikely gene regulatory networks, the edges in protein interaction networks are usually undirected because only the presence or absence of an interaction between two proteins is detected.

The main problem in construction of protein network is the low quality of high-throughput data such as Y2H and AP-MS. Many evaluation studies have reported discrepancies between data sets, large error rates, lack of overlap and contradictions between experiments[30-33]. For the measure of reliability of each PPI, Jansen *et al.*[34] proposed to compute a likelihood ratio for each protein pair, gene $i$ and gene $j$. For example, let $y_{ij}(k)$ and $G_p$ be an element of $Y$ that shows a genomic feature of protein pair (gene $i$ and gene $j$) and an undirected graph (PPI network), $k=1$ respectively, while represents an experiment corresponding to Y2H. The value of $P(Y|G_p)$ is 0 or 1 according to absence or presence of the protein pair of gene $i$ and gene $j$ respectively, in the Y2H experiment.

Thus, the reliability of the PPI between gene *i* and gene *j* is given by the likelihood ratio as below

$$L(i,j)=P(y_{ij}(1),...y_{ij}(N)|\text{pos})/P(y_{ij}(1),...y_{ij}(N)|\text{neg}),$$

where 'pos' and 'neg' are the positive and negative sets of protein pairs constructed in advance and $N$ is the number of experiments such as Y2H and AP-MS. If each experiment is conditionally independent, the likelihood can be written as

$$L(i,j)=\frac{P(y_{ij}(1)|\text{pos})}{P(y_{ij}(1)|\text{neg})}\times\cdots\times\frac{P(y_{ij}(N)|\text{pos})}{P(y_{ij}(N)|\text{neg})}$$

Under a given undirected graph $G_p$, the likelihood of PPI information $Y$ can be computed by a binary Markov network model[35],

$$P(Y|G_p)=\frac{1}{Z_y}\prod_{e(i,j)\in G_p}L(i,j),$$

where $e(i,j)$ is the undirected edge between gene *i* and gene *j*, $Z_y$ is the normalizing constant. The most reliable graph structure $G_p$ or PPI networks is obtained when the value of $P(Y|G_p)$ is maximized.

Chiang *et al.*[36] characterized the error statistics in PPI networks of *Saccharomyces cerevisiae* with three traits using a direct graph model for bait to prey systems and a multinomial error model in all published large-scale datasets: the set of tested interactions, artifacts that lead to false-positive or false-positive observations and estimates of the stochastic error rates that affect the data. That is, there are three types of relationships between protein pairs of an experimental data set: tested with an observed interaction, tested with no observed interaction and untested. This method can benefit the design of future protein interaction experiments. The reliability in PPI can be increased with other genomic data such as gene expression. Nariai *et al.*[37] proposed a statistical method for estimating gene regulatory networks and PPI networks simultaneously based on microarray data, PPIs and protein localizations, essentiality phenotypes and functional categories by unifying Bayesian networks and Markov networks.

## Metabolic Networks

In the past, metabolism was considered as a combination of distinct pathways such as glycolysis, citrate cycle, urea cycle and many others. However, all these pathways are connected to each other. The cell utilizes the metabolic network to generate energy and to make the cellular components that are necessary for its growth and survival. For a completely sequenced genome, reconstruction of metabolic network can be started with the E.C. number (Enzyme Commission number) of each gene. If the EC number is unavailable in the original annotation, a reaction data base such as KEGG LIGAND[38] can be used for the search of the EC number for each gene. The reaction list based on EC number can be converted to a connection matrix representing substrate-product pairs. Thus, a genome-wide metabolic network model can be reconstructed for well-annotated genomes.

However, most annotation efforts fail to assign function to 40-60% of sequences and large numbers of sequences may have non-specific annotations, which results in 'missing enzymes' or 'missing genes' in reconstructed metabolic network. Green and Karp[39] has developed a program called PathoLogic that efficiently combines sequence homology and pathway-based evidence via Bayesian network to identify candidates for filling pathway holes in Pathway/Genome databases. This program uses a set of sequences encoding the required activity in other genomes to identify candidate proteins in the genome interest, and then evaluates each candidate by using a simple Bayes classifier to determine the probability that the candidate has the desired function. Thus, given the parameter vector of a particular candidate enzyme, one can determine the posterior probability of the candidate having the desired functions. When the posterior probability of the candidate having function exceeds a threshold, the

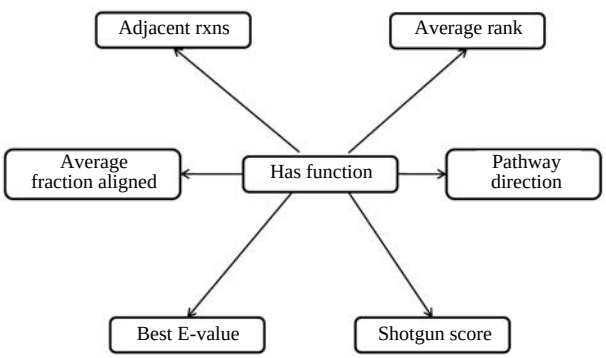



**Figure 6.** The structure of the Bayesian classifier used in PahoLogic program[39]. Description of each term is given as follow: Has-function is true if the protein has the function required the pathway hole, false if it does not; Shotgun score is the number of query sequence whose BLAST output included the candidate sequence; Best E-value is the negative log of the E-value for the best alignment of the candidate with a query sequence; Average rank is the averaged rank of the candidate sequence in the BLAST output lists; Average fraction aligned is the average of each alignment length normalized by the length of the query sequence; Pathway direction is true if the hit in the same direction as another gene in the same pathway; Adjacent-rxns is true if the hit is adjacent to one of the genes coding the enzyme for an adjacent reaction in the pathway.

candidate is considered as 'has-function' enzyme catalyzing the desired reaction (Figure 6).

Geng *et al.*[40] adopted the same approach to identify the list of candidate enzyme, but employed an additional model consisting of a mixture of $k$ radial basis functions (RBFs) and a linear term to predict whether these candidates are has-function or no-function enzymes. They calculated the model order and the parameters using a reversible jump Markov-chain-Monte-Carlo (MCMC) technique to avoid the difficulty in analytic integration of high-dimension of nonlinear functions. The principle of MCMC is to draw random samples from an ergodic Markov chain whose equilibrium distribution is the target posterior distribution. Since both approaches are based primarily on sequence information, they have limitation for enzyme with an extremely divergent sequence.

Once the metabolic networks are constructed, the next step is to study the dynamics of metabolites on them. The dynamic behavior of metabolic networks cannot be easily predicted due to lack of kinetic parameters involved in a number of biochemical reactions in the cell. Instead, the steady-state functionality of genome-scale metabolic networks has been easily described by constraint-based models[41], which represent the metabolic network via a series of physico-chemical constraints including stoichiometric network connectivity with the small number of parameters. These models can also be used to obtain a particular flux distribution by finding optimal distribution given a particular objective function using flux balance analysis (FBA)[42,43]. The FBA approach applied to genome-scale constraint-based metabolic models can be expressed as:

$$\max\ c^T v$$
$$\text{subject to } Sv=0\ (0\le v_i\le v_i^{\max})$$

where $v$ represents flux distribution vector, $c$ represents vector of objective coefficient, and $S$ denotes stoichiometric matrix. The metabolic flux distribution vector $v$ can be obtained with linear optimization. It should be remembered that a complete metabolic network shows all the possible modes of material flows in the cell. This indicates that all parts in metabolic networks might not be active. Thus, it is necessary to identify the correct active reactions to be included in the model from a larger set of possible enzymatic reactions based on comparison between model predictions and experimental data.

Herrgård *et al.*[44] proposed a method called optimal metabolic network identification (OMNI) for determining the active reactions in a genome-scale metabolic network based on a limited number of experimentally measured fluxes. This method uses bi-level optimization problem. That is, the outer optimization problem searches through a set of reactions to include in the model, while the inner optimization problem produces a flux distribution as a solution to a FBA problem. Let $y$ binary vector indicating whether a reaction is part of the model or not, the OMNI can be expressed as follow:

$$y^{opt}=\underset{y}{\operatorname{argmin}} \sum_{i\in M} w_i \left| v_i^{opt}-v_i^{\exp} \right|$$

$$\text{subject to } \begin{cases} v^{opt}=\operatorname{argmax}_v\ c^T v \\ \quad \text{subject to } Sv=0 \\ \qquad 0\le v_j\le v_j^{\max}\ j\in F \\ \qquad 0\le v_k\le v_k^{\max} y_k\ k\in D \\ v_l=v_l^{\exp}\ l\in E \\ v_{biomass}^{opt}\ge v_{biomass}^{\min} \\ y_k=\{0,1\}\forall k\in D \\ \sum_{k\in D}(1-y_k)=K \end{cases}$$

The variables $w$, $F$, $D$ and $K$ represent weight vector for measured flux, set of reactions that are fixed, set of reactions that can be deleted from the model and number of reaction deletions allowed in the model, respectively. This method is applied to intra-cellular flux data for five experimentally evolved *E. coli* strains, and identified specific bottleneck reactions in the metabolic model[44]. Activation of only specific pathways from all possible metabolic reactions in a given organism may be due to regulatory effects working under particular conditions, as well as uncertainties about specific cofactors.

## Signal Transduction Networks

A signal transduction pathway was previously considered to be a linear cascade including PPIs, protein modifications and small signaling molecules such as $Ca^{2+}$, lipids or other second messengers[45]. For example, the activated platelet-derived growth factor (PDGF) receptor can activate different signaling cascades: (i) a MAPK signaling cascade, (ii) phospholipase C gamma, (iii) phosphoinositide-3-kinase and Akt and (iv) can trigger a pathway that finally results in stabilization of $\beta$-catenin[46]. These linear pathway models might be enriched by negative as well as positive feedback loops. The negative feedback loops can limit the strength or duration of a signal, while positive feedback loops create an ultra-sensitive activation of the pathway and a bi-stable behavior. Based on detail-

ed analysis of Wnt related signaling pathways, Kestler *et al.*[45] suggested that Wnt signaling operates within an interwoven Wnt signaling network rather than functioning as independent linear cascades.

The organization of signaling pathways as networks indicates a possibility of cross-talk that might occur with other signaling pathways. Currently, the construction of signal transduction networks at the genome-scale is still problematic due to the lack of experimental data. Thus, the signal transduction networks at the genome-scale are approximated by ortholog abstraction, with species-specific differences between homologous molecules from various species being ignored. Just as metabolic networks describe the potential pathways to be used for carrying out metabolic tasks, these signal transduction networks describe potential pathways to be employed for regulation of gene expression in eukaryotic cells. One of the representative signal transduction databases, the TRANSPATH database, has used this approximation for supplement of incomplete data in the pathways[47].

Microarray studies using cell assays with external interventions into the signaling process also allow for the systemic analysis of the pathways. Froehlich *et al.*[48] proposed a Bayesian method for reconstructing signaling pathways from secondary effects, which were observed on microarray data after silencing genes of interest via RNAi. They distinguished between silenced genes (*S*-genes) and genes showing a downstream effect (*E*-genes). Each *E*-gene is attached to a single *S*-gene. Knocking down the *k*th specific *S*-gene ($S_k$) interrupts signal flow in the downstream pathway, and an effect on the *E*-genes attached to $S_k$ and all *S*-genes depending on $S_k$ is expected. Thus, the outcomes of experiments with $n$ knock-downs and $m$ *E*-genes in total can be summarized in an $m \times n$ data matrix $D$. According to Bayes' formula a specific network hypothesis $\Phi=\{0,1\}^n \times \{0,1\}^n$ can be scored as:

$$P(\Phi|D) \propto P(D|\Phi)P(\Phi)$$

Note that $\Phi_{ij} \in \{0,1\}$ depending on whether the edge $i \rightarrow j$ presents or not. If the $P(\Phi_{ij})$ has a Laplacian distribution with parameter $\lambda$, it can be represented as:

$$P(\Phi_{ij}|\lambda)=\frac{\lambda}{2}\exp(-\lambda|\Phi_{ij}-\hat{\Phi}_{ij}|)$$

Thus, we can write down the log-posterior of $P(\Phi|D)$ as below:

$$\begin{aligned}\log P(\Phi|D) &\propto \log P(D|\Phi)+\log P(\Phi) \\ &\propto \log P(D|\Phi)-\lambda\sum_{i,j}|\Phi_{ij}-\hat{\Phi}_{ij}|\end{aligned}$$

The $\lambda$ specifies the trade-off between the model's fit data and prior assumptions. For example, setting $\lambda \rightarrow \infty$ corresponds to completely trusting the prior while $\lambda=0$ leads a maximum likelihood estimate, i.e., complete trust in data. The Akaike Information Criterion (AIC) can be used for optimal choice of $\lambda$,

$$AIC(\lambda,\Phi_{opt})=-2\log P(D|\Phi_{opt})+2d(\lambda,\Phi_{opt})$$

where $d(\lambda,\Phi_{opt})$ denotes the number of free parameters (the number of unknown edges) in the network structure $\Phi_{opt}$. The optimal value of $\lambda$ is obtained when the AIC is minimized.

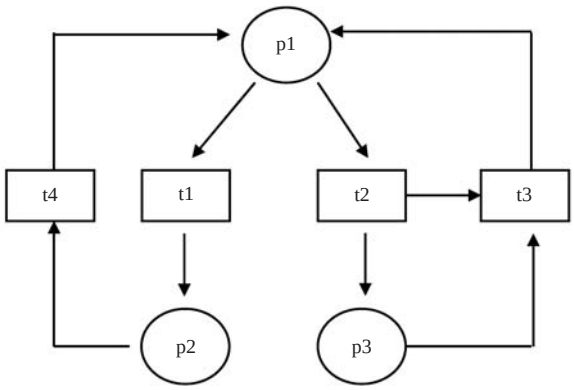



**Figure 7.** A Petri net consisting thee places (p1, p2, p3) and four transitions (t1, t2, t3, t4). The place can represent biological elements such as proteins and metabolites while the transition can represent biochemical reactions in biological networks.

The dynamic behavior of the signal transduction network can be described by various models such as ordinary differential equations and Petri nets. While ordinary differential equations have been mainly used as the techniques for quantitative modeling and simulations, Petri nets have been employed for qualitative modeling and analysis of various biological pathways since many theoretical investigations on Petri nets such as structural analysis of systems[49]. A Petri net is a directed-bipartite graph with two different types of nodes: places and transitions. Bipartite denotes that they consist of two types of nodes called places $P=\{p_1,...,p_n\}$ and transitions $T=\{t_1,...,t_n\}$, and directed arcs, which are weighted by natural numbers and connect only nodes of different types. Places usually represent the passive system elements such as states or biological species like proteins or metabolites, while transitions denote active system elements such as events or activation/deactivation or chemical reactions. Places are depicted as circles and transitions as rectangles (Figure 7).

The standard Petri net models are discrete while the hybrid Petri net (HPN) comprises generally discrete as well as continuous parts. Since Matsuno *et al.*[50] have demonstrated that HPN has high potential to model and simulate biological pathways, many biological path-

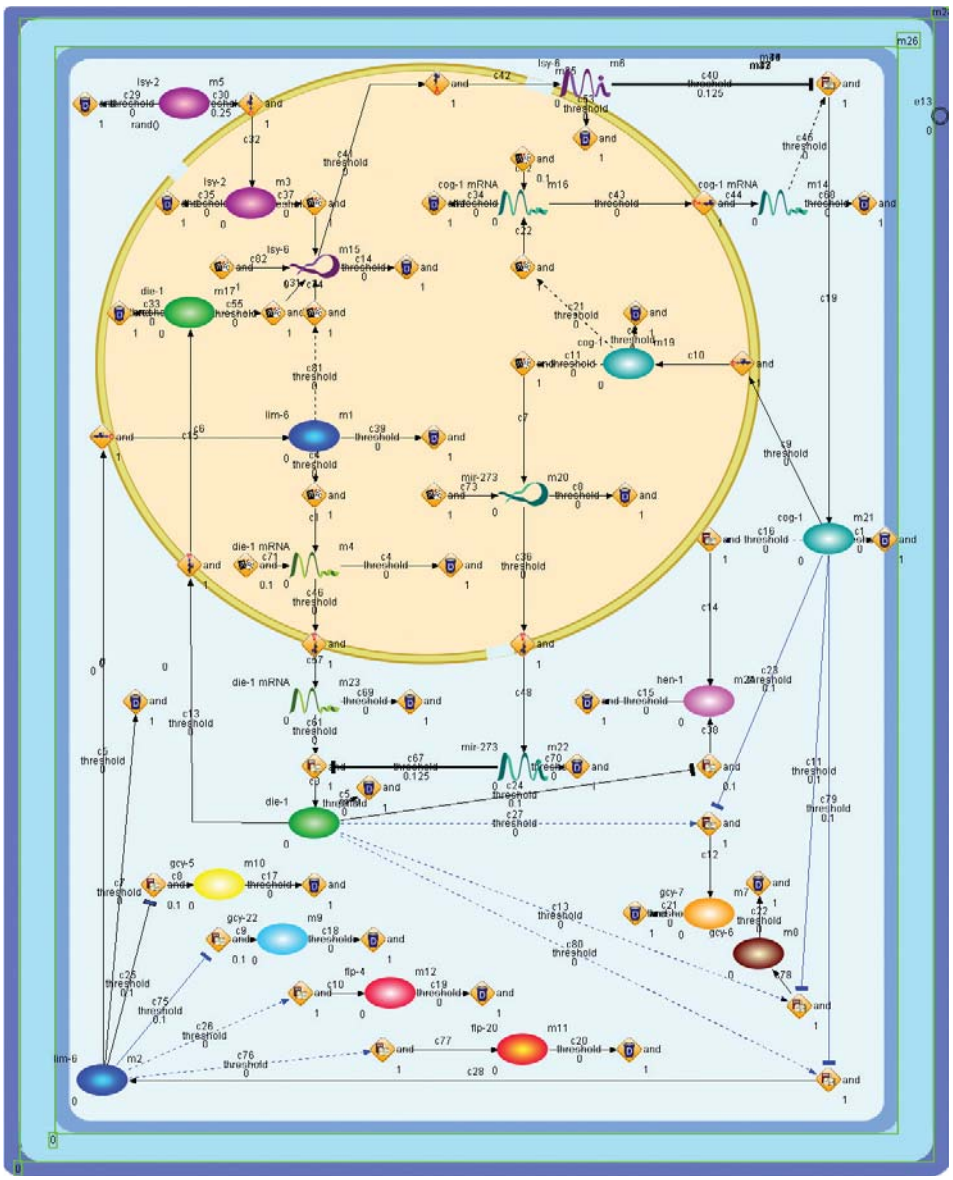



**Figure 8.** HFPNe model of the ASEL/ASER in Cell Illustrator. The model is available at http:www.csml.org/csml-models/.

ways have been created with technique of HPN or its extension: apoptosis induced by protein Fas[51], Notch-Delta signaling pathway in Drosophila[52], gemcitabine chemotherapic drug pathway[53], Huntington's disease[54], role of interleukin-6 in the fate of haematopoietic stem cells[55], Raf-1 kinase inhibitor protein on the extracellular signal regulated kinase[56]. Nagasaki *et al.*[57] proposed a powerful Petri net architecture hybrid functional Petri net with extension (HFPNe) which involves all the functions of existing high-level Petri nets. That is, each ordinary Petri net, stochastic Petri net, colored Petri net, and HPN can be treated as subset of the HFPNe.

A colored Petri net assigns data values to the tokens and expressions are attached to the arcs, which defines the constraints on the token values in the input places to enable the transitions, and defines the token values produced by the firings in the output places[58]. A stochastic Petri net takes into account uncertainty attached to data. Cell Illustrator (http://www.cellillustrator.com/) is a pathway simulation tool which employs the HFPNe as a basic architecture. Figure 8 shows a screen shot of Cell Illustrator which displays a double-negative feedback loop (DNFL) of *lsy*-6 and *mir*-273 miRNAs that determine the ASE cell fates in *C. elegans*, i.e., whether the cell will be ASE left (ASEL) or ASE right (ASER). In this figure, Petri net elements of places and transitions have been changed to pictures of biological images which reflect the roles of these elements, which makes pathway models with HFPNe

more familiar to biologists. By this DNFL model, Saito *et al.*[59] demonstrated miRNAs could be effectively handled with the HFPNe architecture. Sackman *et al.*[60] demonstrated that the Petri net approach can be used to build a discrete model which reflect provably the qualitative biological behavior without any knowledge of kinetic parameters in the mating pheromone response pathways of *Saccharomyces cerevisiae*. This suggests that the Petri net approach would be prevail in modeling and analysis of large and complex signal transduction pathways which including lots of missing kinetic data.

## Conclusions

The completion of genome sequences and subsequent high-throughput mapping of molecular networks have allowed the study of biology from the network perspective[61]. In addition, emerging results have indicated that cellular function is a contextual attribute of strict and quantifiable patterns of interactions between the myriad of cellular constituents in spite of the importance of individual molecules. Thus, the network approach to biology would be the main framework to understand biological systems. We surveyed statistical approaches to construction of biological networks from high-throughput data. The fundamental idea behind these approaches is that models that faithfully capture the relationship between biological elements have predictive capacity and can be used to gain insight about system-wide properties such as steady-state behavior or responses to perturbations or specific stimuli. The main drawback of statistical approach is that the network structure inferred depends on statistical models and data. Nevertheless, the statistical approach to network modeling has an inherent potential of rapid network construction for organisms that are relatively uncharacterized from high-throughput data.

## Acknowledgements

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2006-353-D00006).

## References

1. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662-1664 (2002).
2. Dhar, P.K., Zhu, H. & Mishra, S.K. Computational approach to systems biology: from fraction to integration and beyond. *IEEE Trans. Nanobiosci.* **3**, 144-152 (2004).
3. Westerhoff, H.V. & Palsson, B.O. The evolution of molecular biology into systems biology. *Nat. Biotechnol.* **22**, 1249-1252 (2004).
4. Prifti, E., Zucker, J.-D., Clement, K. & Henegar, C. FunNet: an integrative tool for exploring transcriptional interactions. *Bioinformatics* **24**, 2636-2638 (2008).
5. Junker, B.H., Koschützki, D. & Schreiber, F. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics* **7**, 219 (2006).
6. Shen-Orr, S.S., Milo, R, Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.* **31**, 64-68 (2002).
7. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824-827 (2002).
8. Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci.* **100**, 11980-11985 (2003).
9. Wolf, D.M. & Arkin, A.P. Motifs, modules and games in bacteria. *Curr. Opin. Microbiol.* **6**, 125-134 (2003).
10. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101-113 (2004).
11. Yu, H. & Gerstein, M. Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci.* **103**, 14724-14731 (2006).
12. Martínez-Antonio, A., Janga, S.C. & Thieffry, D. Functional organization of *Escherichia coli* transcriptional regulatory network. *J. Mol. Biol.* **381**, 238-247 (2008).
13. Albert, R. Scale-free networks in cell biology. *J. Cell Sci.* **118**, 4947-4957 (2005).
14. Aldana, M., Balleza, F., Kauffman, S. & Resendiz, O. Robustness and evolvability in genetic regulatory networks. *J. Theor. Biol.* **245**, 433-448 (2007).
15. Hecker, M., Lambeck, S., Toepfer, S., van Someren, E. & Guthke, R. Gene regulatory network inference: data integration in dynamic models-A review. *BioSystems* doi:10.1016/j.biosystems (2009).
16. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. & Nolan, G.P. Casual protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523-529 (2005).
17. Sachs, K., Itani, S., Fitzgerald, J., Wilie, L. & Schoeberl, B. Learning cyclic signaling pathway structures while minimizing data requirements. *Pac. Symp. Biocomput*. 63-74 (2009).
18. Kauffman, S.A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437-467 (1969).
19. Martin, S., Zhang, Z., Martino, A. & Faulon, J.L. Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics* **23**, 866-874 (2007).
20. Li, P., Perkins, E.J., Gong, P. & Deng, Y. Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. *BMC Bioinformatics* **8**(suppl. 7), S13 (2007).
21. Lähdesmäki, H., Hautaniemi, S., Shmulevich, I. & Yli-

Harja, O. Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. *Signal Processing* **86**, 814-834 (2006).
22. Hashimoto, R.F. *et al.* Growing genetic regulatory networks from seed genes. *Bioinformatics* **20**, 1241-1247 (2004).
23. Pal, R., Datta, A., Bittner, M.L. & Dougherty, E.R. Intervention in context-sensitive probabilistic Boolean networks. *Bioinformatics* **21**, 1211-1218 (2005).
24. Li, H. & Zhan, M. Systematic intervention of transcription for identifying network response to disease and cellular phenotypes. *Bioinformatics* **22**, 96-102 (2006).
25. Wu, F.-X. Gene regulatory network modeling: a state-space approach. *Int. J. Data Min. and Bioinform.* **2**, 1-14 (2008).
26. Chen, C.T. Linear system theory and design. 3rd ed., Oxford University Press, New York (1999).
27. Hirose, O. *et al.* Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics* **24**, 932-942 (2008).
28. Chen, K.C., Wang, T.Y., Tseng, H.H., Huang, C.Y. & Kao, C.Y. A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. *Bioinformatics* **21**, 2883-2890 (2005).
29. Climescu-Haulica, A. & Quirk, M.D. A stochastic differential equation model for transcriptional regulatory networks. *BMC Bioinformatics* **8**(suppl. 5), S4 (2007).
30. Hazbun, T.R. & Fields, S. Networking proteins in yeast. *Proc. Natl. Acad. Sci.* **98**, 4277-4278 (2001).
31. Thomas, A., Cannings, R., Monk, N. & Cannings, C. On the structure of protein-protein interactions networks. *Biochem. Soc. Trans.* **31**, 1491-1496 (2003).
32. Poyatos, J. & Hurst, L. How biologically relevant are interaction based modules in protein networks. *Genome Biol.* **5**, R93 (2004).
33. Gagneur, J., David, L. & Steinmetz, L. Capturing cellular machines by systematic screens of protein complexes. *Trends Microbio*. **14**, 336-339 (2006).
34. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-453 (2003).
35. Segal, E., Wang, H. & Koller, D. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **19**(suppl. 1), i264-i271 (2003).
36. Chiang, T., Scholtens, D., Sarkar, D. & Gentleman, R. Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biol.* **8**, R186 (2007).
37. Nariai, N., Tamada, Y., Imoto, S. & Miyano, S. Estimating gene regulatory networks and protein-protein interactions of Saccharomyces cerevisiae from multiple genome-wide data. *Bioinformatics* **21**(suppl. 2), ii206-ii212 (2005).
38. Kanehisa, M. *et al.* From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* **34**, 354-357 (2006).
39. Green, M.L. &Karp, P.D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**, 76 (2004).
40. Geng, B. *et al.* Comparison of reversible-jump Markov-chain-Monte-Carlo learning approach with other methods for missing enzyme identification. *J. Biomed. Inform.* **41**, 272-281 (2008).
41. Price, N.D., Reed, J.L. & Palsson, B.O. Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nat. Rev. Microbiol*. **2**, 886-897 (2004).
42. Bonarius, H.P.J., Schmid, G. & Tramper, J. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends Biotechnol*. **15**, 308-314 (1997).
43. Kauffman, K.J., Prakash, P. & Edwards, J.S. Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14**, 491-496 (2003).
44. Herrgård, M.J., Fong, S.S. & Palsson B.*Φ.* Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS. Comput. Biol.* **2**, e72 (2006).
45. Kestler, H.A., Wawra, C., Kracher, B. & Kühl, M. Network modeling of signal transduction: establishing the global view. *BioEssays* **30**, 1110-1125 (2008).
46. Yang, L., Lin, C. & Liu, Z.R. P68 RNA helicase mediates PDGF-induced epithelial mesenchymal transition by displacing Axin from beta-catenin. *Cell* **127**, 139-155 (2006).
47. Choi, C. *et al.* Consistent re-modeling of signal pathways and its implementation in the TRANSPATH database. *Genome Inform*. **15**, 244-254 (2004).
48. Froehlich, H., Fellmann, M., Sueltmann, H., Poustka, A. & Beissbarth, T. Large scale statistical inference of signaling pathways from RNAi and microarray data. *BMC Bioinformatics* **8**, 386 (2007).
49. Matsuno, H., Li, C. & Miyano, S. Petri net based descriptions for systematic understanding of biological pathways. *IEICE Trans Fundam Electron Commun Comput. Sci.* **E89-A**, 3166-3174 (2006).
50. Matsuno, H., Doi, A., Nagasaki, M. & Miyano, S. Hybrid Petri net representation of gene regulatory network. *Pac. Symp. Biocomput*. 338-352 (2000).
51. Matsuno, H. *et al.* Biopathway representation and simulation on hybrid functional Petri net. *In Silico Biology* **3**, 389-404 (2003).
52. Matsuno. H. *et al.* Boundary formation by Notch signaling in Drosophila multicellular systems: experimental observations and a gene network modeling by genomic object net. *Pac. Symp. Biocomput.* 152-163 (2003).
53. Peleg, M.P., Rubin, D. & Altman, R.B. Using Petri net tools to study properties and dynamics of biological systems. *J. Am. Med. Inform. Assoc*. **12**, 181-199 (2005).
54. Nagasaki, M., Doi, A., Matsuno, H. & Miyano, S. Computational modeling of biological processes with Petri net-based architecture. *In Bioinformatics and*

*Technologies ed. Y.P. Chen.* 179-242 (2005).
55. Troncale, S., Tahi, F., Campard, D., Vannier, J.P. & Guespin, J. Modeling and simulation with hybrid functional Petri nets of the role of interleukin-6 in human early haematopoiesis. *Pac. Symp. Biocomput.* 427-438 (2006).
56. Gilbert, D. & Hiner, M. From Petri nets to differential equations-An integrative approach for biochemical network analysis. *Proc. ICATPN.* **LNCS 4024**, 181-200 (2006).
57. Nagasaki, M., Doi, A., Matsuno, H. & Miyano, S. A versatile Petri net based architecture for modeling and simulation of complex biological process. *Genome Informatics* **15**, 180-197 (2004).
58. Kristensen, L., Christensen, S. & Jensen, K. The practitioner's guide to coloured Petri nets. *Int. J. Software Tools Technol. Transfer* **2**, 98-132 (1998).
59. Saito, A., Nagasaki, M., Doi, A., Ueno, K. & Miyano, S. Cell fate simulation model of gustatory neurons with microRNAs double-negative feedback loop by hybrid functional Petri net with extension. *Genome Inform.* **17**, 100-111 (2006).
60. Sackman, A., Heiner, M. & Koch, I. Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics* **7**, 482 (2006).
61. Han, J.-D. Understanding biological functions through molecular networks. *Cell Res.* **18**, 224-237 (2008).