Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2008, Article ID 248747, 12 pages
doi:10.1155/2008/248747

Research Article

Recovering Genetic Regulatory Networks from Chromatin Immunoprecipitation and Steady-State Microarray Data

Wentao Zhao,[1] Erchin Serpedin,[1] and Edward R. Dougherty[2]

[1] *Electrical and Computer Engineering Department, Texas A&M University, College Station, TX 77843, USA*
[2] *The Translational Genomics Research Institute (TGen), 400 North Fifth Street, Suite 1600, Phoenix, AZ 85004, USA*

Correspondence should be addressed to Erchin Serpedin, serpedin@ece.tamu.edu

Received 28 November 2007; Accepted 20 May 2008

Recommended by Z. Wang

Recent advances in high-throughput DNA microarrays and chromatin immunoprecipitation (ChIP) assays have enabled the learning of the structure and functionality of genetic regulatory networks. In light of these heterogeneous data sets, this paper proposes a novel approach for reconstruction of genetic regulatory networks based on the posterior probabilities of gene regulations. Built within the framework of Bayesian statistics and computational Monte Carlo techniques, the proposed approach prevents the dichotomy of classifying gene interactions as either being connected or disconnected, thereby it reduces significantly the inference errors. Simulation results corroborate the superior performance of the proposed approach relative to the existing state-of-the-art algorithms. A genetic regulatory network for *Saccharomyces cerevisiae* is inferred based on the published real data sets, and biological meaningful results are discussed.

1. Introduction

Currently, one of the most important research problems in molecular biology and bioinformatics consists of finding out the mechanisms that govern the gene regulations, which are considered to play fundamental roles in the operation of all processes taking place in living cells. Learning the structure and machinery of gene regulations opens up the possibility for understanding and controlling the functioning of organisms at the molecular level, and for designing intelligent therapies and drugs. In a biological process such as cell cycle or environmental response, a gene's product, the protein, can serve as a transcription factor of a target gene by binding to the target gene's regulatory region on chromatin and affect its transcription. The protein can also influence another gene's expression in subsequent stages, for example, through splicing or translation. Alternatively, these protein-gene relationships can be viewed as gene-gene interactions, and are modeled in general as genetic regulatory networks.

Recent years have witnessed a number of different frameworks for modeling genetic regulatory networks, ranging from fine-scale modeling at the molecular level in terms of partial differential equations and stochastic equations, to large scale modeling at the gene and protein-level in terms of Boolean and probabilistic Boolean networks, and (dynamic) Bayesian networks; see, for example, [1–6] and their toolboxes [7–9]. The small scale modeling techniques are used to capture the detailed biochemical aspects of molecular interactions and are in general very computational demanding. On the other side, the large-scale models provide a global vision of the interactions among the constituent elements of genetic regulatory networks and are generally represented in terms of graphs.

In the middle of 1990s, the birth of DNA microarrays equipped the industry with the capability to measure simultaneously the concentration of genome-wide mRNA expressions. The gene expression data produced thereafter by gene chips have attracted extensive research on the inference of genetic regulatory networks based on various network models [10–18]. There are two types of DNA microarray data sets: time series (or time dependent) and time independent (also called steady-state or single point time series) data sets. In general, the time-independent gene expression profiles are capable of recovering steady-state attractors, but fail to recover the direct and oriented (temporal regulating) relationships. On the other side, time

series data sets can improve the inference greatly in contrast to time-independent data sets [13]. However, the financial costs, ethnical concerns, and implementation issues prevent collecting beneficial time series data. Recent statistics show that about 70% of published data are time independent [19]. Therefore, the steady-state analysis is highly valuable despite the difficulty of making accurate inference of temporal relationships.

Inference of gene regulatory networks based solely on the information provided by microarray data is limited by a number of factors: number of available microarrays, quality of data samples, experimental noise, and errors (cross-hybridizations). It is also known that post-transcriptional modifications and transcripts that are present at low levels are generally not detectable by microarrays. Since the gene activity is measured by the mRNA level, the underlying assumption is that there is a significant correlation between the mRNA level and the amount of protein associated with mRNA. However, the magnitude of such a correlation varies significantly depending on the type of protein involved. Therefore, a combined approach which, besides gene expression data, exploits additional data sources is likely to enhance the inference process.

The advent of in vivo chromatin immunoprecipitation (ChIP) assays has enabled to test whether a protein acting as a transcription factor binds to a specific DNA segment. Hence, ChIP assays serve as a promising mechanism to examine the regulatory relationships. In ChIP experiments, the protein is immobilized on the chromatin, then the chromatin is broken into DNA fragments, and the DNA-protein complexes are immunoprecipitated by using antibodies corresponding to the tested protein. Afterwards the DNA bound by the protein in question can be isolated and identified by using a cDNA microarray chip. The whole process is also called a ChIP-chip experiment, and inherits several disadvantages. The protein to be tested has to possess a specific antibody, which might not be synthesized, discovered, or exist. In addition, the transcriptional regulation is a complex process that is expressed in several different aspects. The binding of the transcription factor to the promoter region of the target gene is the most pristine mode. Especially for eukaryotic organisms, some regulatory bindings take place at a region far away from the regulated gene. This fact makes the binding information questionable for determining the regulation relationships. Furthermore, the experimental results are represented by p-values and the determination of the binding relationship is achieved through threshold comparison. However, the selection of the p-value threshold introduces a dilemma. A high threshold not only identifies the most probable binding relationships but also might miss many true relationships with lower p-values, while a low threshold infers more relationships, among which more might be false alarms. A good tradeoff is not easy to make. Besides, the cost factor has also to be considered. Generally, ChIP-chip experiments are very expensive and testing thousands of proteins is not affordable.

A combination of both steady-state microarray data and ChIP-chip data might help in making more accurate inferences. Intuitively, these two different types of data complement the shortcomings of each other. This motivates us to propose a Bayesian approach to analyze jointly both data sets and to establish a confidence measure of gene interactions. The proposed scheme possesses six key features which make it different from the existing algorithms. First, gene expression data in steady-state are considered, while time course data are used in other works like [11, 13, 20]. Second, most of the current schemes recover a unique genetic network represented by a graph which best fits the observed data in a certain metric, while the proposed approach determines the posterior probabilities for all gene-pair interactions and avoids to make a dichotomous decision that classifies each gene interaction as being either connected or disconnected. The proposed approach can be easily transformed into a dichotomous scheme by only preserving the highly probable gene interactions. Third, the underlying structural model is assumed to be a directed cyclic graph, which allows cycles (feedback loops) and directed acyclic graphs are treated as special cases. This contrasts to Bayesian networks, which are directed acyclic graphs. Feedback loops are a common network motif in biological processes and their function is to yield the necessary redundancy and stability for the system [1]. Therefore, methods based on Bayesian networks, for example, [21–23], lose their validity in the inference of cyclic graphs. Fourth, the proposed approach assumes continuous-valued variables, and this prevents the information loss incurred by data quantization. This represents an advantage compared with the discrete-valued networks such as [21–23]. Fifth, the proposed connectivity score is oriented and has a very clear meaning, in the sense of posterior probabilities, while the existing scores based on the mutual information [14, 18, 24] are vague and lack orientation information. Sixth, in the proposed approach the system kinetics are assumed to be nonlinear, while linear models are commonly utilized for the purpose of simplification [12, 15]. Besides, the proposed scheme establishes a general framework whose components can be customized to fit the nature of the underlying biological system.

The rest of the paper is organized as follows. Section 2 discusses the graphical model and system dynamics that govern the genetic expressions. Section 3 translates the p-values of ChIP-chip experiments into regulation probabilities and formulates the inference algorithm through Bayesian analysis. In Section 4, the proposed algorithm and other three schemes are simulated on a set of artificial networks. Performance comparisons illustrate that the proposed algorithm exceeds in terms of several metrics. The robustness of kinetics model is also discussed via simulations. Realistic data sets are exploited in the proposed inference framework and a genetic network is presented to account for the genetic response to environmental changes. Finally, Section 5 concludes the paper with remarks on possible future works.

2. Methods

Genetic regulatory networks can be represented by a parameterized graph $(\mathbf{G}, \boldsymbol{\Theta})$, where $\mathbf{G}$ and $\boldsymbol{\Theta}$ stand for the graph structure and parameter set, respectively. The graph structure qualitatively explains the direct gene interactions,

while the parameter set quantitatively describes the system kinetics.

2.1. Structural Model

The graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$ is employed to map gene interactions at transcriptional level, where $\mathbf{V}$ denotes the set of vertices (genes) and $\mathbf{E}$ stands for the set of edges (regulation relationships). If gene X regulates gene Y, graphically such a relation is represented in terms of an oriented edge $X \rightarrow Y$, where X is a parent of Y and Y is considered a child of X. All genes that present incidence edges with gene X represent the set of parental genes of X, and are compactly denoted in terms of the notation Π_X. If two genes X and Y interact with each other but the regulation orientation cannot be determined, an undirected edge is laid between the two genes as X-Y, which means both orientations are possible. A sequence of consecutive-oriented edges constitutes a directed path. If there is no directed path which starts and ends at the same vertex, in other words, the graph contains no loops, the graph is called a directed acyclic graph (DAG). DAGs lie at the basis of Bayesian networks, which are commonly employed to model causal relationships [25].

General directed graphs (with possibly cycles) will serve as our structural model since they are consistent with the features exhibited by biological systems, in which loops account for system redundancy and stability. Given the graph structure $\mathbf{G}$, the parent set Π is specified for any gene X. For conciseness, the subscript X associated with some variables is omitted in the analysis procedure when the context has clearly specified the gene in question. Next, we discuss the system kinetics and parameters defined in $\boldsymbol{\Theta}$.

2.2. System Kinetics

The system kinetics represents the dynamics that governs the gene mRNA concentrations in terms of gene-gene interactions. It can be modeled by a set of differential equations (DEs). A simplified form is a set of linear DEs. However, we accept the more complex model which was employed previously by [16, 17] since it is much more realistic and accounts for the expression saturation. Given a gene X, its parent set Π can be further partitioned into two disjoint subsets: the activator set A and the repressor set R, that is, $\Pi = A \cup R$ and $A \cap R = \phi$. The kinetics of gene X can be explained by the following differential equation:

$$\frac{dx}{dt} = -\lambda x + \frac{\delta + \sum_{i=1}^{|A|} a_i^{\alpha_i}}{1 + \sum_{i=1}^{|A|} a_i^{\alpha_i} + \sum_{j=1}^{|R|} r_j^{\gamma_j}}, \tag{1}$$

where x is the concentration of gene X's transcriptional product, namely, mRNA. In this paper, to simplify the exposition, the gene name and its expression are used interchangeably. The changing rate of gene X is controlled by its activating and repressing parents, denoted individually by $a_i \in A$ and $r_j \in R$. α and γ serve as the regulating factors corresponding to each activator and repressor. α and γ assume positive values, and hence can be modeled by a gamma distribution with shape and scale parameters (κ, β). Here we can unbiasedly assume that the activators and repressers share the same gamma distribution for their regulation factors. Other light-tail distributions, such as Weibull and lognormal distributions, could also be employed. However, since gamma distribution is popular in modeling the reaction rate or molecular concentration [26], the gamma distribution is chosen here. λ stands for the gene degradation rate and the time scale can be properly chosen in order to normalize λ to the unit value ($\lambda = 1$). δ represents the expression baseline rate, taht is, the expression rate for X when there is neither activator nor repressor regulating the target gene X. Suppose that y represents the observation of x, then y assumes the form $y = x + \varepsilon$, where ε incorporates all noise sources and is modeled by an additive Gaussian random variable with zero mean and variance σ^2.

As the response to environmental changes or incitations, a mature biological system always converges to a certain steady-state, in which all genes stay in equilibrium and do not change their expressions. In this context, the periodic processes, for example, cell cycle and circadian rhythm, are excluded from our research interest. By setting $dx/dt = 0$ and $\lambda = 1$, the observation y of the steady-state gene expression for gene X can be expressed as

$$y = \frac{\delta + \sum_{i=1}^{|A|} a_i^{\alpha_i}}{1 + \sum_{i=1}^{|A|} a_i^{\alpha_i} + \sum_{j=1}^{|R|} r_j^{\gamma_j}} + \varepsilon. \tag{2}$$

Given a parent structure Π for gene X, the parameters in Θ can be summarized as follows.

(1) For each parent $\pi \in \Pi$, a binary variable is demanded to specify whether the parent is an activator or repressor, that is, $\mathbf{1}_A(\pi)$, where $\mathbf{1}$ is the indicator function and it assumes the value 1 when $\pi \in A$, and 0 otherwise. It can be modeled by a Bernoulli random variable with known success probability ρ.

(2) For each activator $a \in A$ and repressor $r \in R$, it is assumed that the regulating factors $\alpha, \gamma \sim \text{Gamma}(\kappa, \beta)$, where κ, β are known.

(3) The baseline parameter δ is usually known.

(4) The noise $\varepsilon \sim N(0, \sigma^2)$, where σ^2 can be set to a specific value or estimated.

It is worth to note that the choice of nonlinear differential equation and parameter priors does not influence the flow of analysis. Our scheme stands for a general framework and the detailed parameters can be easily customized to fit different scenarios. There are various mathematical models for system kinetics, such as [27–29]. The kinetics in 1 is chosen as our dynamic model because it possess the property of saturation, a key idea of Michaelis-Menten kinetics [29]. Besides, it is fairly simple and it also takes account of most other biological properties. Therefore, in the simulation of the real data set, we are assuming the proposed kinetics is true.

3. Inference Method

Consider a system composed of n genes indexed by $\{1, 2, \ldots, n\}$. ChIP-chip experiments can be conducted to

examine whether gene i's corresponding protein binds gene j's regulatory region. Usually this regulatory sequence is a promoter region which is located within 600 base pairs upstream of the coding region of gene j. The experimental results are represented in terms of p-values. In the first step, it is necessary to translate the p-value p into the probability of existence of a regulation relationship from gene i to gene j, which is denoted as $\mathcal{P}(i\rightarrow j \mid p)$. This probability will act as the prior knowledge to integrate gene expression data.

3.1. Incorporating ChIP-Chip Data

The p-value is within the range of $[0, 1]$. After studying the properties of the microarray data, Allison proposed to exploit mixed Beta distribution to model the p-value [30]. If the transcription factor i regulates gene j, it is assumed that the ChIP-chip experiment produces a p-value p which conforms to a Beta distribution with parameters (ϕ, ζ),

$$f(p \mid i \longrightarrow j) = \frac{p^{\phi-1}(1-p)^{\zeta-1}}{B(\phi,\zeta)}, \tag{3}$$

where $f(\cdot)$ stands for the probability density function and $B(\cdot,\cdot)$ represents the beta function. On the other hand, if i does not regulate j, the p-value assumes a different Beta distribution with parameters (ψ, ξ):

$$f(p \mid i \nrightarrow j) = \frac{p^{\psi-1}(1-p)^{\xi-1}}{B(\psi,\xi)}. \tag{4}$$

Based on the knowledge provided by established and verified genetic networks, one can infer a prior knowledge about the probability of connectivity between arbitrary genes, denoted as $\eta(i\rightarrow j)$, for all i, j. Such statistics regarding the network connectivity can be found in the open literature, for example, the data sets for yeast [31], and Drosophila [32]. By applying Bayes theorem, we obtain

$$\mathcal{P}(i \longrightarrow j \mid p) = \frac{\eta B(\psi,\xi)p^{\phi-1}(1-p)^{\zeta-1}}{\eta B(\psi,\xi)p^{\phi-1}(1-p)^{\zeta-1} + (1-\eta)B(\phi,\zeta)p^{\psi-1}(1-p)^{\xi-1}}. \tag{5}$$

For simplicity, a uniform distribution can be alternatively employed to account for the p-value when $i \nrightarrow j$. In this case, $\psi = 1$, $\xi = 1$, and (5) takes the form

$$\mathcal{P}(i \longrightarrow j \mid p) = \frac{\eta p^{\phi-1}(1-p)^{\zeta-1}}{\eta p^{\phi-1}(1-p)^{\zeta-1} + (1-\eta)B(\phi,\zeta)}. \tag{6}$$

The determination of ϕ and ζ depends on the experimental knowledge of the accuracy of selecting p-value thresholds. In the first step, a p-value threshold p_t is imposed, then the validity of all bindings with p-values less than p_t is corroborated by biological experiments. In this way, we can gain knowledge of the probability $\mathcal{P}(i\rightarrow j \mid p < p_t)$, which can be written in the form of

$$\begin{aligned}\mathcal{P}(i \longrightarrow j \mid p < p_t) &= \frac{\eta\mathcal{P}(p < p_t \mid i \longrightarrow j)}{\eta\mathcal{P}(p < p_t \mid i \longrightarrow j) + (1-\eta)\mathcal{P}(p < p_t \mid i \nrightarrow j)} \\ &= \frac{\eta\int_0^{p_t} p^{\phi-1}(1-p)^{\zeta-1}dp}{\eta\int_0^{p_t} p^{\phi-1}(1-p)^{\zeta-1}dp + p_t(1-\eta)B(\phi,\zeta)}.\end{aligned} \tag{7}$$

Some works in the literature, for example, [33], have made the observation that at a p-value threshold of 0.001, the frequency of false positives is 6%–10%, that is, $\mathcal{P}(i \nrightarrow j \mid p < p_t) \in [6\%, 10\%]$. Taking into account these special points, we can determine the pair (ϕ, ζ) in a small range. In our case, $\phi \approx 0.1$ and $\zeta \approx 100$. Finally, a table can be set up to map the p-value into the edge existence probability, which can be computed only once. It is an overhead for the computational system but it does not assume much computational resource in the runtime.

3.2. Exploiting Steady-State Gene Expression Data

Assume that m observations of expression vector are obtained and stored in matrix $D^{n\times m}$. Next, we develop a computational approach to establish the posterior probability of the regulation $i\rightarrow j$, that is, the probability of the existence of the edge $i\rightarrow j$, which is represented by $\mathcal{P}(i\rightarrow j \mid D, p)$. This posterior can be obtained through integration over the whole parental gene set and parameter space for gene j:

$$\begin{aligned}\mathcal{P}(i \longrightarrow j \mid D, p) &= \sum_{\Pi_j}\int_{\Theta_j} f(i \longrightarrow j, \Pi_j, \Theta_j \mid D, p)\,d\Theta_j \\ &= \sum_{\Pi_j}\int_{\Theta_j} \mathbf{1}_{\Pi_j}(i) f(\Pi_j, \Theta_j \mid D, p)\,d\Theta_j,\end{aligned} \tag{8}$$

where the function $\mathbf{1}_{\Pi_j}(i)$ is the indicator function, which takes 1 if $i \in \Pi_j$ and 0 otherwise. Applying Bayes theorem, $f(\Pi_j, \Theta_j \mid D, p)$ can be expressed as

$$\begin{aligned}f(\Pi_j, \Theta_j \mid D, p) &= \frac{f(D \mid \Pi_j, \Theta_j, p) f(\Pi_j, \Theta_j \mid p)}{f(D \mid p)} \\ &= \frac{f(D \mid \Pi_j, \Theta_j) f(\Pi_j, \Theta_j \mid p)}{f(D)} \\ &= \frac{f(D \mid \Pi_j, \Theta_j) f(\Pi_j, \Theta_j \mid p)}{\sum_{\Pi_j}\int_{\Theta_j} f(D \mid \Pi_j, \Theta_j) f(\Pi_j, \Theta_j \mid p)\,d\Theta_j} \\ &= \frac{f(D_j \mid D_{\bar{j}}, \Pi_j, \Theta_j) f(\Pi_j, \Theta_j \mid p)}{\sum_{\Pi_j}\int_{\Theta_j} f(D_j \mid D_{\bar{j}}, \Pi_j, \Theta_j) f(\Pi_j, \Theta_j \mid p)\,d\Theta_j},\end{aligned} \tag{9}$$

where D_j denotes the observations of gene X_j, and $D_{\bar{j}}$ represents the collection of all the observations pertaining to all genes excluding those of gene X_j. $f(\Pi_j, \Theta_j \mid p)$ denotes the probability density of the high-dimensional parental model given the observation of ChIP-chip data. $f(D_j \mid D_{\bar{j}}, \Pi_j, \Theta_j)$ stands for the gene expression likelihood given the parental values and the graphical model. It is a Gaussian distribution with known variance and mean determined by the first part of (2). The second equality in (9) holds because we believe the ChIP-chip experiment and steady-state gene expression measurements are independent. By plugging (9) into (8), it can be inferred that

$$\mathcal{P}(i \longrightarrow j \mid D, p) = \frac{\sum_{\prod_j} \int_{\Theta_j} \mathbf{1}_{\Pi_j}(i) f(D_j \mid D_{\bar{j}}, \Pi_j, \Theta_j) f(\Pi_j, \Theta_j \mid p) d\Theta_j}{\sum_{\prod_j} \int_{\Theta_j} f(D_j \mid D_{\bar{j}}, \Pi_j, \Theta_j) f(\Pi_j, \Theta_j \mid p) d\Theta_j}. \tag{10}$$

The integrations at the numerator and denominator of (10) cannot be generally performed in a closed-form expression. However, the Monte Carlo methods enable to numerically evaluate the posterior probabilities. We can generate Monte Carlo samples based on the model probability density $f(\Pi, \Theta \mid p)$ and the integration can be obtained by averaging over these samples. Then the posterior probabilities can be estimated by

$$\mathcal{P}(i \longrightarrow j \mid D, p) \approx \frac{\sum_{\prod_j, \Theta_j} \mathbf{1}_{\Pi_j}(i) f(D_j \mid D_{\bar{j}}, \Pi_j, \Theta_j)}{\sum_{\prod_j, \Theta_j} f(D_j \mid D_{\bar{j}}, \Pi_j, \Theta_j)}. \tag{11}$$

Assuming that the selection of a parent as an activator is performed in an independent manner, and that the selection of the regulation factor value is also performed independently, the model probability density $f(\Pi, \Theta \mid p)$ can be further expanded by using the chain rule

$$f(\Pi, \Theta \mid p) = f(\Theta \mid \Pi)\mathcal{P}(\Pi \mid p) = \prod_{i=1}^{|A|} [\rho f(\alpha_i)] \prod_{j=1}^{|R|} [(1-\rho) f(\gamma_j)] \mathcal{P}(\Pi \mid p). \tag{12}$$

Equation (12) conveys the idea that the random samples of graphical models can be sequentially created and processed. First the network structure is created based on the binding probability of gene regulation obtained in the ChIP-chip experiment, then each parent is randomly assigned to represent an activator or repressor, and finally regulation factors are generated.

3.3. Algorithm Formulation

Our computational procedure can be briefly formulated in terms of Algorithm 1, where the Matlab coding convention is used to write the pseudocode. There exist n genes in the system. An $n \times n$ matrix is created to represent the p-values produced in the ChIP-chip experiment. We collect m steady-state gene expression samples. The output entry C_{ij} stands for $\mathcal{P}(i \rightarrow j \mid D, p)$, and M denotes the number of Monte-Carlo iterations. Lines 1 and 2 deal with the ChIP-chip experimental data and translate p-values into the binding probabilities by using (5). The results are stored in matrix B. Lines 3 and 4 perform the preprocessing of the gene expression data. Let $Y_{(1)}, Y_{(2)}, Y_{(3)}, \ldots, Y_{(m-2)}, Y_{(m-1)}, Y_{(m)}$ be the values of a specific gene expression in ascending order. The smallest two values, $Y_{(1)}$, $Y_{(2)}$, and the largest two values, $Y_{(m-1)}$, $Y_{(m)}$, are treated as outliers and discarded. The dynamic range is defined as $R = Y_{(m-2)} - Y_{(3)}$. The gene expressions are normalized as follows: the smallest two samples are assigned the null value and the largest two samples are assigned the unit value; the intermediary samples $Y_{(i)}$ are normalized as $(Y_{(i)} - Y_{(3)})/R$; if there is a missing sample, it is recovered through interpolation by gene's mean expression. Lines 12 through 16 implement the numerator of (11), and Line 17 computes the denominator of (11).

The algorithm can be easily reorganized into a parallel form so that we can exploit efficiently the distributed computational resources. The entries of output matrix C represent the posterior probabilities of regulation relationships between any two genes. It is directional (asymmetrical), and it possesses a clear probabilistic meaning compared with other vague connectivity metrics, for example, mutual information. It grants the biologists the flexibility first to examine the most significant interactions, then to proceed with less evidenced edges. Therefore, it is advantageous relative to a purely dichotomous scheme, in which genes are treated as being either connected or disconnected. A probability threshold can be imposed to change the algorithm into a dichotomous classifier. Since the posterior probability has a universal meaning, this threshold can be easily selected, usually within the range of [0.3–0.9]. A tradeoff has also to be made for different performance metrics.

4. Results

The simulation consists of two parts. In the first part, artificial networks are created and the performance of the proposed algorithm is compared with other representative algorithms available in the literature, namely the relevance network (RN) method [14], Chow-Liu algorithm [24], and ARACNE [18]. In the second part, the algorithm is tested on the real *Saccharomyces cerevisiae* (budding yeast) data set and a biologically meaningful genetic network is inferred for the genetic response to environmental changes.

4.1. Simulation on Artificial Networks

The proposed algorithm is compared with other three algorithms to evaluate its capability of recovering genetic networks based on gene expression data alone. The relevance network (RN) model [14] represents a robust inference method based on gene expression profiles. In the first step, it computes the mutual information between any two genes X and Y, denoted as $I(X; Y)$. Then it suggests two genes X and Y to be relevant if their mutual information assumes a larger

(1) Input ChIP-chip data set $p^{n\times n}$;
(2) Translate p-values to construct the binding probability matrix $B^{n\times n}$.
(3) Input gene expression data set $D^{n\times m}$;
(4) Normalize the expression data so that each expression is within the range of $[0, 1]$;
(5) Initialize n, $L = 0^{1\times n}$, $C = 0^{n\times n}$;
(6) *for* $k = 1$ to M *do*
(7) Randomly create a directed graph and the adjacency matrix J based on B;
(8) *for* $i = 1$ to n *do*
(9) For gene i's parents specified in $J(:, i)$, randomly assign them to be activators or repressers;
(10) For each parent, randomly create their regulation factor α or γ;
(11) $l \Leftarrow \text{likelihood}(D_i \mid D_{\bar{i}}, \Pi_i, \Theta_i)$;
(12) *for* $j = 1$ to n *do*
(13) *if* $j \in \Pi_i$ *then*
(14) $C_{j,i} = C_{j,i} + l$;
(15) *end if*
(16) *end for*
(17) $L(i) = L(i) + l$;
(18) *end for*
(19) *end for*
(20) $\forall i, j, C_{j,i} = C_{j,i}/L_i$;
(21) *return* C.

Algorithm 1: Inference of connectivity significance.

value than a prespecified threshold and it lays down an undirected edge as X-Y. Hence, RN measures the significance of gene interactions in terms of mutual information between the gene expressions and produces an undirected cyclic graph. Chow-Liu algorithm [24] approaches the inference problem by finding the maximum spanning tree in which the edge weights stand for the mutual information. However, it loses validity if the underlying model is a cyclic graph. In addition, when the graph is densely connected, this scheme might falsely miss too many edges. ARACNE algorithm [18] exploits the data processing inequality (DPI). It starts with a fully connected graph and a predefined mutual information threshold. Whenever the mutual information between two genes X and Y, that is, $I(X; Y)$, is less than a threshold, it disconnects the two genes. Next, in the preliminary graph if there exists Z so that $I(X; Y) < \min(I(X; Z), I(Y; Z))$, then it disconnects X and Y. In our simulations, we resort to an already available but efficient Matlab toolbox [34] to estimate the mutual information.

4.1.1. Performance Definition

Before making performance comparisons, we define inference errors and performance metrics. Because RN, Chow-Liu, and ARACNE algorithms all construct undirected graphs, we have to disregard the orientation information inferred by the proposed algorithm. The synthetic and inferred graphs are represented by $G(V, E)$ and $\hat{G}(V, \hat{E})$, respectively. The two graphs share the same set of vertices but differ in the set of edges.

There are two types of inference errors. The type-1 errors are false positives (FP) and are also called false alarms. If the inference algorithm determines an interaction for two vertices X and Y in the inferred graph, denoted as X-$Y \in \hat{E}$, but there is no such edge in the synthetic graph, that is, X-$Y \notin E$, then an FP is produced. The number of FPs, represented by N_{FP}, can be counted as follows:

$$N_{\text{FP}} = \sum_{\forall X,Y} \left((X - Y \in \hat{E}) \bigcap (X\text{-}Y \notin E) \right), \tag{13}$$

where $\bigcap$ stands for the logic and operator. The type-2 errors are false negatives (FN) and also named misses. If the inference does not discover the connectivity X-Y which resides in the synthetic network, an FN is generated. The number of FNs, depicted by N_{FN}, is obtained by

$$N_{\text{FN}} = \sum_{\forall X,Y} \left((X\text{-}Y \in E) \bigcap (X\text{-}Y \notin \hat{E}) \right). \tag{14}$$

Correct inference can also be divided into two categories. If X-$Y \in \hat{E}$ and X-$Y \in E$, the correctness is defined as a true positive (TP). Its summation, annotated by N_{TP}, is

$$N_{\text{TP}} = \sum_{\forall X,Y} \left((X\text{-}Y \in \hat{E}) \bigcap (X\text{-}Y \in E) \right). \tag{15}$$

On the other hand, if X-$Y \notin \hat{E}$ and X-$Y \notin E$, this correctness is called a true negative (TN). The number of TNs, represented by N_{TN}, is defined as follows:

$$N_{\text{TN}} = \sum_{\forall X,Y} \left((X\text{-}Y \notin \hat{E}) \bigcap (X\text{-}Y \notin E) \right). \tag{16}$$

Different performance metrics are proposed in the literature. Three most popular of them are considered here. The first metric, referred to as the Hamming distance, is the summation of all the inference errors and is given by

$$\text{Hamming distance} = N_{\text{FP}} + N_{\text{FN}}. \tag{17}$$

The Hamming distance is widely accepted as a good measure of the distance between two graphs.

The second metric is called the sensitivity, and is defined as

$$\text{Sensitivity} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}}. \tag{18}$$

The sensitivity describes the inference algorithm ability to identify the regulation relationships among genes. The third metric is called the specificity, and it assumes the form

$$\text{Specificity} = \frac{N_{\text{TN}}}{N_{\text{TN}} + N_{\text{FP}}}. \tag{19}$$

The specificity represents the inference algorithm's capability to avoid falsely connecting two unrelated genes.

4.1.2. Simulation on the Proposed Kinetics

A set of artificial networks are created based on the system dynamic equation (1). Each network has 30 vertices and 60 oriented edges. Such a network scale is selected for the consideration of the computational resources and the biological network that we are going to infer. The steady-state data are sampled by emulating the gene knockout experiment. A gene's expression is mandatorily forced to 0 while all other genes are free to change their expressions. The initial values of the system are randomly generated. When the system converges to the equilibrium, a Gaussian noise $N(0, 0.03)$ is added and a few samples are obtained. All genes are shut down one by one. An extra in silico experiment is performed and no genes are shut down. These samples correspond to the wild type strain.

Different numbers of steady-state samples were generated based on the adopted system kinetics. The transcription factor is assumed to be an activator or repressor with equal probability, that is, $\rho = .5$. The baseline parameter $\delta = 0.5$ and the gamma parameters of regulation factors are ($\kappa = 16$, $\beta = 0.0625$) so that the regulation factor has a unit mean. Chow-Liu algorithm creates a spanning tree; therefore, it preserves only 29 edges, while the original synthetic network possesses 30 vertices and 60 edges. In order to make comparisons, we tune the parameters for the other three schemes so that the number of inferred edges is around 30. For the RN method, we keep the 30 edges with the highest mutual information. For ARACNE, the mutual information threshold is adjusted. In our proposed algorithm, the posterior probability thresholds are changed in the range of $[0.3, 0.9]$ so that approximately 30 edges are obtained. It has to be noted that RN, ARACNE, and Chow-Liu algorithms only preserve interactions but disregard the interaction orientation. Therefore, in order to make consistent comparisons, we have to sacrifice the orientation information offered by the proposed algorithm. Besides, these three schemes have no capability of processing ChIP-chip data. Therefore, we have to configure the proposed algorithm such that any two nodes are associated with a small prior probability of connection (0.1). This reflects the fact that the connection between two arbitrary nodes in the graph is very unlikely, but not impossible. This also exemplifies how the algorithm works in the absence of the ChIP-chip data.

Figure 1(a) compares the performance in terms of Hamming distance for the four schemes assuming different sample sizes. The proposed method provides much better inference accuracy because it achieves the lowest Hamming distance. Larger sample size rewards a better inference precision. Chow-Liu's algorithm and ARACNE do not perform well. This can be attributed to the assumption of the network. Our synthetic networks actually are cyclic networks in order to reflect the real world scenario. However, cycles in the network ruin the inference precisions of Chow-Liu and ARACNE. Figure 1(c) illustrates the impact of sample size on the sensitivity. The proposed scheme outperforms the other three schemes. The sensitivities of all algorithms are less than 0.5. This is mainly due to the constraint that we pose on the number of inferred edges, that is, 30 edges. If we relax the posterior probability threshold, the sensitivity will be improved by sacrificing the specificity. Figure 1(e) depicts specificity for all schemes. All of them have high specificities, which are all greater than 0.90. The proposed scheme still exceeds. This high specificity is mainly due to the stringent constraint posed on the number of inferred edges. When considering the orientation of the edges, we find that 90% true positives inferred by the proposed algorithm are actually oriented correctly. This represents a big advantage of the proposed algorithm compared with the other three schemes.

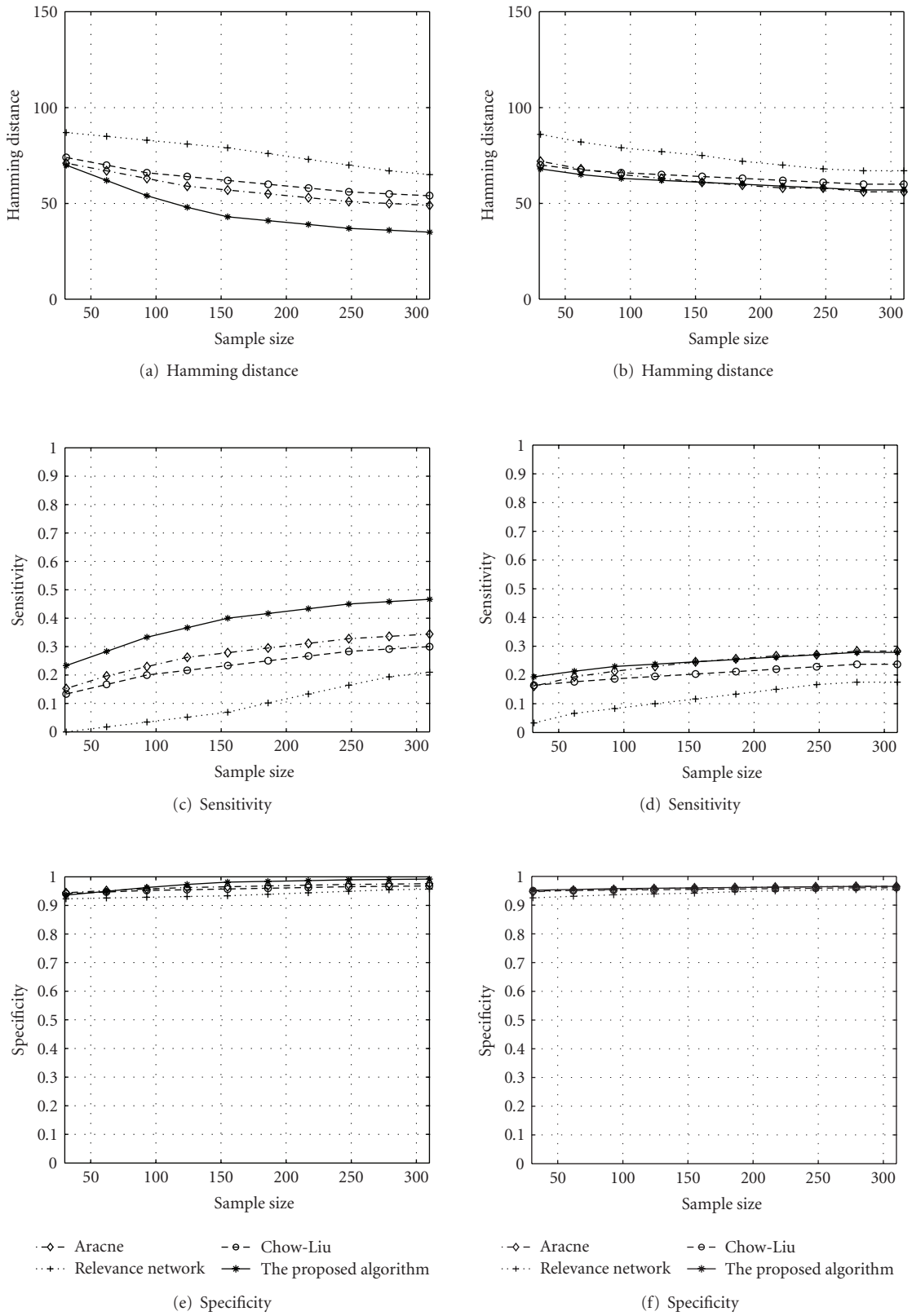
4.1.3. Robustness of Inference

In the previous simulations, the proposed inference algorithm assumes the system dynamic as depicted by (1). Actually, for different biological processes, there exist various mathematical models which achieve tradeoffs between the sophistication of the underlying molecular reaction and the simplification of the formula description (see [27, 29] for model comparisons). Savageau [28] proposed an alternative mathematical model to account for the gene control and various forms of coupling among elementary gene circuits. This model can be denoted as

$$\frac{dx}{dt} = \lambda_A \prod_{i=1}^{|A|} a_i^{\alpha_i} - \lambda_R \prod_{j=1}^{|R|} r_j^{\gamma_j}, \tag{20}$$

where two new symbols λ_A and λ_R are activation and degradation coefficients and all other symbols share the same meanings as in (1).

Although the proposed inference framework can "plug and play" with different models, it is still necessary to



(a) Hamming distance (b) Hamming distance

(c) Sensitivity (d) Sensitivity

(e) Specificity (f) Specificity

Figure 1: Performance comparison in terms of Hamming distance, sensitivity and specificity. Figures in the left column illustrate results based on the same kinetics model employed in both data synthesization and network inference, while figures in the right column represent results based on different kinetics models employed in the simulation process. The Monte Carlo iterations are fixed at 1,000,000 for the proposed algorithm. Thresholds for different algorithms are selected to produce around 30 inferred edges.

examine its robustness against the underlying model. We evaluate this model dependence by the following steps: configure the model as 13 and create a set of synthetic data, then apply the proposed algorithm based on the dynamic equation (1), finally determine the performance metrics for different algorithms and compare the results with those in the previous section.

The simulation results are plotted in Figures 1(b), 1(d), and 1(f). Each figure corresponds to a different performance metric. All algorithms exhibit different values for performance values. This shows that the inference is dependent on the particular data sets and their underlying model. Compared with other three schemes, the proposed algorithm still achieves good performance in terms of three metrics. However, the advantage of the proposed algorithm are not significant now. ARACNE, Chow-Liu, and relevance method do not degenerate much. This attributes mainly to the nonparametric property of these three schemes. The persistent good performance of the proposed algorithm is due to the fact that both dynamic models have to convey the basic properties of the gene interaction kinetics, such as the activation and repression effects and the coupling of the circuitry.

4.2. Simulation on Saccharomyces Cerevisiae Data Sets

Saccharomyces cerevisiae (yeast) has been extensively studied in the literature of molecular biology because it is a unicellular eukaryotic organism, which shares similar cell structure with plants and animals. Also, yeast presents a short life cycle, which makes the experiments to be easily conducted. Lee et al. [33] performed the ChIP-chip experiment, in which 141 transcription factors were tested for binding intergenetic regions corresponding to 6270 genes. The gene expression data were published by Mnaimneh et al. [35], who created promoter shut-off strains for 2/3 of all essential genes. The data set contains 215 steady-state cDNA microarray samples. The model parameters are assumed the same as artificial networks.

The intracellular signalling pathway in response to environmental changes has been conserved through evolution. Therefore, a study of this biological subsystem on the *Saccharomyces cerevisiae* might help to decipher the cell survival mechanism of other organisms. We select 30 genes which are annotated to participate in the stress response process. The given ChIP-chip experiment did not provide full prior knowledge between any two genes (nodes in the graph). We believe that, among these genes, there are some genes whose protein products may also serve as transcription factors. Therefore, if the binding between two genes was not tested in the ChIP-chip experiment, a small probability value 0.1 is assigned as the prior knowledge. The proposed inference algorithm leads to the genetic network illustrated in Figure 2.

The inferred genetic regulatory network shows strong proneness toward a scale-free network instead of a random network. Some genes possess especially high degree of connectivity. Three hub genes CIN5, HSF1, MSN4 already connect with more than 60% of all selected genes. Each of them has a connectivity degree not less than 8 while on average each gene in the network is connected with no more than 4 genes. These hub genes constitute the backbone of the network and they are potential control targets. This scale-free property is advantageous in maintaining the system robustness because a failure in one subsystem will not be propagated to the whole body.

Multiple works, for example [36], have identified MSN4 and MSN2 as two of the most important genes in the response to environmental changes. A recent work [37] recognized the functionality of another crucial gene HSF1, which is a heat shock transcription factor and functions in a different domain than the one corresponding to MSN2/4. Our inferred network confers this experimental result by showing that HSF1 and MSN2/4 regulate different set of genes except a weak connectivity between HSF1 and MSN4. MSN2/4 are not conserved in humans, while HSF genes have been preserved for various organisms such as Drosophila melanogaster, chickens, and mammals. Therefore, a study of the HSF1 pathway opens up the possibility of understanding the mechanism that governs the survival of normal cells under austere conditions.

CIN5 (YAP4) and YAP6 are two genes that play key roles in controlling the resistance to drugs, for example, cisplatin [38]. CAD1(YAP2), CIN5, YAP1, and YAP6 share a structure motif called basic leucine zipper (bZIP) and they are located closely in the network. However, they are not neighboring the other two bZIP genes: YAP5 and YAP7. It is hypothesized that although they have similar molecular structures, their biological functionalities are in distinct domains.

Several edges, discovered by imposing a stringent p-value threshold 0.001 to the location data, were persevered in our inferred network. Actually, these connections constitute a small portion of the proposed network, and they are CIN5→MSN1, CIN5→YAP6, CIN5→ROX1, YAP1→YAP6, MAC1→CUP9, CUP9→YAP6, and HAL9→MSN4. Various evidences are found to corroborate the recovered interactions, which can not be obtained by employing a stringent p-value for the location data. For example, YAP5 is recovered to directly regulate STE50. This regulation relationship has also been reported in the work of Horak [39]. The relationship between MSN2 and SCH9 is studied in [40] in the context of extending the life span.

It is worthwhile to note that gene expression data mainly provide statistical relationships among genes, while location data offer physical binding interactions at the molecular level. By combining the two data sources, we are aiming to refine the inferred network to be biologically more meaningful. However, it also runs at a risk of confusing statistical regulatory relationships with real binding interactions. When such a case occurs, the proposed algorithm is capable of constraining the interacting genes within the same biological process and common functional relationships. A related discussion about the meaning of inferred network can also be found in [41].

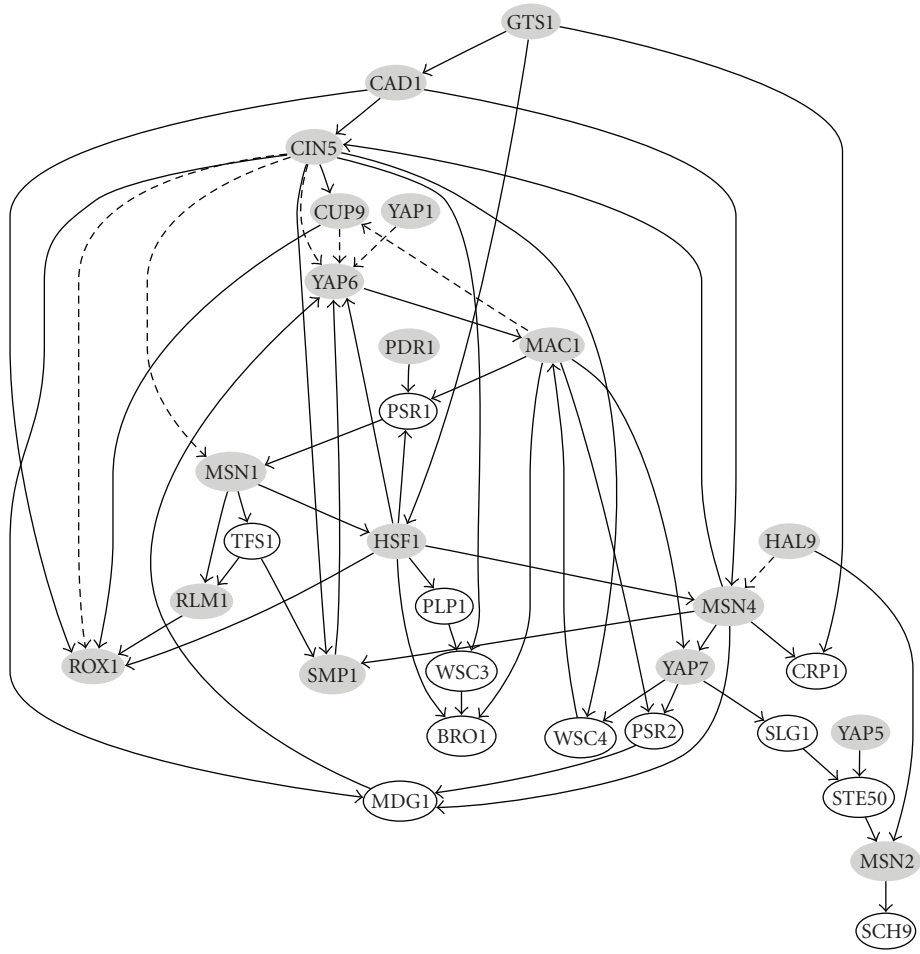


Figure 2: Recovered genetic regulatory network for yeast stress response. The Monte Carlo iterations are 1,000,000. Dashed edges represent interactions preserved by using ChIP-chip data alone under the p-value threshold 0.001. Shadowed vertices are transcription factors tested in the ChIP-chip experiment.

5. Conclusions

A novel algorithm is proposed to recover the genetic regulatory networks in the light of knowledge of transcriptional kinetics, ChIP-chip, and gene microarray data. The analysis is based on the Bayesian methodology and Monte Carlo techniques. The proposed scheme is useful to compensate the shortcomings of the utilization of only one data set alone. Our in silico experiments corroborate that the algorithm outperforms in specificity, sensitivity and Hamming distance relative to three state-of-the-art schemes. A budding yeast genetic regulatory network is proposed to account for the stress response.

There are possible extensions to our current scheme. An analysis of the error estimation is desired for the Monte Carlo simulation in order to determine the appropriate number of iterations. Several other knowledge sources are to be integrated into the current framework. For example protein-protein interactions are useful to identify cobinding regulations. Genome sequencing data have been utilized to find regulatory motifs. Protein structure knowledge can be exploited to categorize the proteins and find similar functionality. A cross-species research is also highly desirable since similar regulation mechanisms are expected to be conserved. If a gene is conserved in both humans and mice, then the knowledge of the gene pathway in the mouse will be an excellent reference for the study of human genetic diseases. We expect a global distributed framework, in which each data source acts as a separate component and its absence does not interfere with the whole computational process.

Acknowledgments

This work was supported by the National Cancer Institute (CA-90301) and the National Science Foundation (ECS-0355227 and CCF-0514644).

References

[1] S. A. Kauffman, "Metabolic stability and epigenesist in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, 1969.

[2] K. Murphy and S. Mia, "Modelling gene expression data using dynamic Bayesian networks," Tech. Rep., Computer Science Division, University of California, Berkeley, Calif, USA, 1999.

[3] P. Sebastiani, M. M. Abad, and M. Ramoni, "Bayesian networks," in *The Data Mining and Knowledge Discovery Handbook*, pp. 193–230, Springer, New York, NY, USA, 2005.

[4] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[5] I. Tabus and J. Astola, "On the use of MDL principle in gene expression prediction," *Journal on Applied Signal Processing*, vol. 2001, no. 4, pp. 297–303, 2001.

[6] H. Öktem, R. Pearson, O. Yli-Harja, D. Nicorici, K. Egiazarian, J. Astola, et al., "A computational model for simulating continuous time Boolean networks," in *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS '02)*, Raleigh, NC, USA, October 2002.

[7] K. Murphy, "Bayes Net Toolbox for Matlab," http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html.

[8] P. Leray, "Structure learning toolbox of Bayesian Networks," http://bnt.insa-rouen.fr/ajouts.html.

[9] N. Friedman and G. Elidan, "Bayesian Network learning tool," http://www.cs.huji.ac.il/labs/compbio/LibB/programs.html#LearnBayes.

[10] A. Rao, A. O. Hero, D. J. States, and J. D. Engel, "Manifold embedding for understanding mechanisms of transcriptional regulation," in *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS '06)*, pp. 3–4, College Station, Tex, USA, May 2006.

[11] S. Liang, S. Fuhrman, and R. Somogyi, "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '98)*, pp. 18–29, Maui, Hawaii, USA, January 1998.

[12] I. T. Luna, Y. Yin, Y. Huang, D. P. R. Padillo, M. C. C. Perez, and Y. Wang, "Uncovering gene regulatory networks using variational Bayes variable selection," in *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS '06)*, pp. 111–112, College Station, Tex, USA, May 2006.

[13] W. Zhao, E. Serpedin, and E. R. Dougherty, "Inferring gene regulatory networks from time series data using the minimum description length principle," *Bioinformatics*, vol. 22, no. 17, pp. 2129–2135, 2006.

[14] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '00)*, pp. 418–429, Honolulu, Hawaii, USA, January 2000.

[15] S. Rogers and M. Girolami, "A Bayesian regression approach to the inference of regulatory networks from gene expression data," *Bioinformatics*, vol. 21, no. 14, pp. 3131–3137, 2005.

[16] J. J. Rice, Y. Tu, and G. Stolovitzky, "Reconstructing biological networks using conditional correlation analysis," *Bioinformatics*, vol. 21, no. 6, pp. 765–773, 2005.

[17] M. K. S. Yeung, J. Tegnér, and J. J. Collins, "Reverse engineering gene networks using singular value decomposition and robust regression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 9, pp. 6163–6168, 2002.

[18] A. A. Margolin, I. Nemenman, K. Basso, et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, supplement 1, S7, pp. 1–15, 2006.

[19] I. Simon, Z. Siegfried, J. Ernst, and Z. Bar-Joseph, "Combined static and dynamic analysis for determining the quality of time-series expression profiles," *Nature Biotechnology*, vol. 23, no. 12, pp. 1503–1508, 2005.

[20] A. Bernard and A. J. Hartemink, "Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '05)*, pp. 459–470, The Big Island of Hawaii, Hawaii, USA, January 2005.

[21] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young, "Combining location and expression data for principled discovery of genetic regulatory network models," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '02)*, pp. 437–449, Lihue, Hawaii, USA, January 2002.

[22] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

[23] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: the combination of knowledge and statistical data," *Machine Learning*, vol. 20, no. 3, pp. 197–243, 1995.

[24] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transaction on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[25] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco, Calif, USA, 1988.

[26] N. Friedman, L. Cai, and X. S. Xie, "Linking stochastic dynamics to population distribution: an analytical framework of gene expression," *Physical Review Letters*, vol. 97, no. 16, Article ID 168302, 4 pages, 2006.

[27] L. F. Wessels, E. P. van Someren, and M. J. Reinders, "A comparison of genetic network models," in *Proceedings of the 6th Pacific Symposium on Biocomputing (PSB '01)*, pp. 508–519, The Big Island of Hawaii, Hawaii, USA, January 2001.

[28] M. A. Savageau, "Rules for the evolution of gene circuitry," in *Proceedings of the 3rd Pacific Symposium on Biocomputing (PSB '98)*, pp. 54–65, Maui, Hawaii, USA, January 1998.

[29] L. Edelstein-Keshet, *Mathematical Models in Biology*, Random House, New York, NY, USA, 1988.

[30] D. B. Allison, G. L. Gadbury, M. Heo, et al., "A mixture model approach for the analysis of microarray gene expression data," *Computational Statistics & Data Analysis*, vol. 39, no. 1, pp. 1–20, 2002.

[31] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès, "Topological and causal structure of the yeast transcriptional regulatory network," *Nature Genetics*, vol. 31, no. 1, pp. 60–63, 2002.

[32] L. Giot, J. S. Bader, C. Brouwer, et al., "A protein interaction map of *Drosophila melanogaster*," *Science*, vol. 302, no. 5651, pp. 1727–1736, 2003.

[33] T. I. Lee, N. J. Rinaldi, F. Robert, et al., "Transcriptional regulatory networks in *Saccharomyces cerevisiae*," *Science*, vol. 298, no. 5594, pp. 799–804, 2002.

[34] A. Ihler, "Kernel density estimation software," http://www.ics.uci.edu/~ihler/code/.

[35] S. Mnaimneh, A. P. Davierwala, J. Haynes, et al., "Exploration of essential gene functions via titratable promoter alleles," *Cell*, vol. 118, no. 1, pp. 31–44, 2004.

[36] A. P. Gasch, P. T. Spellman, C. M. Kao, et al., "Genomic expression programs in the response of yeast cells to environmental changes," *Molecular Biology of the Cell*, vol. 11, no. 12, pp. 4241–4257, 2000.

[37] D. L. Eastmond and H. C. M. Nelson, "Genome-wide analysis reveals new roles for the activation domains of the *Saccharomyces cerevisiae* heat shock transcription factor (Hsf1) during the transient heat shock response," *Journal of Biological Chemistry*, vol. 281, no. 43, pp. 32909–32921, 2006.

[38] T. Furuchi, H. Ishikawa, N. Miura, et al., "Two nuclear proteins, Cin5 and Ydr259c, confer resistance to cisplatin in *Saccharomyces cerevisiae*," *Molecular Pharmacology*, vol. 59, no. 3, pp. 470–474, 2001.

[39] C. E. Horak, N. M. Luscombe, J. Qian, et al., "Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*," *Genes & Development*, vol. 16, no. 23, pp. 3017–3033, 2002.

[40] P. Fabrizio, F. Pozza, S. D. Pletcher, C. M. Gendron, and V. D. Longo, "Regulation of longevity and stress resistance by Sch9 in yeast," *Science*, vol. 292, no. 5515, pp. 288–290, 2001.

[41] A. J. Hartemink, "Reverse engineering gene regulatory networks," *Nature Biotechnology*, vol. 23, no. 5, pp. 554–555, 2005.