Chapter 19
A Linear Programming Framework for Inferring Gene Regulatory Networks by Integrating Heterogeneous Data

Yong Wang
Academy of Mathematics and Systems Science, China

Rui-Sheng Wang
Renmin University, China

Trupti Joshi
University of Missouri, USA

Dong Xu
University of Missouri, USA

Xiang-Sun Zhang
Academy of Mathematics and Systems Science, China

Luonan Chen
Osaka Sangyo University, Japan

Yu Xia
Boston University, USA

ABSTRACT

There exist many heterogeneous data sources that are closely related to gene regulatory networks. These data sources provide rich information for depicting complex biological processes at different levels and from different aspects. Here, we introduce a linear programming framework to infer the gene regulatory networks. Within this framework, we extensively integrate the available information derived from multiple time-course expression datasets, ChIP-chip data, regulatory motif-binding patterns, protein-protein

DOI: 10.4018/978-1-60566-685-3.ch019

interaction data, protein-small molecule interaction data, and documented regulatory relationships in literature and databases. Results on synthetic and real experimental data both demonstrate that the linear programming framework allows us to recover gene regulations in a more robust and reliable manner.

INTRODUCTION

Cells efficiently carry out molecular synthesis, energy transduction, and signal processing across a range of environmental conditions by gene networks, which we define broadly as networks of interacting genes, proteins, and metabolites. Microarray technologies enable the simultaneous measurement of all RNA transcripts in a cell, producing tremendous amounts of gene expression data from different research groups. For instance, the Stanford Microarray Database (SMD) has deposited data for 70,113 experiments, from 341 labs and 56 organisms, as of 2007 (Demeter et al., 2007). Thus there is a pressing need for the development of sophisticated algorithms for reverse-engineering gene networks. So far, many computational algorithms have been developed to analyze gene expression profiles to detect dependencies among genes over different conditions.

Generally speaking, there are two strategies for studying the relationships among genes. The "physical (direct) interaction" approach seeks to identify true physical interactions between regulatory proteins and their binding promoters to reconstruct the so-called transcriptional regulatory network (R. S. Wang, Wang, Zhang, & Chen, 2007). The second strategy, the "genetic (indirect) interaction" approach seeks to identify regulatory influences between RNA transcripts to reconstruct the so-called gene regulatory network (Y. Wang, Joshi, Zhang, Xu, & Chen, 2006). Thus, in general, the regulator transcripts may exert their effects indirectly through the action of proteins, non-coding RNA, metabolites, and the cell environmental factors. An advantage of the influence strategy is that the model can implicitly capture regulatory mechanisms at the protein and metabolite level that are not physically measured (Gardner & Faith, 2005). In this study we focus on the inference problem for gene regulatory networks. The detailed descriptions on the first strategy, i.e. inferring transcriptional regulatory networks, can be found in (R. S. Wang et al., 2007).

So far, a wide variety of approaches have been proposed to infer gene regulatory networks from time-course data or perturbation experiments (De Hoon, Imoto, Kobayashi, Ogasawara, & Miyano, 2003; Dewey & Galas, 2001; Friedman, 2004; Gardner, di Bernardo, Lorenz, & Collins, 2003; Holter, Maritan, Cieplak, Fedoroff, & Banavar, 2001; Husmeier, 2003; Nachman, Regev, & Friedman, 2004; Tegner, Yeung, Hasty, & Collins, 2003). These approaches include discrete models of Boolean networks and Bayesian networks, and continuous models of neural networks and difference/differential equations. A common challenge for all these models is the scarcity of the data, since a typical gene expression dataset consists of relatively few time points (often less than 20) with respect to a large number of genes (generally over thousands). In other words, the number of genes far exceeds the number of time points for which data are available, making the problem of determining gene regulatory network structure a difficult and ill-posed one (D'Haeseleer, Liang, & Somogyi, 2000).

On the other hand, there are many heterogeneous data sources closely related to gene regulatory networks. These data sources provide rich information for depicting complex biological processes in cellular systems at different levels and from different aspects. It is necessary and important to understand gene expression and regulation through mining these data sources. Currently high-throughput microar-

ray technologies have produced tremendous amounts of gene expression data from different labs. At the same time, a large amount of protein-based data exist such as ChIP-chip, protein-protein interaction, and protein-small molecule interaction, which can also provide valuable information. Even though each experiment provides only limited information, these data are increasingly accumulated over many species and can be freely accessed from public databases and individual websites. It is therefore valuable and challenging to integrate gene expression data with other protein-based data generated by different research groups. If such large amounts of data from different experiments or conditions are combined and further exploited in an integrative and systematic manner, the scarcity of data can be greatly alleviated and the more accurate reconstruction of gene regulatory networks can be expected.

To address these challenges, we proposed a novel method to combine multiple time-course microarray datasets from different conditions for inferring gene regulatory networks (Y. Wang, Joshi, Zhang et al., 2006). The proposed method, called GNR (Gene Network Reconstruction tool), is based on linear-programming (LP) and a decomposition procedure. The method ensures the derivation of the network structure that is most consistent with all datasets. As a result, the method not only significantly alleviates the problem of data scarcity, but also markedly improves the prediction reliability. We tested GNR using both simulated data and experimental data in yeast and *Arabidopsis*. The result demonstrates the effectiveness of GNR in terms of predicting new gene regulatory relationships.

Different experimental technologies measure different aspects of a biological system, typically with different systematic biases. For example, current high-throughput assays are usually associated with high false-negative and false-positive rates. Thus, microarray data alone have a limited utility in inferring gene regulatory networks. From the viewpoint of systems biology, the integration of data from different sources provides an effective strategy to deal with this issue by reinforcing consistent and reliable observations and removing inconsistent and noisy ones. Moreover, because different experimental technologies provide different types of insights into a biological system, the integration of multiple data types offers the most comprehensive information about a particular cellular process (Hwang et al., 2005). For example, gene perturbation experiments (e.g., knockouts or RNA interference) may indicate relationships between genes due to direct or indirect genetic interactions. In contrast, chromatin immunoprecipitation chip data may reveal direct protein–DNA interactions or cofactor associations with bound transcription factors. Combining them together with microarray data provides a much more detailed view of the regulatory network than either alone.

In this chapter, we introduce a new computational strategy to infer gene regulatory networks based on linear programming. The main advantage of our strategy is to recover gene regulations in a robust and reliable manner by including all the available information derived from multiple expression datasets at different conditions and time points, motif-occurrence, ChIP-chip data, protein-protein interaction, protein-small molecule interaction, published literature and databases, and knockouts or RNA interference experiments. Furthermore, we can incorporate external inputs or perturbations such as small molecules into the formulation so that molecular targets (genes) can be identified in a systematic way.

The chapter is organized as follows: Firstly, the heterogeneous data sources for deriving gene regulatory relationships are briefly summarized. Secondly, we group the existing prior information into hard and soft constraints, describe the gene regulatory network by linear differential equations, and introduce a linear programming model to integrate data. Thirdly, both synthetic data and real experimental data are used to demonstrate the effectiveness and efficiency of our method. Finally, future research directions are discussed.

HETEROGENEOUS DATA SOURCES

Organisms use dynamic interactions of hundreds of genes to adapt to changes in the environment. To unravel this regulatory complexity, multiple technologies have been developed to detect the dependencies among genes, generating large amounts of heterogeneous data (Joyce & Palsson, 2006). These data depict the living cell from different aspects and angles. Here we give a brief summary of the existing data sources related to gene regulation relationships and their characteristics.

Multiple Time-Course Expression Data

DNA microarray experiments are usually classified based on the type of array used in the experiment (cDNA and oligonucleotide arrays) or according to the organism that is profiled. From the viewpoint of gene regulatory network modeling, we distinguish between static and time series experiments. In static expression experiments, a snapshot of the expression of genes in different samples is measured. In time series expression experiments, a temporal process is measured at various time intervals. Another important difference between these two types of data is that while static data from a sample population (e.g. ovarian cancer patients) are assumed to be independently and identically distributed, time series data exhibit a strong autocorrelation between successive points.

Since many biological systems are dynamic systems, temporal profiles of gene expression levels during a given biological process can often provide more insights into how gene expression levels evolve in time and how genes are dependent among each other during a given biological process. One important feature of such time-course gene expression data is the possible dependency of gene expression levels across time points for a given gene. In addition, as gene expression levels evolve over time, time intervals can be an important factor that affects the gene expression levels. Methods which can preserve the time sequence and the time dependence of the observed data are needed for analyzing the time-course gene expression data.

Due to the limitation of experimental technologies, a typical single time-course gene expression dataset consists of relatively few time points (often less than 20). On the other hand, multiple gene expression data generated by different groups on many species are increasingly available and accessible from public databases or websites. By combining and exploiting such large amounts of data from different experiments or conditions in an integrative and systematic manner, we can expect a more accurate reconstruction of the gene regulatory networks. It is worth mentioning that simply arranging multiple time-course datasets into a single expression profile dataset is inappropriate due to data normalization issues and lack of temporal relationships among these datasets.

ChIP-Chip Data

Protein-DNA interactome data concerns the interactions between proteins and DNA, particularly between transcription factors and their target promoters. They fundamentally define the transcriptional regulatory network of the cell. The recently developed ChIP-chip methodology involves the chromatin immunoprecipitation of an epitope-tagged transcription factor (TF) bound to DNA fragments containing target promoters, followed by the hybridization of those amplified DNA fragments to an intergenic microarray. Currently large amounts of ChIP-chip data in yeast and other organisms are publicly available. For example, genome-wide location data performed in yeast by (Harbison et al., 2004; Lee et al.,

2002) contain information regarding the binding of 204 regulators to their respective target genes in rich medium, and can be downloaded from their websites (http://web.wi.mit.edu/young/regulatory_code/ and http://web.wi.mit.edu/young/regulatory_network/).

ChIP-chip data have the advantage that they provide a direct biochemical link between TFs and promoters and have the potential to identify targets without knowing the activating conditions. From this viewpoint, ChIP-chip data are a very important source of information for analyzing direct transcriptional regulatory interactions.

Regulatory Motif Occurrence Data

We can also use the genome sequence data to infer regulatory relationships by systematically analyzing gene upstream regions in the genome to identify potential regulatory elements (also known as regulatory binding motifs). These motifs, often represented as regular expressions, were transformed into the corresponding weight matrices. We can then simply count the occurrences of regular expression-type patterns with the goal of identifying possible gene regulatory relationships. The weight matrices corresponding to these motifs are subsequently used to screen all intergenic sequences. The higher the score of a motif hit in a gene, the more likely it will be a regulatory relationship (Brazma, Jonassen, Vilo, & Ukkonen, 1998).

Protein-Protein Interaction Data

Proteins are the products of gene transcription and translation, and they play important roles in a cell. Protein-protein interactions occur in many cellular processes, such as signaling cascades and enzyme-complex formation. Identifying all functional protein-protein interactions is important for understanding the structure and function of the integrated cellular network. Currently, a lot of experimental protein-protein interaction data are available on the web (http://www.thebiogrid.org/).

Protein-protein interaction data can be roughly classified into two classes: physical and genetic interactions. There are many methods for mapping physical and genetic interactions. From BioGRID (Breitkreutz et al., 2008), the physical methods include affinity capture MS, two-hybrid, affinity capture western, and reconstituted complex, whereas the genetic methods include synthetic lethality, synthetic growth defect, epistatic miniarray profile, dosage rescue, and phenotypic enhancement. Here we would like to illustrate in detail genetic interaction relationships. For example, synthetic lethality is a genetic phenomenon in which two non-lethal mutations yield a lethal phenotype when combined. This phenomenon signifies the existence of genetic interactions between the two affected genes. Hence, genetic interactions may overlap with direct physical interactions or indirect logical interactions between genes as shown by perturbation experiments (e.g., knockouts or RNA interference).

Protein-Small Molecule Interaction Data

Small molecules can be used to dissect diverse biological processes, such as cellular metabolism, signal transduction and intracellular protein trafficking (Alaoui-Ismaili, Lomedico, & Jindal, 2002). Recently, the proliferation of web-based chemical databases has made information about an increasing number of compound structures and their biological properties publicly available. Among these databases are ChemBank, ZINC, PubChem, ChemDB, ChemMine, ChEBI, and DrugBank. Small molecule and protein

binding data are also abundant. For example, the DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information (Wishart et al., 2007). BindingDB currently contains 20,000 experimentally determined protein–ligand complexes from the literature and PDB (Liu, Lin, Wen, Jorissen, & Gilson, 2007). Binding MOAD is a database of 9,836 protein–ligand crystal structures (Benson et al., 2007). STITCH contains interactions for over 68,000 chemicals and over 1.5 million proteins in 373 species (Kuhn, von Mering, Campillos, Jensen, & Bork, 2008). These data provide useful information for the interactions between the gene regulatory network inside the cell and the environmental factors outside the cell.

Literature and Database Data

More reliable sources for gene regulatory relationships are from the literature and curated databases. For example, YEASTRACT (Yeast Search for Transcriptional Regulators And Consensus Tracking) is a curated repository of more than 12,500 regulatory associations between transcription factors and target genes in *Saccharomyces cerevisiae* (Teixeira et al., 2006), based on more than 900 bibliographic references. The information in YEASTRACT is updated regularly to match the recent literature on yeast regulatory networks. Since the regulatory relationships from literature and databases are usually generated by small-scale experiments, they are believed to be of high quality compared to large-scale experiments.

Co-Expression Relationships from Compendium Data

In addition to time-course data, the steady state gene expression data are also available in the databases. They can be assembled into gene expression profile or compendium data and used to extensively analyze the gene co-expression relationship. These microarray profile data are very useful in our derivation of gene regulatory network in two ways.

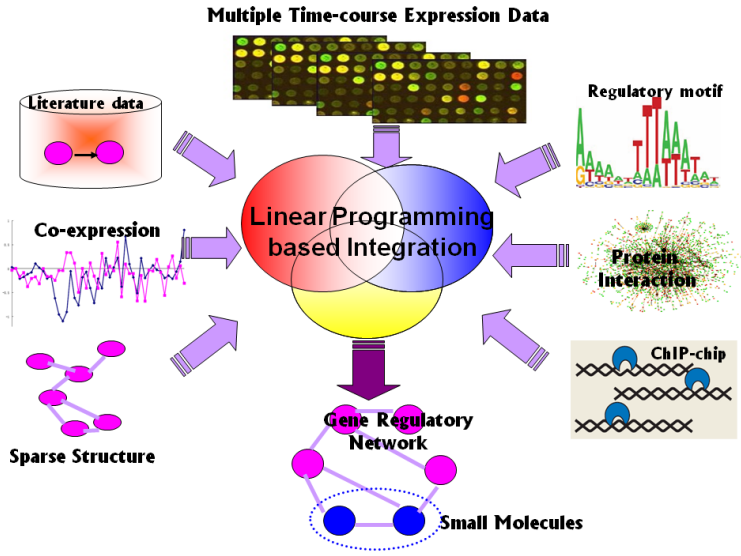
On one hand, gene expression data can be used to find co-expressed gene pairs (which display high correlation coefficient or mutual information score amongst different expression experiments). Over the past few years, several lines of evidence suggest that co-expressed genes possessing similar expression patterns across a set of steady states are likely to encode proteins that participate in the same metabolic pathway, form a common structural complex, or might be regulated by the same mechanism (Butte, Tamayo, Slonim, Golub, & Kohane, 2000). At the same time, diverse regulatory mechanisms may be responsible for the observed co-expression relationships.

On the other hand, gene expression data can be used to pick out the gene pairs which do not possess any co-expression relationships. That is to say, we can use large scale co-expression analysis in different conditions to reveal gene pairs which correlate weakly in terms of their expression level across various conditions. These identified pairs can be used as non-coregulatory samples approximately in our network inference model.

Prior Information about the Network Structure

In addition to various experimental data sources, we can also incorporate prior information about the network structure. For example, from the viewpoint of topology, it is commonly believed that gene regulatory network is sparse in nature, i. e. each gene is only genetically affected by a limited number

Figure 1. The graphic depiction of the strategy to integrate heterogeneous data using a linear programming framework



of genes (Gardner et al., 2003; Yeung, Tegner, & Collins, 2002). Furthermore, some people argue that the gene regulatory network possesses common properties of complex networks such as small world and scale free (Gustafsson, Hornquist, & Lombardi, 2005). It is straightforward to incorporate this prior information into our inference model. The main idea here is to make the gene regulatory network sparse so that it is biologically plausible. Such a strategy has been widely used (Gustafsson et al., 2005; Y. Wang, Joshi, Xu, Zhang, & Chen, 2006; Y. Wang, Joshi, Zhang et al., 2006; Yeung et al., 2002). For instance, a heuristic manipulation of sparseness is used in the procedure of the network reconstruction by computational analysis on a series of time points (Gardner et al., 2003; Yeung et al., 2002). A sparse scheme is performed by specifying the average number of connections for every gene in (Gardner et al., 2003). Another strategy is to use additional information from the microarray analysis and from the published literature to reduce the size of the problem and increase the reliability of the results (Nariai, Tamada, Imoto, & Miyano, 2005).

LINEAR PROGRAMMING FRAMEWORK FOR DATA INTEGRATION

Figure 1 illustrates the scheme of our proposed method. The time-course datasets of microarray experiments from different conditions or perturbations are collected. A gene regulatory network is described by ordinary differential equations (ODE). To infer the relationships between genes, the co-expression relations from time-course datasets and previously known regulations from the heterogeneous sources are collected as prior information, which are converted to hard and soft constraints respectively. In the end, the most consistent gene regulatory network is obtained with a linear programming-based algorithm.

Linear Differential Equations for Gene Regulatory Network

In general, a genetic network can be expressed by a set of nonlinear differential equations. Almost all of the existing approaches for gene regulatory network inference use linear or additive models, primarily due to the complex structures of biological systems and the scarcity of data (R. S. Wang et al., 2007; Y. Wang, Joshi, Xu et al., 2006; Y. Wang, Joshi, Zhang et al., 2006). Furthermore, linear equations can capture the main features of the network near the steady state, and can provide a good starting point for further modeling and analysis.

A common experimental technique for elucidating genetic network architecture is microarray measurements after different perturbations to the cell. An external perturbation means an experimental treatment that can alter the transcription rate of the genes in the cell. An example of perturbation is the alteration of the environment, treatment of the cell with a chemical compound, or genetic perturbation involving over- or under-expression of particular genes. Recent developments in large-scale genomic technologies enable researchers to measure gene expression profiles at multiple time points following perturbation of the genes of interest. We will extend the linear differential equation model to reconstruct gene regulatory networks and identify compound targets by considering the external perturbations outlined in this chapter. The model is based on relating the changes of gene transcript concentrations to each other and to the external perturbations.

Assume that there are N microarray datasets $X^1, X^2 \ldots, X^N$ with $m_1, m_2, \ldots, m_N$ time points respectively for one organism. These time-course datasets may be measured under various environments or stimuli by different labs. Let us first consider one time-course dataset with m time points. A linear differential equation can be used to represent the rate of synthesis of a transcript as a function of the concentrations of other transcripts in a cell and the external perturbations:

$$\frac{dx(t)}{dt} = Jx(t) + Pc(t), \qquad t = t_1, t_2, ..., t_m \tag{1}$$

where $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T \in R^n$, $x_i(t)$ is the expression level (mRNA concentrations) of gene i at time point t. $J=(J_{ij})_{n\times n}$ is an n×n connectivity matrix with elements J_{ij} representing the effect of gene j on gene i with a positive, zero, or negative sign, indicating activation, no interaction, and repression, respectively. $P=(P_{ij})_{n\times s}$ is an n×s matrix representing the effect of the s perturbations or s small molecules on x, and $c(t) \in R^s$ represents the external perturbations with s compounds at time t (In principle, the external perturbation can be of virtually any type. For example, an external environmental factor, a small molecule, an enzyme, a microRNA, or a post-translationally modified protein). A non-zero element P_{ij} of P implies that the i-th gene is a direct target of the j-th perturbation or compound. Identifying P is an important first step towards biological function discovery of small molecules and drug design.

We can rewrite Equation (1) in a compact form for all time points of one dataset by matrix notation:

$$\frac{d\mathbf{X}}{dt} = J\mathbf{X} + P\mathbf{C} \tag{2}$$

where $\mathbf{X}=(\mathbf{x}(t_1),...,\mathbf{x}(t_m))$ and $d\mathbf{X}/dt=(d\mathbf{x}(t_1)/dt,...,d\mathbf{x}(t_m)/dt)$ are n×n matrices with the first derivative of mRNA concentration $dx_i(t_j)/dt=[x_i(t_{j+1})-x_i(t_j)]/[t_{j+1}-t_j]$ for i=1,...,n; j=1,...,m. Although the forward differ-

ence approximation here is utilized for numerical computation of dx/dt, backward or other difference approximation methods can be applied similarly. Suppose that there are s external perturbation compounds, then $\mathbf{C}=(\mathbf{c}(t_1),...,\mathbf{c}(t_m))$ is an s×m matrix representing the s perturbations. The unknowns to be calculated are connectivity matrix J and P.

Equation (2) can be reformulated as:

$$\frac{d\mathbf{X}}{dt} = [J, P]\begin{bmatrix}\mathbf{X}\\ \mathbf{C}\end{bmatrix} \tag{3}$$

We then apply Singular Value Decomposition (SVD) to $[\mathbf{X}^T\mathbf{C}^T]$:

$$\begin{bmatrix}\mathbf{X}\\ \mathbf{C}\end{bmatrix}^T_{m\times(n+s)} = \mathrm{U}_{m\times(n+s)}\mathrm{S}_{(n+s)\times(n+s)}\mathrm{V}^T{}_{(n+s)\times(n+s)} \tag{4}$$

where U is a unitary m×(n+s) matrix of left eigenvectors, $\mathbf{S}$=diag(s_1, …, s_{n+s}) is a diagonal (n+s)×(n+s) matrix containing the (n+s) eigenvalues, and V^T is the transpose of a unitary (n+s)×(n+s) matrix of right eigenvectors. We can then obtain a specific solution of each dataset with the smallest L_2 norm for the Jacobian matrices J and P:

$$[\overline{J}, \overline{P}] = \frac{dX}{dt} US^{-1}V^T \tag{5}$$

where S^{-1}= diag($1/s_1$, …, $1/s_{n+s}$) and $1/s_i$ is set to be zero if s_i=0.

Similarly, we can infer N networks from N datasets respectively:

$$[\overline{J^k}, \overline{P^k}] = \frac{dX^k}{dt} U^k S^{k-1}V^{kT} \tag{6}$$

where the superscript k=1,...,N is the index of the k-th dataset. Note that without explicit normalization, J^k for each dataset is already a normalized matrix for different experiments with different time intervals due to the form of Equation (5).

Thus, the general solution of the Jacobian matrix $J^k = (J^k{}_{ij})$ and $P^k = (P^k{}_{ij})$ for each dataset k is expressed by

$$[J^k, P^k] = [\overline{J^k}, \overline{P^k}] + Y^k V^{kT} \tag{7}$$

Equation (7) represents all possible networks that are consistent with each microarray dataset, depending on arbitrary variables Y^k. $Y^k=(Y^k{}_{ij})$ is an n×(n+s) matrix, where $Y^k{}_{ij}$ is zero if $s^k{}_j \neq 0$ and is otherwise an arbitrary bounded scalar coefficient, i.e., $|Y^k{}_{ij}| \leq M$, where M is a given positive constant. In the next subsection we will explain how to construct the most consistent gene regulatory network [J P] from all $[J^k\ P^k]$ by determining Y^k, k=1,…, N.

Hard and Soft Constraints

Before formally introducing the linear programming based integration framework, we briefly categorize the prior information. In our differential equation model we use the Jacobian matrix J to represent the gene regulatory relationships. The regulatory relationships can be directed, signed, and weighted. For example, element J_{ij} represents an effect of gene j on gene i, while J_{ji} represents an effect of gene i on gene j. Thus the influence between gene i and gene j is directed. Furthermore, a sign associated with J_{ij} represents a specific role of regulation. For example, if the sign of J_{ij} is positive, gene j is the activator of gene i. On the other hand, if the sign of J_{ij} is negative, gene j is the repressor of gene i. Furthermore the associated weight (the absolute value) of element J_{ij} indicates how strong the regulatory interaction is. Obviously, a zero weight of J_{ij} indicates no interaction between two genes.

Thus, existing prior information about regulatory relationships can be roughly classified as follows:

- **Undirected.** Given a gene pair, we only know that there is a regulatory interaction between them, but the information about regulator and target gene is unavailable. For example, protein-protein interactions occur at the protein level instead of gene level, and they provide us with some hints that there exist certain relationships between two genes but no directional information. The (non-) co-expression relationships also belong to this class.
- **Directed and un-signed.** In this class, we know that there is a directed regulatory interaction but we do not know if it is an activation or repression regulation. For example, the ChIP-chip data and regulatory motif occurrence data tell us about the transcriptional regulation relationship, i.e. a transcription factor binds to the promoter region of a target gene and possibly influences its expression level, but the activating or repressing information is not available.
- **Directed and signed.** In this class, we know more about the regulation, both the regulation direction and the activation or repression role. Literature and the existing databases provide such reliable information. Also we can derive such information from the GO functional annotations. For example, activation relation can be obtained by selecting those regulatory relations such that the regulator is either an activator or co-activator in GO function annotation. Similarly the set of repression relation can be obtained by selecting those regulatory relations such that the regulator is either a repressor or a co-repressor.

In practical implementation, we can simply treat the undirected relationship as two directed and un-signed relationships (for example, the undirected relationship between A and B can be decomposed into two directed relationships: A to B and B to A). After this treatment, there are essentially two kinds of prior information: directed signed and directed unsigned. Available information from heterogeneous sources can be incorporated into our linear programming framework as soft and hard constraints, depending on the certainty of the information. The hard constraints include the directed and signed relationships. Their signs must be guaranteed and weight should be inferred. The soft constraints include the directed and un-signed relationships and their signs and weights are determined in the integration process.

Let us compare our method with traditional machine learning methods in terms of prior information incorporation. From the viewpoint of machine learning, the reliable information (gold standard positive and negative data) should be treated in a supervised way, i.e. they are labeled as positive or negative samples which are used to train the classifier. In our model, hard constraints similarly ensure that such

reliable prior information (directed and signed) is properly learned. The difference is that our method ensures that the reliable prior information must appear in the final results while gold standard data in machine learning methods are allowed to be incorrectly classified. On the other hand, the unreliable information (unlabeled data) in machine learning should be used in a semi-supervised way, i. e. they are taken as unlabeled samples which can provide useful information about sample distribution. In our model, soft constraints ensure that the useful prior information (directed and un-signed) is extracted while inaccurate information is filtered.

Next, we represent the hard constraints and soft constraints in matrix forms. Let the gold-standard directed and signed relationships be $K=(K_{ij})_{n\times n}$, which is an n×n matrix representing the known gene regulation information with signs. If the element K_{ij} is nonzero, it means that gene j has regulatory effect on gene i (activation or repression depends on the sign of K_{ij}, as determined by reliable biological experiments). The values for matrix K are set based on known information. Even though it is better to provide the quantitative strength of the known regulatory interactions in K, the vast majority of these are qualitative instead of quantitative in the databases or literature. In other words, one may know that gene i activates gene j, but the quantitative relationship is generally unavailable to depict how strong the activation is. In this case, we will decide final regulatory relationship from gene j to gene i from an LP-based algorithm by setting $K_{ij} >0$ or $K_{ij} <0$ as a hard constraint in the linear programming model. Here K serves as the gold standard positive data in the machine learning nomenclature, the difference is that we require the prior information in K to be correctly reflected in the final network structure.

There exists a second type of noisy prior regulatory information where the activation/repression role is unknown. We can represent such noisy information by soft constraints and store them into matrix $U=(U_{ij})_{n\times n}$, which is an n×n matrix representing the known gene regulation information without weights or signs. If the element U_{ij} is not zero, it means that gene j probably has regulatory effect on gene i (activation or repression is unknown and should be determined by data integration), and 0 if otherwise. We will incorporate U_{ij} into an LP-based algorithm as a soft constraint in our linear programming model by making all gene pairs for which U_{ij} is not zero free of regularization in the optimization process. If small molecule-protein interaction data are available, they can be incorporated by extending matrix U to n×(n+s) in a similar way.

In addition, we will treat the non-regulation relationship separately and store them into matrix $E=(E_{ij})_{n\times n}$, which is an n×n matrix and represents the known gene non-regulation information. If the element E_{ij} is zero, it means that gene i does not regulate gene j. Here E serves the similar role as the gold standard negative data in the machine learning meaning, the difference is prior information in E must be reflected in the final network structure. Because "gold standard" non-regulation relationships from biological experiments are often not published, negative examples need to be chosen with care. One possible selection method is to pick out the non-co-expression relationships from comprehensive expression compendium. The underlying assumption is that high quality non-regulatory relationships can be generated by considering pairs of genes whose expressions correlate weakly across various conditions. This can be further improved by combining several non-co-expression relationship detection methods together or using strict cutoffs. We will incorporate E_{ij} by using $E_{ij}=0$ as a hard constraint in our linear programming model.

In the following, we will discuss how to incorporate existing prior information into the inference of whole network by the LP-based algorithm.

Linear Programming Model for Data Integration

Assume that there are multiple microarray datasets for one organism, each of which corresponds to its own general solution in Equation (7). The next step is to find a consistent and also biologically plausible solution by determining variables Y^k, k=1,…, N. In (Y. Wang, Joshi, Zhang et al., 2006), we developed a method by exploiting L_1 norm in the formulation of the objective function to infer a sparse and consistent gene network. In this chapter, in addition to small molecule perturbations, we further consider the directed and signed regulatory relationship information K, directed and unsigned regulatory information U, and non-regulation relationships E. These new types of prior information are expected to improve the reliability of the inferred network and reduce the computational complexity.

Specifically, according to Equation (7), N networks can be separately inferred from N time-course datasets:

$$[J^k, P^k] = \frac{dX^k}{dt} U^k S^{k-1} V^{kT} = [\overline{J^k}, \overline{P^k}] + Y^k V^{kT} \tag{8}$$

where the superscript k=1,...,N is the index of the k-th dataset. Next, we will derive a sparse network structure $L=[J, P]=(L_{ij})_{n\times(n+s)}$ that is most consistent with $L^k=[J^k, P^k]= (L^k_{ij})_{n\times(n+s)}$ for k=1,...,N, as well as consistent with the directed and signed, directed and unsigned regulations, and non-regulatory relationships between genes. Mathematically the problem can be formulated as:

$$\begin{aligned} \min_{Y^1,Y^2,\ldots,Y^N,L} \quad & \sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n+s} \omega_k \mid L_{ij} - L^k_{ij} \mid + \lambda \sum_{(i,j)\in\{(i,j)\mid K_{ij}=0 \text{ or } U_{ij}=0\}} \mid L_{ij} \mid \\ s.t. \quad & L_{ij} > 0 \quad \text{if} \quad K_{ij} > 0 \qquad i,j \in \{1,2,\ldots,n\} \\ & L_{ij} < 0 \quad \text{if} \quad K_{ij} < 0 \qquad i,j \in \{1,2,\ldots,n\} \\ & L_{ij} = 0 \quad \text{if} \quad E_{ij} = 0 \qquad i,j \in \{1,2,\ldots,n\} \end{aligned} \tag{9}$$

where L^k_{ij} is a function of Y^k, and $Y=(Y^1,...,Y^N)$. The objective function has two terms. The first term is a matching term which forces the matching of L and L^k, whereas the second term is a sparseness term which forces L to be sparse as a result of the minimization of the sum of L_1 norm.λis a positive parameter, which balances the matching and sparseness terms in the objective function. Here the soft constraints are added into the objective function in an implicit way, by removing the related sparseness terms of the objective function in Equation (9). The hard constraints are added as inequality or equality constraints in an explicit way. The first and second constraints are used to add the directed and signed information, and the third one is used to incorporate the non-regulatory relationship information.

The variables in (9) are L_{ij} and all of nonzero Y^k_{ij}. ω^k is a positive weight coefficient for the k-th dataset and $\sum^N_{k=1}\omega^k =1$. Since different datasets may have different data qualities (e.g., different technologies, the number of repeats in measurements, etc.), the weight coefficient is used to represent the reliability of each dataset. The optimization problem (9) is an LP with L_1 norm, which is a well-studied problem. It is known that L_1 gives a more robust answer compared with L_2. The L_1-norm is more robust to outliers than the L_2-norm and does not penalize large deviations as much as the L_2-norm. As a result, the L_1-norm pays less attention to the parts of the regulatory interactions that are very different, and focuses more on the parts of the regulatory interactions that are conserved. As a result, this measure

is less sensitive toward noise and more robust towards outliers. Generally the optimal solution of (9) sets as many $|L_{ij}- L^k_{ij}|$ and $|L_{ij}|$ to zero as possible, thus ensuring a consistent and sparse structure for the inferred gene regulatory network.

As discussed previously, most documented regulation information is qualitative rather than quantitative. Therefore, we add the first and second inequality constraints of Equation (9) as hard constraints according to its activation or repression role stored in matrix K, and the strength of regulation is decided from the optimization algorithm. For example, add $L_{ij} <0$ if a repression relationship is known as $K_{ij} <0$ and derive the value of L_{ij} from the optimization process. In addition, the corresponding gene pair is removed from the second term (regularization term) in the objective function. We can also add the equality constraints $L_{ij}=0$ to Equation (9) to take into account the non-regulatory data. In addition, we also encode prior information in U as soft constraints in the following way. Specifically, for a gene pair where U_{ij} is non-zero (meaning that there probably exists regulatory relationship between the gene pair), we implement a soft constraint by removing the corresponding element of L_{ij} from the second term of objective function so that it is not subject to regularization in the optimization process. In this way, these regulatory interactions may be present in the optimal solution with signs and weights learned from the optimization process. In cases where the optimization process assigns zero weight to a gene pair, we assume that the prior information is probably noisy and is therefore ignored by the algorithm. The final result depends on the consistency of this information with microarray datasets or other prior information. It is reasonable since this prior information may or may not be correct, and therefore should be further filtered.

In Equation (9), each one of the matrices L, Y^1, Y^2, …, Y^N has almost n^2 variables. Thus the total number of variables is about n^3. For a gene regulatory network with 100 genes, even without prior information and other variables such as drug targets, the LP problem has 1,000,000 variables. To solve Equation (9) efficiently, a decomposition algorithm is used based on the special structure of Equation (9). This is done by iteratively solving the following two subproblems. We first fix L to solve N small-sized matching subproblems LP^1, LP^2, …, LP^N, followed by updating L by solving Equation (9) with fixed Y^1, Y^2, …, Y^N from the N subproblems. The procedure is repeated until convergence. The two decomposed subproblems are described in detail as follows.

- Subproblem-1: Set $L^k(q)=L^k(q-1)+ Y^k(q)V^{kT}$. At iteration q, obtain $Y^k_{ij}(q)$ by solving subproblems LP^1, LP^2, …, LP^N,

$$\min_{Y^1,Y^2,\ldots,Y^N} \quad \sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n+s}\omega_k \mid L_{ij}(q-1) - L^k_{ij}(q) \mid + \lambda \sum_{(i,j)\in\{(i,j)|K_{ij}=0 \text{ or } U_{ij}=0\}} \mid L_{ij}(q-1) \mid \qquad (10)$$

where $L_{ij}(q-1)$ is fixed.

- Subproblem-2: At iteration q, obtain $L_{ij}(q)$ by solving the following LP with all of $Y^k_{ij}(q)$ and $L^k(q)$ fixed from Subproblem-1,

$$\begin{aligned}
\min_{L} \quad & \sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n+s}\omega_k \mid L_{ij}(q) - L_{ij}^{k}(q) \mid + \lambda \sum_{(i,j)\in\{(i,j)\mid K_{ij}=0 \text{ or } U_{ij}=0\}} \mid L_{ij}(q) \mid \\
s.t. \quad & L_{ij}(q) > 0 \quad \text{if} \quad K_{ij} > 0 \qquad i,j \in \{1,2,...,n\} \\
& L_{ij}(q) < 0 \quad \text{if} \quad K_{ij} < 0 \qquad i,j \in \{1,2,...,n\} \\
& L_{ij}(q) = 0 \quad \text{if} \quad E_{ij} = 0 \qquad i,j \in \{1,2,...,n\}
\end{aligned} \tag{11}$$

Although the solution depends on λ, λ is the only parameter that needs to be tuned. The procedures of solving Equations (10) and (11) and the choice of parameter λ are similar to (Y. Wang, Joshi, Zhang et al., 2006).

Compared with un-constrained LP model in (Y. Wang, Joshi, Zhang et al., 2006), the constraints in the above LP provide a consistent way to integrate all kinds of prior information. Specifically we incorporate reliable signed, noisy unsigned, and non-regulatory data in a systematic way. Given the nature of the gene regulatory network inference problem is under-determined (In other words, the number of variables far exceeds the equations for which variables are related), the proper incorporation of the prior information from heterogeneous data sources improves the reconstruction accuracy.

It should be noted that the above methodology has three advantages in terms of both model and algorithm. Firstly, the variables L_{ij} in Equation (9) include not only the connectivity matrix of genes which represents the effect of activation, no interaction and repression, but also the connectivity matrix of perturbations which represents the effect of the small-molecule perturbations on genes. This is very important because our method is able to properly identify the target genes of perturbations and thus has the potential to be applied to the drug design and mechanism of action discovery of molecules. Secondly, the objective function has both sparse and non-sparse terms. The non-sparse term is used to represent the interactions or effects among genes or between external inputs and genes based on the noisy prior information or experimental data. In this way, the soft constraints are considered and added in a consistent manner. Thirdly, the new model can improve reconstruction accuracy by introducing hard constraints on "gold-standard" prior information.

From the algorithmic and computational efficiency aspect, Equation (9) is a constrained L_1 linear approximate problem, in contrast to the linear regression model of (Y. Wang, Joshi, Zhang et al., 2006). For the first subproblem, an efficient primal algorithm can be designed by taking advantage of the special structure of the linear programming formulation of the L_1 problem; for the second subproblem, it can be decomposed as a series of constrained and unconstrained small-scale linear programming (Y. Wang, Joshi, Zhang et al., 2006) and the problem can be solved efficiently.

The data integration strategy in this chapter is different from the supervised inference methods (T. Kato, Tsuda, & Asai, 2005) which adopt the kernel matrix representation of networks and integrate different biological data in a simple weighted sum. In this chapter the gene regulations are derived from time-course data by differential equations instead of similarity evaluation in kernel matrix. Specifically, the "gold standard" prior information is expressed as hard constraints or soft constraints (i.e., sparse term in L_{ij} of the objective function) in the LP formulation, depending on the certainty or reliability of the information. Thus, we can obtain the most consistent solution among multiple datasets by satisfying those constraints. In particular, in our prior information learning framework, our method ensures all of the hard constraints to hold, and prefers the soft constraints to hold, but the regulatory interactions corresponding to soft constraints may or may not hold depending on their consistency with other data. Therefore, if prior information is not reflected in the optimal solution, it is because this prior informa-

tion is inconsistent with the microarray datasets and other information. As such, the proposed algorithm can also filter out the noise in prior information based on the requirement of consistency among all data sources.

RESULTS

In this section, we first report a simulated numerical example to validate our method. Then we apply our method to a real experimental data to reconstruct yeast gene regulatory network. We show that our method is effective in recovering the network connectivity from integrated data sources. Importantly, with supervised information, our method can infer the network structure and further identify the compound targets in a more accurate and reliable manner.

Simulated Example

The first example is a small simulated network to demonstrate the usefulness of data integration and prior information in the network inference method. We constructed a small regulatory network with six genes governed by:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \\ \dot{x}_4(t) \\ \dot{x}_5(t) \\ \dot{x}_6(t) \end{bmatrix} = \begin{bmatrix} -1.0 & 0.0 & 0.01 & 0.0 & 0.03 & 0.03 \\ 0.2 & -1.2 & 0.0 & 0.4 & -0.05 & 0.0 \\ 0.0 & 0.0 & -1.0 & 0.0 & 0.0 & -0.05 \\ 0.0 & -0.05 & 0.0 & -1.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2 & 0.0 & -1.2 & 0.0 \\ 0.0 & 0.03 & 0.0 & -0.01 & 0.0 & -1.0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \\ x_5(t) \\ x_6(t) \end{bmatrix} + \begin{bmatrix} 2.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{bmatrix} u_0 \tag{12}$$

where x_i reflects the expression level of the gene-i for i=1,...,6. One perturbation (s=1) is applied to the first gene, which is indicated by $P=[2.0,0.0,0.0,0.0,0.0,0.0]^T$. P has all its elements equal to 0 except the element for the gene that is the direct target of the perturbation. u_0 contains the detailed information about the perturbation, which can be either time-independent or time-dependent.

We generate four time-course datasets in different conditions. Every dataset has five time points and the time points are equally-spaced from the start to the end. These datasets differ in the choice of perturbation and time step. The first dataset is obtained by taking perturbation $u_0=1$ as a constant and the time step is 0.1. The second dataset is also obtained by taking $u_0=1$ as a constant but the time step is 0.15. For the third dataset, perturbation varies with time and gradually increase from $u_0=1$ to $u_0=2$, and the time step is 0.2. The fourth dataset is obtained without perturbation and with time step 0.2. The initial values of the system are randomly generated from [1.0, 1.1] and the Gaussian noise is added to the data matrix with zero mean and fixed standard deviation $\sigma=0.2||X||$, where $||X||$ is the L_∞ norm of the data matrix X. In the following, we will show that the datasets can be combined together to infer the gene regulatory network by our method. The parameter λ is set to 0.1 to make the inferred network sparse. The supervised information K is denoted by the following matrix,

$$K = \begin{bmatrix} -1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.03 \\ 0.0 & 0.0 & 0.0 & 0.4 & -0.05 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 \\ 0.0 & -0.05 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} \tag{13}$$

where nonzero elements will be added in the LP as constraints.

This simulated example illustrates the three advantages of our method for reconstructing gene regulatory networks. Firstly, our method can identify more correct regulatory relationships among genes. The numerical results are depicted in Figure 2, which shows the true network and the reconstructed networks without and with supervised information, respectively. In the case without supervised information, 4 edges are identified correctly out of 7 predicted nonzero edges. In contrast, in the case with supervised information, 11 edges are identified correctly out of 16 predicted nonzero edges. Thus the prediction accuracy is improved from 57.14% to 68.75%.

Secondly, the inferred network by our method is quantitatively more accurate. We use the following indices E_1 and E_2 to assess the prediction accuracy:

$$E_1 := \sum_{i=1}^{n} \sum_{j=1}^{n} \left| J_{ij}^{T} - J_{ij}^{R} \right| \tag{14}$$

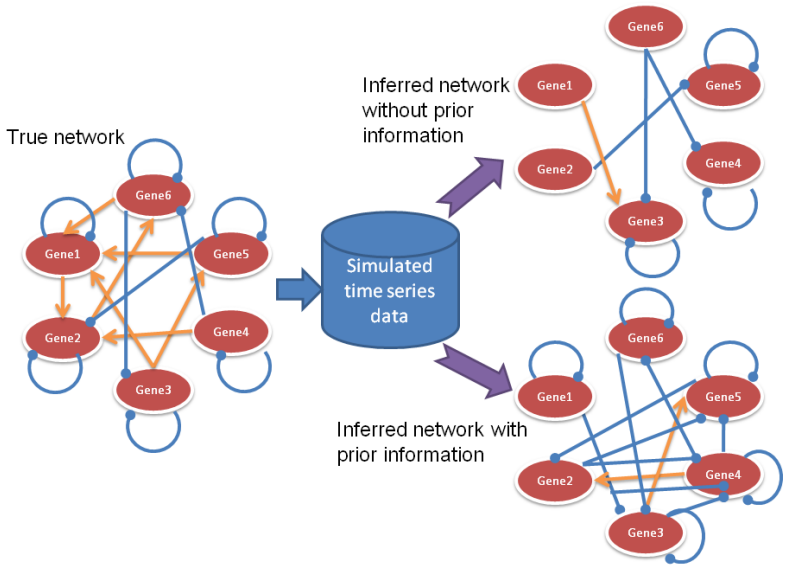
$$E_2 := \sum_{i=1}^{n} \sum_{j=1}^{n} \left(J_{ij}^{T} - J_{ij}^{R} \right)^2 \tag{15}$$

where J_{ij}^{T} and J_{ij}^{R} are interaction strength from gene-j to gene-i for the true and inferred networks, respectively. We found that adding supervised information reduces the inference error. For example, E_1 decreases by 0.7139 (excluding the error reduction 1.73 due to the knowledge of K) and E_2 decreases by 1.4857 (excluding the error reduction 1.20 due to the knowledge of K).

Thirdly, our method makes more accurate predictions about the targets genes of perturbation. This is very important as our method has the potential to be applied to the drug design and function discovery of molecules. According to the computational results, the inferred perturbation vector is P = [0.35, -0.08, 0.09, 0.0, 0.0, 0.0] without supervised information, whereas the prediction results are improved to P = [1.25, 0.0, -0.01, 0.0, 0.0, -0.03] with supervised information. These results show that our method can correctly identify the first gene to be the direct target of the applied perturbation, and the knowledge of the supervised information can helps reduce inference error.

There are two reasons for the accurate inference by our method. The first reason is the contribution of multiple datasets. By combining the time-course datasets of different types and in different perturbation conditions, more information are utilized and the problem of high dimensionality is significantly alleviated (refer to (Y. Wang, Joshi, Zhang et al., 2006) for details). The second reason is the contribution of prior information. Due to the scarcity of gene expression data and the high-dimensionality of the gene network parameter space, the problem of gene network inference is fundamentally under-determined. The supervised information help reduce the intrinsic dimensionality of the search space dramatically, thus making the inferred network more accurate both qualitatively and quantitatively.

Figure 2. Regulatory network reconstruction for the simulated example with 6 genes. Red arrows represent activation, and blue arcs represent repression.



Combining ChIP-Chip and Expression Data to Infer Gene Regulatory Network in Yeast

We combined gene expression data with ChIP-chip data to infer gene regulatory network in yeast. As mentioned above, the ChIP-chip methodology involves the chromatin immunoprecipitation of an epitope-tagged TF bound to DNA fragments containing target promoters, followed by the hybridization of those amplified DNA fragments to an intergenic microarray (Lee et al., 2002). ChIP-chip data have the advantage that they provide a direct biochemical link between TFs and promoters and have the potential to identify targets without knowing the activating conditions. From this viewpoint, ChIP-chip data are an important source of information for direct transcriptional regulatory interactions. In this example we show the network reconstruction accuracy can be improved by incorporating TF DNA binding data (ChIP-chip data) into our model as prior information (soft constraints). We tested our method using the public time-series microarray data for cell cycle studies in *Saccharomyces cerevisiae* which are obtained from the Stanford Microarray Database (Demeter et al., 2007). We collected 4 datasets with different conditions (Response to Elutriation, 14 time points; Response to CDC15, 24 time points; Response to alpha factor fkh1, fkh2, 13 time points; Response to fkh1, fkh2,13 time points). Among all the yeast genes, 145 of them have changes of 2 fold up or down in at least 20% of the expression level across all datasets.

We added the TF-DNA binding data as prior information to infer the gene regulatory network in a more reliable manner. In (Lee et al., 2002), a genome-wide location analysis experiment was performed for 106 yeast TFs. From their supporting website (http://jura.wi.mit.edu/young_public/regulatory_network/), we downloaded the TF-target gene interactions and TF-TF interactions as prior information. As a summary, there are 75 TFs (in the list of 106 TFs of (Lee et al., 2002)) in the 145 gene list, and there

are a total of 161 known interactions including 93 known TF-TF interactions and 68 TF-target gene interactions. Furthermore, we manually selected 22 interactions with known activating or repressing conditions by checking the GO database.

Then we apply our method to these real experimental data in yeast. There are two kinds of prior information. One is the 22 interactions with known activating or repressing conditions which can be directly added as the hard constraints in the LP model. The remaining 139 known interactions without activating or repressing information are taken as soft constraints for which we simply remove their corresponding sparse terms from the objective function of the LP model.

When $\lambda=0$, we obtained 622 interactions. All the 161 known interactions are correctly inferred. Among them, 22 interactions with known activating or repressing conditions are correctly inferred and the activating or repressing conditions of the remaining 139 interactions are predicted by our method. Among the newly inferred 461 edges, validation results by YEASTRACT database (Teixeira et al., 2006) show that there are 11 documented interactions and there are 66 edges identified as potential interactions, for which transcription factors have at least 1 binding site in the promoter regions of their target genes.

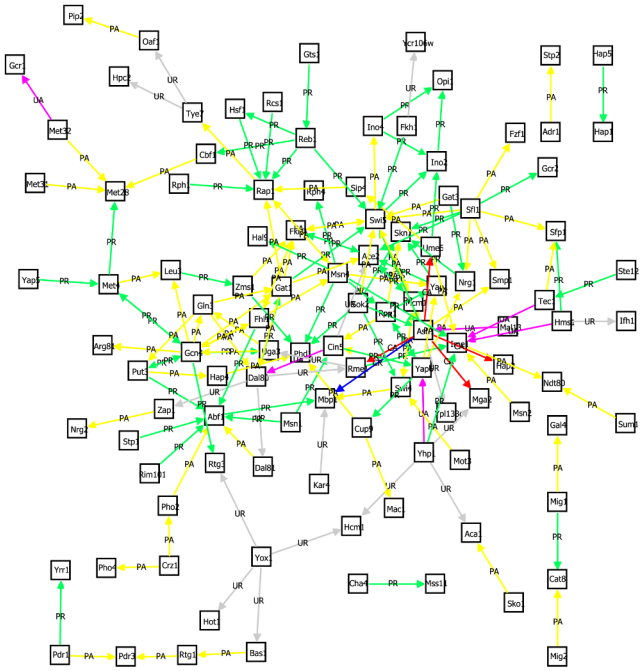
When we set the parameter $\lambda=0.1$ to make the inferred network sparse, a total of 219 interactions were inferred with predicted activating or repressing conditions. Again all the 161 known interactions (Among them, 22 interactions have known activating or repressing conditions) are correctly inferred. The validation results by YEASTRACT show that in the newly inferred 58 edges, 5 are documented and 11 are identified as potential interactions. In Figure 3, we draw the reconstructed gene regulatory network when $\lambda=0.1$ (self regulatory interactions are not shown).

In this experiment, we combine the time-series microarray data and genomic location data to infer whether a regulator acts as an activator or repressor. Generally, we can use genomic location data to infer the presence of regulators at promoters, but we cannot determine the type of TF-target gene interactions. By further combining gene expression data, our method can infer not only the existence of regulatory interactions between TFs and target genes, but also the sign of the regulation (positive or negative). From this example, we can also see that computational method is complementary to experimental methods, e.g., it provides information whether a TF is an activator or repressor by its regulatory role based on the dynamic behavior of the gene expression.

CONCLUSION

Our proposed data integration and network reconstruction method in this chapter not only improves the reliability of the inferred gene regulatory network, but also can be applied to drug design and many other areas of biomedical research and bioengineering. Specifically, we propose to combine computational analysis of multiple microarray datasets and other types of biological experiments together for inferring gene regulatory network and further identifying small molecule targets of perturbation experiments. The proposed algorithm is mainly based on linear programming framework with the variables representing the regulatory relationships among genes and small molecule-protein interactions. Available information from heterogeneous data sources is incorporated into the LP as constraints. They can be divided into two classes according to data reliability. For example, the regulatory relationships mined from literature and databases are more reliable and we know exactly the regulator gene, target gene, and the activation or repression role. Hence, they can be treated as hard constraints in our linear programming which will be strictly reinforced. On the other hand, the co-expression data, ChIP-chip data and protein-protein

Figure 3. Inferred gene regulatory network when λ=0.1 by combining expression data with TF-DNA binding data. Self regulatory interactions are not shown. Known activations are shown in yellow with label 'PA'. Known repressions are shown in green with label 'PR'. Newly inferred activations that are subsequently confirmed are shown in red with label 'CA'. Newly inferred repressions that are subsequently confirmed are shown in blue with label 'CR'. Newly inferred activations and repressions that are yet to be confirmed are shown in pink with label 'UA' and gray with label 'UR', respectively.



interaction data are generally noisy and the related pairs often are unsigned, meaning that the activation or repression role is unknown. In this case, we treat them as soft constraints, which may or may not be satisfied. In this way, our linear programming model provides a flexible prior information learning framework. It finds the most consistent gene regulatory network by balancing among heterogeneous data sources. One major advantage of the proposed method is that it theoretically ensures the derivation of the most consistent network with respect to the available datasets or information, thereby alleviating the problem of data scarcity and improving the reliability. In addition, this algorithm allows us to infer small molecule targets by integrating perturbation experiments, and holds the promise for applications in drug design and other areas in biomedical engineering.

FUTURE RESEARCH DIRECTIONS

With rapid advances of various high-throughput experimental techniques, more and more biological data are increasingly available. Thus it is now possible to quantitatively study regulation interactions in a systematic way. Generally speaking, there are three kinds of regulatory relationships among the

regulators (transcriptional factors and cofactors) and target genes. They are the relationships between target genes, the relationships between regulators, and the relationships between regulators and target genes. Network reconstruction aims to reveal regulatory mechanisms by inferring these relationships from biological data.

The mapping of the gene regulatory network—the set of interactions among all genes in the genome—is one of the most difficult tasks in molecular biology. For example, there are 6000 genes in yeast, and as a result there are at least 18 millions parameters to be determined in our linear differential model. In contrast, the mapping of the transcriptional regulatory network — the set of all physical interactions among transcriptional factors and their target genes — has much less parameters from the computational viewpoint. There are about 200 transcriptional factors in yeast and 6000 target genes and thus there are about 1.2 million parameters to be determined. Compared to the above two tasks, the reconstruction of the transcriptional factor interaction network is perhaps the easiest. Since there are about 200 transcriptional factors in yeast, we only need to determine about 20,000 parameters. It is well known that the transcription factor sub-proteome is very important for gene regulation and especially difficult for experimental characterization. Hence, the computational methodology to predict these regulatory subnetworks among TFs is crucial. In the case of the transcription factor interactome, transcriptional regulation in eukaryotes occurs through the coordinated action of multiple transcription factors. So combinatorial regulation is a primary mechanism for achieving fine-tuned transcriptional control, is an important component of the mechanisms of action for many biologically active small molecules, and holds the promise to reveal the complexity of gene regulation mechanisms (Balaji, Babu, Iyer, Luscombe, & Aravind, 2006; Bluthgen, Kielbasa, & Herzel, 2005; Chang, Wang, & Chen, 2006; M. Kato, Hata, Banerjee, Futcher, & Zhang, 2006).

The linear differential equation model in this chapter makes the important assumption that the structure of the regulatory network is stationary, and does not `rewire' under the environmental conditions for those different datasets. This means that the change of environmental conditions is assumed to alter the level of gene expression instead of the network structure. Obviously this is not true in reality. One of the future research directions is to reverse engineer the network architecture from time-series microarray data based on a nonlinear differential equation model, which will capture the complex and nonlinear properties in gene regulatory process but will involve more parameters.

Data integration is still a very challenging problem. A complex network reconstruction methodology needs high resolution datasets so as to accurately infer the network structure. Here high-resolution data mean high-quality time-course microarray data which are expected to capture the dynamic behavior of the gene regulatory networks and also the conditional responsive transcription factor-DNA and protein-protein interaction data. As a result, sophisticated data integration techniques play a key role.

Recently Faith et al. assembled 445 Escherichia coli microarrays to address this issue and demonstrated an unsupervised network inference method, called context likelihood of relatedness (CLR), which uses transcriptional profiles of an organism across a diverse set of conditions to systematically determine transcriptional regulatory interactions (Faith et al., 2007). By generating a compendium of microarrays, they showed that it is possible to infer a high-precision regulatory map and simultaneously obtain rich data on condition-specific regulation. The strategy here is simple: it assembles all the microarray datasets to a profile or compendium and applies algorithms based on correlation coefficients or mutual information measure, such as Relevance network (Butte et al., 2000), ARACNe (Margolin et al., 2006), and CLR (Faith et al., 2007) to find the co-expression or co-regulation relationships. This strategy can be potentially enhanced in several ways. First, the existing methods treat a set of time-course data

points independently and ignore the dynamic property of gene regulation process. Second, the existing methods can only identify whether two genes have regulatory relationships, but cannot provide the detailed information about regulatory roles such as activation or repression. The existing methods can be improved by considering as much dynamic information as possibly when integrating time-course microarray datasets.

We believe that data integration and network reconstruction should be conducted in a simultaneous way. We should determine the data integration parameters and the network structure parameters together. In this way, the solution will be expected to be globally optimal and the most consistent. In this chapter we provided such a model. It can be further extended to a more general model to assign weights to different sources of data. In the future, we will extend the current work of revealing the complex mechanisms of transcriptional control in two ways. First, regulatory network reconstruction can be greatly improved by the integration of more diverse genomic datasets such as sequence, protein structure, gene expression, TF-DNA interaction, non-coding RNA-mRNA interaction, protein-protein interaction, and metabolic reaction data. Second, transcriptional regulatory processes can be more accurately modeled by taking into account cooperativity among individual proteins, nonlinearity, and dynamic behaviors.

ACKNOWLEDGMENT

YW, RSW, XSZ, and LC are supported by JSPS and NSFC under JSPS-NSFC collaboration project. YW is also supported by National Natural Science Foundation of China under Grant No.10701080 and No. 10801131.YX is supported by a Research Starter Grant in Informatics from the PhRMA Foundation.

REFERENCES

Alaoui-Ismaili, M. H., Lomedico, P. T., & Jindal, S. (2002). Chemical genomics: discovery of disease genes and drugs. *Drug Discovery Today*, *7*(5), 292–294. doi:10.1016/S1359-6446(02)02185-2

Balaji, S., Babu, M. M., Iyer, L. M., Luscombe, N. M., & Aravind, L. (2006). Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *Journal of Molecular Biology*, *360*(1), 213–227. doi:10.1016/j.jmb.2006.04.029

Benson, M. L., Smith, R. D., Khazanov, N. A., Dimcheff, B., Beaver, J., & Dresslar, P. (2007). Binding MOAD, a high-quality protein ligand database. *Nucleic Acids Research*, *36*(Database issue), D674–D678. doi:10.1093/nar/gkm911

Bluthgen, N., Kielbasa, S. M., & Herzel, H. (2005). Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Research*, *33*(1), 272–279. doi:10.1093/nar/gki167

Brazma, A., Jonassen, I., Vilo, J., & Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Research*, *8*(11), 1202.

Breitkreutz, B. J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., & Livstone, M. (2008). The BioGRID interaction database: 2008 update. *Nucleic Acids Research*, *36*(Database issue), D637. doi:10.1093/nar/gkm1001

Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., & Kohane, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 12182–12186. doi:10.1073/pnas.220392197

Chang, Y.-H., Wang, Y.-C., & Chen, B.-S. (2006). Identification of transcription factor cooperativity via stochastic system model. *Bioinformatics (Oxford, England)*, *22*(18), 2276–2282. doi:10.1093/bioinformatics/btl380

D'Haeseleer, P., Liang, S., & Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics (Oxford, England)*, *16*(8), 707–726. doi:10.1093/bioinformatics/16.8.707

De Hoon, M. J. L., Imoto, S., Kobayashi, K., Ogasawara, N., & Miyano, S. (2003). Inferring gene regulatory network from time-ordered gene expression data of bacillus subtilis using differential equations. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 17–28.

Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., & Maier, D. (2007). The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Research*, *35*(Suppl 1), D766–D770. doi:10.1093/nar/gkl1019

Dewey, T. G., & Galas, D. J. (2001). Dynamic models of gene expression and classification. *Functional & Integrative Genomics*, *1*(4), 269–278. doi:10.1007/s101420000035

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., & Cottarel, G. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, *5*(1), e8. doi:10.1371/journal.pbio.0050008

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, *303*(5659), 799–805. doi:10.1126/science.1094068

Gardner, T. S., di Bernardo, D., Lorenz, D., & Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, *301*(5629), 102–105. doi:10.1126/science.1081900

Gardner, T. S., & Faith, J. J. (2005). Reverse-engineering transcription control networks. *Physics of Life Reviews*, *2*(1), 65–88. doi:10.1016/j.plrev.2005.01.001

Gustafsson, M., Hornquist, M., & Lombardi, A. (2005). Constructing and analyzing a large-scale gene-to-gene regulatory network-Lasso-Constrained inference and biological validation. [TCBB]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *2*(3), 254–261. doi:10.1109/TCBB.2005.35

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., & Danford, T. W. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, *431*, 99–104. doi:10.1038/nature02800

Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V., & Banavar, J. R. (2001). Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(4), 1693. doi:10.1073/pnas.98.4.1693

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics (Oxford, England)*, *19*(17), 2271–2282. doi:10.1093/bioinformatics/btg313

Hwang, D., Rust, A. G., Ramsey, S., Smith, J. J., Leslie, D. M., & Weston, A. D. (2005). A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(48), 17296–17301. doi:10.1073/pnas.0508647102

Joyce, A. R., & Palsson, B. O. (2006). The model organism as a system: integrating Omics data sets. *Nature Reviews. Molecular Cell Biology*, *7*(3), 198–210. doi:10.1038/nrm1857

Kato, M., Hata, N., Banerjee, N., Futcher, B., & Zhang, M. Q. (2006). Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biology*, *5*, R56. doi:10.1186/gb-2004-5-8-r56

Kato, T., Tsuda, K., & Asai, K. (2005). Selective integration of multiple biological data for supervised network inference. *Bioinformatics (Oxford, England)*, *21*(10), 2488–2495. doi:10.1093/bioinformatics/bti339

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., & Bork, P. (2008). STITCH: Interaction networks of chemicals and proteins. *Nucleic Acids Research*, *36*(Database issue), D684. doi:10.1093/nar/gkm795

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., & Gerber, G. K. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, *298*(5594), 799–804. doi:10.1126/science.1075090

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: A Web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, *35*(Suppl 1), D198–D201. doi:10.1093/nar/gkl999

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., & Dalla Favera, R. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7*(Suppl 1), S7. doi:10.1186/1471-2105-7-S1-S7

Nachman, I., Regev, A., & Friedman, N. (2004). Inferring quantitative models of regulatory networks from expression data. *Bioinformatics (Oxford, England)*, *20*(Suppl 1), i248–i256. doi:10.1093/bioinformatics/bth941

Nariai, N., Tamada, Y., Imoto, S., & Miyano, S. (2005). Estimating gene regulatory networks and protein-protein interactions of Saccharomyces cerevisiae from multiple genome-wide data. *Bioinformatics (Oxford, England)*, *21*(Suppl 2), ii206–ii212. doi:10.1093/bioinformatics/bti1133

Tegner, J., Yeung, M. K., Hasty, J., & Collins, J. J. (2003). Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(10), 5944. doi:10.1073/pnas.0933416100

Teixeira, M. C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A. R., & Mira, N. P. (2006). The YEASTRACT database: A tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Research*, *34*(Suppl 1), D446–D451. doi:10.1093/nar/gkj013

Wang, R. S., Wang, Y., Zhang, X. S., & Chen, L. (2007). Inferring transcriptional regulatory networks from high-throughput data. *Bioinformatics (Oxford, England)*, *23*(22), 3056–3064. doi:10.1093/bioinformatics/btm465

Wang, Y., Joshi, T., Xu, D., Zhang, X., & Chen, L. (2006). *Supervised inference of gene regulatory networks by linear programming*. (. *LNCS*, *4115*, 551.

Wang, Y., Joshi, T., Zhang, X. S., Xu, D., & Chen, L. (2006). Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics (Oxford, England)*, *22*(19), 2413. doi:10.1093/bioinformatics/btl396

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., & Tzur, D. (2007). DrugBank: A knowledgebase for drugs, drug actions, and drug targets. *Nucleic Acids Research*, *36*, D901–D906. doi:10.1093/nar/gkm958

Yeung, M. K., Tegner, J., & Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(9), 6163. doi:10.1073/pnas.092576199

ADDITIONAL READING

Alon, U. (2007). *An introduction to systems biology: Design principles of biological circuits*. Chapman & Hall/CRC.

Alvarez-Buylla, E. R., Beniez, M., & Daila, E. B., Chaos, Espinosa-Soto, C., & Padilla-Longoria, P. (2007). Gene regulatory network models for plant development. *Current Opinion in Plant Biology*, *10*(1), 83–91. doi:10.1016/j.pbi.2006.11.008

Bansal, M., Gatta, G. D., & di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics (Oxford, England)*, *22*(7), 815–822. doi:10.1093/bioinformatics/btl003

Bonneau, R., Facciotti, M. T., Reiss, D. J., Schmid, A. K., Pan, M., & Kaur, A. (2007). A predictive model for transcriptional control of physiology in a free living cell. *Cell*, *131*(7), 1354–1365. doi:10.1016/j.cell.2007.10.053

Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., & Baliga, N. S. (2006). The inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology*, *7*, R36. doi:10.1186/gb-2006-7-5-r36

Bornholdt, S. (2005). Systems biology: Less is more in modeling large genetic networks. *Science*, *310*(5747), 449–451. doi:10.1126/science.1119959

di Bernardo, D., Thompson, M. J., Gardner, T. S., Chobot, S. E., Eastwood, E. L., & Wojtovich, A. P. (2005). Chemogenomic profiling on a genomewide scale using reverse-engineered gene networks. *Nature Biotechnology*, *23*, 377–383. doi:10.1038/nbt1075

Driscoll, M. E., & Gardner, T. S. (2006). Identification and control of gene networks in living organisms via supervised and unsupervised learning. *Journal of Process Control*, *16*(3), 303–311. doi:10.1016/j.jprocont.2005.06.010

Ernst, J., Vainas, O., Harbison, C. T., Simon, I., & Bar-Joseph, Z. (2007). Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, *3*(74), 1–13.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, *303*(5659), 799–805. doi:10.1126/science.1094068

Gustafsson, M., Hornquist, M., & Lombardi, A. (2005). Constructing and analyzing a large-scale gene-to-gene regulatory network-lasso-constrained inference and biological validation. [TCBB]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *2*(3), 254–261. doi:10.1109/TCBB.2005.35

Hartemink, A. J. (2005). Reverse engineering gene regulatory networks. *Nature Biotechnology*, *23*, 554–555. doi:10.1038/nbt0505-554

Hayete, B., Gardner, T. S., & Collins, J. J. (2007). Size matters: Network inference tackles the genome scale. *Molecular Systems Biology*, *3*, 77. doi:10.1038/msb4100118

Markowetz, F., & Spang, R. (2007). Inferring cellular networks: A review. *BMC Bioinformatics*, *8*(Suppl 6), S5. doi:10.1186/1471-2105-8-S6-S5

Miyano, S. (2003). Use of gene networks for identifying and validating drug targets. *Journal of Bioinformatics and Computational Biology*, *1*(3), 459–474. doi:10.1142/S0219720003000290

Palsson, B. O. (2006). *Systems biology: Properties of reconstructed networks*. New York: Cambridge University Press.

Pan, Y., Durfee, T., Bockhorst, J., & Craven, M. (2007). Connecting quantitative regulatory-network models to the genome. *Bioinformatics (Oxford, England)*, *23*(13), i367. doi:10.1093/bioinformatics/btm228

Quach, M., Brunel, N., & d'Alche-Buc, F. (2007). Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference. *Bioinformatics (Oxford, England)*, *23*(23), 3209. doi:10.1093/bioinformatics/btm510

Reiss, D., Baliga, N., & Bonneau, R. (2006). Integrated biclustering of heterogeneous genomewide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, *7*(1), 280. doi:10.1186/1471-2105-7-280

Rosenfeld, S. (2007). Stochastic cooperativity in non-linear dynamics of genetic regulatory networks. *Mathematical Biosciences*, *210*(1), 121–142. doi:10.1016/j.mbs.2007.05.006

Schlitt, T., & Brazma, A. (2007). Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, *8*(Suppl 6), S9. doi:10.1186/1471-2105-8-S6-S9

Soranzo, N., Bianconi, G., & Altafini, C. (2007). Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: Synthetic vs. real data. *Bioinformatics (Oxford, England)*, *23*(13), 1640. doi:10.1093/bioinformatics/btm163

Sperling, S. (2007). Transcriptional regulation at a glance. *BMC Bioinformatics*, *8*(Suppl 6), S2. doi:10.1186/1471-2105-8-S6-S2

Steinke, F., Seeger, M., & Tsuda, K. (2007). Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models. *BMC Systems Biology*, *1*(51).

Sun, L., Jiang, L., Li, M., & He, D. (2006). Statistical analysis of gene regulatory networks reconstructed from gene expression data of lung cancer. *Physica A: Statistical Mechanics and its Applications, 370*(2), 663-671.

Wang, Y., Zhang, X.-S., & Chen, L. (2009). A network biology study on circadian rhythm by integrating various omics data. *OMICS: A Journal of Integrative Biology*, *13*(4), 313–324. doi:10.1089/omi.2009.0040

Zhang, H., Pu, J., & Zhang, J. (2006). Construction of gene regulatory networks based on gene ontology and multivariable regression. *Proceedings of the 2006 IEEE International Conference on Mechatronics and Automation* (pp. 1324-1328).

Zhao, W., Serpedin, E., & Dougherty, E. R. (2006). Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics (Oxford, England)*, *22*(17), 2129. doi:10.1093/bioinformatics/btl364

Zou, M., & Conzen, S. D. (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics (Oxford, England)*, *21*(1), 71–79. doi:10.1093/bioinformatics/bth463