The Pennsylvania State University
The Graduate School
Eberly College of Science

INFERENCE OF GENE REGULATORY NETWORK BASED ON GENE EXPRESSION DYNAMICS IN RESPONSE TO ENVIRONMENTAL SIGNALS

A Dissertation in
Statistics
by
Yaqun Wang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2015

The dissertation of Yaqun Wang was reviewed and approved* by the following:

Rongling Wu

Distinguished Professor of Public Health Sciences and Statistics

Dissertation Co-Advisor, Co-Chair of Committee

Runze Li

Verne M. Willaman Professor of Statistics

Dissertation Co-Advisor, Co-Chair of Committee

Zhibiao Zhao

Associate Professor of Statistics

Donna Coffman

Research Associate Professor of Health and Human Development

Aleksandra B. Slavkovic

Associate Professor of Statistics

Associate Head for Graduate Studies of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

Thousands of genes are encoded on the genome and their products play important roles to cell survival, phenotypic characteristics of organisms and adaptive behaviors of organisms when environment changes. Detecting of particular sets of genes whose expressions are adaptive in response to environmental signals and identification of dynamic gene regulatory networks (GRN) can help us to understand the mechanistic base of gene-environment interactions and gene-gene interactions in a systematic way. However, it is a challenging work to analyze gene expression across two-dimensional spaces, time and environmental state.

In this dissertation, we develop a functional clustering framework based on a mixture model to analyze time-course gene expression. The mathematical aspects of gene expression dynamics have been captured by Legendre polynomial and the impact of environment on gene expression has been considered jointly. We outline a number of quantitative testable hypotheses about the patterns of dynamic gene expression in changing environments and gene-environment interactions causing developmental differentiation. The method is illustrated with simulation studies and application on a real data set from a rabbit hemodynamic study.

In addition, we propose two models for inference of GRN based on gene expression. We reform the Dynamic Bayesian Network (DBN) model for identification of GRN to overcome its limitation that evenly spaced measurements is required. The reformed model can accommodate to any possible irregularity and sparsity of time-course expression data by adaptively fitting gene expression curves, followed by a step of interpolating data at missing time points before conducting of DBN analysis. We also develop an ordinary differential equation (ODE) model to reconstruct GRNs based on functional clustering of genes. A set of ordinary differential equations are constructed to quantify the dynamic of GRN and the regulatory effects including positive and negative regulation are identified in a regression setting

by using Smoothly Clipped Absolute Deviation (SCAD)-based variable selection. Both GRN models are equipped with unique power to integrate gene expression data from multiple environments and, therefore, provides an unprecedented tool to elucidate a comprehensive picture of GRN. By analyzing real data sets from a surgical study and through extensive simulation studies, the new models have been well demonstrated for their usefulness and utility.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Rongling Wu, for his excellent guidance, encouragement and help for the past several years at the Pennsylvania State University. His infectious enthusiasm and immense talent have been major driving forces through my Ph.D. studies. His mentor-ship is paramount in building the solid foundation for my future academic endeavors. He has helped me to open a brand new area of my life.

I would like to deeply thank my co-advisor, Dr. Runze Li. He has helped and supported me at every stages during my Ph.D. studies. His excellent courses, publications and brilliant ideas inspired me from time to time. My research could not be carried out in the manner it has been done without his help and encouragement.

I would show my gratitude to Professors Zhibiao Zhao and Donna Coffman, for their serving on my dissertation committee, as well as their suggestions and insightful comments on my work.

I also want to extend the thanks to my colleagues, faculty and staff at the Department of Statistics, to this great program, for making my time at Penn State one of the most memorable periods of my life.

Last but not least, I would like to gratefully thank my parents, my wife and my kids for their love and support all the time.

This dissertation research is supported by National Institutes of Health (NIH) grants 1U10HL098115, 1U01HL119178 and UL1TR000127. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH.

Chapter 1
Introduction

The genome is crucial to living processes such as cell survival, response of cells to the environmental signals and cell differentiation toward the phenotypic formation of organisms. Benefiting from the discovery of the molecular structure of deoxyribonucleic acid (DNA) in 1953 by Watson and Crick (Watson et al., 1953), people have made great progress in understanding the mechanism of storing genetic information in the genome as well as how these information be utilized to produce proteins which is the biological foundation of a organism. In addition, the relationships between the genome information and certain phenotypic characteristics have been explored more and more.

The development of high-throughput technologies, such as DNA microarrays and proteomics platforms, has made it possible to ask and address many fundamental but difficult questions in molecule biology (Schena et al., 1995; Shalon et al., 1996; Lockhart and Winzeler, 2000). These technologies have increasingly played a pivotal role in measuring gene expression and studying biological functions by linking differential pattern of expression with environmental signal changes. It also enable ones to study the interactions between genes through regulatory network based on gene expression.

1.1 Gene expression

A gene is a segment of DNA that encodes function and is contained in the nucleus of every cell. The DNA molecules lie in linear order along microscopic bodies called chromosomes which contain many genes. Watson and Crick have found the well known double-helix, a double-stranded structure, in which each DNA molecule organizes itself. One of the most important functions of DNA is to produce proteins. The central dogma of molecular biology explains the flow of genetic information within a biological system from DNA to proteins with two steps (Crick et al., 1970). In the first step (Transcription), DNA is transcribed into Messenger ribonucleic acid (mRNA) and in the second step (Translation), proteins are synthesized using a template based on the information in mRNA (Figure 1.1).

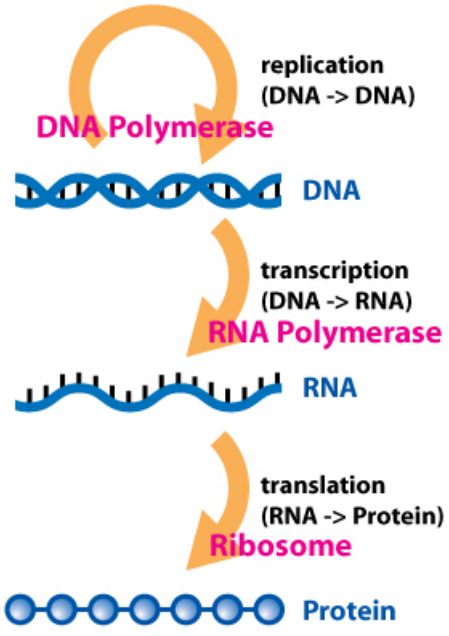


Figure 1.1: Central dogma of molecular genetics (from wikipedia.org)

Gene expression level, which is a metric of gene activeness, can be measured through the amount of mRNA produced during transcription. It determines to

what degree the information from a particular gene is used in the synthesis of its functional product protein. In the mid 1990s, the technology of Microarray is developed to detect differences in mRNA levels of thousands of genes simultaneously This technology is based on the principles of that every DNA strand is capable of recognizing complementary sequences through base pairing (Lipshutz et al., 1999). Brown and Botstein (1999) proposed a two-color microarray method to compare gene expression levels in two biological samples. They use a robotic arrayer to print arrays of thousands of discrete DNA sequences on glass microscope slides and then use different fluorescent dyes to label the samples. After the samples being mixed and hybridized with the arrayed DNA spots, microscope is used to determine the fluorescence measurements and decide the ratio which reflect the relative gene expression levels of the two samples. A two-color microarray image is shown in Figure 1.2.

Another popular microarray technology is Affymetrix® which is a type of high density synthetic oligonucleotide arrays (Lipshutz et al., 1999). With this approach, high-density arrays of synthetic oligonucleotides is designed directly using sequence information. On a small glass surface, hundreds of thousands of different oligonucleotides could be contained with the GeneChip® probe arrays. Photolithography is utilized in this technology to allow the construction of arrays with extremely high information content. Probe is designed based on complementarity to the gene whose expression is to be measured and is unique relative to other genes. Probe redundancy is very important in this approach to detect real signals from those due to non-specific or semi-specific hybridization. Mismatch (MM) control probes are used in the measuring process as specificity controls and they are identical to their perfect match (PM) partners with only a single base difference in a central position. Gene expression levels is obtained by analyzing of scanned images of microarray

slides. An Affymetrix image is shown in Figure 1.3.

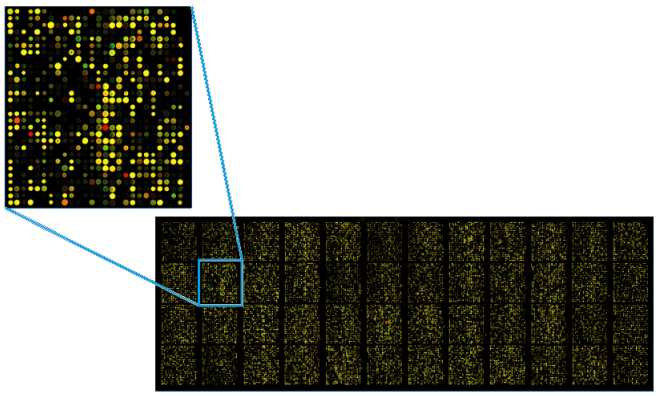


Figure 1.2: Image of microarray with enlarged inset to show detail. (from wikipedia.org)

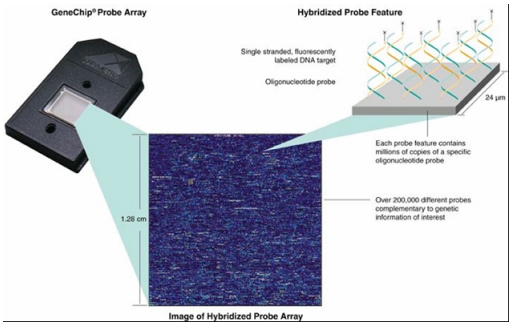


Figure 1.3: Image of Affymetrix microarray.

Gene expression levels measured using microarray need to be normalized, so that they are comparable across chips and conditions. Through normalization, expression levels could be centered and standardized and the unwanted noise and systematic bias could be reduced(Figure 1.4).

1.2 Gene regulation

In biology, it is an important task to identify the genetic causes behind phenotypic traits of organisms. Traditionally, people want to establish the connection between

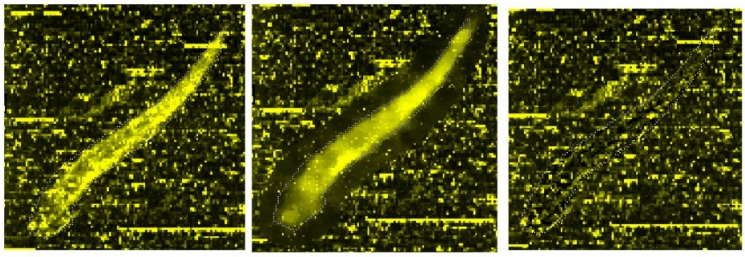


Figure 1.4: Array artifacts are corrected by image process

phenotype and specific DNA regions and consider the middle parts between phenotype and genotype as a black-box. However, we need to uncover the black-box to understand the mechanisms underlying the formation of phenotypic characteristics. Fig 1.5 shows biological pathways from DNA to phenotypic traits and processes of characteristics formation. It indicates that there are several procedures exist in the pathways including transcription, translation and biosynthesis. Elements such as DNA, mRNA, protein and even metabolic involve in the process of traits formation. The final phenotypes of an organism depend not only on genotype but also on the interaction between these elements. We consider the elements and their interactions consist of a regulation system as shown in Fig 1.6. It can be observed that genes interaction with others through their products such as protein and metabolic. For example, gene 1 has interaction with gene 2 through protein 1 which is a product of gene 1. There are different levels of regulation within this system. Though a couple of biochemical networks for this system could be considered such as metabolic network and protein network, gene network is an excellent abstraction of whole system with interactions between genes only (Brazhnik et al., 2002). In a gene network, the changes of expression level of genes are considered to be affected by the expression level of the others (as shown in Fig 1.6 with dashed lines) and the network could be reconstructed accordingly based on the gene expressions.

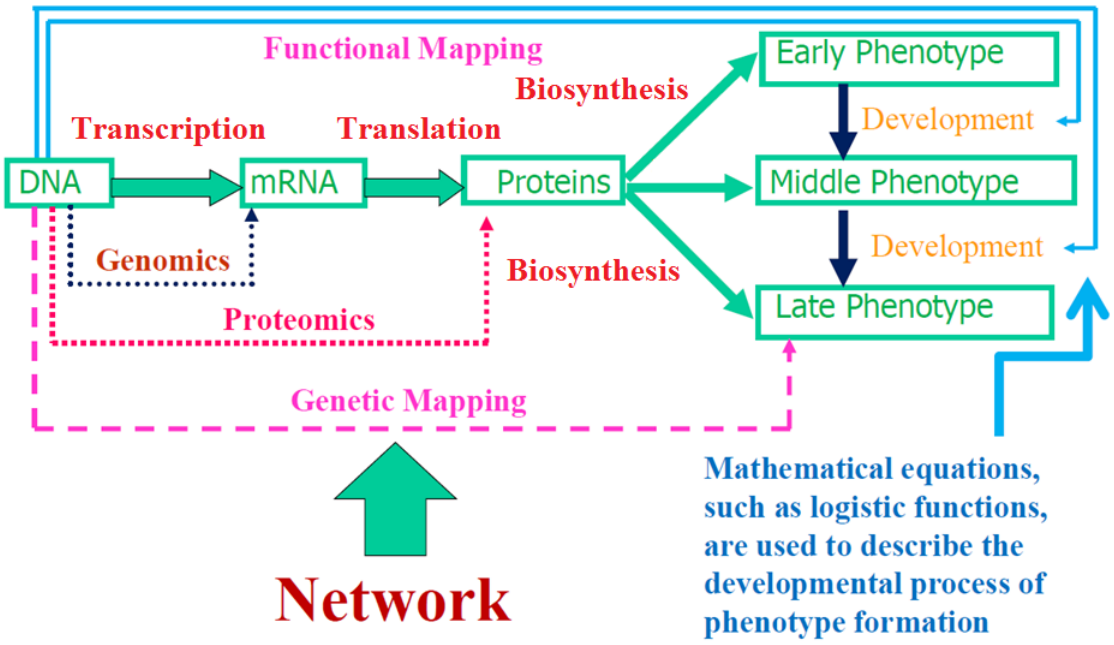


Figure 1.5: Biological Pathways and Processes of Trait Formation

1.3 Gene environment interaction

In biology, there is a widespread phenomenon that organisms cope with biotic and abiotic environments by controlling gene expression to harness the complement of active proteins. When the environment alternates between discrete states, organisms will stimulate their regulatory system through adjusting gene transcription rates to best adapt to the environment. In a study of bacterial evolution, McAdams et al. (2004) found that bacterials living in complex environments and have correspondingly complex sensor-response-control subsystems which enable them to adaptive gene expression as well as Regulatory complexity when environment changes. Seshasayee et al. (2006) further pointed out that the changes in the environment can be sensed by bacterial in the way of detecting extracellular signals. The change of signals of metabolite concentrations, pH levels, oxygen or water availability and surface contact eventually affects the transcriptional regulatory systems and pattern

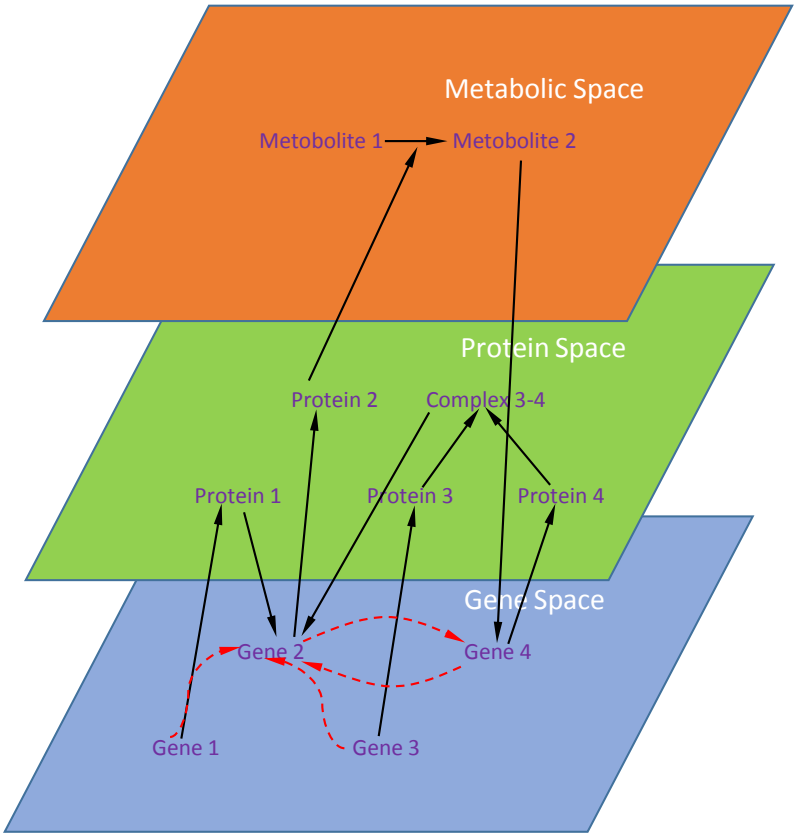


Figure 1.6: Biological Regulation System(redrawn form Brazhnik et al. (2002))

and abundance of transcription which leads to the physiological and morphological changes that enable organisms to survival with effective adaption. Wittkopp (2007) have addressed the questions of how and why cellular and organismal functions differ among environments, among individuals, and among species. They confirmed that environmental cues can create differences in gene expression without any genetic differences. In a gene-environment interaction study (Smith and Kruglyak, 2008), the genetic and molecular basis of variation in gene expression was examined between two yeast strains grown in two different conditions. It have been observed

that 2,037 transcripts showed significant strain-condition interaction effect. Li et al. (2006) studied the differential expression due to change of temperatures and concluded that heritable differences in plastic responses of gene expression are largely regulated in trans. By detecting the difference in the pattern of gene expression trajectories between discrete environments, we will be in a better position to study interactions between genes and environments and dictate a comprehensive map of gene-environment relationships.

1.4 Contributions of this dissertation

This dissertation developed statistical models for analyzing of gene expression in response to environment signals. Aiming at the adaptive gene expression dynamics under discrete environments, we successfully integrate environmental factor into clustering framework working on time-course gene expression data. To elucidate a comprehensive picture of gene regulatory network, we reform the dynamic Bayesian network model to overcome its limitation that evenly spaced measurements is required. The reformed model can accommodate to any possible irregularity and sparsity of time series data. Also, we proposed an ordinary differential equation model by dealing with the continuous gene expression measurements directly to reconstruct gene regulatory network. Both GRN models are equipped with unique power to integrate gene expression data from multiple environments and enable researcher to compare the change of regulation due to environment signals. Gene expression studies equipped with these sophisticated statistical models will play a more important role in identifying the interaction of gene and environment.

This dissertation is organized as follows. In Chapter 2, we develop a functional clustering framework based on a finite mixture model. We use Legendre polynomial to quantify the mathematical aspects of gene expression dynamics and consider

impact of environment on gene expression jointly. An EM algorithm has been developed for parameter estimation. Simulation studies demonstrate its advantages over traditional single environment analysis.

In Chapter 3, we employ dynamic Bayesian network to identify gene regulation under different environments. However, this approach requires expression data measured at even time intervals. In practice, time points at which gene expression is recorded are usually uneven-spaced, determined on the basis of distinct phases of biological processes. We reform DBN modeling to accommodate to any possible irregularity and sparsity of time series data. Application of this model on real data sets and extensive simulation studies have been conducted in this chapter.

In Chapter 4, we consider a model to work on the continuous gene expression data directly to avoid the loss of information in the process of DBN inference when doing data discretization. A ordinary differential equation model is presented in this chapter with the detailed three steps: clustering gene into functional groups, variable selection to detect significant regulation effects and analyzing gene functions. Both simulations and real data analysis are performed.

Finally, we summary the finding of our work and discuss future research in Chapter 5.

Chapter 2

Functional clustering of gene expression dynamics in response to environmental signals

2.1 Introduction

The development of high-throughput technologies, such as DNA microarrays and proteomics platforms, has made it possible to ask and address many fundamental but difficult questions in developmental biology and biomedicine. For example, variation in the pattern of gene expression may point to unique physiological or pathological properties of individual cells, organs, or organisms that cannot be observed readily or directly (Arbeitman et al., 2002; Rustici et al., 2004a). By screening approximately 1000 proteins in individual cancer cells, Cohen et al. (Cohen et al., 2008) detected a subset of proteins whose expression displays different dynamic patterns between seemingly identical cancer cells that actually have different fates. To identify distinct patterns of gene expression dynamics from a flood of microarray and chip data, powerful computational tools for clustering

genes or proteins based on their dynamic profiles have become essential. In the past decade, enormous efforts have been made to develop computational methods for cataloguing gene expression dynamics and use these distinct patterns to assess developmental functions and mechanisms of biological phenomena (Holter et al., 2001; Zhao et al., 2001; Ramoni et al., 2002; Park et al., 2003; Bar-Joseph et al., 2003; Luan and Li, 2003; Ernst et al., 2005; Ma et al., 2006; Inoue et al., 2007; Müller et al., 2008; Kim et al., 2008, 2010). There is also a pressing need for computational approaches to cluster analysis of dynamic gene expression that interacts with the environment, because the activation and expression of many genes is environment-contingent (Smith and Kruglyak, 2008). However, to our best knowledge, no literature has reported such specific approaches.

In biology, there is a widespread phenomenon that organisms cope with biotic and abiotic environments by controlling gene expression to harness the complement of active proteins (McAdams et al., 2004; Seshasayee et al., 2006; Wittkopp, 2007). When the environment alternates between discrete states, organisms will stimulate their regulatory system through adjusting gene transcription rates to best adapt to the environment. For this reason, by detecting the difference in the pattern of gene expression trajectories between discrete environments, we will be in a better position to study interactions between genes and environments and dictate a comprehensive map of gene-environment relationships. Traditional approaches for studying gene-environment interactions are based on quantitative trait locus (QTLs) mapping usually with experimental crosses. Significant gene-environment interactions are identified if specific QTLs are detected, through statistical tests, to display different effects between environments (Zhao et al., 2004b,a). More recently, gene transcript abundance has been used to study gene-environment interactions in many organisms such as yeast (Smith and Kruglyak, 2008; Landry et al., 2006)

and worms (Li et al., 2006), but all these studies are limited to gene expression data measured at single time points of development.

The purpose of this chapter is to describe a general framework for identifying environment-specific clusters of gene expression. The use of gene expression dynamics to understand gene-environment interactions is highly informative because of its capacity to identify development-related genes. However, this is a challenging work in terms of gene clustering across two-dimensional spaces, time and environmental state. The framework described in this article integrates developmental and environment-dependent programs of gene expression. Mathematical aspects of gene expression dynamics are implemented into a mixture model setting by considering the impact of environment on gene expression. The patterns of gene expression related to specific physiological functions can be parsimoniously modeled using a set of mathematical parameters (Kim et al., 2008; de Lichtenberg et al., 2005). Thus, by estimating the parameters that determine mathematical functions, the pattern of how genes change their level of expression over time and environment can be estimated and tested. The results from these models, therefore, can better be interpreted in a biologically sensible way. In addition, the framework considers the intrinsic structure of time-dependent correlations based on an optimal statistical process, which increases the power of detecting significantly differentiated patterns. To demonstrate its usefulness and utilization in practice, we use this framework to analyze a real data set from a rabbit hemodynamic study in which gene expression is observed in two distinct blood flow environments (Jiang et al., 2004; Fernandez et al., 2004). We evaluated the advantages of this tool by performing simulation studies. The simulation results illustrated that the tool has favorable statistical properties and can be used in any environment-dependent gene expression data.

2.2 Gene-clustering framework

2.2.1 Statistical model

Suppose there are n genes each measured at T time points in L environments. Let $\mathbf{y}_i^l = (y_i(t_1^l), ..., y_i(t_T^l))$ denote the gene expression data for gene i in environment l. Combining all the environments, we have $\mathbf{y}_i = (y_i(t_1^1), ..., y_i(t_T^1); ...; y_i(t_1^L), ..., y_i(t_T^L))$. If these genes are grouped into J clusters, this means that any one of genes (i) is assumed to arise from one (and only one) of the J possible clusters. Thus, the phenotypic value of gene i expressed at time t_τ^l in environment l is written as

$$y_i(t_\tau^l) = \sum_{j=1}^{J} \xi_{ij}\mu_j(t_\tau^l) + \sum_{c=1}^{C} \beta_c x_{ic} + e_i(t_\tau^l) \tag{2.1}$$

where ξ_{ij} is an indicator for gene i, defined as 1 if this gene belongs to cluster j and 0 otherwise, $\mu_j(t_\tau^l)$ is the mean of all genes belonging to cluster j at at time t_τ^l in environment l, x_{ic} is the value of covariate c $(c = 1, ..., C)$ for gene i, β_c is the effect of covariate c, and $e_i(t_\tau^l)$ is the residual assumed to follow a Gaussian distribution with mean zero and variance $\sigma^2(t_\tau^l)$. For longitudinal data, residual errors at different time points may be correlated with covariance $\sigma(t_{\tau_1}^{l_1}, t_{\tau_2}^{l_2})$ $(l_1, l_2 = 1, ..., L, l_1 \neq l_2; \tau_1, \tau_2 = 1, ..., T, \tau_1 \neq \tau_2)$. The residual variances and covariance comprise a $(TL \times TL)$ covariance matrix $\mathbf{\Sigma}$.

The distribution of gene expression data is expressed as the J-component mixture probability density function, i.e.,

$$\mathbf{y}_i \sim f(\mathbf{y}_i; \omega, \boldsymbol{\mu}, \mathbf{\Sigma}) = \sum_{j=1}^{J} \omega_j f_j(\mathbf{y}_i; \boldsymbol{\mu}_j, \mathbf{\Sigma}), \tag{2.2}$$

where $\omega = (\omega_1, \cdots, \omega_J)$ is a vector of mixture proportions which are non-negative

and sum to unity; $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_J)$ contains the mean vector of cluster j; and $\boldsymbol{\Sigma}$ contains residual variances and covariances among T time points over L environments which are common for all clusters. The probability density function of cluster j, $f_j(\mathbf{y}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, is assumed to be multivariate normally distributed with TL-dimensional mean vector

$$\begin{aligned} \boldsymbol{\mu}_j &= \Big(\mu_j(t_1^1) + \sum_{c=1}^{C} \beta_c x_{ic}, \ldots, \mu_j(t_T^1) + \sum_{c=1}^{C} \beta_c x_{ic}; \ldots; \\ &\qquad \mu_j(t_1^L) + \sum_{c=1}^{C} \beta_c x_{ic}, \ldots, \mu_j(t_T^L) + \sum_{c=1}^{C} \beta_c x_{ic}\Big) \end{aligned} \tag{2.3}$$

and covariance matrix $\boldsymbol{\Sigma}$. Notice that $\boldsymbol{\mu}_j$ contains gene-specific covariate effects.

The likelihood based on a mixture model containing J clusters can be written as

$$L(\boldsymbol{\Theta}|\mathbf{y}) = \prod_{i=1}^{n} \sum_{j=1}^{J} [\omega_j f_j(\mathbf{y}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma})], \tag{2.4}$$

where $\boldsymbol{\Theta}$ is a vector of unknown parameters including the mixture proportions, cluster-specific mean vectors, and covariance.

Different from traditional treatments, we will incorporate mathematical and statistical models to fit the mean-covariance structures. Instead of estimating all elements in the vectors and covariance, we estimate the mathematical and statistical parameters that model the mean-covariance structures. Thus, the question of clustering gene expression dynamics becomes how to find a set of parameters (arrayed in Θ_{μ_j}) that models cluster-specific expression profiles in a biologically and statistically meaningful way and to find a set of parameters (arrayed in Θ_v) that models the covariance structure both parsimoniously and flexibly.

2.2.2 Structural modeling of mean vectors

Since the transcript levels of DNA microarrays generally vary in a time course, we may use mathematical and statistical models to approach their dynamic changes. Below is a list of approaches for modeling time-dependent gene expression profiles:

2.2.2.1 Parametric modeling

For clock gene cases, the amount of mRNAs within the cell division cycle changes periodically, coincident with the cell cycle, which helps to maintain proper order during cell division or to conserve limited resources. The oscillation of cell cycle-regulated genes can be mathematically described by periodic Fourier functions or other functions (Spellman et al., 1998; Whitfield et al., 2002; Shedden and Cooper, 2002; Breeden, 2003; Rustici et al., 2004b; Ahdesmäki et al., 2005). The Fourier function can be approximated by its first K term, expressed as

$$F_K(t) = \alpha_0 + \sum_{k=1}^{K} \left(\alpha_k \cos\left(\frac{2\pi kt}{\tau}\right) + \beta_k \sin\left(\frac{2\pi kt}{\tau}\right) \right). \tag{2.5}$$

The coefficients α_k and β_k determine the times at which the expression level achieves maximums and minimums, α_0 is the average expression level of the gene, and τ specifies the periodicity of the regulation. From equation (3), the mean expression value of gene cluster j at time t_τ^l in environment l is expressed as

$$\mu_j(t_\tau^l) = F_K(t_\tau^l; \Theta_{\mu_j})$$

where

$$\Theta_{\mu_j} = \{\alpha_{0j}^1, \alpha_{1j}^1, \ldots, \alpha_{Kj}^1, \beta_{1j}^1, \ldots, \beta_{Kj}^1, \tau_j^1; ...; \alpha_{0j}^L, \alpha_{1j}^L, \ldots, \alpha_{Kj}^L, \beta_{1j}^L, \ldots, \beta_{Kj}^L, \tau_j^L\}$$

denotes the vector of Fourier parameters of the first K orders. Thus, by estimating the parameters that define the periodic curves for individual clusters, we can determine the differences in the temporal pattern of gene expression (see ref. (Kim et al., 2008)).

There are many other biologically well-justified curves that can be used to model gene expression dynamics. These include sigmoid equations for gene expression related to biological growth (von Bertalanffy, 1957; Richards, 1959; West et al., 2001; Guiot et al., 2003, 2006), triple-logistic equations for gene expression related to human body growth (Li et al., 2009), bi-exponential equations for gene expression related to HIV dynamics (Perelson et al., 1996), sigmoid Emax models for gene expression related to pharmacodynamic response (Ahn et al., 2010), biological thermal dynamics (Kingsolver and Woods, 1997), aerodynamic power curves for gene expression related to bird flight (Tobalske et al., 2003; Lin et al., 2006), hyperbolic curves for gene expression related to photosynthetic reaction (Wu et al., 2007) etc.

2.2.2.2 Nonparametric modeling

If time-varying expression of genes does not obey an explicit mathematical function, nonparametric approaches, such as kernel estimators or B-splines, can be used (Luan and Li, 2003; Daub et al., 2004; BORGWARDT et al., 2006). Kernel estimators are based on local polynomial regression, whereas smoothing splines use a piece-wise polynomial function. As shown in Silverman (Silverman et al., 1984), kernel smoothing and smoothing-spline smoothing are asymptotically equivalent for independent data and splines are higher-order kernels. More recently, Legendre orthogonal polynomials (LOP) have been used to model dynamic changes of complex traits that do not fit a specific mathematical curve (Lin and Wu, 2006; Cui et al.,

2006, 2008; Yang et al., 2007). Since the LOP are orthogonal to each other and integrate to 0 in the interval [-1,1], nonparametric estimates derived from this approach display favorable asymptotic properties (Marie and Pranab, 1985; McKay, 1997). The LOP have been successfully used to model time-varying phenotypic or genetic changes for many complex traits, such as milk production (Meyer, 2000) and plant height growth (Lin and Wu, 2006; Cui et al., 2006). As will be seen from an example below, It should be equally useful for modeling the dynamic pattern of gene expression profiles in a time course.

2.2.2.3 Semiparametric modeling

If gene expression spans multiple distinct stages (see (Müller et al., 2008)), at some of which the expression values follow a parametric form but at others of which they do not, we can implement a semiparametric model that combines the parsimony and biological relevance of parametric approaches and the flexibility of nonparametric approaches. In Cui et al. (Cui et al., 2006), such a semiparametric approach was used to model the growth process and death process of tiller number in a lifetime of rice. A similar semiparametric approach can also be implemented in the clustering framework of multi-stage gene expression dynamics. This will enable us to study dynamic changes of gene expression by relating its temporal profiles from different developmental stages.

2.2.3 Structural modeling of covariance

Unstructured estimate of a longitudinal covariance matrix may be highly unstable for large matrices. This, in conjunction with the fact that the covariance among repeated measures over time has an inherent structure (Diggle et al., 2002), implies that structuring a covariance matrix with few parameters may be crucial for

parsimonious and efficient parameter estimates in dynamic gene clustering. An extreme of covariance structuring is compound symmetry and autoregression of order one, but this may be far from the true covariance, leading to severe bias. The best covariance estimator should be at the balance between its variance and bias. Below, we list several commonly used approaches for covariance structure.

2.2.3.1 ARMA(p,q)

The autoregressive moving-average process, ARMA(p,q) (Box et al., 2013), is flexible to provide a robust estimate of gene expression covariance structure (Li et al., 2009). The zero-mean residual error $e_i(t_\tau^l)$ in environment l (1) is generated according to the following process

$$e_i(t_\tau^l) = \eta_\tau^l + \sum_{b=1}^{p} \varphi_b^l e_i(t_\tau^l - t_b^l) + \sum_{b=1}^{q} \theta_b^l \eta_{\tau-b}^l \tag{2.6}$$

where $\varphi_1^l, \ldots, \varphi_p^l$ and $\theta_1^l, \ldots, \theta_q^l$ are unknown parameters and $\{\eta_\tau^l\}$ is a sequence of independent and identically distributed normal random variables with zero mean and variance σ_l^2. The ARMA(p,q) model parameters are arrayed in $\Theta_v = \{\varphi_1^1, \ldots, \varphi_p^1, \theta_1^1, \ldots, \theta_q^1, \sigma_1^2; ...; \varphi_1^L, \ldots, \varphi_p^L, \theta_1^L, \ldots, \theta_q^L, \sigma_L^2\}$. The merit of the ARMA model includes the existence of closed forms for the estimates of the inverse and determinant of the structured covariance matrix (Haddad, 2004; Brockwell and Davis, 2009), which enhances computational efficiency.

By various constraints, the ARMA model can be reduced to a simple autoregressive (AR) model and structured antedependence (SAD) model (Zimmerman et al., 2001). Although the first-order AR and first-order SAD models use only two parameters, the latter is more flexible than the former since the latter allows the variance and correlation to change over time. The SAD model has been successfully incorporated in functional mapping of dynamic complex traits (Zhao et al., 2005).

For the ARMA model, it is important to determine its optimal order to model covariance structure. A model selection procedure based on penalized likelihood criteria, such as AIC and BIC, can be established to determine the most parsimonious approach.

2.2.3.2 Kernel smoothing

The kernel smoothing method has been used to estimate longitudinal covariances (Fan and Yao, 2003). The advantage of this method lies in its flexibility to specify any form of covariances and asymptotic properties. Under the homogeneous assumption, the covariance of gene expression between any two time points $t^l_{\tau_1}$ and $t^l_{\tau_2}$ in environment l is written as a function of time interval, i.e., $\text{Cov}\{t^l_{\tau_1}, t^l_{\tau_2}\} = f(|t^l_{\tau_1} - t^l_{\tau_2}|)$. Kernel smoothing describes this covariance by

$$f(|t^l_{\tau_1} - t^l_{\tau_2}|) = \frac{\frac{1}{n}\sum_{\tau_1=1}^{T}\sum_{\tau_2=1}^{T} K\left\{\frac{t^l_{\tau_1} - t^l_{\tau_2} - |t^l_{\tau_1} - t^l_{\tau_2}|}{h}\right\}(y_i(t^l_{\tau_1}) - \bar{y}^l_i)(y_i(t^l_{\tau_2}) - \bar{y}^l_i)}{\sum_{\tau_1=1}^{T}\sum_{\tau_2=1}^{T} K\left\{\frac{t^l_{\tau_1} - t^l_{\tau_2} - |t^l_{\tau_1} - t^l_{\tau_2}|}{h}\right\}} \quad (2.7)$$

where n is the total number of genes, T is the total number of time points, $K(\dot{})$ is a kernel function, h is a bandwidth, and $\bar{y}^l_i$ is the mean of gene i, given by $\bar{y}^l_i = \frac{1}{T}\sum_{\tau=1}^{T} y_i(t^l_\tau)$. One of the mostly used kernel functions is Gaussian kernel, i.e., $K(d) = \exp(-d^2)$. A variety of statistical methods have been developed to choose an optimal kernel and appropriate bandwidth (see (Fan and Yao, 2003)).

2.2.3.3 Modeling covariance over time and environment

In this article, we consider gene expression dynamics for multiple environments. The approaches described above are used to model time-dependent covariances of gene expression separately for each environment. The covariance structure over time and environment is then modeled by taking the product of purely temporal

and environmental covariances. This so-called separable approach is simple but has many undesirable properties since it does not allow environment-time interactions. We can implement a nonseparable stationary model (see Cressie and Huang (1999); Gneiting (2002); Gneiting et al. (2007)) to structure time-environment covariance of gene expression. A nonseparable covariance is not expressed as a Kronecker product of two matrices like separable structures can. The main significance of the covariance in this context is in providing a better characterization of the random process to obtain optimal kriging or prediction of unobserved portions of it. More recently, Yap et al. (Yap et al., 2011) has successfully incorporated Cressie and Huang's (Cressie and Huang, 1999) nonseparable model to estimate the covariance of photosynthetic rate over temperature and irradiance within the framework of functional mapping aimed to identify genes for dynamic traits.

2.2.4 Estimation and Tests

A hybrid EM-simplex algorithm was implemented to estimate the parameters, $\boldsymbol{\Theta}$, contained in the likelihood (2.4). The EM algorithm provides a platform for estimating the proportions of different clusters, within which the simplex algorithm is embedded to estimate base vectors for each cluster and the covariance-structuring parameters. This can be described as follows:

In the E step, we define and estimate the posterior probabilities of gene i, with which it belongs to a particular expression pattern j, by

$$\Omega_{j|i} = \frac{\omega_j \prod_l^L f_j(\mathbf{y}_i; \boldsymbol{\mu}_j^l, \boldsymbol{\Sigma}_l)}{\sum_{j'=1}^J \left[\omega_{j'} \prod_l^L f_{j'}(\mathbf{y}_i; \boldsymbol{\mu}_{j'}^l, \boldsymbol{\Sigma}_l)\right]}. \tag{2.8}$$

In the M step, the proportion of expression pattern j is calculated by

$$\omega_j = \frac{\sum_{i=1}^{n} y_i \Omega_{j|i}}{\sum_{i=1}^{n} \Omega_{j|i}}. \tag{2.9}$$

Suppose we model gene expression dynamics with Lengendra polynomial and model the covariance the ARMA process (Kim et al., 2008; Li et al., 2009), then the base mean vector in $\mathbf{\Theta}_{\mu_\mathbf{j}}$ and $\mathbf{\Theta}_\mathbf{v}$ can be estimated in the M step as follows.

$$\hat{u}_j = \frac{\sum_{i=1}^{n} \hat{p}_{ij} M' \hat{\Sigma}^{-1} y_i}{\sum_{i=1}^{n} \hat{p}_{ij} M' \hat{\Sigma}^{-1} M} \tag{2.10}$$

$$M = \begin{pmatrix} P_0(t_1^*) & P_1(t_1^*) & \cdots & P_r(t_1^*) \\ P_0(t_2^*) & P_1(t_2^*) & \cdots & P_r(t_2^*) \\ \vdots & \vdots & \vdots & \vdots \\ P_0(t_m^*) & P_m(t_1^*) & \cdots & P_r(t_m^*) \end{pmatrix} \tag{2.11}$$

where $P_r(t_m^*)$ is the Lengendra polynomial with order r at time point t_m^* which is the normalized version of t_m.

$$\hat{\Sigma} = \hat{\sigma}^2 \hat{R} = \hat{\sigma}^2 \begin{pmatrix} 1 & \hat{\rho}^{t_2-t_1} & \hat{\rho}^{t_3-t_1} & \cdots & \hat{\rho}^{t_m-t_1} \\ \hat{\rho}^{t_2-t_1} & 1 & \hat{\rho}^{t_3-t_2} & \cdots & \hat{\rho}^{t_m-t_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}^{t_m-t_1} & \hat{\rho}^{t_m-t_2} & \hat{\rho}^{t_m-t_3} & \cdots & 1 \end{pmatrix} \tag{2.12}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{J} \hat{p}_{ij} (y_i - \hat{\mu}_{ij})' \hat{R}^{-1} (y_i - \hat{\mu}_{ij})}{mn} \tag{2.13}$$

$$\hat{\rho} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{J}\hat{p}_{ij}\left[(1/(1-\hat{\sigma}^2))\hat{\mu}'_{ij}\hat{R}\hat{\mu}_{ij} + \hat{\sigma}\sum_{\tau=2}^{m}\hat{\mu}_{ij}^2(t_{i\tau}) - \sum_{\tau=1}^{m}\hat{\mu}_{ij}(t_{i\tau})\mu_j(t_{i\tau+1})\right]}{(m-1)n\hat{\sigma}^2} \quad (2.14)$$

The framework for clustering gene expression dynamics over multiple environments allows the test of many biologically meaningful hypothesis tests. First, an optimal number of gene clusters in terms of their different expression dynamics over all environments can be determined using AIC or BIC approaches (see the example shown below). Second, we need to determine the optimal number of gene clusters in a specific environment. For two given patterns j and j', they may be identical in an environment, although different for all the L environments. This can be tested by

$$.H_0 : \mathbf{u}_j^l = \mathbf{u}_{j'}^l \text{ vs. } H_1 : \mathbf{u}_j^l \neq \mathbf{u}_{j'}^l, \text{ for } j < j' = 1, \ldots, J \quad (2.15)$$

If the H_0 is accepted for two given patterns, this means that the optimal number of patterns in environment l is $J-1$. By performing this pairwise test for all gene clusters, this approach allows the identification of the optimal number of expression patterns for environment l.

Third, we can test the significance of gene-environment interactions. This can be done by testing

$$H_0 : \mathbf{u}_j^l = \mathbf{u}_j^{l'} \text{ vs. } H_1 : \mathbf{u}_j^l \neq \mathbf{u}_j^{l'}, \text{ for } l < l' = 1, \ldots, L \quad (2.16)$$

If the H_0 is rejected for two given environments, this means that expression pattern j displays significant gene-environment interactions. This test provides a quantitative way to study the relationship between genes and environments.

2.3 Worked example

2.3.1 Data analysis

The new tool is demonstrated by analyzing a data set of microarray genes associated with response to vein bypass grafting designed to treat arterial occlusive disease. The data were obtained using a rabbit bilateral vein graft construct, as previously described in Jiang et al. (2004). New Zealand White rabbits (3.0–3.5 kg) were treated by bilateral jugular vein interposition grafting and unilateral distal carotid artery branch ligation to create two distinct flows, i.e., two different environments. Through ligation of the internal carotid and three of the four primary branches of the external carotid artery, an immediate 6-fold difference in blood flow between the right and left vein grafts was obtained. A segment of the vein was retained at the time of implantation for baseline morphometric measurements. Vein grafts were harvested at 1, 3, 7, 14, 28, 90 and 180 days after implantation. Expression of 14,958 microarray genes was recorded for each of these time points under both treatments, high flow and low flow. Other parameters related to hemodynamic behavior, such as graft flow rate, intraluminal pressure, mean circumferential wall stress, and shear stress, were also measured or estimated.

An initial step is the selection of an appropriate model that fits the dynamic change of gene expression over. Figure 2.1 shows the plotting of 10 randomly selected genes expressed over time in the two treatments, from which we found, we found that there is a great variability in gene expression trajectories, of which some are curvaceous while others are quite flat. Thus, we used a flexible nonparametric approach based on Legendre orthogonal polynomials (LOP) to model gene expression dynamics. Let $\mathbf{P}_r(t^*) = [P_0(t^*), P_1(t^*), ..., P_r(t^*)]$ denote a family of LOP with a particular order r derived from a special differential equation, where t^* is a scaled

time with a range $[-1, 1]$. Let $\mathbf{u}_{jr}^{l} = [u_{j0}^{l}, u_{j1}^{l}, ..., u_{jr}^{l}]$ denote a vector of base values for cluster j in environment l. Then, time-varying mean values for cluster j in environment l in equation (2.2). can be expressed as a linear combination of $\mathbf{u}_{jr}^{l}$ weighted by the family of LOP, i.e.,

$$\mu_{j}^{l}(t^{*}) = \mathbf{P}_{r}(t^{*})\mathbf{u}_{jr}^{l}. \tag{2.17}$$

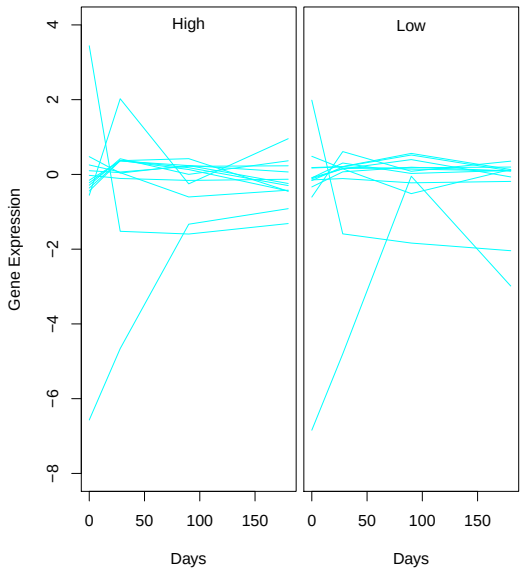


Figure 2.1: Trajectories of expression for ten genes randomly chosen from those associated with response to vein bypass grafting in rabbits in two treatments, high flow and low flow.

Our task now is to estimate the base vector $\mathbf{u}_{jr}^{l}$ from the given data. The variance of expression among genes seems to be broadly consistent over time points, suggesting that the first-order AR model may fit the data. To combine the expression data from the two flows, we used a separable model to structure the covariance over time and environment. In Figure 2.2, a plot of BIC values is illustrated against varying numbers of gene clusters under different LOP orders, from which we chose

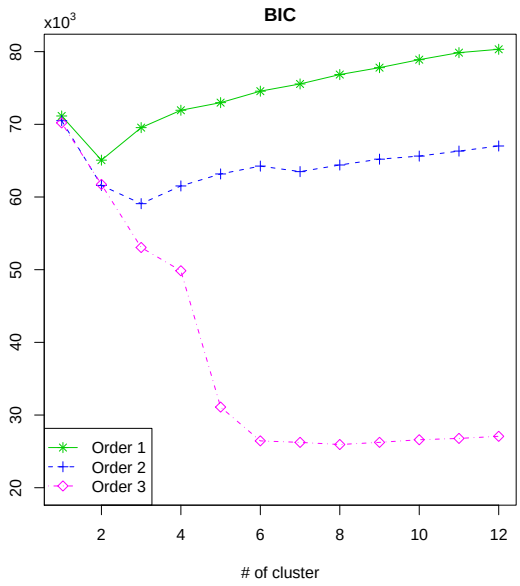


Figure 2.2: Plot of BIC values calculated for expression trajectories of different gene clusters over cluster number and LOP order.

eight clusters and three orders that provide a best combination for curve fitting. Implementing this combination, we estimated the expression trajectories of each of the eight gene clusters for both high and low flows. We need to detect if any of these eight clusters, labeled from A to H (see Supplementary Figure S1), overlap in a flow. Pairwise tests using hypothesis test (2.15). indicate that no pair of clusters is identical for expression trajectories in each flow ($P < 0.0001$). This result suggests that the optimal number of clusters should be eight for both flows. Table 2.1 gives the estimated proportions and their standard errors of each cluster from these genes.

By calculating the posterior probabilities of each gene that belongs to different clusters using Equation (2.8), we can determine the most likely cluster of this gene. Thus, we can draw gene expression trajectories for all genes that belong to a particular cluster A-H, separately for the two flows and the mean trajectory of the cluster for each flow using the estimates of curve parameters (Figure 2.3

Table 2.1: Estimated proportions of gene clusters and standard errors (in parentheses) estimated by resampling for 14958 genes associated with response to vein bypass grafting in rabbits under two different treatments, high flow and low flow. The significance of gene-environment interactions for each cluster is also given

Cluster	A	B	C	D	E	F	G	H
Proportion	0.0116 (0.0019)	0.1023 (0.0068)	0.3354 (0.0056)	0.3831 (0.0075)	0.1134 (0.0062)	0.0359 (0.0033)	0.0100 (0.0012)	0.0083 (0.0010)
P-value	< 0.0001	> 0.100	> 0.250	< 0.0001	> 0.400	< 0.0001	< 0.001	< 0.0001

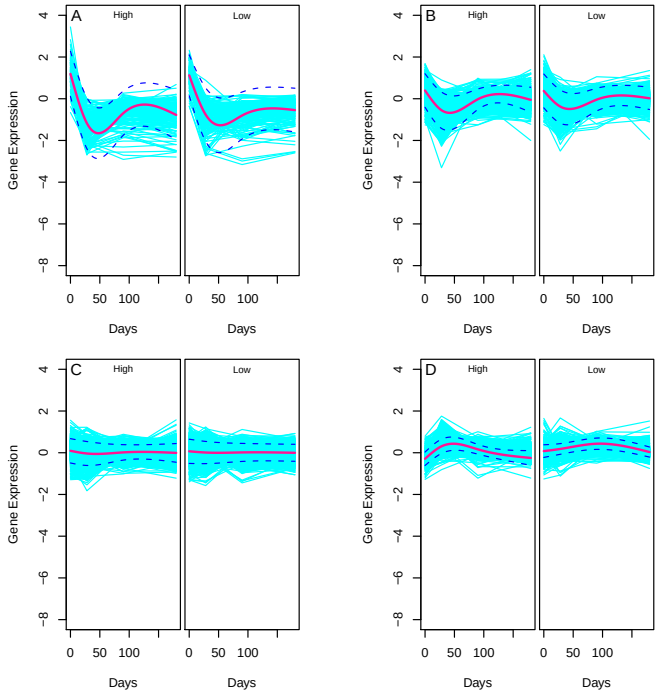


Figure 2.3: Expression trajectories of individual gene clusters A-D under high (H) and low flows (L)

and 2.4). The 95% confidence intervals of each estimated trajectory are generally within the variation of temporal gene expression profiles among individual genes, suggesting that our estimates are reasonably accurate. In general, most individual gene expression trajectories display similar time-varying trends between the two flows, but marked discrepancies in expression trajectories were detected for some particular clusters.

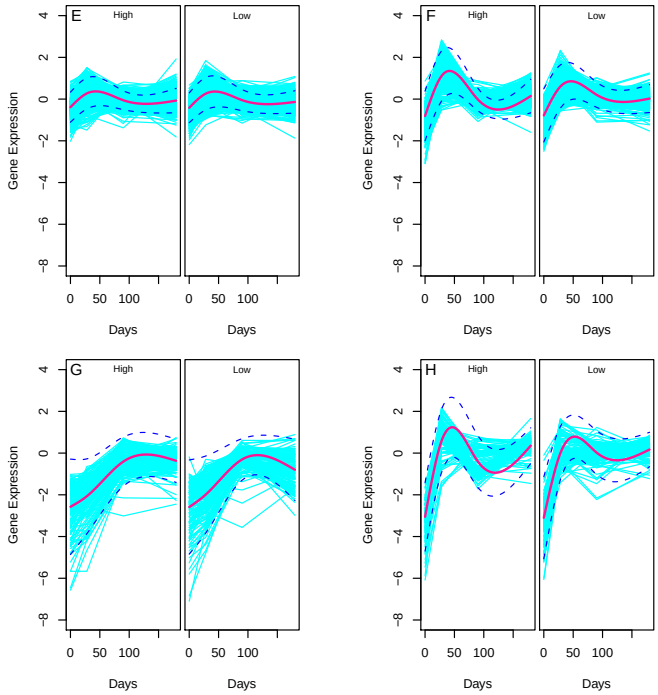


Figure 2.4: Expression trajectories of individual gene clusters E-H under high (H) and low flows (L)

Hypothesis test (2.16) allows the detection of gene-environment interactions in expression profiles over high and low flows. Table 2.1 gives the results of significance tests for gene-environment interactions. Except for clusters B, C and E, all other clusters are expressed differently between high and low flows. To better show treatment-dependent expression differences, we draw the mean trajectories of each cluster from high and low flows in the same plot (see Supplementary Figure S2). The expression of cluster A has the highest level right after implantation, decreases drastically within 25–35 days and then increases gradually. This cluster displays a more pronounced change of expression over time in high flow than low flow. Cluster B has a similar trend of gene expression profiles although its time-varying change is milder compared to cluster A. It appears that clusters C and D have a minimum level of expression throughout experimental time. Clusters E, F and H

are expressed at a low level in the beginning of treatment, reach a peak at Day 50 after implantation. Cluster H changes its expression level over time most abruptly, followed by clusters F and E. The expression of cluster G increases over time monotonously until Days 100–120, after which its expression decreases although it is more striking in low flow than high flow.

2.3.2 Simulation

We performed simulation studies to examine the statistical properties of the clustering model. The expression data were simulated by mimicking the structure of the real data analyzed above. A total of 4800 genes were assumed to include eight different clusters in two environments specified by mean trajectories as shown in Figure 2.3 and 2.4. These genes each have an expression trajectory over 4 time points as the sum of the mean trajectory of the underlying cluster and residual errors whose covariance structure follows the first-order AR model, but assuming the value of variance that triples the estimated variance. The proportions of eight clusters in Table 2.1 were used to simulate gene expression profiles.

Simulated data were analyzed by the model. Based on BIC values, the model can detect the correct number of clusters and the correct order of LOP. The model estimates the proportions of clusters A-H precisely Also, expression trajectories of each gene cluster can be reasonably well estimated (Figure 2.5 and 2.6), despite a tripled variance used. This shows that the results from the real data set analyzed by the model are convincing from a statistical point of view.

An additional simulation was conducted to test the power of the model to detect gene-environment interactions and its false positive rates (FPR). Consider Patterns A, D, F, G and H obtained from rabbit gene expression data which display a certain level of gene-environment interactions. By repeating the simulation and estimation

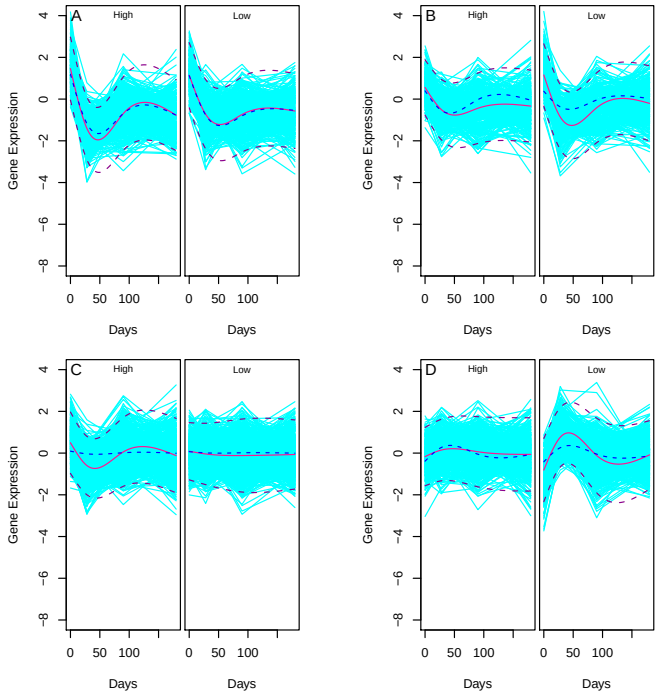


Figure 2.5: Comparison of estimated (dashed) and true (solid) expression trajectories for clusters A-D under two hypothesized conditions 1 and 2. The simulated data mimicked the structure of the rabbit data, but assuming an error variance that triples the estimated variance.

procedure 100 times, we detected the numbers of cases in which hypothesis tests for gene-environment interactions using test (2.16). are significant are 85–97 for clusters A, D, F, G and H, respectively. To test the FPR, we used the same parameters for clusters B, C, and E to simulate gene expression data for high and low flows. Of 100 simulation replicates, less than 5 times were detected to be significant. This suggests that the FPR of our tool is acceptably low.

2.4 Discussion

There is a pressing need for computational tools that can unravel the developmental machinery of time-dependent gene expression profiles, despite a vast body of

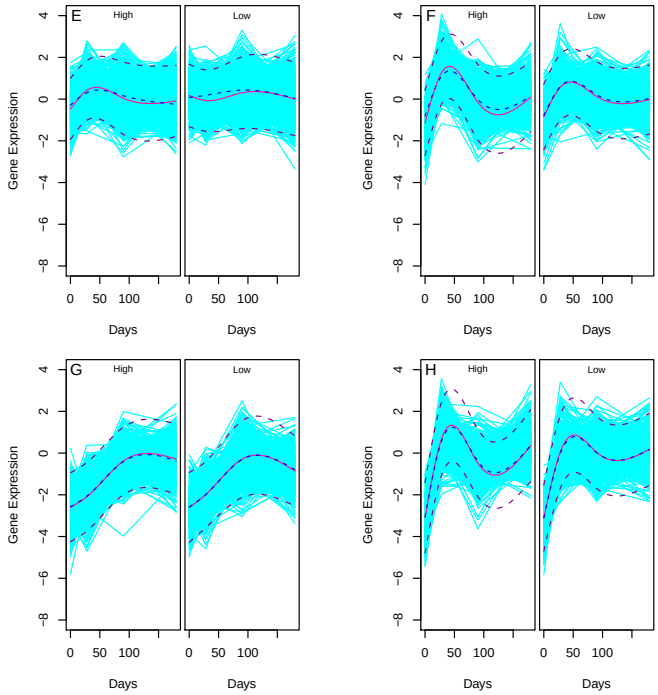


Figure 2.6: Comparison of estimated (dashed) and true (solid) expression trajectories for clusters E-H under two hypothesized conditions 1 and 2.

literature presenting these tools (Holter et al., 2001; Zhao et al., 2001; Ramoni et al., 2002; Bar-Joseph et al., 2003; Ernst et al., 2005; Ma et al., 2006; Inoue et al., 2007). One significant lack is the unavailability of models for analyzing gene-environment interactions for gene expression dynamics. A few pioneering studies have demonstrated the capacity of gene transcript abundance to comprehend the genetic architecture of gene-environment interactions (Smith and Kruglyak, 2008; Landry et al., 2006; Li et al., 2006).

In this article, we describe a computational tool that can test gene-environment interactions on a genomic level using dynamic gene expression data. The developmental dynamics of cells, organs, or organisms are related with many fundamental phenomena in biology, such as growth and phenotypic plasticity. Our understanding of how developmental dynamics is regulated through a balance of gene and

environment helps to reveal the mechanistic origins of these phenomena. The new tool may provide an important means for analyzing temporal expression data and construct a map of gene-environment interactions. As an example, we used Legendre-based nonparametric fitting to model dynamic changes of gene expression within a mixture-model framework. One advantage of the Legrendre approach lies in its flexibility for curve fitting (Marie and Pranab, 1985), computational efficiency, and avoidance of knot choice essential for B-splines. For gene expression data with explicit curves, such as periodic transcriptional profiles, robust mathematical equations with better biological relevance and better parsimony can be used.

Gene-environment interactions important to understand biology can now be tested in a quantitative way by the tool presented. By applying it to a real data set of microarray genes associated with response to vein bypass grafting between high and low flows in rabbits, this tool identified eight distinct patterns of expression trajectories in a time course. Each of these patterns may be related with a particular biological function operational in hemodynamic processes. Although many patterns display a similar trend of time-varying expression between high and low blood flows, many of them are essentially different based on hypothesis tests by our tool. A further molecular study may help identify specific biochemical pathways related to each of these different gene expression patterns. In a recent study, Cohen et al. (Cohen et al., 2008) detected the discrepancy of dynamic trajectories for a few proteins, which corresponds to cell death or survival, between seemingly identical cells. The tool presented should gain more quantitative insights into the distinction of seemingly trivial differences in genomic and proteomic dynamic studies.

While modeling gene expression dynamics jointly over time and environment in the real example, we assumes no interactions between time and environment. Although it facilitates our modeling and computing, this assumption may not

be realistic in some cases. A general model should consider the dependence of gene expression profiles in different times and environments by non-separable covariance structuring approaches (see Cressie and Huang (1999); Gneiting (2002); Gneiting et al. (2007); Yap et al. (2011) for examples). Different approaches for covariance structure based on nonparametric or semiparametric models should be incorporated and compared using penalized likelihood criteria. In our software package for environment-dependent functional clustering, we implement many of these approaches for users to choose the most parsimonious one given their particular data sets. The computer code for the tool developed is available at Penn State Center for Statistical Genetics web site, http://statgen.psu.edu.

Chapter 3

Inference of gene regulatory network through advanced Dynamic Bayesian Network

3.1 Introduction

Thousands of genes on the genome encode the products essential for cell division and differentiation toward the phenotypic formation of organisms. How the properties of these products, including abundance, mutual interactions, and temporal pattern, determine the process of life is governed by regulatory networks of genes. A gene regulatory network (GRN) is formed by a set of genes in a cell which interact with each other through their RNA and protein products and regulated by the transcription factors that activate the expression of particular genes (Brazhnik et al., 2002). Knowledge about the structure and organization of GRN can help us identify the causal regulations involved in metabolic and physiological processes within cells. With the availability of high-throughput data, increasing efforts have been made to reconstruct GRN by developing either model based or machine learning

based approaches (Barabási et al., 2011; Zhu et al., 2012; Zhang et al., 2013; Wang et al., 2013). These approaches have played an important role in referring the complex regulatory mechanisms that underlie biological functions and phenotypic characteristics (Gerstein et al., 2012; Hurley et al., 2012). To better separate direct regulations from indirect ones among genes within a GRN, Zhang et al. (2014) proposed a concept of conditional mutual inclusive information and implemented it into a computing algorithm for quantifying the mutual information between two genes given a third one.

Given that life is a dynamic process (de Lichtenberg et al., 2005), a considerable body of modeling studies have begun to reconstruct dynamic GRN from expression data measured across a time and space scale (Li et al., 2011). The formation of any biological characteristics activated by developmental signals is contingent on dynamic changes of gene expression. For example, in flowering plants, embryogenesis undergoes three distinct phases, asymmetric cell divisions to establish apical-basal polarity (early phase), the initiation of major organs and primordia (intermediate phase) and the mature embryo (late phase) (De Smet et al., 2010). By genome-wide profiling of gene expression during a complete developmental process from the zygote to the mature embryo in Arabidopsis thaliana, Xiang et al. (2011) constructed stage-specific regulatory networks, which provide an important foundation for understanding the dynamic pattern of pathway interactions during embryogenesis. The application of stage-specific regulatory networks to study the genetic underpinnings of trait development has now become a routine approach in a wide range of biological areas from plant biology to cancer biology (Zhang et al., 2014; Yosef et al., 2013).

Approaches for reconstructing dynamic GRN from time series gene expression data have been well developed, including dynamic Boolean networks and proba-

bilistic Boolean networks (Akutsu et al., 2000; Martin et al., 2007) and dynamic Bayesian networks (Murphy et al., 1999; Friedman et al., 2000; Zou and Conzen, 2005; Ogami et al., 2012; Godsey, 2013) among others. By integrating expression data measured at multiple time points, these approaches have been used to infer the temporal change of the structure and topological features of multiple interactions within genomic networks during a period of biological process. However, they may suffer the limitation of being unable to manipulate sparse, unevenly-spaced expression data which are quite popular in practice. On the other hand, there has been increasing recognition of using multiple different experiments to reconstruct a comprehensive GRN, in which data were rarely measured at the same schedule (Hecker et al., 2009; Greenfield et al., 2010). As a consequence, the statistical issue of simultaneous use and modeling of irregular data from different experiments should be addressed.

In this chapter, we present and validate a computational procedure for dynamic GRN reconstruction from sparse, irregular gene expression data by interpolating those missing points in time series measurements. The idea of interpolation used to model GRN is not new. Wessels et al. (2001) and Bansal et al. (2006) proposed cubic interpolation for GRN modeling. Yu et al. (2004) devised a linear interpolation method for dynamic Bayesian network construction. By implementing a parametric (such as Fouries series approximation) or nonparametric (such as Legendre orthogonal polynomials) function whose optimal order is determined by information criteria, our interpolation approach is adaptive, assuring the best function to fit a given expression dataset and, thus, capturing dynamic features of genes precisely. Different from the previous work, we integrate functional clustering into the DBN modeling framework by which to infer GRN based on functional clusters of genes. Functional clustering classifies gene profiles into distinct categories

according to their similarity, and estimates a functional nonlinear curve for the mean dynamic expression of genes within the same cluster. By interpolating missing data based on the functional curve, an evenly-spaced, regular time series data can be obtained from which DBN is used to infer GRN among gene clusters. Our approach can handle any dynamic gene expression data, regardless of its sparsity and irregularity, thereby providing a broader application in computational biology.

This chapter is organized as following. In section 2, we describe the detailed of this model which contains four steps. Then in section 3 we applied the proposed procedure on Rabbit microarray data to construct the regulatory network. In section 4 we conduct simulation studies for model validation. At last, we have a discussion section to close the chapter.

3.2 Methods

3.2.1 Dynamic Bayesian network modeling

Consider a hypothetical gene network (Fig. 1.6), in which three different levels of regulation exist: gene, protein and metabolic. Here we assume that genes do not directly affect each other but interact through the action of their specific products, proteins, metabolites, or protein-metabolite complexes. Gene 2 is regulated by the protein product of the gene 1 and by the complex 3-4 formed by the products of gene 3 and gene 4. The regulation of gene 4 is made by the metabolite 2 which in turn is produced by protein 2. Based on these webs of regulation, we can construct a gene network which describes how one gene interact with others (denoted by dashed lines in Fig. 1.6).

A Bayesian network (BN) approach derived from the combination of graph theory and probability theory can be used to yield topologies or qualitative networks

of interactions between the genes. A BN is considered as a directed acyclic graph $G(X, E)$, where X is a set of nodes, x_is, which are random variables representing genes' expression and E is a set of edges which indicate the dependencies between nodes (Aluru, 2005). The nodes follow conditional probability mass function $P(x_i|Pa(x_i))$, where $Pa(x_i)$ is the set of parents of node x_i. The Markov assumption is encoded implicitly in a Bayesian network; i.e. each nodes is independent of its non-descendants given its parents. Therefore, the joint distribution of all nodes can be decomposed down to the conditional distributions of the nodes as (3.1).

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} P(x_i|Pa(x_i)) \tag{3.1}$$

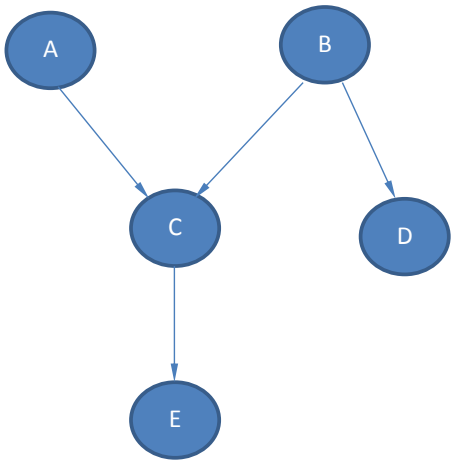


Figure 3.1: A Bayesian network

A sample of Bayesian network is shown in Figure 3.1, under the Markov

assumption, we have

$$P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|B)P(E|C) \tag{3.2}$$

To handle dynamic gene expression, dynamic Bayesian network (DBN) is developed by taking into account the time components, i.e. two copies of the same BN are used to model a state transition of gene network from time t to time $t + 1$. In a DBN as shown in Figure 3.2, the state of A is affected by B and itself but the state at a previous time.

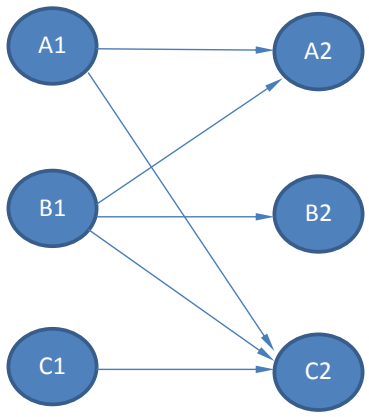


Figure 3.2: A Dynamic Bayesian network

The DBN approach for reconstruction of GRN based on gene expression data includes the following steps (Zou and Conzen, 2005):

1. Discretizing the expression levels.

   The expression levels for all genes are discretized as 1 (down-regulation) or 2 (up-regulation) by comparing with baseline gene expression level.

2. Realigning expression levels for potential regulator and target genes.

   Expression levels for potential regulator and target genes will be realigned according to the transcriptional time lag which is defined as the time that it take for the regulator gene to express its protein product and the transcription of the target gene to be affected by this regulator protein (Zou and Conzen, 2005). Suppose we have two hypothetical genes, gene A and its potential target gene B, their expression levels are measured at six evenly spaced time points $t_1 - t_6$. We use $A_{t_1}, A_{t_2}, ...$ and $B_{t_1}, B_{t_2}, ...$ to denote them respectively. If we decide the time lag is one time unit, then A_{t_1} will be aligned with B_{t_2}, A_{t_2} will be aligned with B_{t_3} and so on.

3. Determining regulators by calculating conditional probabilities and marginal likelihood scores.

   In this step, conditional probabilities (target gene give potential regulators) and marginal likelihood scores will be calculated using the realigned expression levels. The potential regulators which have highest marginal likelihood score will be selected as regulators.

Assume two genes whose expression levels are shown in figure 3.3. We followed the three steps of DBN as described above to discretize the expression levels (using fold change 1.2 as the cutoff point), realign them (using one time unit as the time lag) and calculate the conditional probabilities of gene B with respect to its potential regulator gene A (Table 3.1). Intuitively, since $P(B = 1|A = 1) = 1$ and $P(B = 2|A = 2) = 0.67$, we would consider gene A as a regulator of gene B. The basic condition of using DBN is that it requires the expression levels measured at evenly spaced time points because the time points are realigned one by one in step 2. If this condition was not satisfied, two issues would arise.

Suppose we do not measure the expression levels of gene A and B at t_4 and t_5

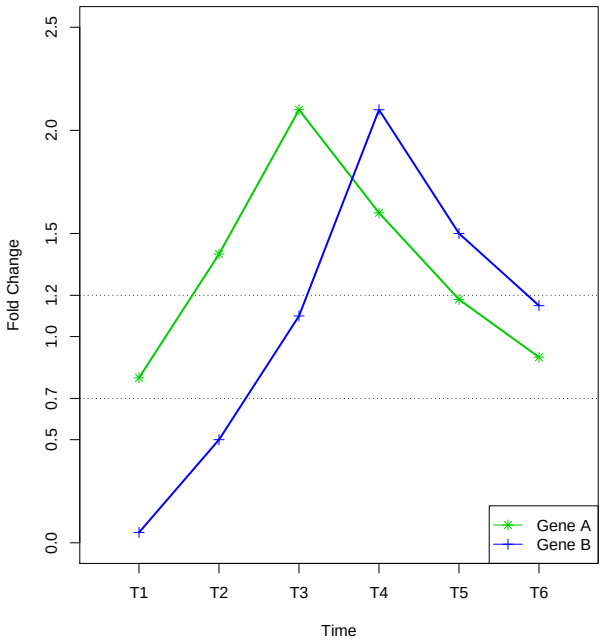


Figure 3.3: Original expression levels for gene A and B

Table 3.1: Conditional probabilities of gene B given gene A

	$A = 1$	$A = 2$
$B = 1$	1	0.33
$B = 1$	0	0.67

(as shown in figure 3.4). This will lead to the two problems as follows:

1. Mismatching.

   In this situation, we could still align A_{t_1} with B_{t_2} as well as A_{t_2} with B_{t_3}. However, we cannot align A_{t_3} with B_{t_4} because it is missing. We cannot align A_{t_3} with B_{t_6} either since the time period between them is much different from that between A_{t_1} and B_{t_2}.

2. Losing information.

   Since expression levels at t_4 and t_5 are missing, we lose those information completely. For gene B, the information we have is misleading since it seems that gene B has no up-regulation at all.

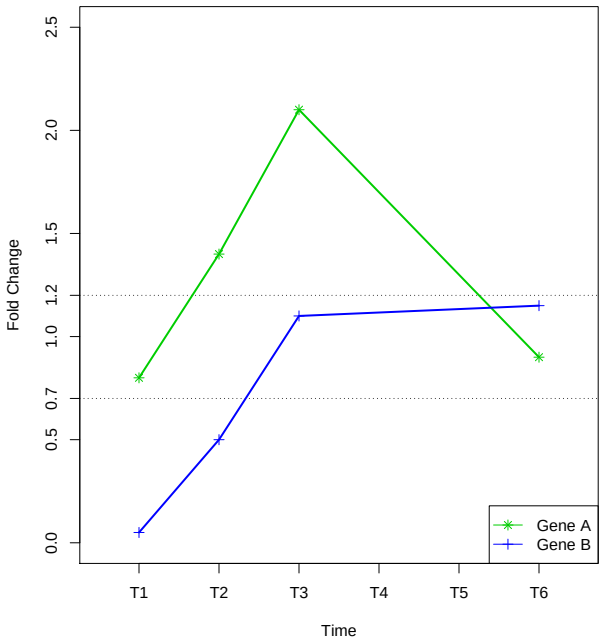


Figure 3.4: Expression levels for gene A and B without t_4 and t_5

Due to these two issues, DBN cannot be employed in this situation. If we applied it anyway, we would only have information from two pairs: A_{t_1} with B_{t_2} and A_{t_2} with B_{t_3}. Conditional probabilities would be $P(B = 1|A = 1) = 1$, $P(B = 1|A = 2) = 1$ and $P(B = 2|A = 2) = 0$ then it is difficult to decide whether or not gene A is a regulator of gene B.

3.2.2 Interpolation by a parametric or nonparametric function

A natural way to solve this problem is to restore the information missed at t_4 and t_5. Indeed, this can be done if sufficient data is observed. For example, for gene B, we would expect up-regulations should happen between t_3 and t_6, as shown in figure 3.5 by a dashed line, if the expression level would develop following the trend of t_1 to t_3. If this dashed line could be estimated as a function of time, it

is straightforward to interpolate the values of expression levels at t_4 and t_5 (as marked in red in figure 3.5) based on such a function. However, if this function is estimated from time-dependent observations of a single gene, we may not eliminate the effect of measurement noises. On the other hand, since many genes have a similar biological function, they should be classified into the same group with an indistinguishable time course expression pattern. These genes can be put together to provide a more precise estimation of functional curve.

Functional clustering, aimed to group those genes of similar function, can serve as a tool to estimate functional curves. Kim et al. (2008, 2010) implemented a Fourier series approximation to model periodic patterns of gene expression, whereas a nonparametric approach based on Legendre orthogonal polynomials (LOP) was developed in Chapter 2 to characterize time-varying expression levels when no explicit parametric function can be used. These approaches consider the mean of a cluster as a representative gene, thereby providing a more stable and accurate interpolation of missing points.

3.2.3 Four-step procedure to reconstruct GRN

In this section, we present a procedure for identifying gene regulatory network based on time course gene expression data in the four following steps:

1. Clustering genes into different groups by parametric or nonparametric functional clustering and estimating the mean function for each cluster;

2. Interpolating missing values in uneven intervals to obtain evenly spaced measurements;

3. Constructing the GRN using the DBN model (Zou and Conzen, 2005) to identify the effects of regulation due to interactions between clusters;

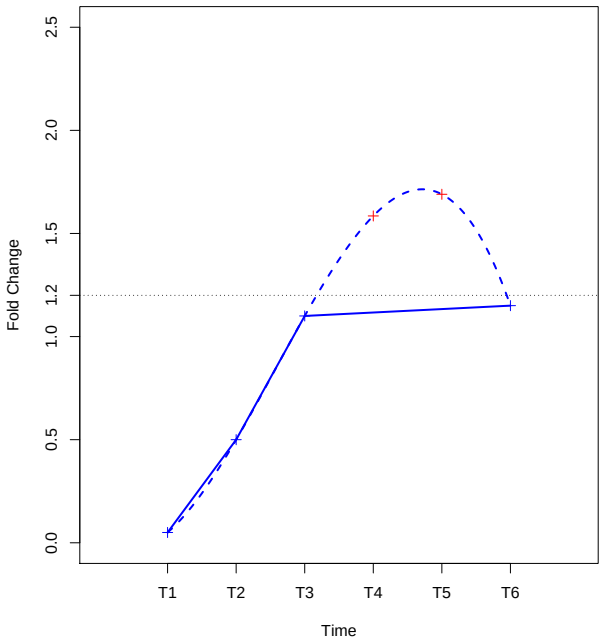


Figure 3.5: Estimated expression levels for gene B at t_4 and t_5

4. Analyzing gene functions by Gene Ontology to explore the biological relevance of gene clusters in the reconstructed regulatory network.

3.2.3.1 Clustering gene into different groups

Consider a high-dimensional set of genes (say n) measured at multiple time points from different experiments. Thus, it is possible that different genes are measured with different time schedules. For a particular process, such as embrogenesis, gene expression levels may be measured more densely in an early stage than late stage, making the time intervals of measurement unevenly-spaced. Overall, we have a sparse, irregular time series gene expression data for GRN reconstruction.

Let $\mathbf{y}_i = (y_i(t_1), \cdots, y_i(t_{T_i}))$ denote a vector of expression levels for gene $i(i = 1, \cdots, n)$ measured at time points $(t_1, \cdots, t_{T_i})$. Note that time points are gene-specific. We assume that these n genes can be classified into m clusters because of their similarity and differences. This can be expressed by a mixture model in

which there are m components. Each gene arises from one and only one of the m possible components. We further assume that $\mathbf{y}_i$ is a realization of a mixture of m multivariate normal distributions with the density function specified as

$$\mathbf{y}_i \sim f_i(\mathbf{y}_i; \boldsymbol{\omega}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \omega_1 f_{1|i}(\mathbf{y}_i; \boldsymbol{\mu}_{1|i}, \boldsymbol{\Sigma}_i) + \cdots + \omega_m f_{m|i}(\mathbf{y}_i; \boldsymbol{\mu}_{m|i}, \boldsymbol{\Sigma}_i) \qquad (3.3)$$

where $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_m)$ is a vector of non-negative proportions for the m possible clusters that sum to unity and $f_{j|i}(\mathbf{y}_i; \boldsymbol{\mu}_{j|i}, \boldsymbol{\Sigma}_i)$ denotes the density function for gene cluster $j(j = 1, \cdots, m)$, a multivariate normal with mean vector $\boldsymbol{\mu}_{j|i} = (\mu_{j|i}(t_1), \cdots, \mu_{j|i}(t_{T_i}))$ and the common $T_i \times T_i$ covariance matrix $\boldsymbol{\Sigma}_i$. Let $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{1|i}, \cdots, \boldsymbol{\mu}_{j|i})$ contain the cluster-specific mean vectors for gene i.

Parametric functional clustering implements an explicit mathematical equation to approximate time-dependent expression. If the genes are periodically regulated (Rustici et al., 2004b), Kim et al. (2008, 2010) used the Fourier series function, showing adequate power to capture the temporal expression pattern of oscillating genes. In the case where no explicit mathematical equation is available, in Chapter 2, we deployed a flexible approach based on the Legendre orthogonal polynomials (LOP) to model gene-specific function curves for each cluster. Both parametric and nonparametric approaches allow to handle the sparsity of time points in gene expression data. Also, by determining the best order of Fourier series or LOP by information criteria, both approaches can provide an optimal function for modeling time series expression levels for each cluster from a given dataset.

Increasing power of functional clustering also results from the parsimonious modeling of the covariance structure by a few number of parameters. Parametric, nonparametric or semi-parametric approaches have been used to model the covariance matrix $\boldsymbol{\Sigma}_i$, each with specific strengths and weakness. Li et al. (2010) proposed a general parametric approach for covariance modeling through a general

autoregressive moving-average process of order (p, q), the so-called ARMA(p, q). These authors derived the EM algorithm to estimate the ARMA parameters that model the covariance structure within a mixture model framework. The orders p and q of the ARMA process that provide the best fit are identified by model selection criteria.

3.2.3.2 Interpolating missing values in uneven intervals

For DBN modeling, we interpolate missing data adaptively to satisfy the requirement of evenly spaced intervals. The mean vectors for each cluster which can be expressed as a function of time have been obtained from step 1. Here, as an example, we describe step 2 by using LOP-based nonparametric functional clustering.

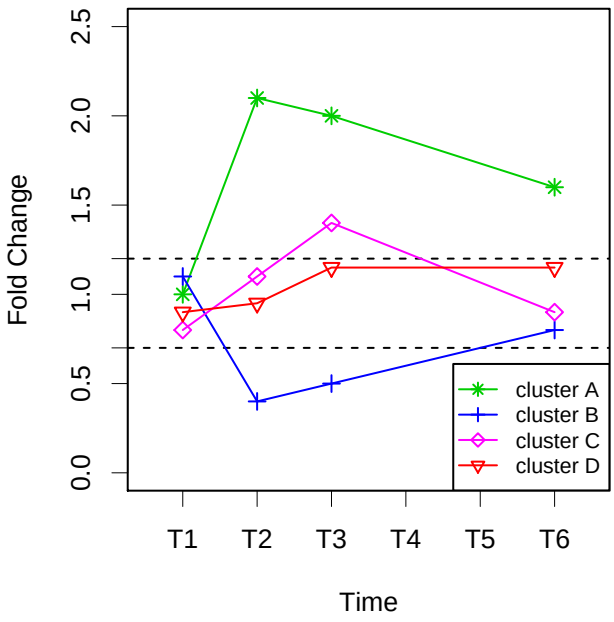


Figure 3.6: Expression levels for clusters A-D

Suppose we have four hypothetical gene clusters A, ..., D whose expressions measured at unevenly spaced time points t_1, t_2, t_3 and t_6 (Fig. 3.6), rather than six evenly spaced time points $t_1 - t_6$ in Zou and Conzen (2005). Let $\boldsymbol{u}_A$, $\boldsymbol{u}_B$, $\boldsymbol{u}_C$ and

$\boldsymbol{u}_D$ denote the base means of these clusters, respectively. According to Chapter 3, their mean vectors $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$, $\boldsymbol{\mu}_C$ and $\boldsymbol{\mu}_D$ are determined by $\mathbf{M}\boldsymbol{u}_A$, $\mathbf{M}\boldsymbol{u}_\mathbf{B}$, $\mathbf{M}\boldsymbol{u}_\mathbf{C}$ and $\mathbf{M}\boldsymbol{u}_\mathbf{D}$ where $\mathbf{M}$ is a 4 by r matrix constructed by LOP (with$r$ being the optimal order of LOP). For example, the mean vector of cluster A is expressed as

$$\boldsymbol{\mu}_A = \begin{pmatrix} \mu_{A_1} \\ \mu_{A_2} \\ \mu_{A_3} \\ \mu_{A_6} \end{pmatrix} = \mathbf{M}\boldsymbol{u}_\mathbf{A} = \begin{pmatrix} P_0(t_1^*) & P_1(t_1^*) & \cdots & P_r(t_1^*) \\ P_0(t_2^*) & P_1(t_2^*) & \cdots & P_r(t_2^*) \\ P_0(t_3^*) & P_1(t_3^*) & \cdots & P_r(t_3^*) \\ P_0(t_6^*) & P_1(t_6^*) & \cdots & P_r(t_6^*) \end{pmatrix} \begin{pmatrix} u_{A_1} \\ u_{A_2} \\ \vdots \\ u_{A_r} \end{pmatrix} \tag{3.4}$$

Where $\mu_{A_1}, \mu_{A_2}, \mu_{A_3}$ and μ_{A_6} are expression values of cluster A at time points t_1, t_2, t_3 and t_6 respectively. t_1^*, t_2^*, t_3^* and t_6^* are the normalized time points using the formula (3.5).

$$t^* = -1 + \frac{2\,(t - t_1)}{t_6 - t_1} \tag{3.5}$$

To interpolate expression values at t_4 and t_5, we first calculate the rescaled time values t_4^* and t_5^* and then insert two rows corresponding to t_4 and t_5 into matrix $\mathbf{M}$, obtaining the interpolated $\boldsymbol{\mu}_A$, denoted as $\hat{\boldsymbol{\mu}}_A$, by the following equation.

$$\hat{\boldsymbol{\mu}}_A = \begin{pmatrix} \mu_{A_1} \\ \mu_{A_2} \\ \mu_{A_3} \\ \hat{\mu}_{A_4} \\ \hat{\mu}_{A_5} \\ \mu_{A_6} \end{pmatrix} = \begin{pmatrix} P_0(t_1^*) & P_1(t_1^*) & \cdots & P_r(t_1^*) \\ P_0(t_2^*) & P_1(t_2^*) & \cdots & P_r(t_2^*) \\ P_0(t_3^*) & P_1(t_3^*) & \cdots & P_r(t_3^*) \\ P_0(t_4^*) & P_1(t_4^*) & \cdots & P_r(t_4^*) \\ P_0(t_5^*) & P_1(t_5^*) & \cdots & P_r(t_5^*) \\ P_0(t_6^*) & P_1(t_6^*) & \cdots & P_r(t_6^*) \end{pmatrix} \begin{pmatrix} u_{A_1} \\ u_{A_2} \\ \vdots \\ u_{A_r} \end{pmatrix} \tag{3.6}$$

Similarly, we can have $\hat{\boldsymbol{\mu}}_B$, $\hat{\boldsymbol{\mu}}_C$ and $\hat{\boldsymbol{\mu}}_D$. As shown in Figure 3.7, all of these gene clusters have evenly spaced time series measurement of gene expression.

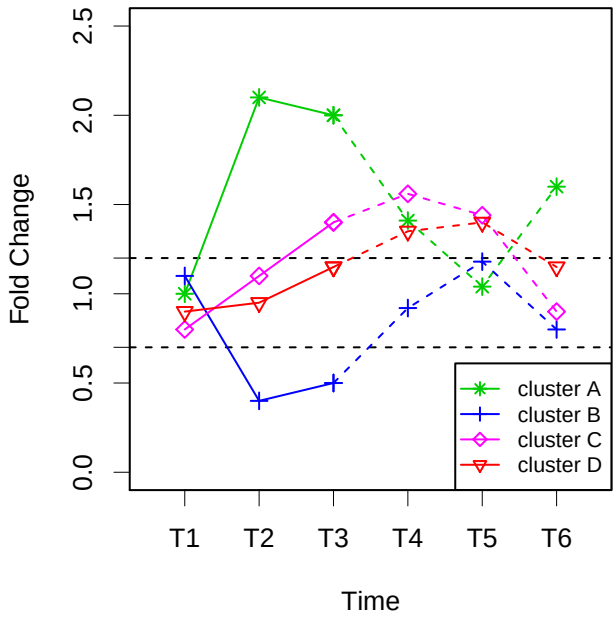


Figure 3.7: Expression levels for clusters A-D after interpolation

3.2.3.3 Constructing the GRN using the DBN model

We follow the improved DBN approach by Zou and Conzen (2005) to take advantage of its high efficiency and accuracy. According to these authors, only those genes

are considered as potential regulators when they have either earlier or simultaneous expression changes (up- or down-regulation) compared to targets. The up-regulation and down-regulation are defined as $\geq$ 1.2-fold and $\leq$ 0.7-fold compared to baseline gene expression. These relatively modest cutoffs are used to avoid missing any genes with potentially important changes in gene expression although these changes could be small. Per these cutoffs, we determine the initial regulation time points of gene clusters A-D, after which the regulators of genes that change later in expression are viewed as those genes that change earlier or simultaneous. As shown in Figure 3.6 , it is obvious that cluster A has initial down-regulation at t_2 while cluster B is also initial regulated at t_2 but down-regulation, Cluster C and D initially change expression at time t_3 andt_4, respectively. Since cluster A, B and C each have an earlier change in expression than cluster D, the former is selected as potential regulators of the latter. Similarly, we can decide potential regulators, cluster A for cluster B and cluster A and B for cluster C.

Based on the determined initial regulation time points, we could also decide the transcriptional time lag between regulator and target genes. We calculate the time difference between the initial regulation time points for potential regulator and its target gene , which is confided as a more accurate estimation of the corresponding transcriptional time lag (Zou and Conzen, 2005). In this way the time lag between cluster D and its potential regulators cluster C is estimated as one units. Similarly, the time lags between cluster D and B, D and A are estimated as two time units. According to the time lags between potential regulators and its target clusters, potential regulators are grouped into different categories with regulators in a category of the same time lag in terms of the target clusters. The reason for this grouping is that different regulators may have different time frames when interacting with targets. By grouping we analyze regulators separately, with

a possibility to identify co-regulators. As an example, cluster D has two groups of potential regulators; one group, including cluster A and B, has the time lag of two time units, and the other group, including cluster C, has the time lag of one time unit. It is here possible that cluster A and B are the co-regulators of cluster D.

After determining potential regulators for each clusters and calculating the corresponding time lags, the DBN framework developed by Murphy et al. (1999), is applied for network inference. Though continuous data can by directly analyzed by DBN, the assumptions of continuous DBN may not be satisfied in certain domain. In particular, continuous models assume additive influence of multiple regulators on a target, it may not be a case in gene regulation. Discrete network is chosen for our data set by discritizing continuous gene expression data. Two categories are used for discretization with 1-fold as cutoff instead of $\geq$ 1.2-fold and $\leq$ 0.7-fold since the relative increase or decrease in expression levels is more important than the absolute expression value during the inference of relationships between potential regulators and targets. Specifically, '2' is assigned if the expression level is equal to or higher then 1-fold; otherwise '1' is assigned. The discretized expression levels for cluster A-D are shown in Table 3.2.

Table 3.2: Discretized expression levels for Cluster A - D

	t_1	t_2	t_3	t_4	t_5	t_6
cluster A	1	2	2	2	2	2
cluster B	2	1	1	2	2	2
cluster C	1	2	2	2	2	1
cluster D	1	1	2	2	2	2

Another import step for DBN algorithm is to align the expression levels for potential regulators and targets according to the relevant transcriptional time lags between them. Suppose the time lag regulators and their target is Δt, then the expression levels of regulators at t_1 will be aligned with the expression level of the

target at $t_1 + \Delta t$. In this way, a $R \times K$ matrix will be constructed for regulators and the target, where R is the number of potential regulators with same time lag plus one which represents the target, and K is the number of time points between $t_1 + \Delta t$ and t_6. As an example, for cluster D, we have calculated the time lag between it and its potential regulator, cluster A, which is two time units in. Therefore, we align the expression level of cluster A at t_1 with the expression level of cluster D at t_3 and have a 2×4 matrix expressed as

$$\begin{pmatrix} 1 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \end{pmatrix} \tag{3.7}$$

where the first row is for cluster A and second row for cluster D.

As seen from above, the potential regulators of cluster D have been classified into two groups according to their time lags: one group of cluster A and B with two time units as the time lag and another group of cluster C with one time unit as the time lag. To identify all possible co-regulators of cluster D, we generate all subsets of each group and examine the relationships between every subsets of co-regulators. For the first group, the possible subsets are {cluster A}, {cluster B} and {cluster A, cluster B} and the subset of second group is {cluster C}. For each subset, a matrix like (3.7) is constructed based on the corresponding time lags and number of regulators. We then calculate the conditional probabilities of target cluster with respect to its regulators based on the matrix containing aligned expression levels. For each target cluster, we calculated marginal likelihood scores for every subsets of potential regulators using their conditional probabilities and the one has highest score is selected as the final regulator for this target.

We use the algorithm proposed by Murphy et al. (1999) for DBN inference. The idea of this algorithm is to select the optimal model that maximizes the following

conditional probability,

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)}, \tag{3.8}$$

where G denotes the network structure and D denotes the observed data.

3.2.3.4 Analyzing gene functions

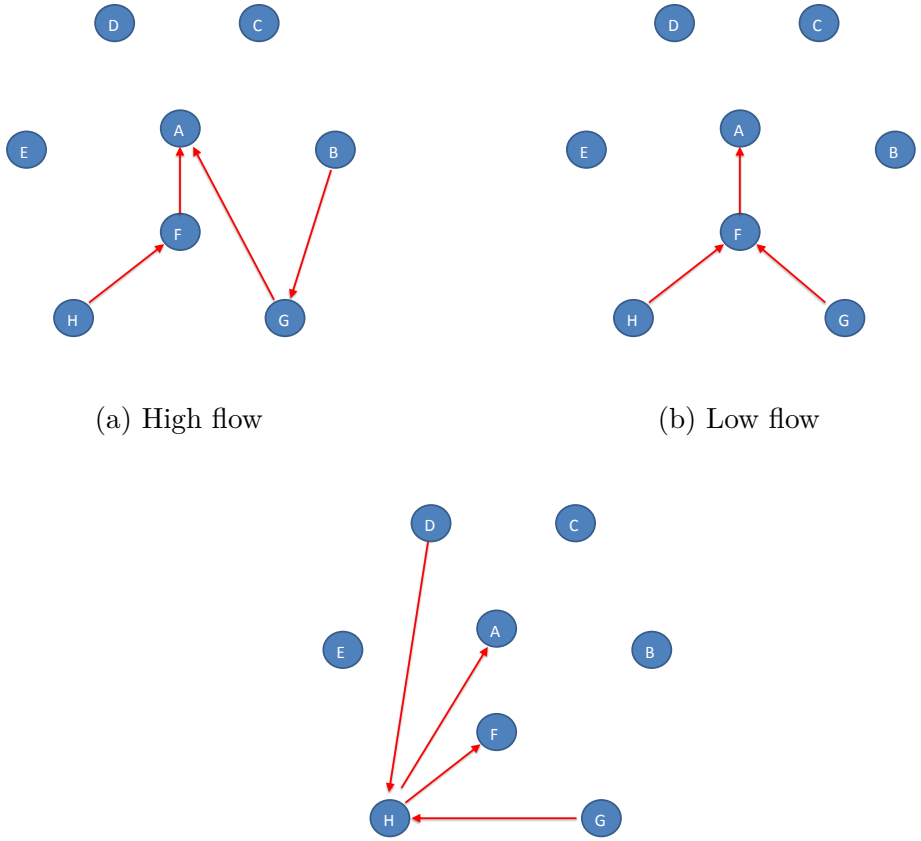
Genes with a similar profile pattern in each cluster usually share the common biological functions. Gene ontology (GO) analysis enable us to figure out what function is shared by genes in a cluster. Therefore, based on the clustering results in step 1 and regulation network established in step 2, we perform function analysis in this step. Different from traditional clustering approaches, functional clustering model developed in Chapter 2 allows environment-dependent expression plasticity to be clustered, producing results directly related to the mechanistic machineries of gene expression induced by environmental signals. Through GO analysis, we can shed more light on the regulation mechanisms underlying cellar and physiological processes.

3.3 Real data analysis

We demonstrated the application of the proposed procedure for GRN reconstruction by analyzing a real data set of time series gene expression from the surgery study of a rabbit bilateral vein graft construct (Jiang et al., 2004; Fernandez et al., 2004). The study involved two different environments, created by two distinct blood flows (differing by 6-fold) in vein grafts for New Zealand White rabbits (weighing 3.0-3.5 kg) resulting from the treatment of bilateral jugular vein interposition grafting and unilateral distal carotid artery branch ligation, respectively. With a segment of the

vein retained at the time of implantation for baseline morphometric measurements, vein grafts were harvested at 2 hours, 28, 90 and 180 days after implantation. Expression of 14,958 genes was recorded for each of these time points under both of treatments, high flow and low flow. By combining the dynamic expression data from the two treatments, we used the LOP-based functional clustering model in Chapter 2 to identify eight gene clusters, denoted as A (0.0116), B (0.0123), C (0.3354), D (0.3831), E (0.1134), F (0.0359), G (0.0100) and H (0.0083), where the numbers in parentheses are the proportions of genes belonging to a particular cluster. These clusters each display different patterns of environment-induced changes in gene expression trajectories. If we treat the mean expression curve for each cluster as a representative profile, the DBN model can be applied to inferring interactions between clusters. Since expression values were not measured at evenly spaced time intervals, we used our adaptive DBN model to reconstruct GRN, respectively, for high and low flows.

Figure 3.8 illustrates three different networks of gene expression under high and low flows and the difference of gene expression between the two flows. It is interesting to see that the structure of GRN is different dramatically between the two flows, although with some extent of similarity. Under high flow, cluster A is regulated jointly by cluster F and G. Meanwhile, cluster F and G are regulated by cluster H and B respectively (Fig. 3.8a). Under low flow, cluster A is regulated only by cluster F, whereas the latter is regulated by two clusters, H and G (Fig. 3.8b). Thus, cluster A is regulated directly by cluster G under high flow, but such a regulation operates through an indirect way under low flow. Under high flow, cluster B plays a role in regulating cluster G, but this regulation role disappears under low flow. For cluster C, D and E, since their expression patterns are relatively flat over time in both environments, with no up- or down-regulation (See Chapter



(a) High flow

(b) Low flow

(c) Difference between high flow and low flow

Figure 3.8: Networks inference for rabbit data set

2), they are not regulated by any clusters and also do not regulate other clusters.

We further made an inference of GRN based on the gene expression plasticity between high and low flows. Gene expression plasticity is defined as the environmentally induced alteration of gene expression, which is a capacity of the organism to respond to its environment. Let $\mu_j^H(t^*)$ and $\mu_j^L(t^*)$ denote the mean expression of cluster j at time t^* under high flow and low flow, respectively. The expression

plasticity of this cluster is defined as

$$\Delta\mu_j(t^*) = \mu_j^H(t^*) - \mu_j^L(t^*) = \mathbf{P}_r(t^*)\left(\mathbf{u}_{jr}^H - \mathbf{u}_{jr}^L\right). \tag{3.9}$$

The regulatory network based on expression plasticity data emphasizes the similarities of gene groups in terms of their pattern of differential expression over two different flows (Figure 3.8c). It was observed that cluster B, C and E which are not expressed differentially between two flows have no regulation effects. This is consistent with Chapter 2's finding that their expression difference is close to zero. On the other hand, the other clusters are heavily involved in the regulation (Fig. 5c) since they have significant differential expression according to hypothesis tests performed in Chapter 2. It appears that cluster H plays a multiple role in affecting the structure of GRN by regulating cluster A and F and by being regulated by cluster D and G. Given this, cluster H links the mutual relationships between discrete clusters D, A, F and G.

3.4 Computer Simulation

Yu et al. (2004) used simulation studies to investigate the influence of interpolation on DBN modeling. Their results showed that DBN can benefit from moderate data interpolation by reducing false positives. Here, we performed computer simulation to evaluate the performance of our adaptive model by answering the following questions: Is LOP-based interpolation better than non-interpolation in the case of missing data? Is LOP interpolation is better than linear-interpolation? What is the difference between even interpolation and uneven interpolation?

For each simulation, we follow the process shown in Figure 3.9.

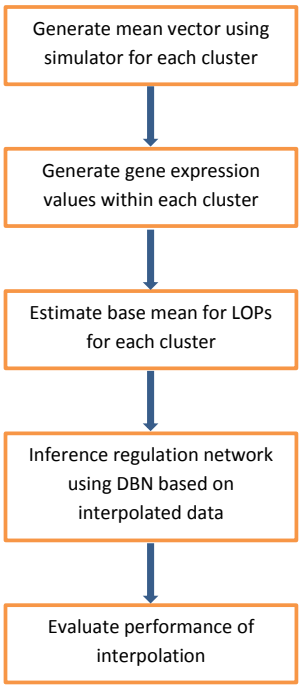


Figure 3.9: The process of each simulation

3.4.1 Simulation process

We generated simulation data of mean vectors for every cluster at every time point with the process as follows:

$$\boldsymbol{Y_{t+1}} = \boldsymbol{Y_t} + \boldsymbol{R}\,(\boldsymbol{Y_t} - \boldsymbol{C}) + \epsilon \tag{3.10}$$

Where $\boldsymbol{Y_t}$ is a vector which denotes the expression levels of means for all clusters in a regulatory network at time t. $\boldsymbol{R}$ is a design matrix used to define the regulatory relationships between clusters. Let r_{ij} denote the entry of $\boldsymbol{R}$ at row i and column j,

then the regulatory relationship between cluster i and cluster j could be interpreted completely by r_{ij}. If $r_{ij} = 0$, then cluster j has no regulation upon cluster i. If $r_{ij} > 0$, cluster j has activation upon cluster i. Otherwise, cluster j has repression upon cluster i. Moreover, the strength of regulation is defined by the magnitude of r_{ij}.

The constitutive expression value for each cluster is denoted by the vector $\boldsymbol{C}$ in 3.10. Similar to Yu et al. (2004), 0 and 100 are set as the minimum and maximum expression values and 50 set as constitutive value for all simulated cluster. Therefore, a cluster with expression value larger than 50 plays the effect in the direction specified in $\boldsymbol{R}$ while a cluster with expression value less than 50 plays the effect in the opposite direction specified in $\boldsymbol{R}$. The ϵ term drawn from a normal distribution plays the role of biological noise.

The above procedure was used to generate mean expression values for each cluster. The expression values for clusters with no regulators (entries of the corresponding row in $\boldsymbol{R}$ are all zero) could be generated in a different way by moving these clusters in a random walk according noise term ϵ. However, we let the trajectory of expression level for a cluster move along a curve specified by an LOP. It is assumed that expression values of genes within each cluster follow a multivariate normal distribution with cluster-specific mean vectors and covariance matrix. We assume that the covariance follows an autoregressive structure described by a correlation and variance (see Chapter 2). Therefore, we generate data for clusters without regulator with a LOPs model then generate data for clusters with regulators by process (3.10).

Let's define $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{iT})^T$ as the mean vector of cluster i, where $1, 2, \cdots, T$ denote time points when expression levels are measured. Then, we have

the relationship between $\boldsymbol{Y}$ and $\boldsymbol{\mu}$ as shown in (3.11).

$$\boldsymbol{Y_t} = \begin{pmatrix} Y_{t1} \\ Y_{t2} \\ \vdots \\ Y_{tn} \end{pmatrix} = \begin{pmatrix} \mu_{1t} \\ \mu_{2t} \\ \vdots \\ \mu_{nt} \end{pmatrix} \tag{3.11}$$

After deciding the regulation relationships in matrix $\boldsymbol{R}$, we first simulate mean vectors for all clusters which have no regulator. For such a cluster, say cluster i, a base mean $\boldsymbol{u}_i$ is specified first, after which the mean vector $\boldsymbol{\mu}_i$ is obtained through LOPs with calculations similar to (3.4). Mean vectors for clusters which have regulators are also generated. At a time point t, the expression values for all clusters are generated by (3.10) based on previously simulated data. We note that the expression values of cluster i is also updated and maybe different from the values gotten from LOPs. Therefore, those values need to be replaced with the ones from LOPs. This procedure is illustrated in (3.12).

$$\boldsymbol{Y_t} = \begin{pmatrix} Y_{t1} \\ Y_{t2} \\ \vdots \\ Y_{ti} \\ \vdots \\ Y_{tn} \end{pmatrix} = \begin{pmatrix} Y_{t1} \\ Y_{t2} \\ \vdots \\ \mu_{ti} \\ \vdots \\ Y_{tn} \end{pmatrix} \tag{3.12}$$

3.4.1.1 Generation of expression levels of genes

We assume that expression values of genes within each cluster follow multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. From Chapter 2, we know that Σ can be determined by σ and ρ in the form of (3.13).

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho^{t_2-t_1} & \rho^{t_3-t_1} & \cdots & \rho^{t_m-t_1} \\ \rho^{t_2-t_1} & 1 & \rho^{t_3-t_2} & \cdots & \rho^{t_m-t_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{t_m-t_1} & \rho^{t_m-t_2} & \rho^{t_m-t_3} & \cdots & 1 \end{pmatrix} \tag{3.13}$$

By specifying the values of σ and ρ, we can have Σ for each cluster. Expression values of genes can be drawn from multivariate normal distribution with parameters Σ and mean vector generated in previous section in a straightforward way.

3.4.1.2 Estimating base means for each clusters

After generating gene expression values for every clusters, they are pooled together and the clustering model we proposed in Chapter 2 is applied for grouping. As shown in Chapter 2, LOP-based functional clustering performs very well in classifying genes into distinct clusters. Here, we focus on the inference of GRN from our procedure. Therefore, we work on the clusters data separately instead of pooled data. i.e. we analysis the data of each cluster with applying a one-component clustering to use EM algorithm for estimating the base mean for LOPs. These means will be used in next step for interpolation.

3.4.1.3 DBN inference

The approach presented in Section 3.2 is used to analysis the structure of regulator network over clusters. Using the base means estimated in previous step, we can interpolate expression values if needed. The reconstructed networks are then compared with true relationships specified by matrix $\boldsymbol{R}$ and the performance of our method will be evaluated based on the comparison results.

3.4.1.4 Performance evaluation

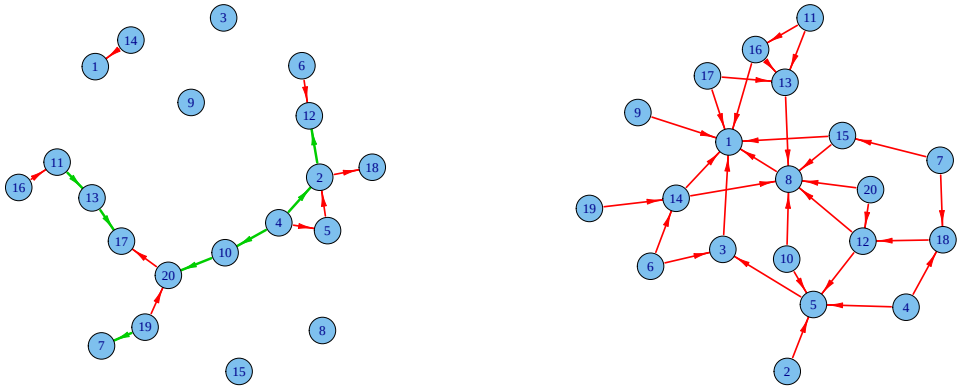
To evaluate the performance of our interpolation method, we define two metrics, positive predictive value (PPV) and false negative rate (FNR) as follows:

$$PPV = \frac{TP}{TP + FP} \tag{3.14}$$

$$FNR = \frac{FN}{TP + FN} \tag{3.15}$$

Where TP denotes true positive (regulatory relationships exist in both reconstructed network and true network), FN denotes false negative (regulatory relationships exist only in true network), TN denotes true negative (regulatory relationships do not exist in either network) and FP denotes false positive (regulatory relationships exist only in reconstructed network). Essentially, PPV is the proportion of TPs in the reconstructed network and FNR is the proportion of TPs which are not identified successfully Therefore, we prefer a higher PPV and a lower FNR.

In each of the six randomly simulated networks, we have 20 genes and about 10



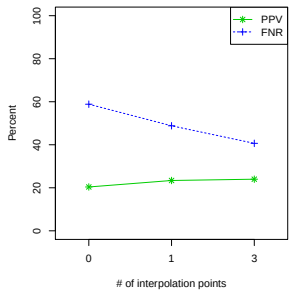
(a) True network (b) Reconstructed network

Figure 3.10: A true nework and the corresponding reconstructed nework

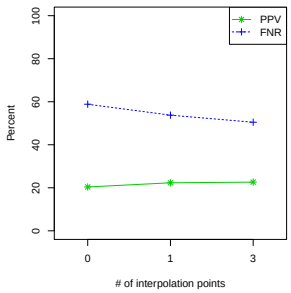
regulatory relationships in each network. For each relationship, we randomly assign a possible regulation strength, 0.05, 0.1, 0.15 or 0.2. A simulated true network is shown in Figure 3.10. For each of the regulatory network, we generate 100 sets of expressions so we have totally 600 simulated networks. By comparing the reconstructed networks with true ones, PPV and FNR were obtained.

3.4.2 Results

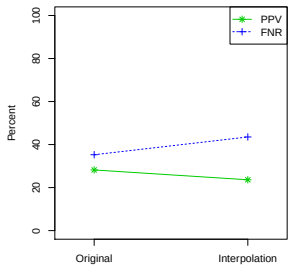
Whether or not LOPs interpolation is helpful for DBN inference? We first pick up the simulation expression values at time point $10, 50, 90, \cdots$ in the simulation run as analogous to a time course microarray expression data. We then interpolate one or three points in each interval with LOP to have two more data set. Therefore, we have 3 data set: $\{\boldsymbol{Y_{10}}, \boldsymbol{Y_{50}}, \boldsymbol{Y_{90}}, \cdots\}$, $\{\boldsymbol{Y_{10}}, \hat{\boldsymbol{Y}}_{\boldsymbol{30}}, \boldsymbol{Y_{50}}, \cdots\}$ and $\{\boldsymbol{Y_{10}}, \hat{\boldsymbol{Y}}_{\boldsymbol{20}}, \hat{\boldsymbol{Y}}_{\boldsymbol{30}}, \hat{\boldsymbol{Y}}_{\boldsymbol{40}}, \boldsymbol{Y_{50}}, \cdots\}$, where we use $\hat{\boldsymbol{Y}}$ to denote interpolated value. From the comparison of a reconstructed network (Figure 3.10b) with true one



(a) LOPs interpolation



(b) Linear interpolation



(c) Unevenly interpolation

Figure 3.11: Results of simulations

(Figure 3.10a), it was observed that most edges in true network have been identified but many false edges have also produced.

We evaluate the overall quality of all reconstructed networks from those simulated data sets. The results in Figure 3.11a show that interpolation does help to reduce the FNR; the FNR for the non-interpolated data is 58.85% while it is 40.66% for the data with 3 interpolation points for each interval. Moreover, interpolation also improved the PPV from 20.38% to 23.99% . Therefore, GRN reconstruction can benefit from interpolation with the LOP.

Whether is LOP interpolation better than linear interpolation? Following the same scenario above, we first pick up the simulated expression values at time point $10, 50, 90, \cdots$ in the simulation run first, after which we interpolate one or three points in each interval with LOPs and linear method separately to have

four more data set.

The results in Figure 3.11b show that LOP interpolation has better performance than linear method; the FNR is also reduced by linear interpolation but only from 58.85% to 50.47% when there are three points interpolated . Liner interpolation also has less improvement of PPV, from 20.38% to 22.63%.

What is the difference between even interpolation and uneven interpolation? In this case, the simulated expression values at time point $10, 20, 30, \cdots$ are picked up in the simulation run. We randomly drop 1 to 3 consecutive points as analogous to unevenly measured expression data.

Suppose we have $\{\boldsymbol{Y_{10}}, \boldsymbol{Y_{20}}, \boldsymbol{Y_{40}}, \boldsymbol{Y_{80}}, \boldsymbol{Y_{110}}, \cdots\}$ after dropping, we interpolate the missing points by LOP to have $\{\boldsymbol{Y_{10}}, \boldsymbol{Y_{20}}, \hat{\boldsymbol{Y}}_{\boldsymbol{30}}, \boldsymbol{Y_{40}}, \hat{\boldsymbol{Y}}_{\boldsymbol{50}}, \hat{\boldsymbol{Y}}_{\boldsymbol{60}}, \cdots\}$. The quality of recoveried network from this dataset will be compared with the one from the original data set of $\{\boldsymbol{Y_{10}}, \boldsymbol{Y_{20}}, \boldsymbol{Y_{30}}, \boldsymbol{Y_{40}}, \boldsymbol{Y_{50}}, \cdots\}$.

The comparison results are shown in Figure 3.11c. Obviously, the quality of reconstructed network form LOP-interpolated data is worse than that from true data. We can see that the FNR raised from 35.27% to 43.53%, while the PPV dropped from 28.20% to 23.63%. Therefore, it is not too worse and the most important thing is that LOP-interpolation make DBN model doable even when the measurements are not evenly spaced which is the case commonly happened in practice.

3.5 Conclusion

Many biological processes, including plant and animal development, disease pathogenesis and surgical recovery, are coordinated by cell-to-cell communications under the regulation of genes. High-throughput measurement techniques have now made it feasible to identify tens of thousands of genes at a time involved in sensing external

cues. The understanding of the relationships between genes and biological functions has become one of the hottest and most promising aspects in contemporary biology (Barabási et al., 2011; Gerstein et al., 2012). However, the dynamic interplay of genes is highly complex and cannot be understood by a simple approach (Brazhnik et al., 2002). The reconstruction of gene regulatory networks has proven to be a valuable tool for identifying the key mechanisms that shape the dynamics of cellular and transcriptional processes (Zhu et al., 2012; Yosef et al., 2013; Hecker et al., 2009).

Modeling of biological regulatory networks regulated by gene expression using dynamic Bayesian networks has been popular since Murphy et al. (1999)'s pioneering work. However, the requirement of evenly spaced measurements limits its widespread application. Time series records of gene expression are usually based on the distinct phases of biological processes (Quint et al., 2012), some of which receives more dense measurements than the others. Furthermore, increasing computational studies tend to integrate gene expression data from different experiments, in order to gain a comprehensive regulatory network underlying a biological phenomenon (Hecker et al., 2009). Because of these, the time series data of gene expression for GRN reconstruction are generally sparse and irregular. Despite tremendous efforts to model sparsely measured gene networks (Yu et al., 2004), a systematical procedure for DBN modeling using such imperfect data has still not been available in the literature.

In this chapter, we reformed DBN modeling by interpolating missing data points based on functional clustering (Kim et al., 2008, 2010). The new model can handle any dynamic gene expression data, no matter they are evenly spaced or not, thereby providing a broader tool in computational biology. The model was used to analyze two time series data sets of gene expression measured for vein bypass

grafts in rabbits that receive two distinct treatments, high and low blood flow. The similarity and difference in the structure and organization of genetic networks can be identified under high and low flow, providing new insights into the mechanisms of how genes regulate each other to determine final phenotypic formation. We have performed extensive simulation studies to demonstrate the practical usefulness and utility of the new model. It should be noted that the functional clustering model we used is under the assumption of independence among different clusters. A general model that does not rely on this assumption has been developed by Zhang (2013). The implementation of Zhang (2013)'s epistatic clustering may glean additional insight into the results of clustering dynamically differentiated genes and GRN reconstruction.

The past decade has witnessed tremendous milestones in high-throughput sequencing and large-scale data generation because of improvement in the accuracy of these techniques and their cost reduction for the required sample size. These developments enable researchers to not only dissect genomes but also unravel the regulatory interactions that allow genomes to regulate cellular structure, function, and behavior. The new model modified from a commonly used network modeling approach will find its widespread application given the popularity of collecting and using high-throughput expression data in human and other model or non-model systems. The model emphasizes on transcriptional data, but can be refined and extended to integrate multiple data types, such as mRNA and microRNA (miRNA) expression data, TF DNA-binding data, and protein interaction data (Bolouri, 2014). Also, the model should be linked to complex phenotypes or diseases within a causal-effect network framework toward identifying phenotype- or disease-causing perturbations. The model can be further perfected to readily determine the time of onset and duration of transcriptional activity and the magnitude of expression

of particular genes. Finally, the nature and topological features of regulatory networks may vary among different individuals, thus the identification and mapping of network-controlling quantitative trait loci (nQTLs) would be important for the prediction of network behavior.

Chapter 4

Inference of gene regulatory network through ordinary differential equation

4.1 Introduction

Cell survival and phenotypic characteristics of organisms are crucially affected by genes encoded on the genome as well as their products. Properties of these products including abundance and temporal pattern, which are governed by gene regulatory networks, play important roles in the process of life. Genes, regulators and regulatory connections between them, together with an interpretation scheme, form a biological regulatory system. Though proteins are usually the regulators, small molecular, like RNAs and metabolites, sometimes also participate in the overall regulation (Aluru, 2005). To abstract all interactions of proteins and metabolites in a living system, we use a gene regulatory network to describe genes acting on other genes (indicated in Figure 1.6 by dashed lines). One can gain not only new insights into the causality of transcriptional and cellular processes

but also the complex regulatory mechanisms that underlie biological function and phenotypic characteristics by reconstruction of gene regulatory network using gene expression data.

In the past decade, increasing efforts have been made to reconstruct GRN by developing either model based or machine learning based approaches based on the availability of high-throughput data (Barabási et al., 2011; Zhu et al., 2012; Zhang et al., 2013; Wang et al., 2013). These approaches have enabled ones to refer the complex regulatory mechanisms that underlie biological functions and phenotypic characteristics (Gerstein et al., 2012; Hurley et al., 2012). Given that life is a dynamic process (de Lichtenberg et al., 2005), a considerable body of models have been proposed successfully to reconstruct dynamic GRN from expression data measured across a time and space scale. Typical models include vector autoregressive and state space model (Shimamura et al., 2009; Kojima et al., 2009), differential equation model (De Jong, 2002; Lu et al., 2011; Wu et al., 2014), dynamic baysian network (Murphy et al., 1999; Zou and Conzen, 2005; Yu et al., 2004) and dynamic boolean network model (Thomas, 1973; Bornholdt, 2008). Most of these approaches such as the vector autoregressive and state space models require intensive time-series data to estimate the model parameters. Also, high computational cost is needed for these model, which limit us to inference only small-scale network.

Dynamic Bayesian network (DBN) modeling has been increasingly used to reconstruct GRN for the temporal pattern of transcriptional interactions in a time course (Murphy et al., 1999; Zou and Conzen, 2005; Yu et al., 2004). But, there is a major problem with DBN that, it requires the expression levels measured at evenly spaced time points. In practice, time points at which gene expression is recorded are usually uneven-spaced, determined on the basis of distinct phases of

biological processes. We proposed a method in Chapter 3 to overcome this limit with interpolation of data points based on functional clustering thereby the new model should be able to find its broader application in computational biology. However, we notice that certain information could be lost in the process of DBN inference after data discretization. Therefore, in this chapter we focus on inference gene regulatory network through approaches which works on continuous data directly based on Ordinary Differential Equation (ODE) (Voit, 2000; Holter et al., 2001; Aluru, 2005; Yeung et al., 2002; Lu et al., 2011).

Ordinary differential equations have been successfully used in the models for reconstruction of gene regulatory network (De Jong, 2002). By constructing differential equations for each individual gene with expression measurements at multiple time points, these approaches have been used to identify gene regulations within genomic networks during a period of biological process. Although the mathematic functions used in differential equations can take any forms to quantify the regulation effects, nonlinear specification of regulation functions requires very high computational cost which limits its application only suitable for small-scale network (Weaver et al., 1999; Sakamoto and Iba, 2001; Spieth et al., 2006; Wu et al., 2014). Wu et al. (2014) proposed a nonlinear ODE model by approximating the nonlinear regulation using spline functions and established its asymptotic properties. However, that model still suffered from the cost of computation, so that the simulation studies were conducted based on a system with only eight coupled ODEs and the model was applied on a real data set with only 58 genes.

Most of ODE models construct differential equations for individual genes. Consequently, they have a common requirement that, a large number of replicates for each gene are needed to reduce measurement noise and estimate the true profile curves. Therefore, they have difficulty to manipulate expression data with few

replicates which are quite popular in practice due to the high cost of experiments. In this chapter, we address this issue by integrating functional clustering approach into the ODE model. The information of multiple genes can be combined together in the procedure of clustering and provide us more power to estimate gene expression trajectories accurately. Lu et al. (2011) developed a five-step ODE model based on nonparametric cluster modeling. Though the five steps formed a comprehensive procedure for analysis of gene expression, more steps may lead to larger accumulative errors and reduce the accuracy of network inference. In our model, gene clustering and mean function estimating are completed simultaneously in one step, so that the model efficiency is improved and potential accumulative errors is reduced. Furthermore, we have explicit forms of mean functions as well as its derivative that help us in identifying the gene regulation effects. Consequently, parameter polish in Lu et al. (2011) is no longer needed.

In addition, we integrate the information of transcriptional time lags into the ODE modeling for the first time. Transcriptional time lag is defined as the difference between the time when the regulator gene to encode its protein product and the time when the transcription of the target gene to be affected by this regulator protein. By realigning the expression of potential regulators and target genes according to the transcriptional time lag, our model have more capability than previous ODE models to detect true regulation relationships because the regulation effects of regulators do not influence the expression of targets immediately. There are little biology support for most of previous models that consider instant regulation effects which are rare in reality. It is also not reasonable to consider effects over fixed time difference such as one time unit because the time lags are variable over pairs of regulators and targets. Moreover, while most of previous works including Lu et al. (2011) and Wu et al. (2014) focus on gene regulatory network in a single environment,

our model is equipped with unique power to integrate gene expression data from multiple environments and enable us to explore the difference of gene regulation effects under distinct environments. Therefore, it provides an unprecedented tool to elucidate a comprehensive picture of gene regulation system.

This chapter is organized as following. In section 2, we describe the detailed of our model which contains three stages. Then in section 3 we applied the proposed procedure on a time-course gene expression data for a surgical study. In section 4 simulation studies are conducted for model validation. At last, we have a discussion section to close the chapter.

4.2 Method

4.2.1 Ordinary Differential Equation Modeling

In a ODE method, the change rate of a gene expression (the derivative of expression) is models as a function of expression values of all involved genes. It describe the dynamic features of gene regulatory network by a directed network graph. A general ODE model (Lu et al., 2011) for gene regulatory network can be written as

$$\frac{d\mathbf{X}(\mathbf{t})}{dt} = \mathbf{F}(t, \mathbf{X}(\mathbf{t}), \boldsymbol{\theta}), \tag{4.1}$$

where $\mathbf{X}(t) = (x_1(t), x_2(t), \cdots, x_q(t))^T$ denotes the vector of gene expression values at time point t for all gene $1, 2, \cdots, q$. $\mathbf{F}$ is a function with parameter θ which is used to describe the regulatory effects, including positive, negative and feedback effects, of other genes on a certain gene i. Though any function, no matter linear or nonlinear, can serve as function $\mathbf{F}$, prior knowledge about biological mechanisms

is needed if **F** is nonlinear function. Also, high computational cost is expected so nonlinear ODE is usually applied on small-scale network (Weaver et al., 1999; Sakamoto and Iba, 2001; Spieth et al., 2006; Wu et al., 2014). Therefore, linear ODE models for network inference is more popular in practice. Lu et al. (2011) applied a simple liner ODE model as shown in (4.2).

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{q} \theta_{ij} x_j(t), \quad i = 1, 2, \cdots, q. \tag{4.2}$$

Where θ_{ij} denote the regulation effects between genes. A five-step procedure was proposed to identify GRN. They first used a nonparametric smoothing-based mixture model to cluster genes into distinct groups. Then, a mixed-effect model was applied to estimate the mean curve and its derivative of each group. The simple linear ODE model based on gene groups had been be considered as a standard regression model in statistics, then the significant regulatory effects had been detected by variable selection method, smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001). Due to the fact that the estimates of mean curves and their derivative may inherit large error, a nonlinear mixed-effects model was performed to refine the estimates of parameters obtained from SCAD.

In practice the number of measurements for each gene is usually much smaller than the number of genes, dimension reduction approaches are needed in this situation due to the curse of dimensionality. In Chapter 2, we proposed a functional clustering framework for identifying environment-specific gene groups. Similar to Lu et al. (2011), we consider that cluster method works as a powerful tool for dimension reduction which is crucial for linear ODE model. Since many genes have similar profile patterns during certain period, they are not distinguishable based on time course expression levels. Our functional clustering method enable us group

those genes with similar profile pattern into different clusters. We can make the network inference based on these clusters instead of individual genes. Since the number of clusters is much smaller than the number of genes, we are able to reduce the dimension of the ODE model dramatically. The ODE model for gene clusters can be written as

$$G'_k(t) = \sum_{j=1}^{J} \beta_{kj} G_j(t), \quad k = 1, 2, \cdots, J; t = t_1, t_2, \cdots, t_T, \tag{4.3}$$

where J is the number of clusters obtained in the clustering procedure, $G_k(t)$ is the mean profile for cluster k which is used as representative profile and $G'_k(t)$ is the corresponding derivative. β_{kj} is used to quantify the regulation effects. Additionally, variable selection approaches could be applied to reduce dimension further. After that, nonzero β_{kj} is considered as significant regulation effect for reconstruction of regulatory network.

4.2.2 Three-stage ODE procedure

In this section, we present a procedure for identifying gene regulatory network based on time course gene expression data with three stages

1. Clustering gene into different groups and estimating the mean functions for each clusters.

2. Detecting significant regulation effects between clusters by establishing an ODE system in a linear regression setting and applying variable selection method to identify regulatory relationships.

3. Analyzing gene functions through Gene Ontology with respect to gene groups in the reconstructed regulatory network.

4.2.2.1 Clustering gene into different groups

In the first stage, we use the clustering framework proposed in Chapter 2 for detecting differential expression over environment to group genes. In this way, we can reduce the dimension of the ODE model significantly and be prepared for the inference in the next stage. The clustering process is reasonable because many genes behave similarly during the experiment period thereby they are not distinguishable based on time course microarray data. In addition, gene within a cluster usually have same biological function. Therefore, they tend to have common regulation effects on other genes.

The clustering framework we proposed in Chapter 2 integrates developmental and environment-dependent programs of gene expression. Mathematical aspects of gene expression dynamics are implemented into a mixture model setting by considering the impact of environment on gene expression. Suppose there are n genes each measured at T time points in L environments. Let $\mathbf{y}_i^l = (y_i(t_1^l), ..., y_i(t_T^l))$ denote the gene expression data for gene i in environment l. Combining all the environments, we have $\mathbf{y}_i = (y_i(t_1^1), ..., y_i(t_T^1); ...; y_i(t_1^L), ..., y_i(t_T^L))$. If these genes are grouped into J clusters, this means that any one of genes (i) is assumed to arise from one (and only one) of the J possible clusters. Thus, the phenotypic value of gene i expressed at time t_τ^l in environment l is written as

$$y_i(t_\tau^l) = \sum_{j=1}^{J} \xi_{ij}\mu_j(t_\tau^l) + \sum_{c=1}^{C} \beta_c x_{ic} + e_i(t_\tau^l) \tag{4.4}$$

where ξ_{ij} is an indicator for gene i, defined as 1 if this gene belongs to cluster j and 0 otherwise, $\mu_j(t_\tau^l)$ is the mean of all genes belonging to cluster j at at time t_τ^l in environment l, x_{ic} is the value of covariate c $(c = 1, ..., C)$ for gene i, β_c is the effect of covariate c, and $e_i(t_\tau^l)$ is the residual assumed to follow a

Gaussian distribution with mean zero and variance $\sigma^2(t_\tau^l)$. For longitudinal data, residual errors at different time points may be correlated with covariance $\sigma(t_{\tau_1}^{l_1}, t_{\tau_2}^{l_2})$ $(l_1, l_2 = 1, ..., L, l_1 \neq l_2; \tau_1, \tau_2 = 1, ..., T, \tau_1 \neq \tau_2)$. The residual variances and covariance comprise a $(TL \times TL)$ covariance matrix $\mathbf{\Sigma}$.

The distribution of gene expression data is expressed as the J-component mixture probability density function, i.e.,

$$\mathbf{y}_i \sim f(\mathbf{y}_i; \omega, \boldsymbol{\mu}, \mathbf{\Sigma}) = \sum_{j=1}^{J} \omega_j f_j(\mathbf{y}_i; \boldsymbol{\mu}_j, \mathbf{\Sigma}), \tag{4.5}$$

where $\omega = (\omega_1, \cdots, \omega_J)$ is a vector of mixture proportions which are non-negative and sum to unity; $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_J)$ contains the mean vector of cluster j; and $\mathbf{\Sigma}$ contains residual variances and covariances among T time points over L environments which are common for all clusters. The probability density function of cluster j, $f_j(\mathbf{y}_i; \boldsymbol{\mu}_j, \mathbf{\Sigma})$, is assumed to be multivariate normally distributed with TL-dimensional mean vector

$$\begin{aligned} \boldsymbol{\mu}_j \;=\; & \Bigg(\mu_j(t_1^1) + \sum_{c=1}^{C} \beta_c x_{ic}, \ldots, \mu_j(t_T^1) + \sum_{c=1}^{C} \beta_c x_{ic}; \ldots; \\ & \mu_j(t_1^L) + \sum_{c=1}^{C} \beta_c x_{ic}, \ldots, \mu_j(t_T^L) + \sum_{c=1}^{C} \beta_c x_{ic}\Bigg) \end{aligned} \tag{4.6}$$

and covariance matrix $\mathbf{\Sigma}$. Notice that $\boldsymbol{\mu}_j$ contains gene-specific covariate effects.

The likelihood based on a mixture model containing J clusters can be written as

$$L(\mathbf{\Theta}|\mathbf{y}) = \prod_{i=1}^{n} \sum_{j=1}^{J} [\omega_j f_j(\mathbf{y}_i; \boldsymbol{\mu}_j, \mathbf{\Sigma})], \tag{4.7}$$

where $\mathbf{\Theta}$ is a vector of unknown parameters including the mixture proportions, cluster-specific mean vectors, and covariance.

Different from traditional treatments, we will incorporate mathematical and statistical models to fit the mean-covariance structures. Specifically, we used AR model for the covariance structures and a flexible nonparametric approach based on LOP to model gene expression dynamics. Then, time-varying mean values for cluster j in environment l in equation (4.5). can be expressed as a linear combination of $\mathbf{u}_{jr}^{l}$ weighted by the family of LOP, i.e.,

$$\mu_j^l(t^*) = \mathbf{P}_r(t^*)\mathbf{u}_{jr}^l. \tag{4.8}$$

We implement an EM algorithm to estimate model parameter and decide the optimal number of clusters and optimal order of LOP by BIC criterion.

4.2.2.2 Detecting significant regulation effects between gene clusters

In this stage, we are detecting the significant regulation effects by using variable selection methods. Similar to Chen and Wu (2008a,b) and Liang and Wu (2008), we construct differential equations as a regression model,

$$y_k(t^*) = \sum_{j=1}^{J} \beta_{kj} x_j(t^*) + \varepsilon_k(t^*), \quad k = 1, 2, \ldots, J; t^* = t_1^*, t_2^*, \ldots, t_T^*. \tag{4.9}$$

where $x_k(t^*) = \hat{G}_k^l(t^*)$, is the representative (mean) profile curve for cluster k in environment l and $y_k(t^*) = \hat{G}_k'^l(t^*)$ is its derivative. When we cluster genes into function groups in the first stage, the estimate of the mean functions of time for each cluster have been also obtained in the form of a linear combination of LOPs,

$$\hat{G}_j^l(t^*) = \mathbf{P}_r(t^*)\hat{\mathbf{u}}_{jr}^l. \tag{4.10}$$

Additionally, its derivative has the form of a linear combination of LOPs' derivative correspondingly,

$$\hat{G}'^{l}_{j}(t^*) = \mathbf{P}'_r(t^*)\hat{\mathbf{u}}^{l}_{jr}. \tag{4.11}$$

The error term $\varepsilon_k(t^*)$ is introduced due to the fact that we plug the estimation of differential equation variables into model (4.3). $t^* = t^*_1, t^*_2, \ldots, t^*_T$ are normalize time points using formula (4.12) since LOP are defined in the interval $[-1, 1]$. The time points t are not restricted to the original experimental time schedule since LOP are continuous in the interval $[-1, 1]$. Therefore, we may interpolate measurements between the original time points. D'haeseleer et al. (1999) and Bansal et al. (2006) have justified that data interpolation can help us to have better estimation of coefficients β_{kj}.

$$t^* = -1 + \frac{2\,(t - t_1)}{t_T - t_1} \tag{4.12}$$

To detecting true regulatory relationships based on biological knowledge, we incorporate the information of transcriptional time lags into the ODE modeling. Transcriptional time lag is defined as the difference between the time when the regulator gene to encode its protein product and the time when the transcription of the target gene to be affected by this regulator protein. There are little biology support for most of previous models that consider instant regulation effects; i.e. transcriptional time lag is 0, because the regulation effects of regulators can not influence the expression of targets immediately. It is also not reasonable to consider effects over fixed time lags such as one time unit since the time lags are usually variable over pairs of regulators and targets. The transcriptional time lag between a regulator and a target gene is decided based on the determined initial regulation

time points (See Chapter 3). We calculate the time difference between the initial regulation time points for potential regulator and its target gene, which is considered as a more accurate estimation of the corresponding transcriptional time lag (Zou and Conzen, 2005). To integrate the time lag information, the time points for potential regulators and targets are realigned. Suppose the time lag between a regulator and its target is Δt, then time point t for the target is aligned with the time point $t_1 - \Delta t$ for the regulator. Consequently, our ODE regression model is expressed as

$$y_k(t^*) = \sum_{j=1}^{J} \beta_{kj} x_j(t_j^*) + \varepsilon_k(t^*), \tag{4.13}$$

where t_j^* is rescaled version of $t_j = t - \Delta t_{kj}$ and Δt_{kj} is the transcriptional time lag between cluster k and cluster j. By realigning the expression of potential regulators and target genes according to the transcriptional time lag, our model have more capability than previous ODE models to detect true regulation relationships.

To identify significant regulation effects in the model (4.13), we apply variable selection method for linear regression. Traditionally, a subset of predictors in a regression model is obtained by forward selection, backward elimination, and stepwise selection, but these approaches are computationally expensive and unstable even when the number of predictors is not large. To overcome the computational disadvantages and theoretical difficulties classical variable selection procedures, alternative approaches have been developed, including ridge regression, bridge regression (Frank and Friedman, 1993), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), elastic net (Zou and Hastie, 2005) and the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001). All these models can be unified in a penalized least squares framework. Suppose we have a

linear regression setting

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N_n(0, \sigma^2 I_n), \tag{4.14}$$

where $\mathbf{y}$ is the vector of response variable, $\mathbf{X} = (\mathbf{x_1}, \cdots, \mathbf{x_n})^T$ is a $n \times d$ design matrix and $\beta = (\beta_1, \cdots, \beta_d)^T$ is the vector of regression coefficients. The penalized least squares can be written as

$$Q(\beta) = \frac{1}{2} \parallel \mathbf{y} - \mathbf{X}\beta \parallel^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|) \tag{4.15}$$

where $p_\lambda(.)$ is a penalty function indexed by the regularization parameter $\lambda > 0$, and λ balances the accuracy of in-sample fit and the parsimony of final model. Fan and Li (2001) argued that three statistical properties should be taken into account for a good penalty function: sparsity, unbiasedness and continuity. Sparsity means the small estimated coefficients can be set to zero automatically to reduce model complexity. Unbiasedness refers that the estimates for the true large coefficients should be unbiased. Moreover, we prefer the resulting estimator to be continuous in the sense that the variable selections should be stable when the observed sample changes slightly. Based on these principles, they suggested the SCAD penalty function in the form of

$$p'_\lambda(|\beta|) = \lambda I(|\beta| \leq \lambda) + \frac{(s\lambda - |\beta|)^+}{s-1} I(|\beta| > \lambda) \tag{4.16}$$

for some $s > 2$ and usually $s = 3.7$ is used. It has been shown that this SCAD penalty function enjoys all the three properties. However, it is challenging to develop numerical algorithms for SCAD estimator due to the singularity and nonconvexity

of the penalty function. Fan and Li (2001) proposed a algorithm based on the local quadratic approximation (LQA) of penalty functions and Newton-Raphson algorithm can be used to optimize the penalized object function but suffered from that a variable could not be added into the final model if it had been excluded at any step in the LQA algorithm. Hunter and Li (2005) developed a perturbed version of LQA to solve this problem but another tuning parameter needs to be introduced into the model and its value cannot be determined easily. To address these issues, Zou and Li (2008) proposed a new unified algorithm based on the local linear approximation (LLA) of SCAD penalty functions as shown below.

$$p_\lambda(|\beta|) \approx p_\lambda(|\beta^{(0)}|) + p'_\lambda(|\beta|^{(0)})((|\beta| - (|\beta^{(0)}|) \tag{4.17}$$

They further proposed using one-step LLA estimator from the LLA algorithm as the final estimates. It has been shown that the one-step LLA estimator can reduce the computation cost dramatically and keep the statistical efficiency meanwhile. It turns out that the least angle regression (LARS) (Efron et al., 2004) can be used to optimize the object function yielded from LLA.

Recently, coordinate descent algorithm have been utilized as an alternative to LARS algorithm when fitting penalized regression models like SCAD (Breheny and Huang, 2011). Breheny and Huang (2011) have established theoretical convergence properties of this algorithm when applying it on SCAD and also shown that it is much faster than competing methods. Here, we employ SCAD model using the coordinate descent algorithm to select significant regulation effects for a certain target and reconstruct regulatory network over clusters according to the identified effects including positive, negative and feedback ones.

4.2.2.3 Analyzing gene functions

Beyond the topology of a gene regulatory network, we are also interested in particular biological functions shared by genes with similar profile pattern within each distinct clusters. These functions can be identified through gene ontology analysis. Therefore, based on the clustering results in stage 1 and regulation network established in stage 2, we perform function analysis in this stage by accessing public database of gene functions such as Gene Ontology Consortium. Further more, since we integrate environmental signals into our model, it is possible to identify the change of regulation relationships between clusters from one environment to another. This help us at a better position for understanding the regulation mechanism in an organism in response to environment signals.

4.3 Real data analysis

In this section, we analyze a real data set of time course gene expression from the surgery study of a rabbit bilateral vein graft construct (Fernandez et al., 2004; Jiang et al., 2004) to demonstrated the application of the proposed procedure for GRN reconstruction.There are two different environments in this study, low and high blood flows (differing by 6-fold) conditions created in vein grafts for New Zealand White rabbits (weighing 3.0-3.5 kg). They are from the treatment of bilateral jugular vein interposition grafting and unilateral distal carotid artery branch ligation, respectively. With a segment of the vein retained at the time of implantation for baseline morphometric measurements, vein grafts, exposed to either high or low flow, were harvested at six time points ranging from 2 hours to 28 days. We have a 35,000 feature microarray chip that covers the entire rabbit genome. There are a number of genes that have been assigned hundreds to thousands of

probe sets on the array. Using a more focused microarray dataset by collapsing all corresponding probe sets into a single gene expression profile makes the most sense. Therefore, we applied our clustering model proposed in Chapter 2 on the condensed data set containing 9272 unique genes by collapsing the 35,000+ probe sets. 29 distinct clusters have been identified according BIC criterion. The profile plot of these clusters is shown in Fig. 4.1 and Fig. 4.2

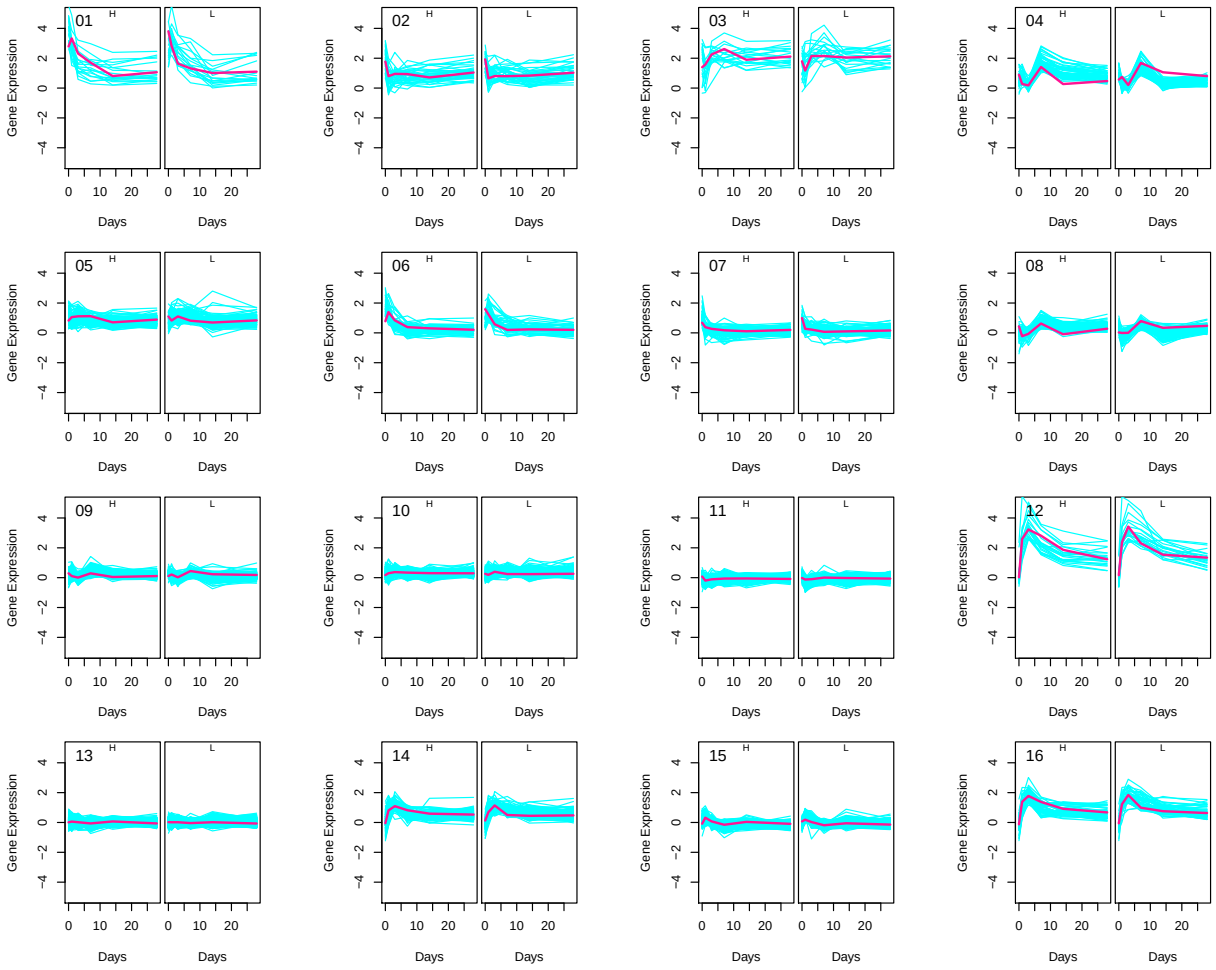


Figure 4.1: Expression trajectories of gene clusters 1 - 16 under high (H) and low flow (L).

It can be seen that these clusters each display different patterns of environment-induced changes in gene expression trajectories. SCAD method with coordinate

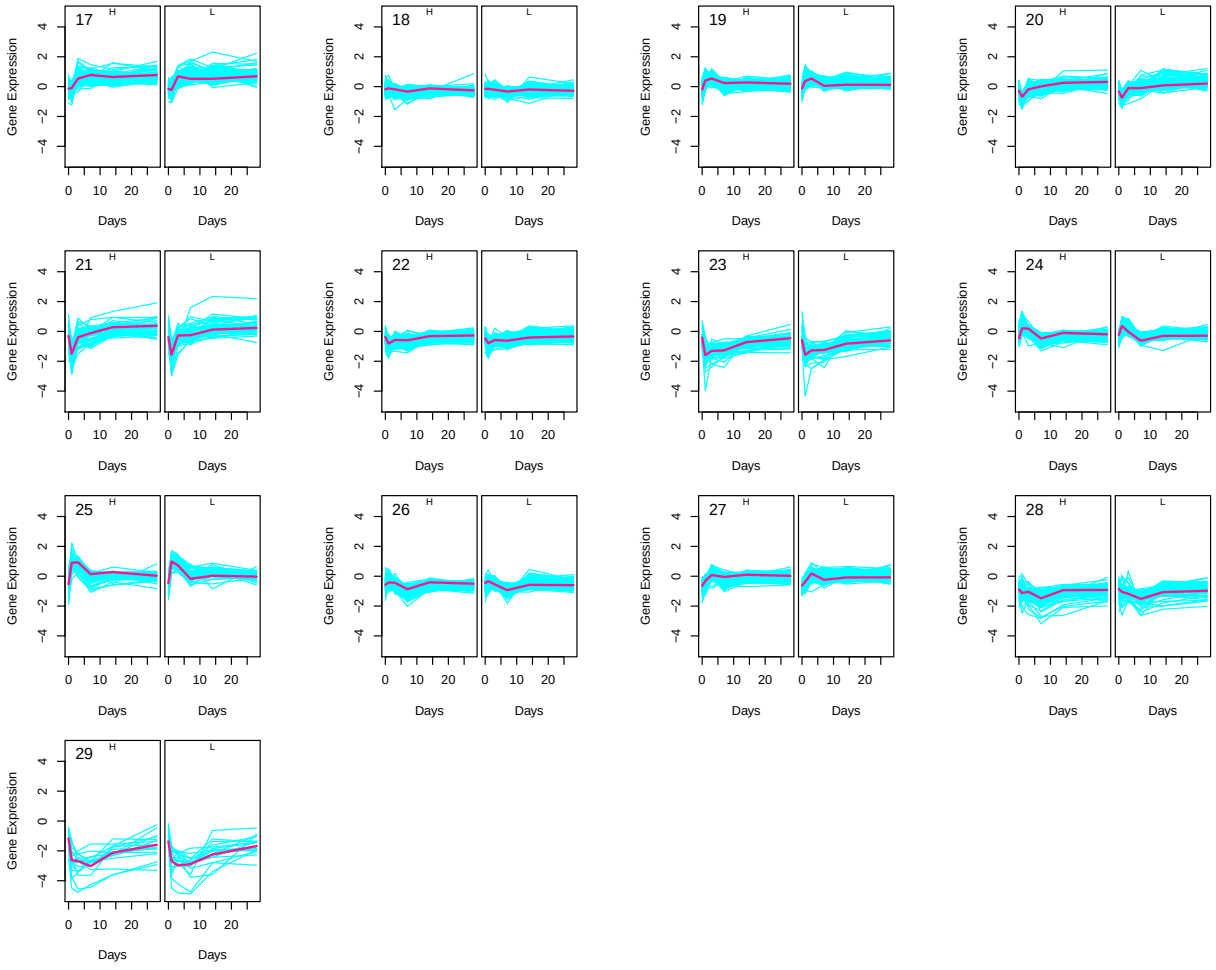
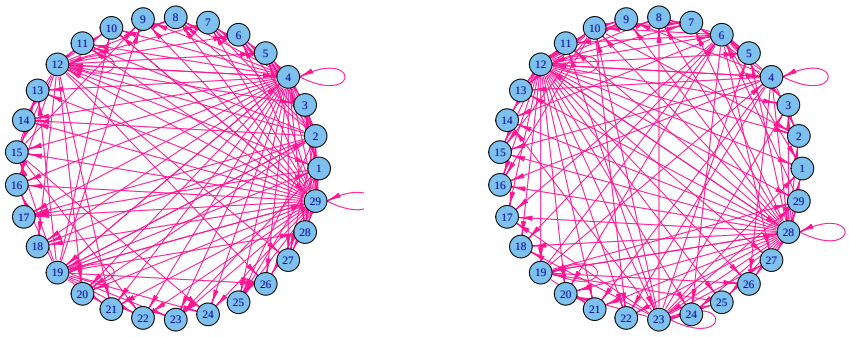


Figure 4.2: Expression trajectories of gene clusters 17 - 29 under high (H) and low flow (L)

descent algorithm is then applied to inference significant regulation effects between clusters for both high and low flows conditions.

Fig. 4.3 illustrates two different networks of gene expression under high and low flows. The nature of sparsity of GRN i.e. a target cluster could only have a few regulator clusters, can be observed in both GRNs. In most case, the number of regulator clusters for a specific target is between two and six. In addition, we can observe that there are some main driving clusters in a GRN to serving as regulators for many other clusters. For example, cluster 4 and cluster 12 are main



(a) High flow

(b) Low flow

Figure 4.3: GRN for high flow (a) and low flow (b).

driving clusters under both high and low flow. It is interesting to see that the structure of GRN is similar between the two flows, although with some extent of difference if we observe carefully. Detailed regulatory effects between 29 clusters are summarized in Table 4.1. We can see that, under low flow condition, there is one more driving cluster 28 (Fig. 4.3b), while it only works as target cluster under high flow. It suggests that cluster 28 is more active under low flow condition which can also be observed in Fig. 4.1. Meanwhile, cluster 29 is more active under high flow condition.

It has been shown that intimal hyperplasia is a main factor leading to the failure of vein bypass graft (Jiang et al., 2004). There are several processes involving in the development of intermal hyperplasia after the vein implantation surgical including endothelial cells (EC) apoptosis, smooth muscle cells (SMC) apoptosis, EC proliferation, SMC proliferation, Proteoglycan synthesis, Collagen synthesis, SMC migration, Matrix degradation and Monocyte influx, among others. We notice that some processes such as EC apoptosis and SMC apoptosis happen within hours after the implantation while some processes like Proteoglycan synthesis starts a

Table 4.1: Regulatory effects between 29 clusters under high and low flow conditions. The regulators of a gene specified in column 1 are listed in colum 2 and 4 for high and low flow conditions, respectively. The regulation targets of a gene specified in column 1 are listed in colum 3 and 5 for high and low flow conditions, respectively.

	High Flow		Low Flow	
Cluster	Regulator	Target	Regulator	Target
1	4,12,19,29	17,23,25	6,12,23,28	
2	4,19,29	3,5,6,7,8,11,12,14,17, 18,19,20,25,27,28,29	4,10,12,28	
3	2,4,12,29		6,12,23,28	
4	4,12,19,29	1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16,17, 18,19,20,21,22,23,24, 25,26,27,28	4,12,19	2,4,6,7,9,12,13,15,16, 21,22,26,28
5	2,4,12,29		6,12,23,28	
6	2,4,12,29		4,12,19,29	1,3,5,8,10,11,14, 17,18,20,24,27
7	2,4,12,29		4,10,12,28	
8	2,4,12,29		6,12,23,28	
9	4,15,19,29		4,10,12,28	
10	4,12,19,29	29	6,12,23,28	2,7,9,12,13,15,16,22
11	2,4,12,29		6,12,23,28	
12	2,4,12,29	1,3,4,5,6,7,8,10,11, 12,14,15,16,18,20, 21,26,27	4,10,12,28	1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16, 17,18,19,20,21,22,23, 24,25,26,27,28,29
13	4,15,19,29		4,10,12,28	
14	2,4,12,29		6,12,23,28	
15	4,12,19,29	9,13,22,24	4,10,12,28	
16	4,12,19,29		4,10,12,28	
17	1,2,4,29		6,12,23,28	
18	2,4,12,29		6,12,23,28	
19	2,4,19,29	1,2,4,9,10,13,15,16,19, 21,22,23,24,26,28,29	12,19,25,28	4,6,19,21,26,28,29
20	2,4,12,29		6,12,23,28	
21	4,12,19,29		4,12,19,28	
22	4,15,19,29		4,10,12,28	
23	1,4,19,29		12,23,28	1,3,5,8,10,11,14,17,18, 20,23,24,25,27,29
24	4,15,19,29		6,12,23,28	
25	1,2,4,29		12,23,28	19
26	4,12,19,29		4,12,19,28	
27	2,4,12,29		6,12,23,28	
28	2,4,19,29		4,12,19,28	1,2,3,5,7,8,9,10,11, 12,13,14,15,16,17,18, 19,20,21,22,23,24,25, 26,27,28,29
29	2,10,19,29	1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16,17, 18,19,20,21,22,23,24, 25,26,27,28,29	12,19,23,28	6

few days after implantation. In addition, it takes several weeks for the process of Matrix degradation to be getting active from the implantation. From Fig. 4.1 we can see that gene clusters 01, 12 and 16 have high expression in the early stage after

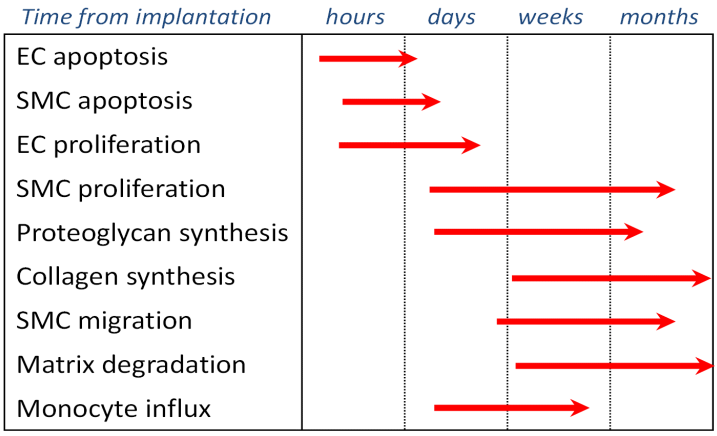


Figure 4.4: Processes related to intermal hyperplasia after bypass implantation surgical

implantation. Therefore, it is reasonable to conclude that they may play important roles to the process of apoptosis. Similarly, gene clusters 04 and 08 could be related to the process of Proteoglycan systhesis which happens a few days later than the implantation. Under low flow condition, cluster 12 is the regulator of cluster 4 (Fig. 4.3) but this relationship does not exist under high flow condition. This may indicate that the relationship between process of apoptosis and the process of proteoglycan systhesis is weaker when the blood flow is low. Gene Ontolgoy Consortium is a good resource for us to study the biological functions of genes in clusters. Also, should the experimental biologist be suggested to explore the genes in the aforementioned clusters and the corresponding functions could be confirmed.

4.4 Simulation study

In this section, we performed computer simulation to evaluate the properties our model. The design of simulation study is based on a randomly generated ODE network (Fig. 4.5). For each of the regulatory relationships, we randomly assigned a possible regulation strength picked from 1, 1.6, 1.8, -1.5 or -1.8. The coupled

ODEs are set up for each of the target clusters.

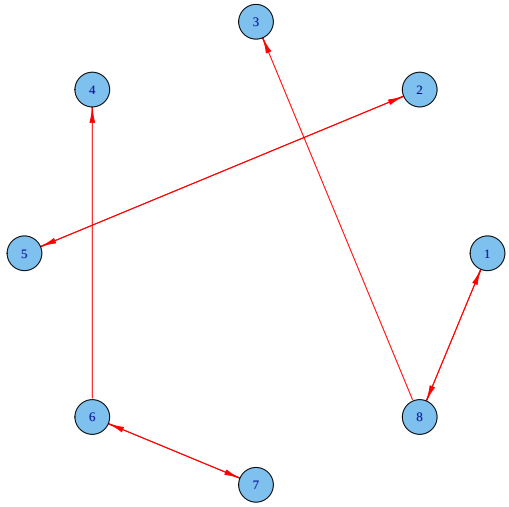


Figure 4.5: A simulated ODE network.

Based on the network shown in Fig. 4.5, the ODEs are set up as following.

$$\frac{dG_1(t)}{dt} = -1.5G_8(t), \tag{4.18}$$
$$\frac{dG_2(t)}{dt} = 1.8G_5(t), \tag{4.19}$$
$$\frac{dG_3(t)}{dt} = -1.5G_8(t), \tag{4.20}$$
$$\frac{dG_4(t)}{dt} = 1.8G_6(t), \tag{4.21}$$
$$\frac{dG_5(t)}{dt} = -1.8G_2(t), \tag{4.22}$$
$$\frac{dG_6(t)}{dt} = -1G_7(t) \tag{4.23}$$
$$\frac{dG_7(t)}{dt} = 1.8G_6(t), \tag{4.24}$$
$$\frac{dG_8(t)}{dt} = 1.6G_1(t). \tag{4.25}$$

After setting the initial values of each state variables $G1$-$G8$ by sampling from

a normal distribution with mean 3 and variance 5, we solve the ODE system (4.18)-(4.25) and generate the mean curves $\bar{G}_k(t)$ for all 8 clusters. For each observation in a cluster, we combine its random departure from the mean trajectory and measurement error with a random variable $\varepsilon_k(t_{ij})$ which follows a normal distribution with mean zero and variance σ^2 and $\sigma^2 = 0.10$ or 1.00. Therefore, a observation is simulated from

$$y_k(t_{ij}) = \bar{G}_k(t_{ij}) + \varepsilon_k(t_{ij}), i = 1, 2, \cdots, T, \quad j = 1, 2, \cdots, q_k, \tag{4.26}$$

where T is the number of time points and we set $T = 15, 25$ or 100, q_k is the number of gene in each cluster.

As shown in Chapter 2, LOP-based functional clustering performs very well in classifying genes into distinct clusters. Here, we focus on the inference of GRN from our procedure. To evaluate the performance of our method, we use the two metrics defined in Chapter 3, positive predictive value (PPV) and false negative rate (FNR).

$$PPV = \frac{TP}{TP + FP} \tag{4.27}$$

$$FNR = \frac{FN}{TP + FN} \tag{4.28}$$

Essentially, PPV is the proportion of TPs in the reconstructed network and FNR is the proportion of TPs which are not identified successfully Therefore, we prefer a higher PPV and a lower FNR.

We generated 100 sets of expressions for the randomly simulated network and

compared the reconstructed networks with true ones to obtain PPV and FNR.

Table 4.2: Simulation results for GRN reconstruction based on functional clustering by using SCAD with coordinate descent algorithm. The values of average of PPV and FNR are obtained from 100 simulated networks for each of different settings of sample sizes and noise levels. The numbers in parenthesis are corresponding standard deviations.

	$T = 15$		$T = 25$		$T = 100$	
Noise level	PPV	FNR	PPV	FNR	PPV	FNR
$\sigma^2 = 0.10$	0.52(0.137)	0.17(0.114)	0.80(0.020)	0.28(0.014)	0.80(0.029)	0.01(0.029)
$\sigma^2 = 1.00$	0.38(0.138)	0.34(0.144)	0.31(0.073)	0.17(0.108)	0.73(0.090)	0.01(0.017)

The simulation results are summarized in Table 4.2. We can see that, when the number of time points is large ($T = 100$) and the noise level is very low ($\sigma^2 = 0.10$), the PPV is high (0.80) and the FNR is low (0.01). When the number of time points is decreased to 25, the PPV is still 0.80 but the FNR is increased to 0.28. When we have the smallest sample size ($T = 15$) and largest noise level $\sigma^2 = 1.00$, the PPV is reduced to 0.38 and the FNR is increased to 0.34. The aforementioned results indicated that our model works great when the sample size is large and noise level is low. It still perform well even when the sample size is reduced and the noise level is getting larger.

4.5 Discussion

Many biological processes including plant and animal development are coordinated by cell-to-cell communication regulated by genes (Chen et al., 2003). High-throughput measurement techniques have now led to the identification of tens of thousands of genes involved in sensing external cues. However, the dynamic interplay between genes is highly complex and cannot be understood by a simple approach (Sivriver et al., 2011). The reconstruction of gene regulatory networks can be a valuable tool for identifying the key mechanisms that shape the dynamics

of cellular and transcriptional processes (Zhu et al., 2012; Hecker et al., 2009).

It is getting popular to model biological regulatory networks regulated by gene expression using Ordinary Differential Equation approach (Lu et al., 2011). However, the reality of lacking of replicates for gene measurements limits its widespread application.

In this chapter, we proposed an ODE model for inference of gene regulatory network based on functional clustering method. This model has several advantage. We cluster genes and estimate the mean functions simultaneously then make network inference on gene clusters instead of individual gene to deal with the problem of small number of replicates. Furthermore, we have explicit forms of mean functions as well as its derivative thereby we can have precise estimation of regression coefficients. Consequently, parameter polish in Lu et al. (2011) model is not needed. Therefore, our model is more efficient with only three steps to establish gene regulatory network: clustering gene into functional groups, model selection to detect significant regulation effects and analyzing gene functions. In addition, we take into account the environmental signals in our clustering framework. This enable us to explore the difference of regulation effects among clusters under distinct environments. Also, the information of transcriptional time lag is incorporated into the model to help detecting true regulation effects.

In the past years, the accuracy of high-throughput sequencing have been improved greatly by the advance of technology and the cost of large-scale data generation have been reduced for the required sample size as well. These developments have enabled ones to unravel the regulatory interactions that allow genomes to regulate cellular structure, function, and behavior based on dissection of genomes. Given the popularity of collecting and using high-throughput expression data in human and other systems, the new model proposed in this chapter will find its

widespread application. Thought the model emphasizes on transcriptional data, it can be refined and extended to integrate multiple data types, such as mRNA and microRNA (miRNA) expression data, TF DNA-binding data, and protein interaction data (Bolouri, 2014). To identify phenotype- or disease-causing perturbations, the model should be linked to complex phenotypes or diseases within a causal-effect network framework.

Chapter 5
Summary and future work

5.1 Summary

In this dissertation, we introduce a clustering framework to integrates developmental and environment-dependent programs of gene expression. We successfully integrate environmental factor into clustering framework that enable us to perform statistical comparison of gene expression in distinct environments. In addition, we reformed DBN modeling for GRN reconstruction by interpolating missing data points based on functional clustering. The new model can handle any dynamic gene expression data, no matter they are evenly spaced or not, thereby providing a broader tool in computational biology. Furthermore, we proposed an ODE model for inference of gene regulatory network dealing with the continuous gene expression measurements directly. This model enjoys several good properties; clustering genes and estimating the mean curves for each cluster are conducted simultaneously, we have explicit forms of mean functions as well as its derivative thereby we can have precise estimation of regression coefficients. Benefiting from our functional clustering framework which integrate environmental conditions into the model, we are able to analyze the change of regulation effects between genes in response to environmental signals.

Extensive simulation studies have been conducted to validate the performance of proposed models and all model have been applied on real data set to demonstrate their usefulness and utility.

The models emphasize on transcriptional data, but can be refined and extended to integrate multiple data types, such as mRNA and microRNA (miRNA) expression data, TF DNA-binding data, and protein interaction data. Also, the model should be linked to complex phenotypes or diseases within a causal-effect network framework toward identifying phenotype- or disease-causing perturbations.

5.2 Future work

5.2.1 Integrating prior knowledge into ODE model

Prior knowledge on biological mechanisms can help to construct differential equations even for nonlinear Weaver et al. (1999); Sakamoto and Iba (2001); Spieth et al. (2006). However, we usually have little such prior information. This is one reason of why linear ODE model is popular. Fortunately, we do obtain some information from clustering. We consider only genes have either earlier or simultaneous expression changes (up- or down-regulation) compared to targets could be potential regulators. In other word, the genes which have a flat expression profile and do not have up-regulation or down-regulation expression could not be a regulator nor a target. Our DBN model has already take this advantage. For the rabbit data in Chapter 2, cluster C, D and E have flat profile patterns. They are not regulator nor target in the gene network constructed by DBN model.

In the ODE model, every cluster has its own regression model so it is likely to identify significant regulation effects for every cluster. We observe this in the results of Chapter 4 (Figure 4.3). We can see every cluster has its regulator but

actually several clusters, such as cluster 11, 13 and 18, have flat profile and do not express during the experiment period. Therefore, we may work on integrate clustering information into ODE model to construct more reasonable differential equations.

Also, we may improve our clustering model to have adaptive orders for each cluster to avoid over-fitting flat clusters with high order of polynomials. Then these flat clusters could be excluded from the regression models automatically.

5.2.2 Working on individual genes instead of gene clusters

There are several advantage to work on gene clusters for GRN study. First, it enable us to make inference about gene expression when there is no or little measurement replicate. Due to the cost of biological experiments, we usually have 2 or 3 measurement replicates for real data sets in the past time. The aim of replicates is to reduce measurement errors. It is hard to make statistical inference based on such a small sample size. With the method of clustering, we group gene into clusters and consider the expression of genes with a cluster are iid samples. Therefore, we can have a larger sample size to estimate more accurate mean curves and then to reconstruct GRN. Second, the dimensionality of problem could be reduced dramatically because the number of cluster is usually much smaller than the number of genes. Consequently, more sparse networks are expected.

However, it is hard to give interpretation for a GRN based on cluster. In a gene regulatory network, a node could be a cluster of genes, because several genes could be regulated by the same set of regulators. However, there are 14,945 genes in the real data set, and the average number of in one cluster is about 515. It is hard to explain that 515 genes can be regulated by a same set of regulators.

With the fast development of high-throughput technology, the cost of experiment

of gene expression is getting cheaper and cheaper. It is possible for us to have data set with more replicates for each gene. Therefore, we are able to work on individual genes rather than gene cluster to make inference of GRN. However, it is common that the number of genes be more than a million. If we apply ODE model with a regression setting, then we would encounter a ultra-high dimensional issues which is very challenging for variable selection models.

Fan and Lv (2008) developed sure independence screening (SIS) for ultra-high dimensional. The dimensionality of the problem could be reduced from a extreme large scale $\tilde{p}$ to a relatively large scale p by using this variable screening technical. Consequently, an existing variable selection method could be applied on the screened data. Under some technical conditions including iid sample, it can be shown that sure independence screening enjoys the sure screening property which ensure that all the important variables will be retained in the the reduced p-dimensional model with asymptotic probability one. In the ODE regression setting, the condition of iid sample does not hold. We believe that SIS still could be applied though some theoretical properties of it need to be verified.

In summary, when we have gene expression data with reasonable number of replicates for each gene, our ODE model still can be used with the help of SIS technology. In this case, we could have better interpretation of the reconstructed network.

5.2.3 Models for next-generation sequence

All of our three models introduced in this proposal works on micro array data. However, next generation technology like RNA sequencing is getting more and more popular in modern gene expression studies. RNA-seq directly counts the number of reads that map to the transcript while microarray measures a transcript by a

continuous variable. Comparing to microarray, RNA-seq has several advantages, including low background noise, large dynamic range of expression levels, and the ability to measure the expression levels of unannotated genomic sequence.

We have been working on discrete RNA-seq data with models based on Poisson, Skellam or Beta distribution. It may be straightforward to adopt the idea of DBN on discrete data due to the fact that even continuous data need to be discretized in DBN model. However, ODE model is originally working on continuous data thereby reasonable transformation is needed to apply it on RNA-seq data. Efforts should be made on normalization, clustering and network models for next-generation sequence data.

Bibliography

Ahdesmäki, M., Lähdesmäki, H., Pearson, R., Huttunen, H., and Yli-Harja, O. (2005). Robust detection of periodic time series measured from biological systems. *BMC bioinformatics*, 6(1):117.

Ahn, K., Luo, j., Berg, A., Keefe, D., and Wu, R. L. (2010). Functional mapping of drug response with pharmacodynamic-pharmcokinetic principles. *Trend Pharmacol Sci*, 31(7):306–311.

Akutsu, T., Miyano, S., and Kuhara, S. (2000). Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8):727–734.

Aluru, S. (2005). *Handbook of computational molecular biology*. CRC Press.

Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W., and White, K. P. (2002). Gene expression during the life cycle of drosophila melanogaster. *Science*, 297(5590):2270–2275.

Bansal, M., Della Gatta, G., and Di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822.

Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S., and Simon, I. (2003). Continuous representations of time-series gene expression data. *Journal of Computational Biology*, 10(3-4):341–356.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

Bolouri, H. (2014). Modeling genomic regulatory networks with big data. *Trends in Genetics*, 30(5):182–191.

BORGWARDT, K. M., VISHWANATHAN, S., and KRIEGEL, H.-P. (2006). Class prediction from time series gene expression profiles using dynamical systems

kernels. In *Pacific Symposium on Biocomputing*, volume 11, pages 547–558. Citeseer.

Bornholdt, S. (2008). Boolean network models of cellular regulation: prospects and limitations. *Journal of the Royal Society Interface*, 5(Suppl 1):S85–S94.

Box, G. E., Jenkins, G. M., and Reinsel, G. C. (2013). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brazhnik, P., de la Fuente, A., and Mendes, P. (2002). Gene networks: how to put the function in genomics. *TRENDS in Biotechnology*, 20(11):467–472.

Breeden, L. L. (2003). Periodic transcription: a cycle within a cycle. *Current Biology*, 13(1):R31–R38.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232.

Brockwell, P. J. and Davis, R. A. (2009). *Time series: theory and methods*. Springer.

Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with dna microarrays. *Nature genetics*, 21:33–37.

Chen, D., Toone, W. M., Mata, J., Lyne, R., Burns, G., Kivinen, K., Brazma, A., Jones, N., and Bähler, J. (2003). Global transcriptional responses of fission yeast to environmental stress. *Molecular biology of the cell*, 14(1):214–229.

Chen, J. and Wu, H. (2008a). Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to hiv-1 dynamics. *Journal of the American Statistical Association*, 103(481):369–384.

Chen, J. and Wu, H. (2008b). Estimation of time-varying parameters in deterministic dynamic models. *Statistica Sinica*, 18(3):987–1006.

Cohen, A. A., Geva-Zatorsky, N., Eden, E., Frenkel-Morgenstern, M., Issaeva, I., Sigal, A., Milo, R., Cohen-Saidon, C., Liron, Y., Kam, Z., et al. (2008). Dynamic proteomics of individual cancer cells in response to a drug. *science*, 322(5907):1511–1516.

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1339.

Crick, F. et al. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Cui, Y., Wu, R., Casella, G., and Zhu, J. (2008). Nonparametric functional mapping of quantitative trait loci underlying programmed cell death. *Statistical applications in genetics and molecular biology*, 7(1).

Cui, Y., Zhu, J., and Wu, R. (2006). Functional mapping for genetic control of programmed cell death. *Physiological genomics*, 25:458.

Daub, C. O., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data. *BMC bioinformatics*, 5(1):118.

De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103.

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727.

De Smet, I., Lau, S., Mayer, U., and Jürgens, G. (2010). Embryogenesis–the humble beginnings of plant life. *The Plant Journal*, 61(6):959–970.

D'haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R. (1999). Linear modeling of mrna expression levels during cns development and injury. In *Pacific symposium on biocomputing*, volume 4, pages 41–52.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of longitudinal data*. Oxford University Press.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Ernst, J., Nau, G. J., and Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, 21(suppl 1):i159–i168.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Yao, Q. (2003). *Nonlinear time series*, volume 2. Springer.

Fernandez, C. M., Goldman, D. R., Jiang, Z., Ozaki, C. K., Tran-Son-Tay, R., and Berceli, S. A. (2004). Impact of shear stress on early vein graft remodeling: a biomechanical analysis. *Annals of biomedical engineering*, 32(11):1484–1493.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620.

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., et al. (2012). Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600.

Gneiting, T., Genton, M., and Guttorp, P. (2007). Geostatistical space-time models, stationarity, separability and full symmetry. *Statistical Methods for Spatio-Temporal Systems*, pages 151–175.

Godsey, B. (2013). Improved inference of gene regulatory networks through integrated bayesian clustering and dynamic modeling of time-course expression data. *PloS one*, 8(7):e68358.

Greenfield, A., Madar, A., Ostrer, H., and Bonneau, R. (2010). Dream4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS one*, 5(10):e13397.

Guiot, C., Degiorgis, P. G., Delsanto, P. P., Gabriele, P., and Deisboeck, T. S. (2003). Does tumor growth follow a âĂIJuniversal lawâĂİ? *Journal of theoretical biology*, 225(2):147–151.

Guiot, C., Delsanto, P. P., Carpinteri, A., Pugno, N., Mansury, Y., and Deisboeck, T. S. (2006). The dynamic evolution of the power exponent in a universal growth model of tumors. *Journal of theoretical biology*, 240(3):459–463.

Haddad, J. N. (2004). On the closed form of the covariance matrix and its inverse of the causal arma process. *Journal of Time Series Analysis*, 25(4):443–448.

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models–a review. *Biosystems*, 96(1):86–103.

Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V., and Banavar, J. R. (2001). Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences*, 98(4):1693–1698.

Hunter, D. R. and Li, R. (2005). Variable selection using mm algorithms. *Annals of statistics*, 33(4):1617.

Hurley, D., Araki, H., Tamada, Y., Dunmore, B., Sanders, D., Humphreys, S., Affara, M., Imoto, S., Yasuda, K., Tomiyasu, Y., et al. (2012). Gene network inference and visualization tools for biologists: application to new human transcriptome datasets. *Nucleic acids research*, 40(6):2377–2398.

Inoue, L. Y., Neira, M., Nelson, C., Gleave, M., and Etzioni, R. (2007). Cluster-based network model for time-course gene expression data. *Biostatistics*, 8(3):507–525.

Jiang, Z., Wu, L., Miller, B. L., Goldman, D. R., Fernandez, C. M., Abouhamze, Z. S., Ozaki, C. K., and Berceli, S. A. (2004). A novel vein graft model: adaptation to differential flow environments. *American Journal of Physiology-Heart and Circulatory Physiology*, 286(1):H240–H245.

Kim, B.-R., McMurry, T., Zhao, W., Wu, R., and Berg, A. (2010). Wavelet-based functional clustering for patterns of high-dimensional dynamic gene expression. *Journal of Computational Biology*, 17(8):1067–1080.

Kim, B.-R., Zhang, L., Berg, A., Fan, J., and Wu, R. (2008). A computational approach to the functional clustering of periodic gene-expression profiles. *Genetics*, 180(2):821–834.

Kingsolver, J. G. and Woods, H. A. (1997). Thermal sensitivity of growth and feeding in manduca sexta caterpillars. *Physiological and Biochemical Zoology*, 70(6):631–638.

Kojima, K., Yamaguchi, R., Imoto, S., Yamauchi, M., Nagasaki, M., Yoshida, R., Shimamura, T., Ueno, K., Higuchi, T., Gotoh, N., et al. (2009). A state space representation of var models with sparse learning for dynamic gene networks. *Genome Informatics*, 22:56–68.

Landry, C. R., Oh, J., Hartl, D. L., and Cavalieri, D. (2006). Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene*, 366(2):343–351.

Li, N., Das, K., and Wu, R. (2009). Functional mapping of human growth trajectories. *Journal of theoretical biology*, 261(1):33–42.

Li, N., McMurry, T., Berg, A., Wang, Z., Berceli, S. A., and Wu, R. (2010). Functional clustering of periodic transcriptional profiles through arma (p, q). *PloS one*, 5(4):e9894.

Li, Y., Álvarez, O. A., Gutteling, E. W., Tijsterman, M., Fu, J., Rikcosh, J. A., Hazendonk, E., Prins, P., Plasterk, R. H., Jansen, R. C., et al. (2006). Mapping determinants of gene expression plasticity by genetical genomics in c. elegans. *PLoS genetics*, 2(12):e222.

Li, Z., Li, P., Krishnan, A., and Liu, J. (2011). Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis. *Bioinformatics*, 27(19):2686–2691.

Liang, H. and Wu, H. (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484).

Lin, M. and Wu, R. (2006). A joint model for nonparametric functional mapping of longitudinal trajectory and time-to-event. *BMC bioinformatics*, 7(1):138.

Lin, M., Zhao, W., and Wu, R. (2006). A statistical framework for genetic association studies of power curves in bird flight. *Biological procedures online*, 8(1):164–174.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature genetics*, 21:20–24.

Lockhart, D. J. and Winzeler, E. A. (2000). Genomics, gene expression and dna arrays. *nature*, 405(6788):827–836.

Lu, T., Liang, H., Li, H., and Wu, H. (2011). High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association*, 106(496).

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics*, 19(4):474–482.

Ma, P., Castillo-Davis, C. I., Zhong, W., and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, 34(4):1261–1269.

Marie, H. and Pranab, K. S. (1985). On sequentially adaptive asymptotically efficient rank statistics: Rank statistics. *Sequential Analysis*, 4(3):125–151.

Martin, S., Zhang, Z., Martino, A., and Faulon, J.-L. (2007). Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics*, 23(7):866–874.

McAdams, H. H., Srinivasan, B., and Arkin, A. P. (2004). The evolution of genetic regulatory systems in bacteria. *Nature Reviews Genetics*, 5(3):169–178.

McKay, M. D. (1997). Nonparametric variance-based methods of assessing uncertainty importance. *Reliability engineering & system safety*, 57(3):267–279.

Meyer, K. (2000). Random regressions to model phenotypic variation in monthly weights of australian beef cows. *Livestock Production Science*, 65(1):19–38.

Müller, H.-G., Chiou, J.-M., and Leng, X. (2008). Inferring gene expression dynamics via functional regression analysis. *BMC bioinformatics*, 9(1):60.

Murphy, K., Mian, S., et al. (1999). Modelling gene expression data using dynamic bayesian networks. Technical report, Technical report, Computer Science Division, University of California, Berkeley, CA.

Ogami, K., Yamaguchi, R., Imoto, S., Tamada, Y., Araki, H., Miyano, S., et al. (2012). Computational gene network analysis reveals tnf-induced angiogenesis. *BMC systems biology*, 6(Suppl 2):S12.

Park, T., Yi, S.-G., Lee, S., Lee, S. Y., Yoo, D.-H., Ahn, J.-I., and Lee, Y.-S. (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, 19(6):694–703.

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. (1996). Hiv-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586.

Quint, M., Drost, H.-G., Gabel, A., Ullrich, K. K., Bönn, M., and Grosse, I. (2012). A transcriptomic hourglass in plant embryogenesis. *Nature*, 490(7418):98–101.

Ramoni, M. F., Sebastiani, P., and Kohane, I. S. (2002). Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences*, 99(14):9121–9126.

Richards, F. (1959). A flexible growth function for empirical use. *Journal of experimental Botany*, 10(2):290–301.

Rustici, G., Mata, J., Kivinen, K., Lió, P., Penkett, C. J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and Bähler, J. (2004a). Periodic gene expression program of the fission yeast cell cycle. *Nature genetics*, 36(8):809–817.

Rustici, G., Mata, J., Kivinen, K., Lió, P., Penkett, C. J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and Bähler, J. (2004b). Periodic gene expression program of the fission yeast cell cycle. *Nature genetics*, 36(8):809–817.

Sakamoto, E. and Iba, H. (2001). Inferring a system of differential equations for a gene regulatory network by using genetic programming. In *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, volume 1, pages 720–726. IEEE.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470.

Seshasayee, A. S., Bertone, P., Fraser, G. M., and Luscombe, N. M. (2006). Transcriptional regulatory networks in bacteria: from input signals to output responses. *Current opinion in microbiology*, 9(5):511–519.

Shalon, D., Smith, S. J., and Brown, P. O. (1996). A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome research*, 6(7):639–645.

Shedden, K. and Cooper, S. (2002). Analysis of cell-cycle gene expression in saccharomyces cerevisiae using microarrays and multiple synchronization methods. *Nucleic Acids Research*, 30(13):2920–2929.

Shimamura, T., Imoto, S., Yamaguchi, R., Fujita, A., Nagasaki, M., and Miyano, S. (2009). Recursive regularization for inferring gene networks from time-course gene expression profiles. *BMC Systems Biology*, 3(1):41.

Silverman, B. W. et al. (1984). Spline smoothing: the equivalent variable kernel method. *The Annals of Statistics*, 12(3):898–916.

Sivriver, J., Habib, N., and Friedman, N. (2011). An integrative clustering and modeling algorithm for dynamical gene expression data. *Bioinformatics*, 27(13):i392–i400.

Smith, E. N. and Kruglyak, L. (2008). Gene–environment interaction in yeast gene expression. *PLoS biology*, 6(4):e83.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297.

Spieth, C., Hassis, N., and Streichert, F. (2006). Comparing mathematical models on the problem of network inference. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 279–286. ACM.

Thomas, R. (1973). Boolean formalization of genetic control circuits. *Journal of theoretical biology*, 42(3):563–585.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tobalske, B., Hedrick, T., Dial, K., and Biewener, A. (2003). Comparative power curves in bird flight. *Nature*, 421(6921):363–366.

Voit, E. O. (2000). *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists*. Cambridge University Press.

von Bertalanffy, L. (1957). Quantitative laws for metabolism and growth. *The Quarterly Review of Biology*, 32(1):217–231.

Wang, J., Chen, B., Wang, Y., Wang, N., Garbey, M., Tran-Son-Tay, R., Berceli, S. A., and Wu, R. (2013). Reconstructing regulatory networks from the dynamic plasticity of gene expression by mutual information. *Nucleic acids research*, 41(8):e97–e97.

Watson, J. D., Crick, F. H., et al. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738.

Weaver, D. C., Workman, C. T., Stormo, G. D., et al. (1999). Modeling regulatory networks with weight matrices. In *Pacific symposium on biocomputing*, volume 4, pages 112–123.

Wessels, L. F., van Someren, E. P., Reinders, M. J., et al. (2001). A comparison of genetic network models. In *pacific Symposium on Biocomputing*, volume 6, pages 508–519.

West, G. B., Brown, J. H., and Enquist, B. J. (2001). A general model for ontogenetic growth. *Nature*, 413(6856):628–631.

Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O., et al. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell*, 13(6):1977–2000.

Wittkopp, P. J. (2007). Variable gene expression in eukaryotes: a network perspective. *Journal of Experimental Biology*, 210(9):1567–1575.

Wu, H., Lu, T., Xue, H., and Liang, H. (2014). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109(506):700–716.

Wu, J., Zeng, Y., Huang, J., Hou, W., Zhu, J., and Wu, R. (2007). Functional mapping of reaction norms to multiple environmental signals. *Genetical research*, 89(01):27–38.

Xiang, D., Venglat, P., Tibiche, C., Yang, H., Risseeuw, E., Cao, Y., Babic, V., Cloutier, M., Keller, W., Wang, E., et al. (2011). Genome-wide analysis reveals gene expression and metabolic network dynamics during embryo development in arabidopsis. *Plant physiology*, 156(1):346–356.

Yang, R., Gao, H., Wang, X., Zhang, J., Zeng, Z.-B., and Wu, R. (2007). A semi-parametric approach for composite functional mapping of dynamic quantitative traits. *Genetics*, 177(3):1859–1870.

Yap, J. S., Li, Y., Das, K., Li, J., and Wu, R. (2011). Functional mapping of reaction norms to multiple environmental signals through nonparametric covariance estimation. *BMC plant biology*, 11(1):23.

Yeung, M. S., Tegnér, J., and Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, 99(9):6163–6168.

Yosef, N., Shalek, A. K., Gaublomme, J. T., Jin, H., Lee, Y., Awasthi, A., Wu, C., Karwacz, K., Xiao, S., Jorgolli, M., et al. (2013). Dynamic regulatory network controlling th17 cell differentiation. *Nature*, 496(7446):461–468.

Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., and Jarvis, E. D. (2004). Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594–3603.

Zhang, J. (2013). Epistatic clustering: a model-based approach for identifying links between clusters. *Journal of the American Statistical Association*, 108(504):1366–1384.

Zhang, X., Liu, K., Liu, Z.-P., Duval, B., Richer, J.-M., Zhao, X.-M., Hao, J.-K., and Chen, L. (2013). Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics*, 29(1):106–113.

Zhang, X., Zhao, J., Hao, J.-K., Zhao, X.-M., and Chen, L. (2014). Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic acids research*, page gku1315.

Zhao, L. P., Prentice, R., and Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proceedings of the National Academy of Sciences*, 98(10):5631–5636.

Zhao, W., Hou, W., Littell, R. C., and Wu, R. (2005). Structured antedependence models for functional mapping of multiple longitudinal traits. *Genetics and Molecular Biology*, 4(1):33.

Zhao, W., Ma, C.-X., Cheverud, J. M., and Wu, R. (2004a). A unifying statistical model for qtl mapping of genotype-sex interaction for developmental trajectories. *Physiol. Genomics*, 19(2):218–227.

Zhao, W., Zhu, J., M., G.-M., and Wu, R. L. (2004b). A unified statistical model for functional mapping of genotype environment interactions for ontogenetic development. *Genetics*, 168:1567–1575.

Zhu, H., Rao, R. S. P., Zeng, T., and Chen, L. (2012). Reconstructing dynamic gene regulatory networks from sample-based transcriptional data. *Nucleic acids research*, 40(21):10657–10667.

Zimmerman, D. L., Núñez-Antón, V., Gregoire, T. G., Schabenberger, O., Hart, J. D., Kenward, M. G., Molenberghs, G., Verbeke, G., Pourahmadi, M., Vieu, P., et al. (2001). Parametric modelling of growth curve data: an overview. *Test*, 10(1):1–73.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509.

Zou, M. and Conzen, S. D. (2005). A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–79.

Vita
Yaqun Wang

423 Thomas Building,
Penn State University,
University Park, PA 16802

Phone: (814) 321-1189
E-mail: yxw179@psu.edu
http://www.personal.psu.edu/yxw179

Education

Ph.D. in Statistics, 2010 - 2015.
Department of Statistics, Pennsylvania State University, USA
Co-Advisers: Dr. Rongling Wu, Distinguished Professor of Public Health Sciences and Statistics
Dr. Runze Li, Verne M. Willaman Professor of Statistics

M.S. in Computer Software, 1994 - 1997
Department of Computer Science and Engineering, Zhejiang University, China

B.S. in Computer Science, 1990 - 1994
Department of Computer Science and Technology, Dalian University of Technology, China

Professional Appointments

Visiting Scholar, 2009 - 2010
Center for Statistical Genetic, Pennsylvania State University, USA

Lecturer of Computer Science, 2004 - 2009
Department of Computer Science and Engineering, China Jiliang University, China

Chief Technology Officer, 1998 – 2004
Huayuan Microcomputer Co. Ltd, China

Research Associate, 1997 - 1998
Institute of Software, Chinese Academy of Sciences, China

Awards

- Outstanding Faculty Award, China Jiliang University, 2008
- The First Place Award of The 7th Lecture Presentation Competition, China Jiliang University, 2006

Research Interests

- Gene Regulatory Network Modeling
- Dynamic Gene Expression and Gene-Environment Interaction Analysis
- Computational Biology and Statistical Genetics
- Data Mining and Machine Learning