This is to certify that the
dissertation entitled

INTEGRATING GENE EXPRESSION AND METABOLIC PROFILES TO OPTIMIZE CELLULAR FUNCTIONS

presented by

ZHENG LI

has been accepted towards fulfillment
of the requirements for the

Ph.D degree in Chemical Engineering

Major Professor's Signature

Aug. 23, 2006

Date

MSU is an Affirmative Action/Equal Opportunity Institution

INTEGRATING GENE EXPRESSION AND METABOLIC PROFILES TO OPTIMIZE CELLULAR FUNCTIONS

By

ZHENG LI

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Chemical Engineering and Materials Science

2006

ABSTRACT

INTEGRATING GENE EXPRESSION AND METABOLIC PROFILES TO OPTIMIZE CELLULAR FUNCTIONS

By

ZHENG LI

With advances in high throughput technology, profiles of gene expressions, proteins and metabolites can be acquired to help elucidate the network of pathways involved in producing a specific phenotype. This dissertation presents a systems approach that was developed to integrate gene expression, metabolic and phenotypic profiles to identify active pathways that confer a phenotype. The approach involves several separate components to i) identify genes that are relevant to a cellular or metabolic process, ii) integrate multi-source information, and iii) reconstruct pathways and networks.

Approaches to identify genes relevant to a phenotype were developed using genetic algorithm coupled partial least squares analysis (GA/PLS) and discussed in Chapter 2. GA/PLS used a log linear model to identify subsets of genes that can best predict a phenotype e.g. a metabolic function. Next we applied Bayesian network analysis to infer network structures from metabolic data, discussed in Chapter 3. Metabolic data was chosen initially because if Bayesian network analysis is able to infer well-known metabolic structures, pathways e.g. TCA cycle and urea cycle, from experimental data, which provided confidence in the ability of this methodology to infer other networks, such as, genetic regulatory networks from gene data. In Chapter 4, we integrated both ideas from the previous two chapters into a Three-stage Integrative Pathway Search (*TIPS*©), which combined methods to identify relevant genes with

network reconstruction. Unlike other approaches, this approach identified the active pathways without requiring interaction measurements or libraries of genetic mutants, and with limited amount of data. The reconstructed network was validated through in silico perturbations studies with published results and further experiments. The framework provided very good predictions of the effect of some, but not all, of the perturbation studies. This may be due in part to Bayesian network analysis' inability to handle transients, such as cycles and feedback loops. Uncovering the additional information from these transients would be valuable in elucidating the mechanism involved in producing a particular phenotype. Therefore, in Chapter 5, we developed a dynamic model using time-series gene expression data. To illustrate the ability of the model, we applied the model to *Escherichia coli K12* (*E. coli)* and *Saccharomyces cerevisiae* (yeast), which were more readily amendable to validatation. Finally, Chapter 6 discussed the improvements to the current *TIPS*© approach namely, to include multiple metabolites, incorporate multi-source information, and dynamic modeling for time-series data.

DEDICATION

TO LULU
FOR HER SUPPORT AND LOVE

ACKNOWLEDGMENTS

First I would like to thank Professor Christina Chan, who gave me guidance throughout my research. Without her knowledge, patience and support this dissertation would have not been possible. She is not only my academic advisor, but also a mentor and a good friend. The knowledge and friendship I gained from her will definitely influence the rest of my life. I would also thank Professor Sarat Dass, his keen insights and valuable discussions on statistics often gave me stimulating ideas in research.

I am grateful to Professor Worden and Professor Dale for taking time to serve on the guidance committee and overseeing this work. Their insightful comments and suggestions have enhanced the technical soundness of this dissertation. I am also grateful to all my friends and colleagues. The discussions with them substantially contributed to my work and broadened my knowledge.

Last but not the least, many thanks go to my family: my wife, my Mom and Dad, and my brother. Without their encouragement and continuous support, I would not be who I am today.

This research was supported by National Science Foundation, Whitaker Foundation, Environmental Protection Agency. I was also partially supported by QBMI award and DeVlieg fellowship for Computational Engineering at Michigan State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ACC: acetyl-CoA carboxylase

ALDH: aldehyde dehydrogenase

ANOVA: analysis of variance

Bcl-2: B-cell lymphoma 2

BSA: Bovine Serum Albumin

BN: Bayesian network analysis

ChIP: chromatin immunoprecipitation

EM: expectation maximization

ETC: electron transport chain

FDA: Fischer's discriminant analysis

FFA: free fatty acid

GA: genetic algorithm

GSEA: gene set enrichment analysis

GST: glutathione-s transferase

IAP: inhibitor of apoptosis protein

ICA: independent component analysis

LDH: lactate dehydrogenase

MBPLS: multi-block partial least squares

NAC: N-acetyl-cysteine

NF-KB: nuclear factor kappa B

PLS: partial least squares analysis

PKC-delta: protein kinase C delta isoform

PKR: double stranded RNA dependent protein kinase

ROS: reactive oxygen species

SCD: stearoyl-CoA desaturase

SOD: superoxide dismutase

SSM: state-space model

TIPS: three stage integrative pathway search system

TNF: tumor necrosis factor

TG: triglyceride

TRAF: TNF receptor associated factor

CHAPTER 1 INTRODUCTION

1.1 Background

A cell is a complex multi-scale biological system, with information flowing between genes, proteins, and metabolites. Genes transcribe to proteins and some proteins are employed as enzymes to catalyze metabolic reactions, while some other proteins are employed as transcription factors. Genes can be transcriptionally regulated by both proteins (transcription factors) and metabolites. Thus, a cell is a complex system regulated by interconnected gene regulatory and signal transduction networks. The development of high throughput technologies, such as cDNA mircoarray and mass spectroscopy, makes it possible to probe a cell system globally and measure their gene expression, protein and metabolite profiles. Due to the quantity of data, there is an increasing reliance on models to interpret the data and understand the underlying mechanisms.

Cellular functions are regulated by networks of genes, proteins, and small molecules, such as metabolites and signaling molecules. When cellular functions are abnormally regulated, disease states may ensue. For example, apoptosis is abnormally regulated in cancer cells, while glucose level is abnormally regulated by insulin in type 2 diabetes. Therefore, to understand the mechanisms of disease development requires knowledge of how cellular function are regulated, which in turn would aid the development of therapies. There are two types of approaches that have been used to reconstruct pathways from high throughput data measurements. The first approach reconstructs pathways from measurements of molecular interactions such as protein-

protein and protein-DNA interactions. The second approach infers the pathways from measurements of cellular states, such as gene expression and metabolic profiles.

With the first approach, gene regulatory [Lee T.I. et al., 2002] and signal transduction [Steffen, M. et al., 2002] networks have been reconstructed using protein-DNA and protein-protein interactions, respectively. Protein-DNA interactions can be measured by chromatin immunoprecipitation (ChIP). Lee T.I. et al. (2002) integrated gene expression with protein-DNA interaction data to reconstruct the transcriptional regulatory network in yeast cell cycle. Similarly, gene expression profiles have been integrated with protein-protein interaction data to construct signal transduction pathways [Steffen M. et al., 2002] while a global molecular interaction network was reconstructed by integrating gene expression with protein-DNA and protein-protein interaction [Ideker T. 2002]. In that case, the gene expression data were used to identify the active pathways or sub-networks. The method assumes that genes involved in the same sub-network/pathway should show similar expression changes. Therefore, the active sub-network is identified by selecting a subset of genes that are connected based upon interaction data and whereby their expression profiles changed significantly.

However the availability of genome-wide protein-protein and protein-DNA interaction measurements is more limited in mammalian systems. To address this problem, we applied the second approach to infer networks from measurements of cellular states, such as gene expression and metabolite profiles, which are observed cellular responses induced by the underlying regulatory networks. A typical approach is to identify co-regulated genes by clustering genes with similar expression. For example, [Segal, E. et al., 2003] built relational probabilistic models to combine gene expression

data with DNA sequence profile or protein-protein interaction data to identify co-expressed genes that share common regulatory motifs or whose gene products interact with each other. [Tanay, A. et al. 2004] developed an approach, SAMBA, to cluster genes based upon the combined information from gene expression, protein-protein interaction, protein-DNA interaction, and growth phenotype data. However, cluster analysis can not identify potential pathways by which the cells are regulated to confer a specific phenotype.

To infer the active pathways from gene expression data, Collins' group has developed a model by assuming a log linear relationship between the expression of a gene, and the genes that regulate it, as well as the external perturbations [di Bernardo, D. et al. 2005]. The gene expression profiles of yeast, under different perturbed states, such as, gene deletion, promoter insertion and drug compound treatment, were obtained from the literature and used to train the model. Then, the gene expression response to a new drug (external perturbation) was obtained and the log linear model was applied to predict the effects induced by a new drug. This method requires a large amount of data for the network reconstruction [di Bernardo, D. et al. 2005].

Therefore, we propose an alternative approach that identifies the active pathways without requiring interaction measurements or libraries of genetic mutants, and with a limited amount of data, namely, by integrating gene expression and phenotypic profiles. The method assumes that information about the regulatory networks is contained within the profiles. Central to our approach is Bayesian Network (BN) analysis, a probabilistic graphic model, which detects implicit as well as explicit connections [Li and Chan, 2004b]. BN can detect indirect influences and unmeasured events and is not susceptible

to the existence of unobserved variables. It has been applied to infer gene regulatory network of yeast cell cycle from gene expression data [Friedman 2004], metabolic subnetworks from metabolic data [Li and Chan, 2004b] and protein signaling pathways from protein activity data [Sachs et al. 2005]. Unlike previous studies, in which the nodes are fairly well established in the literature and the BN approach was applied to infer the connections between the nodes, this study aims to identify the relevant nodes, which are unknown *a priori* to be activated by the environment or important to a cellular process of interest. The BN approach is computationally inefficient when applied to large networks with thousands of genes. In other words, it is more difficult to infer the structure of large networks using BN analysis. Therefore, in our approach we address this shortcoming as follows: since only part of the network is typically perturbed (active) when the cells are subjected to an environmental stress, methods to reduce the set of genes to a smaller and more relevant subgroup will be valuable. BN analysis can then be used to reconstruct the active sub-network from the smaller group of genes to reveal which pathways are induced by the external stimuli or environment.

As proof-of-concept, the framework was applied to identify pathways that are involved in palmitate and tumor necrosis factor (TNF)-α induced cytotoxicity in human liver cells. The level of cytotoxicity was measured by lactate dehybrogenase (LDH) release, which is a measure of cell death. Elevated levels of free fatty acids (FFAs) and TNFα have been shown to be involved in the pathogenesis of liver disorders, such as fatty liver disease and steatohepatitis [Felber 2002, Kobayashi 1998, Tilg 2001, Watada 2003]. Quantification of the genetic responses of hepatocytes to physiological actions of

these factors helped to identify the pathways involved in conferring cytotoxicity (e.g., producing a cellular phenotype with a high level of LDH release).

1.2 Organization of the Dissertation

In the following chapters, I will talk about techniques to select important genes relevant to a cellular function. In chapter 2, genetic algorithm coupled partial least square analysis (GA/PLS) is introduced and applied initially to experimental data to of primary rat hepatocyte culture system to select genes relevant to cellular functions such as urea production and TG accumulation. In chapter 3, Bayesian network analysis is introduced and applied to reconstruct regulatory networks from experimentally obtained metabolic data to illustrate its capabilities. Chapter 4 discusses the Three-stage Integrative Pathway Search (*TIPS*©) framework, which combines several techniques e.g. GA/PLS, independent component analysis (ICA) and BN analysis, and its application to a HepG2 cell culture system to infer pathways important in regulating the cytotoxic phenotype. These pathways were subsequently validated through both a literature survey and further experiments. In chapter 5, multisource data including gene regulatory network structure information and gene expression data were integrated in a dynamic state space model to extract underlying information of transcription factor activities. Chapter 6 discusses several extensions to improve the *TIPS*© framework, namely, by incorporating multiple metabolites and multi-source information into a dynamic model

1.3 Contribution of the Dissertation

This dissertation is one of the first endeavors to integrate phenotypic, metabolic and gene expression profiles with the objective to identify active pathways that regulate the cellular functions. It contributes in the following aspects:

1) Introduced GA/PLS to select a subset of genes relevant to cellular functions and was applied to experimental systems.

2) Proposed a Bayesian framework to reconstruct biological pathways. Metabolic fluxes and gene regulatory network were reconstructed from experimental data.

3) Independent component analysis was applied to extract pathways relevant a cellular function.

4) *TIPS*© framework integrated GA/PLS, ICA and BN to infer active pathways relevant to a phenotype. *TIPS*© approach was applied to a HepG2 cell system and meaningful pathways were extracted and verified.

5) A dynamic, state space model was introduced to infer transcription factor activities from gene expression data.

CHAPTER 2 IDENTIFYING GENES RELEVANT TO A CELLULAR FUNCTION

2.1 Introduction

As a step towards uncovering the interplay between genes and metabolic functions, the objective of this chapter is to develop a framework capable of selecting a subset of genes that can i) quantitatively predict metabolic functions based upon their expression levels; ii) be used to reconstruct regulatory pathways. Since metabolic functions are determined in part by the levels and activities of enzymes in the network, which in turn are regulated in part by their gene expression, it is reasonable to suggest that metabolic functions could be predicted by the expression level of a subset of relevant genes. To test this assumption, we developed a genetic algorithm coupled partial least squares analysis (GA/PLS) framework to identify a subset of genes that can predict metabolic function(s). The relevance of the gene subset is assessed by reconstructing their regulatory roles in the metabolic network with aid from the current literature knowledge. Gene expression and metabolic data were experimentally obtained for a hepatocellular system.

This concept of identifying a subset of genes that can predict metabolic function(s) is analogous to selecting an optimal subset of wavelengths in constructing a spectroscopic calibration model to predict chemical concentrations from signal measurements, e.g. absorbance value [Banglore 1996]. Full spectrum data (signal measurements) consists of absorbance values at different wavelengths, which are easily collected in a matter of seconds, but only a subset of the spectral data is essential to building a calibration model. The optimal subset of wavelengths used to build the

calibration model changes as the chemicals being measured change [Leardi 2000]. By analogy, a full range of gene expression data is also easily collected, but only a subset of these genes are relevant to a particular cellular function. This subset of genes changes as the cellular function being predicted changes. In the current study, genetic algorithm (GA) was used to identify a subset of genes that are most influential or relevant to a metabolic function. Then a partial least squares analysis (PLS) model was subsequently constructed to predict the metabolic function based upon the expression level of the selected genes. We reviewed the biological function of the selected genes to elucidate their potential role in regulating the metabolic function. Finally, we evaluated the model robustness by artificially adding noise to the gene expression data and comparing the subset of genes selected before and after noise addition.

2.2 Methodology of GA/PLS

The GA/PLS framework is illustrated in Figure 2.1. In brief, the whole dataset is separated into three independent groups, namely training, monitoring, and testing data sets, in order to avoid over-fitting. First, GA randomly selects a subset of genes from the training data set. The expression levels of these genes are subsequently used to construct a PLS model to predict a metabolic function, e.g. intracellular triglyceride (TG) or urea synthesis. The accuracy of the PLS model is evaluated using both the training and monitoring datasets. The accuracy of the PLS model prediction is used to assess the fitness of the subset of genes selected. Genetic algorithm continues to search for a gene subset until the accuracy of the PLS model prediction is maximized. The testing data set is used to validate the prediction ability of the final GA/PLS model.

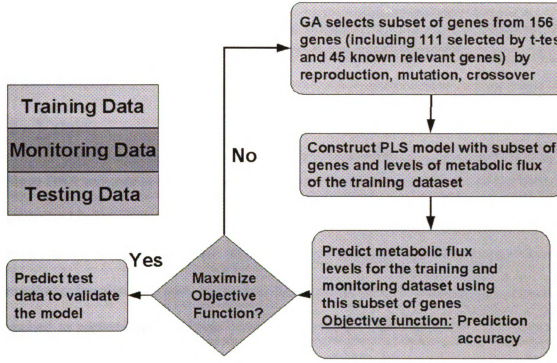


FIGURE 2.1: An overview of the GA/PLS feature selection algorithm. The data was separated into three independent subsets: training set, monitoring set, and testing set. GA randomly selects a subset of genes from the training data set. A PLS model was used to predict metabolic function based upon the expression of the selected genes. The prediction accuracy of the PLS model was used as the fitness value of the gene subset. The gene subsets with high fitness values were selected into the next iteration and subjected to crossover and mutation to introduce diversity into the population. The process terminates when the objective function reaches its maximum or when the termination condition (e.g., maximum number of iterations) is satisfied. The final model is validated with a testing data set.

2.2.1 Assumption of GA/PLS

We approximate the relation between the metabolic function and expression level of this subset of genes with a log-linear model:

$$\frac{Met(treatment)}{Met(control)} = \prod_{i=1}^{n} \left(\frac{Gene(treatment)_i}{Gene(control)_i}\right)^{C(i)} \tag{2.1}$$

where *Met(treatment)* and *Met(control)* are the metabolic function for the treated and control cultures, respectively; $Gene(treatment)_i$ and $Gene(control)_i$ are the expression level of gene *i* for the treated and control cultures, respectively.

Denoting Y as $\log_2(\frac{Met(treatment)}{Met(control)})$ and X_i as $\log_2(\frac{Gene(treatment)_i}{Gene(control)_i})$, equation (2.1) is transformed to:

$$Y = \sum_{i=1}^{n} C(i)X_i \qquad (2.2)$$

Since DNA mircoarray data are typically measured with respect to a reference level, we applied a log-linear model, which works well when data are presented as relative levels. Furthermore, a log-linear model allows some of the nonlinear relationships between metabolic function and gene expression to be captured. Log-linear models have been applied to approximate nonlinear processes in biochemical systems [Ni and Savageau, 1996; Ni and Savageau, 1996]. In this study the coefficients *C(i)* in equation (2.2) are determined by partial least square (PLS) analysis and the genes, $Gene_i$, are selected by GA/PLS as described below.

2.2.2 Partial Least Squares Analysis

PLS [Wold 1984, Geladi 1986] is a statistical approach capable of modeling a large number of variables using a relatively small set of observations. It circumvents typical problems associated with highly correlated and collinear nature of experimental data by projecting the data onto a few independent latent factors. These latent factors simplify the complex and diverse relationships by capturing the variable interactions contained in the original data into a new set of fewer unobserved/latent variables. We used this approach to map the levels of gene expression (***X***) to a metabolic function (***Y***), to gain an understanding of the interplay between a hepatic function and the gene expression profile. The PLS algorithm determines, based upon a nonlinear iterative partial least

squares (NIPALS) approach, a set of orthogonal projection axes ***W***, henceforth called PLS-weights, and sample projections T. For direct projection of the samples, $W^* = (W(P^T*W)^{-1})$ is used:

$$T = XW^* \tag{2.3}$$

Then, regression coefficients β in equation (2.4) are obtained by regressing Y onto the sample projections T:

$$Y = T\beta \tag{2.4}$$

With a PLS factors, the PLS model is:

$$\hat{Y} = XW_a^*\beta = T_a\beta = \sum_{j=1}^{a} t_j\beta_j \tag{2.5}$$

$$\beta = \left(T_a^T T\right)^{-1} T_a^T Y \tag{2.6}$$

2.2.3 Genetic Algorithm

GA is an optimization method based upon the theory of natural selection. A more complete discussion of GA can be found in [Goldberg 1989]. Here we demonstrate a feature selection algorithm that combines GA and PLS (see Figure 2.1). GA begins with an initial, randomly selected population. Each ***individual*** in the ***population*** is a potential solution to the optimization problem and its ***fitness*** to the optimization problem is evaluated by a ***fitness function***. The population evolves over ***generation*** by way of three genetic operators: ***reproduction, crossover, and mutation***. The process terminates when the objective function reaches its maximum or when the termination condition (e.g., maximum number of iterations) is satisfied. The algorithm includes four main steps: initialization, evaluation, reproduction and crossover, and mutation.

Step 1: Initialization

In GA, an ***individual*** is a string of length *N*, schematically shown in Figure 2.2. The first *N-1* values of the string are binary with value 1/0 representing the inclusion or exclusion of a gene in the PLS model and the N^{th} value represents the number of latent variables in the PLS model. A ***population*** is a collection of different individuals, i.e. a set of different strings (e.g., genes) of the same length. The initial population is created randomly in a user specified bounds of the *N* variables in the string. For example, the bounds of the first *N-1* bits of the string in this current study were [0,1], thus the individuals in the initial population were assigned randomly with values 0 or 1. The N^{th} bit of the string, i.e. number of latent variables, is randomly assigned a value between [1,10].

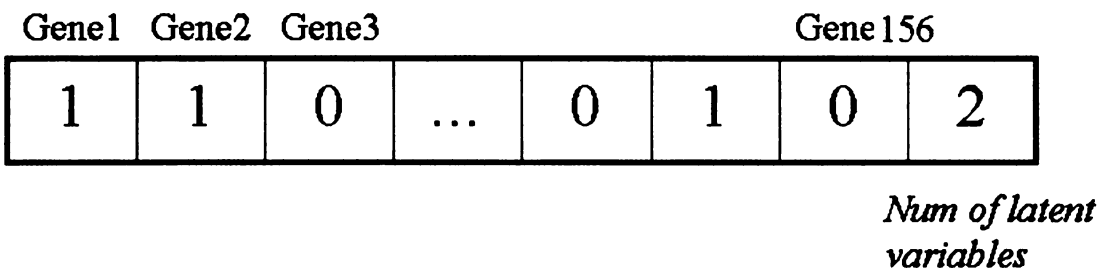


FIGURE 2.2: Representative of the string in the GA/PLS algorithm. 156 genes were selected by t-test and addition of known relevant genes. The first 156 elements of the string contain a binary 0/1 data representing the exclusion or inclusion of the genes in the PLS model. The last bit of the chromosome represents the number of latent variables in the PLS model.

Step 2: Evaluation

Each individual in the population is evaluated with a fitness function to determine how well they fit or improved the optimization problem. The optimization problem in our case is to find a subset of genes that can be used to construct a PLS model to predict metabolic functions with a minimal prediction error and number of latent variables. The fitness

function is defined by equation 2.7. Similar fitness function has been used in [Bangalore 1996] to select wavelengths that optimize calibration models.

$$fitness = \frac{1}{\sum (y_i - \hat{y}_i)^2 + LV^w} \tag{2.7}$$

where y_i is the measured metabolic flux value, $\hat{y}_i$ is the PLS predicted metabolic flux value, LV is the number of latent variables in the PLS model, and w is a weighting factor to establish an optimal balance between prediction accuracy and the model size (number of PLS latent variables). w normalizes the model size to the same scale as the prediction error terms. An optimal value of w is critical to the success of the GA/PLS model. Too large a w results in a model with too few factors to explain the gene expression data. Too small a w causes the PLS model to include too many factors and become essentially a curve fit. An optimal w, guided by a GA search, allows one to find a model that will provide the most accurate prediction with a minimum number of latent variables. To determine an optimal w value, different w values, ranging from 0 to 0.5, were used to search for a PLS model. The different models were compared using two criteria, the *mean square error of prediction* and the *number of PLS latent variables*. An optimal value of w was determined by balancing these two criteria to obtain a reasonable prediction error with a minimum number of latent variables. The optimal value for w was determined to be 0.4 for predicting intracellular TG accumulation and 0.45 to predict urea synthesis, shown in Figures 2.3 and 2.4, respectively.

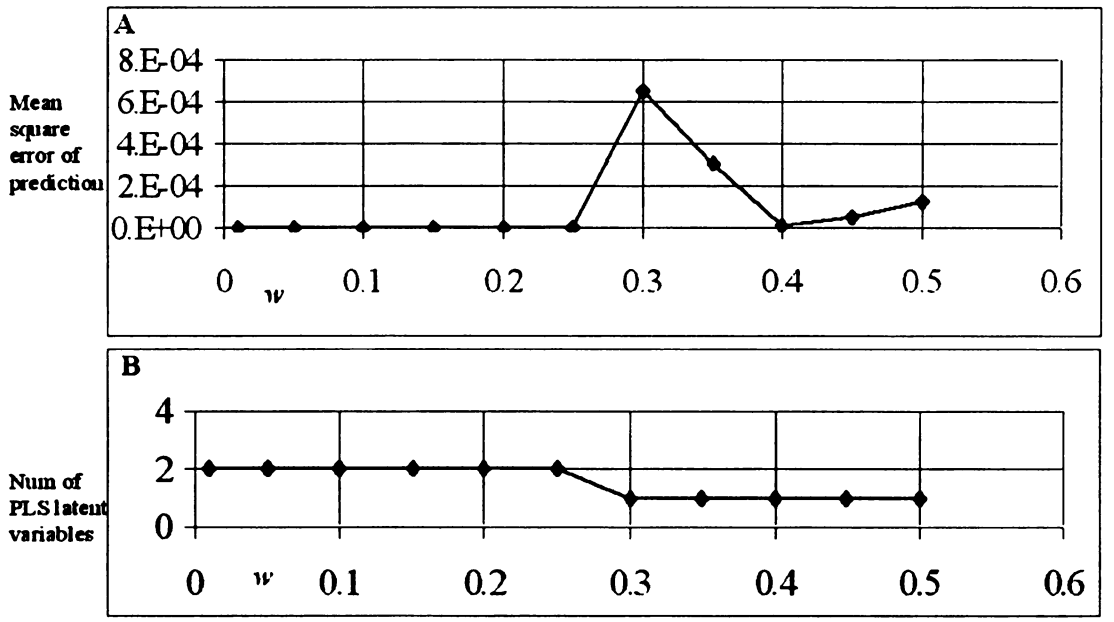


FIGURE 2.3: Determining *w* value for selecting TG genes. An optimal value for *w* in equation 2.7 was determined by balancing the two criteria of GA/PLS performance: the mean square error of the PLS model prediction (shown in 3A), and the number of latent variables in the PLS model (shown in 3B). Different values of *w* values ranging from 0 to 0.5 was tested while monitoring the above two criteria. An optimal value of 0.4 for *w* was found for intracellular TG. At this *w* value, both the number of latent variables and prediction error reach their minimum levels.

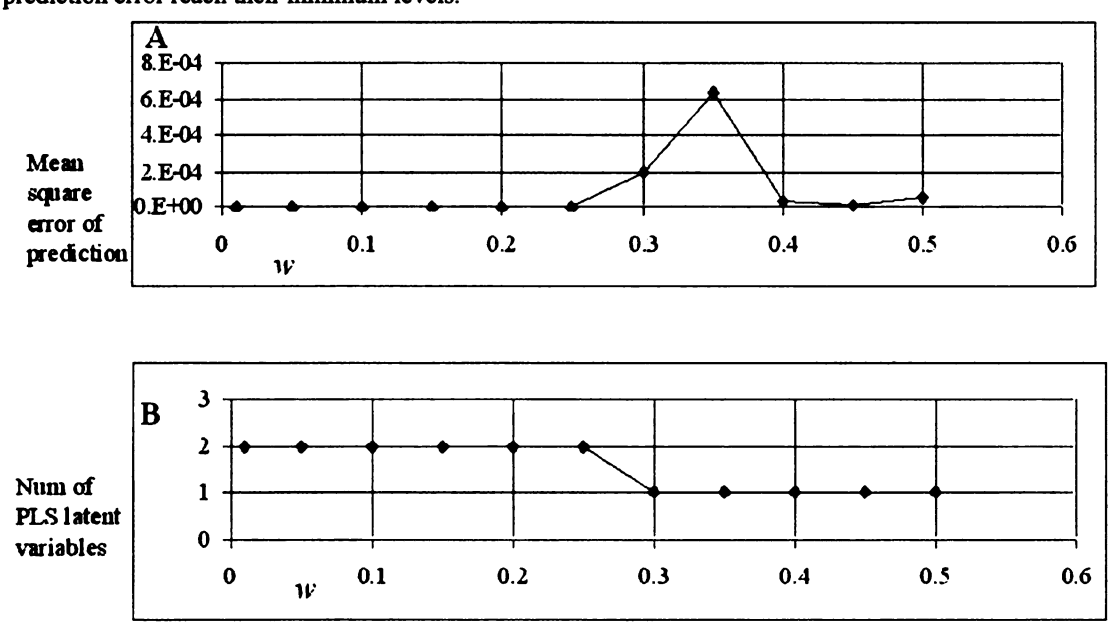


FIGURE 2.4: Determining *w* for selecting urea synthesis genes. An optimal value for *w* in equation 2.7 was determined by balancing the two criteria in GA/PLS performance: the mean square error of the PLS model prediction (shown in 4A), and the number of latent variables in the PLS model (shown in 4B). Different values of *w* ranging from 0 to 0.5 were tested while monitoring the above two criteria. An optimal value of 0.45 for *w* was found for urea synthesis. At this *w* value, both the number of latent variables and prediction error reach their minimum levels.

Step 3: Reproduction

After the evaluation step, each individual in the population is assigned a probability value according to its fitness value, which determines whether it will be selected to the next iteration or ***generation*** in GA. We used roulette selection [Holland 1975] to determine the probability of selecting an individual into the next iteration:

$$p_i = \frac{f_i}{\sum f_i} \tag{2.8}$$

where the fitness f_i ,$i=1,2,...,N$ of each individual is calculated using the fitness function defined in equation 2.7.

Step 4: Crossover and Mutation

Crossover and mutation was applied to introduce diversity into the population, which ensures the escape from local optimum in the GA search algorithm. In crossover, two distinct individuals were selected randomly from the population and some portions of the strings are exchanged with a probability P_c. In mutation, one or more bits of the string are altered with probability P_m. In this study, we applied an arithmetic crossover with a probability of 0.6 and a boundary mutation with probability of 0.05 (see reference [Houck 1995] for details on the crossover and mutation methods applied in this chapter).

2.2.4 Gene Subset Selection Optimization

GA/PLS selects different subsets of genes to predict the same metabolic flux given different initial populations. In other words, different subsets of genes can give the same or nearly same degree of accuracy in the prediction of the metabolic flux based upon their gene expression level. Therefore, a strategy was developed to identify a group of genes

from multiple possible solutions. The strategy involved running the GA/PLS model with different initial populations and counting the frequency of appearance of each gene in the multiple solutions. The initial population size ranged from 30 to 100 individuals and each individual contained a set of different genes. GA/PLS was run 14 times with different initial population of individuals. A gene was included in the final subset if it was selected by the GA/PLS model in more than half of the runs. Therefore, the genes that appeared more than 8 times as a solution in the GA/PLS model were selected into the final gene subset. Similar methods have been applied for wavelength selection in spectroscopic calibration model [Leardi 2000].

2.3 Application of GA/PLS to Rat Hepatocypte Culture System

To test the ability of GA/PLS in identifying relevant genes of a cellular function, GA/PLS was applied to the gene expression and metabolic profiles obtained from primary rat liver cell cultures. GA/PLS selected a subset of genes relevant to cellular functions of urea production and triglyceride accumulation. Then a partial least squares analysis (PLS) model was subsequently constructed to predict the metabolic function based upon the expression level of the selected genes. We reviewed the biological function of the selected genes to elucidate their potential role in regulating the metabolic function. Finally, we evaluated the model robustness by artificially adding noise to the gene expression data and comparing the subset of genes selected before and after noise addition.

2.3.1 Experimental System

The gene expression and metabolic data were obtained from cultured primary rat hepatocytes [Chan 2002, 2003 a,b,c]. Hepatocytes were isolated from adult female Lewis rats. The isolated hepatocytes were cultured in a collagen sandwich configuration and incubated in standard hepatocyte culture medium containing either 0.5 U/ml insulin (high insulin) or 50 μU/ml insulin (low insulin). The hepatocytes were cultured in this fashion for at least 6 days prior to plasma exposure. This interval is considered the pre-conditioning period. The six-day-old-sandwiched hepatocyte cultures were subsequently exposed to unsupplemented, or amino acid supplemented plasma solution for an additional seven days [Chan 2003 a,b]. The four combinations of pre-conditioning and plasma supplementation evaluated were as follows. They were i) low insulin pre-conditioned and unsupplemented plasma (denoted as *[L,0]* thereafter), ii) low insulin pre-conditioned and amino acid supplemented plasma (denoted as *[L,A]*), iii) high insulin pre-conditioned and unsupplemented plasma (denoted as *[H,0]*), and iv) high insulin pre-conditioned and amino acid supplemented plasma (denoted as *[H,A]*) cultures. The metabolites were measured using HPLC and biochemical assays described elsewhere [Chan 2002, 2003 a,b,c]. A model for hepatocyte metabolism was created based on the known stoichiometry of the hepatic metabolic network. Metabolite measurements were coupled to Metabolic Flux Analysis (MFA) to obtain the intracellular fluxes [Chan 2002a]. The gene expression profiles of the samples were collected using Affymetrix chips. All expression profiles were generated using total RNA, with *in vitro* transcription yielding biotinylated cRNA for hybridization to Affymetrix Rat UG34A GeneChip array.

The chips were analyzed at the Brigham and Women's Hospital in Boston. The expression data can be accessed at www.msu.edu/~lizheng1/microarray.zip.

2.3.2 Data Preprocess

Expression levels of 8,799 genes were measured using Affymetrix chips and a subset of these genes were differentially expressed across the range of experimental conditions in this study. Student t-test was used to identify the subset of genes that were differentially expressed between the high insulin pre-conditioned (including *[H,0]*, *[H,A]*) and the low insulin pre-conditioned (including *[L,0]*, *[L,A]*) groups. In t-test, normalized distance between the high insulin and low insulin groups is calculated as follows:

$$D = (m_h - m_{lh}) / \sqrt{\frac{s_h^2}{n_h} + \frac{s_{lh}^2}{n_{lh}}}, \tag{2.9}$$

where (m_h,m_{lh}) is the population mean value of the high insulin and low insulin groups, respectively, (s_h,s_{lh}) is the population variances, and (n_h, n_{lh}) is the number of the samples in the population. It is known that D follows approximately a student distribution. The two populations are considered different when D exceeds a certain threshold value, which depends on the confidence level selected. Using student t-test, 179 genes were identified as significantly different with a 95% confidence level. The 179 genes were checked manually to exclude irrelevant ones, for example, genes expressed specifically in other tissues, e.g., brain were removed. Finally, 111 genes remained for further GA/PLS analysis and are listed in Table 2.1.

In addition to the genes selected by t-test, many other genes encode the enzymes that are involved in the metabolic network defined in reference [Chan 2003 a] and are relevant to heptocellular function, but were not selected by the t-test. We included these

informative genes, namely, 45 genes that encode the enzymes or transcription factors involved in the metabolic network detailed elsewhere [Chan 2003 a]. These genes are listed in Table 2.2.

Table 2.1: 111 genes selected after t-test

No.	Accession number	Gene Name
1	AFFX_Rat_GAPDH_3	Glyceraldehydes-3-phosphate-dehydrogenase (GAPDH) (GAPDH, EC 1.2.1.12)
2	U93306	U93306 Rattus norvegicus VEGF receptor-2/FLK-1 mRNA
3	AF000139	25-hydroxyvitamin D 1-hydroxylase (CYP1) mRNA, complete cds
4	D90048	RATATPB2S Rat Na+, K+ -ATPase (EC 3.6.1.3) beta2 subunit gene and 5 flank
5	E03344	E03344cds cDNA sequence of peroxisome forming factor
6	J01435	mitochondrial cytochrome oxidase subunits I,II, III genes,
7	J05156	cholesterol side-chain cleavage enzyme mRNA (P450SCC), complete cds
8	K00750	cytochrome c nuclear-encoded mitochondrial gene and flanks
9	K00996	cytochrome p-450e (phenobarbital-induced) mRNA, 3 end
10	K01721	RATCYP45A Rat P-450(1) variant (phenobarbital-inducible) cytochrome 3 end, flank
11	L27129	stress activated protein kinase gamma isoform, mRNA, 5 end
12	M20131	M20131cds RATCYP45I Rat cytochrome P450IIE1 gene, complete cds
13	M80784	M80784 RATTGFBET R.norvegicus type III TGF-beta receptor mRNA, complete cds
14	J03960	J03960 Rat 5-lipoxygenase mRNA, complete cds
15	M88592	M88592 Rattus rattus peroxisome proliferator activated receptor (PPAR) mRNA,
16	M95058	M95058completeSeq Rattus rattus steroid 5-alpha-reductase 2 Mrna
17	U12187	U12187 Rattus norvegicus ras-related protein (rad) mRNA
18	U39207	U39207 Rattus norvegicus cytochrome P450 4F5 (CYP4F5) mRNA
19	U11681	U11681 Rattus norvegicus rapamycin and FKBP12 target-1 protein (rRAFT1) mRNA
20	U48596	U48596 Rattus norvegicus MAP kinase kinase kinase 1 (MEKK1) mRNA
21	U68544	cyclophilin D mRNA, nuclear gene encoding mitochondrial protein
22	U40004	U40004 Rattus norvegicus cytochrome P450 pseudogene (CYP2J3P2) mRNA
23	D83538	D83538 Rat mRNA for 230kDa phosphatidylinositol 4-kinase
24	AF012891	AF012891 Rattus norvegicus frizzled related protein frpAP mRNA
25	S78284	S78284 bcl-xshort=apoptosis inducer [rats, ovary, mRNA Partial, 537 nt]
26	U18982	fos-related antigen 2 (fra-2) mRNA, complete cds
27	U39943	cytochrome P450 monooxygenase (CYP2J3) mRNA, complete cds
28	Y00156	mRNA for hepatic microsomal UDP-glucuronosyltransferase (UDPGT)
29	A09811	A09811cds R.norvegicus mRNA for BRL-3A binding protein
30	AB004278	AB004278 Rat mRNA for protocadherin 2, partial cds
31	AB009372	AB009372 Rattus norvegicus mRNA for Lysophospholipase, complete cds

Table 2.1 (continued)

32	AB010154	PKN mRNA for serin/threonine protein kinase expressed in hippocampus
33	AB011068	AB011068 Rattus rattus mRNA for type II iodothyronine deiodinase, complete cds
34	AB011530	AB011530 Rattus norvegicus mRNA for MEGF4, complete cds
35	AF001417	AF001417 Rattus norvegicus zinc finger protein mRNA, complete cds
36	AF058795	AF058795 Rattus norvegicus GABA-B receptor gb2 mRNA, complete cds
37	AF082126	AF082126 Rattus norvegicus aryl hydrocarbon receptor (AHR) mRNA
38	AF085693	G protein-coupled receptor kinase-associated ADP ribosylation factor
39	AF093773	cytosolic malate dehydrogenase (Mdh) mRNA, complete cds
40	AJ000347	mRNA for 3 (2),5 -bisphosphate nucleotidase
41	D10891	mRNA for metabotropic glutamate receptor mGluR5, complete cds
42	D13978	D13978 Rattus sp. mRNA for argininosuccinate lyase, complete cds
43	D16101	D16101 RATSGLT1 RAT mRNA for sodium/glucose cotransporter
44	D63411	D63411 RATMPR Rat mRNA for mitochondrial precursor receptor, complete cds
45	D79215	D79215 Rattus norvegicus mRNA for FGF-10, complete cds
46	J01435	mitochondrial cytochrome oxidase subunits I,II, III genes
47	J04807	J04807mRNA RATINSIIA Rat insulin II gene mRNA, 3 end
48	L04672	L04672 RATGPCRCTR Rattus rattus G protein-coupled receptor mRNA, complete cds
49	L13151	L13151cds RATGAPX Rat GTPase-activating protein (GAP) gene, complete cds
50	L27061	L27061 RATPHOSE Rattus norvegicus phosphodiesterase mRNA, 3 end
51	M81784	M81784 RATKCAB Rattus norvegicus K+ channel mRNA, sequence
52	S48325	S48325 diabetes-inducible cytochrome P450RLM6 [rats, liver, mRNA Partial, 1093 nt]
53	S52878	intestinal 15 kda protein=fatty acid-binding protein homolog [rats, ileum, mRNA, 460 nt]
54	S59893	S59893 La=autoantigen SS-B/La {3 region, clone K51} [rats, mRNA, 106 nt]
55	S68736	myosin heavy chain [rats, CCl4-cirrhotic liver fat-storing cell line, mRNA, 2924 nt]
56	S76404	beta -HKA=H,K-ATPase beta-subunit [rats, Genomic, 8983 nt, segment 2 of 2]
57	U04934	(CD-1) clone Kc1 Na-Ca exchanger mRNA, partial cds
58	U37099	small GTP-binding protein (rab3c) mRNA, partial cds
59	U75923	U75923UTR#1 SEG_RNTRNAIS3 Rattus norvegicus isoleucyl tRNA synthetase mRNA
60	U78090	potassium channel regulator 1 mRNA, complete cds
61	X06889	X06889cds RNRAB3 Rat ras-related mRNA rab3
62	X54250	X54250mRNA RRATBP2 Rat mRNA for zinc finger protein AT-BP2, partial cds
63	X62660	X62660mRNA RRGTS8 R.rattus mRNA for glutathione transferase subunit 8
64	X65296	X65296cds RRESHVEL R.rattus mRNA for carboxylesterase (Es-HVEL)
65	X74833	X74833cds RNACRB1 R.norvegicus mRNA for acetylcholine

Table 2.1 (continued)

		receptor beta-subunit
66	Y09453	Y09453cds RNY09453 R.norvegicus mRNA for calcium channel gamma subunit
67	Y10258	Y10258cds RSKET Rattus sp. mRNA for DNA binding protein, KET
68	Z11932	Z11932cds RRVASOV2M R.rattus mRNA for vasopressin V2 receptor
69	E02468	E02468cds DNA sequence of rat TNF
70	E01884	E01884cds DNA sequence coding for rat IL-1-beta(interleukin-1 beta)
71	E01789	E01789cds cDNA sequence coding for rat C-kinase type-II (beta-2)
72	M74494	M74494 Rat sodium/potassium ATPase alpha-1 subunit truncated isoform mRNA,
73	M59211	M59211 Rat potassium channel Kv3.2b mRNA
74	M64033	M64033 Rat secretin gene, complete cds
75	L22558	adenylyl cyclase-activated serotonin receptor (5-HT7) mRNA
76	M27433	M27433 Rattus norvegicus germinal histone H4 gene
77	M22253	M22253 Rattus norvegicus sodium channel I mRNA
78	U61772	U61772 Rattus norvegicus merlin (NF2) mRNA
79	L10072	L10072 Rattus norvegicus 5-hydroxytryptamine receptor (5HT5a) mRNA
80	AF041246	AF041246 Rattus norvegicus orexin receptor-2 mRNA
81	U44750	NAD-dependent 15-hydroxyprostaglandin dehydrogenase mRNA
82	J04629	J04629 Rat (Na+, K+)-ATPase-beta-2 subunit mRNA
83	M60617	M60617 Rat CCAAT binding transcription factor-B subunit (CBF-A11) mRNA
84	J05031	J05031 Rat isovaleryl-CoA dehydrogenase (IVD) mRNA
85	J03170	J03170 Rat liver specific transcription factor (LF-B1) gene
86	M60655	M60655 Rat alpha-1B adrenergic receptor mRNA
87	U35775	U35775 Rattus norvegicus gamma-adducin mRNA
88	L46791	L46791 Rattus norvegicus cholesterol esterase mRNA
89	U39044	U39044 Rattus norvegicus cytoplasmic dynein intermediate chain 2A mRNA
90	X12459	X12459 Rat mRNA for argininosuccinate synthetase (EC 6.3.4.5)
91	X66494	X66494 R.norvegicus CHOT1 mRNA
92	X80477	X80477 R.norvegicus P2X mRNA
93	X86086	X86086 R.norvegicus RNA for annexin VI
94	X62528	X62528 R.rattus mRNA for ribonuclease inhibitor
95	X04979	X04979 Rat gene for apolipoprotein E /cds=(23,958)
96	X91234	X91234 R.norvegicus mRNA for 17-beta hydroxysteroid dehydrogenase type 2
97	U57362	U57362 Rattus norvegicus collagen XII alpha 1 (Col12a1) mRNA
98	U69673	U69673 Rattus norvegicus protein tyrosine phosphatase 20 mRNA
99	D49980	D49980 Rat gene for 8-oxo-dGTPase
100	U53859	U53859 Rattus norvegicus calpain small subunit (css1) mRNA
101	D84418	D84418 Rat mRNA for chromosomal protein HMG2
102	M18467	M18467 Rattus norvegicus mitochondrial aspartate aminotransferase mRNA
103	D78359	D78359 Rat drs (a gene down-regulated by v-src) mRNA
104	J04486	J04486 Rat insulin growth factor-binding protein mRNA
105	AB009636	AB009636 Rattus norvegicus mRNA for phosphoinositide 3-kinase
106	M15883	M15883 Rat clathrin light chain (LCB2) mRNA
107	AF003008	AF003008 Rattus norvegicus Mxi1 protein (rMxi1) mRNA
108	AB006137	AB006137 Rattus norvegicus FTA mRNA for alpha 1,2-fucosyltransferase
109	D90109	D90109 Rat mRNA for long-chain acyl-CoA synthetase (EC 6.2.1.3)

Table 2.1 (continued)

110	J02646	J02646 Rat translational initiation factor (eIF-2) alpha subunit mRNA
111	M14053	M14053 Rat glucocorticoid receptor mRNA

Table 2.2: 45 genes involved in the metabolic network

No.	Accession number	Gene Name
1	L37333	glucose-6-phosphatase (G6Pase) mRNA, complete cds
2	X58865	X58865mRNA Rat PFK-L mRNA for liver phosphofructokinase
3	U32314	U32314 Rattus norvegicus pyruvate carboxylase mRNA
4	K03243	Phosphoenolpyruvate carboxykinase (GTP) gene, exons 1-3
5	U07177	U07177 Rattus norvegicus lactate dehydrogenase-C (LDH-C) mRNA
6	M54926	M54926 Rat lactate dehydrogenase A mRNA, 3 end
7	U07181	U07181 Rattus norvegicus lactate dehydrogenase-B (LDH-B) mRNA
8	U75393	succinyl-CoA synthetase alpha subunit mRNA,
9	J04473	J04473 Rat mitochondrial fumarase mRNA, complete cds
10	K03041	ornithine carbamoyltransferase mRNA
11	J02720	J02720 Rat liver arginase mRNA, complete cds
12	Z27513	gene for carbamoylphosphate synthase I, exon 38
13	D10354	D10354 RATAAT Rat mRNA for alanine aminotransferase
14	J03865	J03865mRNA RATSDH22 Rat serine dehydratase (SDH2) gene, 3 flank
15	X07467	X07467 Rat mRNA for glucose-6-phosphate dehydrogenase (Gd, EC 1.1.1.49)
16	M16235	M16235 Rat hepatic lipase mRNA
17	X78593	X78593 R.norvegicus mRNA for glycerol-3-phosphate dehydrogenase
18	L46791	L46791 Rattus norvegicus cholesterol esterase mRNA, complete cds
19	J05446	J05446 Rat glycogen synthase mRNA, complete cds
20	M76767	M76767 RATFASA Rattus norvegicus fatty acid synthase mRNA, complete cds
21	J05029	Rat long chain acyl-CoA dehydrogenase (LCAD) mRNA, complete cds
22	D38448	D38448 Rat mRNA for 88kDa-diacylglycerol kinase (DGK-III), complete cds
23	AF017251	AF017251 Rattus norvegicus phospholipase D (PLDs) mRNA, complete cds
24	X12752	X12752 Rat gene for DNA binding protein C/EBP
25	M34238	CCAAT binding transcription factor-B subunit (CBF-B) mRNA, complete cds
26	X54423	X54423cds RNHNF1 Rat mRNA for hepatic nuclear factor one (HNF1)
27	X57133	X57133mRNA RNHNF4 Rat mRNA for hepatocyte nuclear factor 4 (HNF4)
28	Y14933	Y14933mRNA RNHNF6B R.norvegicus mRNA for hepatocyte nuclear factor 6 beta
29	X55955	X55955 Rat mRNA for hepatocyte nuclear factor 3A (HNF-3A)

Table 2.2 (continued)

30	L09647	hepatocyte nuclear factor 3a (HNF-3 beta) mRNA, complete cds
31	K03486	K03486 RATPKC32 Rat protein kinase C type III mRNA, 3 region
32	M15523	protein kinase C-family related mRNA, partial cds, clone RP16
33	M55417	protein kinase C-gamma (PRKC-gamma) gene, exon 1
34	X07286	X07286cds RNPKCAR Rat mRNA for protein kinase C alpha
35	X07287	X07287cds RNPKCG Rat mRNA for protein kinase C gamma
36	X54096	X54096 Rat mRNA for lecithin-cholesterol acyltransferase (EC 2.3.1.43) (LCAT)
37	X53588	X53588 Rat mRNA for glucokinase, alternatively spliced GK2 (EC 2.7.1.1)
38	AB004329	AB004329 Rattus norvegicus mRNA for acetyl-CoA carboxylase, complete cds
39	X55286	X55286 R.norvegicus mRNA for HMG-CoA reductase
40	U03763	U03763cds RRU03763 Rattus rattus phospholipase mRNA, complete cds
41	L03294	L03294 Rattus norvegicus lipoprotein lipase mRNA, complete cds
42	L06040	L06040 Rattus norvegicus 12-lipoxygenase mRNA, complete cds
43	AF071495	type II pneumocyte CD36-related class B scavenger receptor (SRB1R) mRNA
44	J05460	J05460 RATCHOL7H Rat cholesterol 7-alpha-hydroxylase mRNA, complete cds
45	S77528	S77528cds rNFIL-6=C/EBP-related transcription factor [rats, Genomic/mRNA, 1759 nt]

2.3.3 Results

GA/PLS was applied to gene expression and metabolic data obtained from 4 culture conditions, *[H,0], [H,A], [L,0], [L,A]*, as described in the Methods. The data was separated into 3 groups, a training (*[H,A], [L,0]*), monitoring (*[H,0]*), and testing (*[L,A]*) set. We investigated the ability of GA/PLS to select a subset of genes able to predict hepatic functions (intracellular TG accumulation and urea synthesis). In addition, the role of the selected genes in regulating the aforementioned functions was reconstructed with the aid of a literature review. Finally, we evaluated the robustness of the GA/PLS model to noise in the gene expression data.

Genes Selected for Intracellular TG

The frequency of each gene selected by GA/PLS in the multiple runs for intracellular TG is plotted in Figure 2.5A. 59 genes were selected with a frequency

greater than 8 and are listed in Table 2.3. The selected genes can be categorized into the following groups: fatty acid and lipid metabolism, transcription factors, signal pathways, electron transport chain and oxidative phosphorylation, and ion transporter. A PLS model with two latent variables was constructed to predict intracellular TG based upon the expression level of the 59 genes. As shown in Figure 2.5B, intracellular TG was predicted well using the GA/PLS model (mean square error is 0.0425).

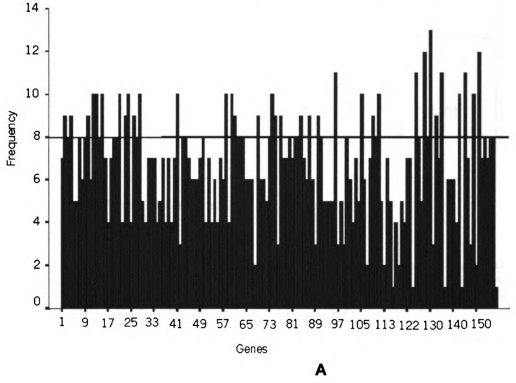


FIGURE 2.5: TG gene selection and metabolic function prediction. 2.5A) TG gene frequency. GA/PLS was run 14 times with different initial populations. The frequency at which each gene appeared in the 14 solutions was counted and plotted. Those genes that appeared with a frequency higher than 8 were selected into the final gene subset (shown in Table 2.3) for prediction and pathway reconstruction

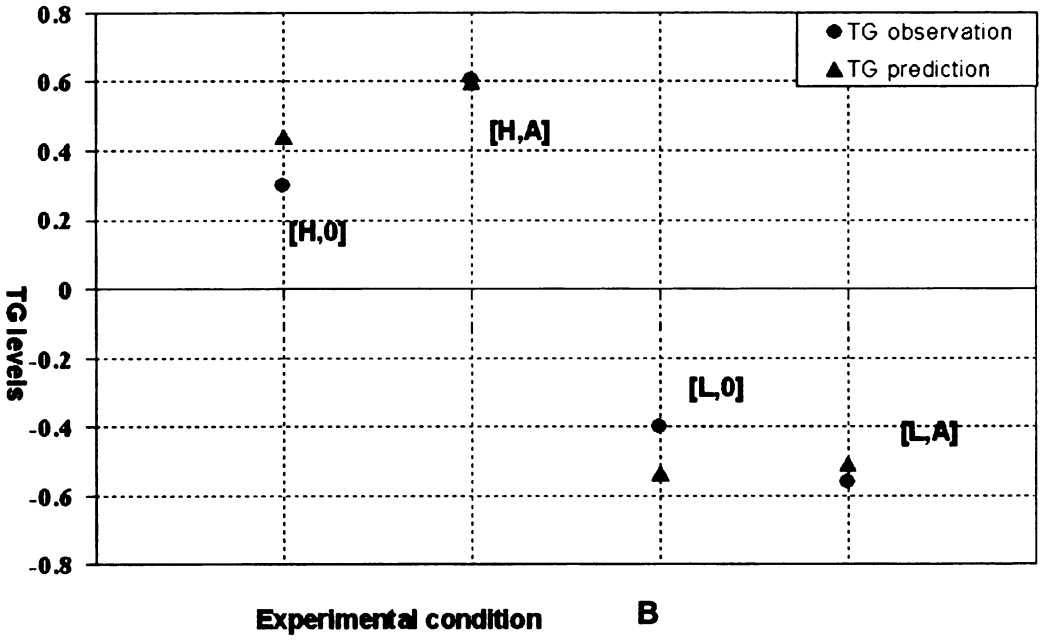


FIGURE 2.5: TG gene selection and metabolic function prediction.. 2.5B) PLS prediction of TG. The level of TG accumulation was predicted based upon the expression level of the genes selected by GA/PLS. [H,A], [L,0] was used for training, [H,0] was used for monitoring, and [H,A] was used for testing and validating the PLS model. The mean square error of the PLS prediction model was 0.0425.

Table 2.3 A. Genes selected for intracellular triglyceride

Functional Category	Accession Number	Gene name	Frequency
Fatty acid and lipid metabolism			
	D90109	mRNA for long-chain acyl-CoA synthetase (EC 6.2.1.3)	8
	J03960	5-lipoxygenase mRNA, complete cds	8
	X04979	gene for apolipoprotein E /cds=(23,958)	11
	M16235	Rat hepatic lipase mRNA	8
	X78593	mRNA for glycerol-3-phosphate dehydrogenase	13
	J05029	long chain acyl-CoA dehydrogenase (LCAD) mRNA, complete cds	9
	AF017251	phospholipase D (PLDs) mRNA, complete cds	11
	AB004329	mRNA for acetyl-CoA carboxylase, complete cds	12
	U03763	phospholipase mRNA, complete cds	8
	L06040	12-lipoxygenase mRNA, complete cds	8
	J05156	cholesterol side-chain cleavage enzyme mRNA (P450SCC), complete cds	8

Table 2.3A (continued)

J05460	Cholesterol 7-alpha-hydroxylase mRNA, complete cds	8
X54096	X54096 Rat mRNA for lecithin-cholesterol acyltransferase (EC 2.3.1.43) (LCAT)	10
Transcription factor		
M88592	Peroxisome proliferator activated receptor(PPAR) mRNA	10
M60617	Rat CCAAT binding transcription factor-B subunit	9
X54250	mRNA for zinc finger protein AT-BP2	8
L09647	hepatocyte nuclear factor 3a (HNF-3 beta) mRNA, complete cds	10
Ion transporter		
D90048	Na+,K+-ATPase(EC3.6.1.3) beta2 subunit gene	9
D16101	mRNA for sodium/glucose cotransporter	8
U78090	Potassium channel regulator 1 mRNA, complete cds	10
M74494	sodium/potassium ATPase alpha-1 subunit truncated isoform mRNA,	8
M59211	Potassium channel Kv3.2b mRNA	10
J04629	(Na+, K+)-ATPase-beta-2 subunit mRNA	8
Electron transport chain Oxidative phosphoration		
K00996	RATCYP45E Rat cytochrome p-450e (phenobarbital-induced) mRNA, 3 end	8
K01721	RATCYP45A Rat P-450(1) variant (phenobarbital-inducible) cytochrome 3 end	9
M20131	RATCYP45I Rat cytochrome P450IIE1 gene, complete cds	10
U39943	cytochrome P450 monooxygenase (CYP2J3) mRNA, complete cds	8
AF000139	25-hydroxyvitamin D 1-hydroxylase (CYP1) mRNA, complete cds	8
Signal pathways		
U93306	VEGF receptor-2/FLK-1 mRNA	9
M80784	type III TGF-beta receptor mRNA, complete cds	10
U11681	Rapamycin and FKBP12 target-1 protein (rRAFT1) mRNA	8
U48596	MAP kinase kinase kinase 1 (MEKK1) mRNA	8
D83538	mRNA for 230kDa phosphatidylinositol 4-kinase	9
D10891	mRNA for metabotropic glutamate receptor mGluR5, complete cds	10
D63411	mRNA for mitochondrial precursor receptor, complete cds	8
U37099	small GTP-binding protein (rab3c) mRNA, partial cds	10
X06889	Rat ras-related mRNA rab3	9
Z11932	mRNA for vasopressin V2 receptor	9
L10072	5-hydroxytryptamine receptor (5HT5a) mRNA	8
M60655	alpha-1B adrenergic receptor mRNA	9
M15523	protein kinase C-family related mRNA, partial cds, clone RP16	11
J04486	insulin growth factor-binding protein mRNA	10
L27061	Phosphodiesterase mRNA 3 end	8
Others		
J02646	Rat translational initiation factor (eIF-2) alpha subunit mRNA	10
D10354	mRNA for alanine aminotransferase	11
J03865	serine dehydratase (SDH2) gene, 3 flank	8
X07467	mRNA for glucose-6-phosphate dehydrogenase (Gd, EC	12

Table 2.3A (continued)

	1.1.1.49)	
AB006137	FTA mRNA for alpha 1,2-fucosyltransferase	9
X12459	mRNA for argininosuccinate synthetase (EC 6.3.4.5)	8
U39044	cytoplasmic dynein intermediate chain 2A mRNA	9
M64033	secretin gene, complete cds	9
M27433	germinal histone H4 gene	9
U44750	NAD-dependent 15-hydroxyprostaglandin dehydrogenase mRNA	8
X62660	mRNA for glutathione transferase subunit 8	8
Y00156	mRNA for hepatic microsomal UDP-glucuronosyltransferase (UDPGT)	10
AF012891	frizzled related protein frpAP mRNA	10
U18982	fos-related antigen 2 (fra-2) mRNA, complete cds	9
U68544	cyclophilin D mRNA, nuclear gene encoding mitochondrial protein	10

Table 2.3 B. Genes selected for urea synthesis

Functional Category	Accession Number	Gene name	Frequency
Urea cycle			
	D13978	mRNA for argininosuccinate lyase	12
	X12459	argininosuccinate synthetase (EC 6.3.4.5)	11
	K03041	Ornithine carbamoyltransferase mRNA	8
	Z27513	gene for carbamoylphosphate synthase I, exon 38	10
	M18467	mitochondrial aspartate aminotransferase mRNA	9
gluconeogensis			
	U32314	Pyruvate carboxylase mRNA	9
	K03243	phosphoenolpyruvate carboxykinase (GTP) gene, exons 1-3	9
	L37333	Glucose-6-phosphatase (G6Pase) mRNA	11
	X58865	PFK-L mRNA for liver phosphofructokinase	9
TCA cycle			
	AF093773	Cytosolic malate dehydrogenase mRNA	9
Transcription factor			
	X12752	gene for DNA binding protein C/EBP	12
	S77528	rNFIL-6=C/EBP-related transcription factor	12
	X55955	Hepatocyte nuclear factor 3A (HNF-3A)	10
Electron transport chain Oxidative phosphoration			
	U39943	cytochrome P450 monooxygenase (CYP2J3) mRNA, complete cds	8
	K01721	RATCYP45A Rat P-450(1) variant (phenobarbital-inducible) cytochrome 3 end, flank	10
	J01435	mitochondrial cytochrome oxidase subunits I,II, III genes,	9
	S48325	diabetes-inducible cytochrome P450RLM6 [rats, liver, mRNA Partial, 1093 nt]	8
	E03344	cDNA sequence of peroxisome forming factor	10

Table 2.3B (continued)

	K00750	Cytochrome c nuclear-encoded mitochondrial gene	9
	S48325	diabetes-inducible cytochrome P450RLM6	8
Signal pathways			
	U12187	ras-related protein (rad) mRNA	10
	L27129	Stress activated protein kinase mRNA	9
	D83538	mRNA for 230kDa phosphatidylinositol 4-kinase	9
	AF085693	GTPase-activating protein (GIT1) mRNA	8
	L04672	G protein-coupled receptor mRNA, complete cds	8
	L13151	GTPase-activating protein (GAP) gene, complete cds	8
	L22558	adenylyl cyclase-activated serotonin receptor (5-HT7) mRNA	9
	AF041246	orexin receptor-2 mRNA	10
	M60655	alpha-1B adrenergic receptor mRNA	8
	AB009636	mRNA for phosphoinositide 3-kinase	9
Others			
	M95058	steroid 5-alpha-reductase 2 mRNA	13
	AF000139	25-hydroxyvitamin D 1-hydroxylase (CYP1) mRNA, complete cds	8
	AF012891	frizzled related protein frpAP mRNA	8
	S78284	bcl-xshort=apoptosis inducer [rats, ovary, mRNA Partial, 537 nt]	8
	Y00156	mRNA for hepatic microsomal UDP-glucuronosyltransferase (UDPGT)	10
	AB004278	mRNA for protocadherin 2, partial cds	8
	AJ000347	mRNA for 3 (2),5 -bisphosphate nucleotidase	9
	D63411	mRNA for mitochondrial precursor receptor, complete cds	11
	J04807	insulin II gene mRNA, 3 end	9
	L27061	Phosphodiesterase mRNA, 3 end	8
	X62660	mRNA for glutathione transferase subunit 8	10
	X65296	mRNA for carboxylesterase (Es-HVEL)	8
	U61772	U61772 Rattus norvegicus merlin (NF2) mRNA	8
	U35775	U35775 Rattus norvegicus gamma-adducin mRNA	10
	X66494	X66494 R.norvegicus CHOT1 mRNA	12
	X86086	X86086 R.norvegicus RNA for annexin VI	8
	X62528	X62528 R.rattus mRNA for ribonuclease inhibitor	9
	U57362	U57362 Rattus norvegicus collagen XII alpha 1 (Col12a1) mRNA	8
	D49980	D49980 Rat gene for 8-oxo-dGTPase	11
	U53859	U53859 Rattus norvegicus calpain small subunit (css1) mRNA	9
	D84418	D84418 Rat mRNA for chromosomal protein HMG2	10
	M15883	M15883 Rat clathrin light chain (LCB2) mRNA	9
	J02646	J02646 Rat translational initiation factor (eIF-2) alpha subunit mRNA	8
	L46791	L46791 Rattus norvegicus cholesterol esterase mRNA, complete cds	8
	X55286	X55286 R.norvegicus mRNA for HMG-CoA reductase	9
	AF071495	type II pneumocyte CD36-related class B scavenger receptor (SRB1R) mRNA	10

To assess the relevance of the genes that were selected by the GA/PLS framework to predict the specified metabolic function(s), we evaluated the role of these genes in connection with these functions by way of a literature review. Many of the genes selected by GA/PLS were closely related to enzymes that are involved in reactions that affect TG levels or its precursors. Gene M16235 encodes hepatic lipase enzyme which reduces TG to glycerol and fatty acids. Gene D90109 and J05029 encode long-chain acyl-CoA synthetase and dehydrogenase enzymes, respectively. These two enzymes catalyze the initial reactions of fatty acid β-oxidization and reduce the level of long chain acyl-CoA which is an important precursor in TG synthesis. Similarly, genes involved in microsomal oxidation and fatty acid synthesis enzymes were also identified. The P450 family of genes (such as M20131), encode microsomal oxidase enzyme, which is involved in the oxidation of fatty acid to α-hydroxy fatty acids.

Conversely, gene AB004329 encodes acetyl-CoA carboxylase enzyme, which catalyzes the committed step in fatty acid synthesis to produce malonyl-CoA from acetyl-CoA. The level of fatty acid affects the level of fatty acyl-CoA, which in turn affects TG accumulation. Phosphatidic acid is another important precursor of TG. Gene AF017251 and U03763 encode the phospholipase D and phospholipase A2 enzymes, respectively, which reduce phosphatidylcholine to phosphatidic acid. Another gene X78593 encodes the glycerol-3-P dehydrogenase enzyme which is involved in the first committed pathway to produce phophatidic acid from dihydroxyacetone-P.

In addition, GA/PLS also identified genes involved in the signal pathways that are relevant to TG accumulation, notably, the phosphodiesterase (L27061) and protein kinase

C (M15523) genes. Insulin activates phosphodiesterase to reduce the cAMP level and thus counteracts glucagon's ability to activate TG lipase to hydrolyze TG. Protein kinase C (PKC) is involved in Ca^{2+} related signal transduction. A high level of intracellular Ca^{2+} activates PKC, which attenuates insulin's action and activates lipolysis [Idris 2001]. Interestingly, three genes (D90048, M74494, J04629) encoding Na^+, K^+-ATPase were also identified. The activities of Na^+, K^+-ATPase are significantly decreased in plasma membranes of dietary obese rats [Izpisua 1989]. The reason for this down-regulation is currently not known. Although, Na^+, K^+-ATPase inhibition has been found to be involved in increasing Ca^{2+} level [Xie 2002].

The model also identified genes encoding transcription factors that regulate lipid metabolism. Gene M88592 encodes the transcription factor PPAR-α, which has been found to regulate the expression of many enzymes involved in fatty acid oxidation and synthesis, e.g. fatty acid synthetase, fatty acid oxidase, etc. [Smith 2002]. It acts as a lipid sensor to modulate cellular capacity for fatty acid oxidation and synthesis. Gene M60617 encodes the CCAAT binding transcription factor. C/EBP family of transcription factors are involved in acute phase and inflammatory response of the liver as well as lipid metabolism. This transcription factor regulates fatty acid synthase (FAS) expression [Roder 1999] and activates genes of cholesterol and fatty acid metabolism with sterol regulatory element-binding protein (SREBP) as a co-regulator [Magana 2000].

Indeed, many of the genes selected by the GA/PLS model are closely related to TG metabolism. Based upon the current knowledge from the literature, we reconstructed the role of these genes and gene products in the regulating TG metabolism. For illustration purposes, only the genes enumerated above and their role in regulating TG

accumulation are shown in Figure 2.6. This GA/PLS framework provides a mechanism to identify potential areas (pathways) for further study in optimizing, in the current study, minimizing cellular TG accumulation.

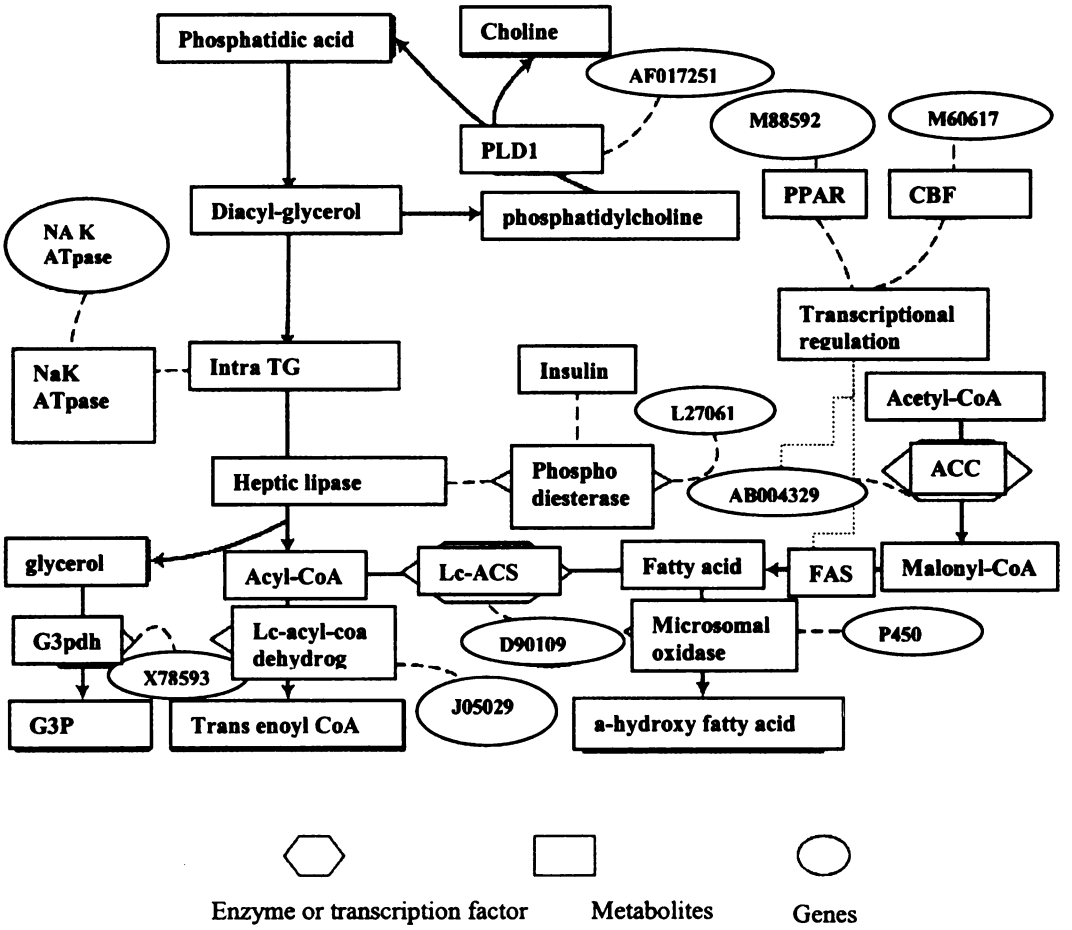


FIGURE 2.6: Pathway reconstruction for TG. The pathways were reconstructed by identifying the biological function of the genes selected by GA/PLS based upon current literature knowledge. The elements of the pathways include gene, enzyme or gene products, and metabolites. The genes include lipid and fatty acid metabolism genes, PPAR-α gene, P450 genes, and Na^+,K^+-ATPase genes. These genes are known in literature to affect TG metabolism. Dotted line indicates an indirect relationship.

Genes Selected for Urea Synthesis

GA/PLS also was applied to select genes relevant to urea synthesis. It selected 57 genes which are listed in Table 2.3B. Based on the expression level of these 57 genes, the

GA/PLS model predicted urea synthesis very well (with a mean square error of 0.0125) as illustrated in Figure 2.7. The role of these genes in regulating urea synthesis was investigated by reviewing the published literature.

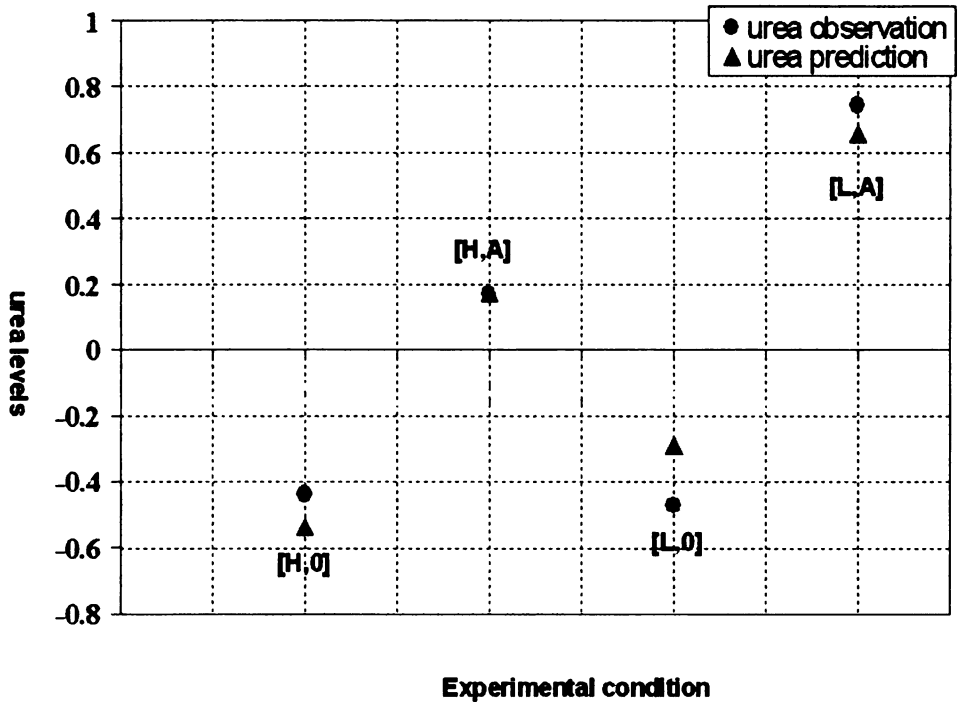


FIGURE 2.7: PLS prediction of urea synthesis. The level of urea synthesis was predicted based upon the expression level of the genes selected by GA/PLS. [H,A], [L,0] was used for training, [H,0] was used for monitoring data, and [H,A] was used for testing and validating the PLS model. The mean square error of the PLS prediction model was 0.0125.

Urea cycle is essential for proper disposal of ammonia in mammals. The cycle is mainly regulated by five enzymes: carbamoyl phosphate synthetase-I (CPS-I; EC 6.3.4.16), ornithine transcarbamylase (OTC; EC 2.1.3.3), argininosuccinate synthetase (AS; EC 6.3.4.5), argininosuccinate lyase (AL; EC 4.3.2.1), and arginase (EC 3.5.3.1). It is encouraging that the genes encoding four of the enzymes CPS-I (Z27513), OTC (K03041), AS (X12459) and AL (D13978) were selected by GA/PLS.

An upstream enhancer required for liver-specific expression and hormonal induction of the rat CPS-I gene has been identified. This enhancer region was revealed to contain DNA elements that bind the transcription factors such as hepatocyte nuclear factor 3 (HNF3) and C/EBP. Detailed transcriptional regulation of urea synthesis has been reviewed in reference (24-26). It is interesting that GA/PLS identified two genes (X12752 and S77528) that encode transcription factors C/EBP and C/EBP-related transcription factors, respectively, and one gene X55955 that encodes the transcription factor HNF3α. In addition, a C/EBP binding element has been found in the OTC and arginase genes [Meijer 2000, Morris 1992, 2002].

GA/PLS also uncovered the coupling between urea cycle, TCA cycle and gluconeogenesis. GA/PLS identified a group of genes involved in gluconeogenesis and TCA cycle that are relevant to urea synthesis. The genes include pyruvate carboxylase (U32314), phospheoenolpyruvate carboxykinase (K03243), glucose-6-phosphatase (G6Pase) (L37333), aspartate aminotransferase (M18467), cytosolic malate dehydrogenase (AF093773) and phosphofructokinase (X58865) genes. Parallel increases in gluconeogenesis and urea synthesis have been found in rat liver during prolonged starvation or by hormonal stimulation [Martin-requero 1992]. The anaplerotic effect of increasing the flux through pyruvate carboxylase may enhance the production of aspartate, which drives urea synthesis by producing more arginosuccinate.

Using the genes highlighted above, we reconstructed the pathways involved in regulating urea synthesis and illustrated in Figure 2.8. Urea synthesis is regulated by genes encoding the enzymes of urea cycle, gluconeogenesis, TCA cycle and transcription factors, such as C/EBP and HNF families.

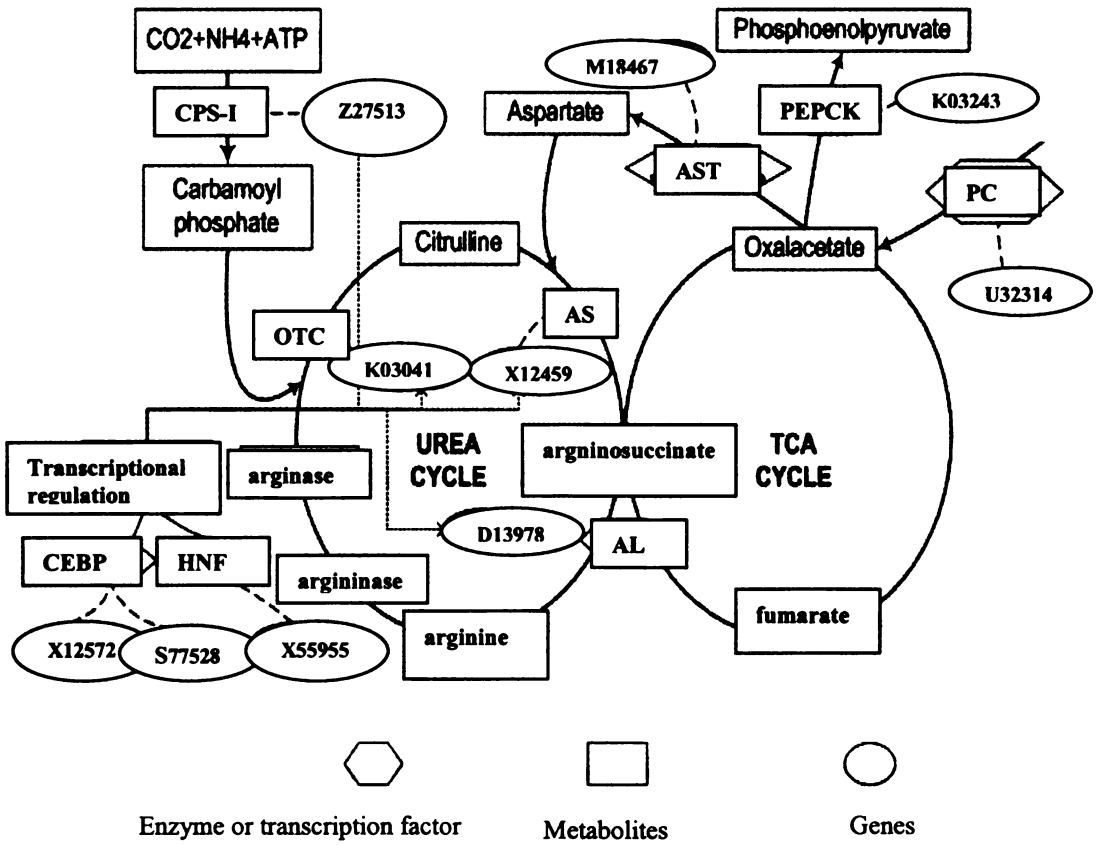


FIGURE 2.8: Pathway reconstruction for urea synthesis. The pathways were reconstructed by identifying the biological function of the genes selected by GA/PLS. The elements of the pathways include gene, enzyme or gene products, and metabolites. The genes include urea cycle genes, C/EBP and HNF genes, gluconeogenic and TCA cycle genes. These genes are known in the literature to affect urea synthesis. Dotted line indicates an indirect relationship.

2.3.4 Model Robustness Evaluation

Gene expression data is typically noisy, therefore, we evaluated the effect of noise addition to the gene expression data on our model results. To test the robustness of the model, 10% noise was artificially added to the gene expression data. Noise in the form of a Gaussian distribution was generated according to the following equation,

$$Noise = N * \sigma_s * R$$

where N is a vector of random numbers of length n following a normal distribution, the σ_s is the standard deviation of the data, and the R is any number between [0,1]. The robustness of the GA/PLS framework was evaluated by comparing the genes selected before and after noise addition.

The genes selected after noise addition are enumerated in Table 2.4A and Table 2.4B for intracellular TG and urea production, respectively. 31 of the 59 genes remain selected for TG and 34 of the 57 genes remain selected for urea synthesis (highlighted in bold, Table 2.4). Of the 59 genes selected for TG before noise addition, 14 were from the supervised group (the 45 genes added after t-test). It is notable that 11 of the 14 genes selected from the supervised group remained the same after noise addition. Similarly, 12 of the 13 genes selected from the supervised group remained for urea synthesis. In the case of TG, most of the fatty acid and lipid metabolism genes such as long-chain fatty acyl-CoA synthetase and dehydrogenase, phospholipase, hepatic lipase, acetyl-CoA carboxylyase genes remain selected. In addition, the CCAAT binding transcription factor, Na^+,K^+ ATPase and P450 family genes were also selected after noise addition. In the case of urea synthesis, the urea cycle enzymes, e.g. AS, OTC, CPS-I, aspartate aminotransferase genes; and transcription factors genes, C/EBP and HNF; and gluconeogenesis enzyme genes, such as pyruvate carboxylase, G6Pase, phosphofructokinase and phosphoenolpyruvate carboxykinase genes remain selected after noise addition. This suggests that GA/PLS framework is robust to a certain level of noise in the gene expression data. Nevertheless, several genes such as the PPAR gene for TG and AL gene for urea synthesis disappeared after noise addition.

Table 2.4 A. Genes selected for intracellular triglyceride after noise addition

Functional Category	Accession Number	Gene name	Frequency
Fatty acid and lipid metabolism			
	J05156	cholesterol side-chain cleavage enzyme mRNA (P450SCC)	9
	D90109	mRNA for long-chain acyl-CoA synthetase (EC 6.2.1.3)	8
	M16235_at	M16235 Rat hepatic lipase mRNA	11
	X78593_at	mRNA for glycerol-3-phosphate dehydrogenase	13
	J05029_s_at	RATACOADA Rat long chain acyl-CoA dehydrogenase (LCAD) mRNA	9
	D38448_at	mRNA for 88kDa-diacylglycerol kinase (DGK-III), complete cds	10
	AF017251_at	phospholipase D (PLDs) mRNA, complete cds	11
	X54096_at	mRNA for lecithin-cholesterol acyltransferase (EC 2.3.1.43) (LCAT)	9
	AB004329_at	mRNA for acetyl-CoA carboxylase, complete cds	12
	L06040_s_at	12-lipoxygenase mRNA, complete cds	9
	X07467_at	mRNA for glucose-6-phosphate dehydrogenase (Gd, EC 1.1.1.49)	11
Transcription factor			
	X54250	RRATBP2 Rat mRNA for zinc finger protein AT-BP2, partial cds	9
	M60617	CCAAT binding transcription factor-B subunit (CBF-A11) mRNA	8
Ion transporter			
	D90048	Na+, K+ -ATPase (EC 3.6.1.3) beta2 subunit gene and 5 flank	11
	U04934	(CD-1) clone Kc1 Na-Ca exchanger mRNA, partial cds	8
	U78090	Potassium channel regulator 1 mRNA, complete cds	10
Electron transport chain Oxidative phosphoration			
	E03344	cDNA sequence of peroxisome forming factor	8
	K00750	cytochrome c nuclear-encoded mitochondrial gene and flanks	9
	K01721	P-450(1) variant (phenobarbital-inducible) cytochrome 3 end, flank	8
	U39207	Cytochrome P450 4F5 (CYP4F5) mRNA	8
Signal pathways			
	U93306	VEGF receptor-2/FLK-1 mRNA	9
	M80784	type III TGF-beta receptor mRNA, complete cds	9
	U48596	MAP kinase kinase kinase 1 (MEKK1) mRNA	8
	D83538	mRNA for 230kDa phosphatidylinositol 4-kinase	8
	A09811	mRNA for BRL-3A binding protein	8
	AF085693	ADP ribosylation factor GTPase-activating protein (GIT1) mRNA	10

Table 2.4A (continued)

	L04672	G protein-coupled receptor mRNA, complete cds	9
	L22558	adenylyl cyclase-activated serotonin receptor (5-HT7) mRNA	8
	M60655	Rat alpha-1B adrenergic receptor mRNA	9
	J04486	insulin growth factor-binding protein mRNA	8
	M14053	Rat glucocorticoid receptor mRNA	8
	M15523_s_at	Rat protein kinase C-family related mRNA, partial cds, clone RP16	9
Others			
	AF000139	25-hydroxyvitamin D 1-hydroxylase (CYP1) mRNA, complete cds	10
	M95058	steroid 5-alpha-reductase 2 mRNA	9
	S78284	bcl-xshort=apoptosis inducer [rats, ovary, mRNA Partial, 537 nt]	9
	U18982	fos-related antigen 2 (fra-2) mRNA, complete cds	9
	Y00156	mRNA for hepatic microsomal UDP-glucuronosyltransferase (UDPGT)	12
	AB010154	PKN mRNA for serin/threonine protein kinase expressed in hippocampus	10
	AB011530	mRNA for MEGF4, complete cds	8
	D13978	mRNA for argininosuccinate lyase, complete cds	8
	D79215	mRNA for FGF-10, complete cds	8
	X62660	mRNA for glutathione transferase subunit 8	9
	X65296	mRNA for carboxylesterase (Es-HVEL)	8
	M27433	germinal histone H4 gene	8
	U44750	NAD-dependent 15-hydroxyprostaglandin dehydrogenase mRNA	9
	J05031	Rat isovaleryl-CoA dehydrogenase (IVD) mRNA	9
	U35775	gamma-adducin mRNA	8
	X66494	CHOT1 mRNA	10
	X86086	RNA for annexin VI	8
	X91234	mRNA for 17-beta hydroxysteroid dehydrogenase type 2	9
	U53859	calpain small subunit (css1) mRNA	8
	J02646	Rat translational initiation factor (eIF-2) alpha subunit mRNA	13

Table 2.4 B. Genes selected for urea synthesis after noise addition

Functional Category	Accession Number	Gene name	Frequency
Urea cycle			
	X12459	mRNA for argininosuccinate synthetase (EC 6.3.4.5)	8
	M18467	mitochondrial aspartate aminotransferase mRNA	9
	K03041	mRNA RATOTCB Rat (Sprague-Dawley) ornithine carbamoyltransferase mRNA	8
	Z27513	gene for carbamoylphosphate synthase I, exon 38	11
gluconeogensis			
	L37333	glucose-6-phosphatase (G6Pase) mRNA, complete cds	9
	X58865	mRNA for liver phosphofructokinase	9
	U32314	pyruvate carboxylase mRNA	12
	K03243	Rat phosphoenolpyruvate carboxykinase (GTP) gene, exons 1-3	12
	X53588	mRNA for glucokinase, alternatively spliced GK2 (EC 2.7.1.1)	11
	J05446	Rat glycogen synthase mRNA, complete cds	9
TCA cycle			
	AF093773	cytosolic malate dehydrogenase (Mdh) mRNA, complete cds	9
Transcription factor			
	M88592	peroxisome proliferator activated receptor (PPAR) mRNA,	8
	J03170	liver specific transcription factor (LF-B1) gene	9
	X12752	gene for DNA binding protein C/EBP	11
	X55955	mRNA for hepatocyte nuclear factor 3A (HNF-3A)	12
	S77528	rNFIL-6=C/EBP-related transcription factor [rats, Genomic/mRNA, 1759 nt]	9
Electron transport chain Oxidative phosphoration			
	K00996	RATCYP45E Rat cytochrome p-450e (phenobarbital-induced) mRNA, 3 end	10
	K01721	RATCYP45A Rat P-450(1) variant (phenobarbital-inducible) cytochrome 3 end, flank	8
	E03344	E03344cds cDNA sequence of peroxisome forming factor	9
	U39943	cytochrome P450 monooxygenase (CYP2J3) mRNA, complete cds	10
Signal pathways			
	U12187	ras-related protein (rad) mRNA	9
	U48596	MAP kinase kinase kinase 1 (MEKK1) mRNA	10
	D83538	mRNA for 230kDa phosphatidylinositol 4-kinase	9
	AF082126	aryl hydrocarbon receptor (AHR) mRNA	8
	AF085693	ADP ribosylation factor GTPase-activating protein (GIT1) mRNA	12
	D10891	mRNA for metabotropic glutamate receptor mGluR5, complete cds	12

Table 2.4B (continued)

	D63411	Rat mRNA for mitochondrial precursor receptor, complete cds	9
	J04807	Rat insulin II gene mRNA, 3 end	10
	U37099	small GTP-binding protein (rab3c) mRNA, partial cds	8
	X06889	ras-related mRNA rab3	10
	Z11932	mRNA for vasopressin V2 receptor	10
	L22558	adenylyl cyclase-activated serotonin receptor (5-HT7) mRNA	11
	L10072	5-hydroxytryptamine receptor (5HT5a) mRNA	8
	J04486	insulin growth factor-binding protein mRNA	8
	AB009636	mRNA for phosphoinositide 3-kinase	9
	M14053	Rat glucocorticoid receptor mRNA	8
	M55417	Rat protein kinase C-gamma (PRKC-gamma) gene, exon 1	10
Others			
	D90048	Rat Na+, K+ -ATPase (EC 3.6.1.3) beta2 subunit gene and 5 flank	9
	M95058	M95058completeSeq Rattus rattus steroid 5-alpha-reductase 2 mRNA	8
	U11681	U11681 Rattus norvegicus rapamycin and FKBP12 target-1 protein (rRAFT1) mRNA	10
	S78284	apoptosis inducer [rats, ovary, mRNA Partial, 537 nt]	8
	U18982	fos-related antigen 2 (fra-2) mRNA, complete cds	11
	Y00156	mRNA for hepatic microsomal UDP-glucuronosyltransferase (UDPGT)	9
	AB004278	mRNA for protocadherin 2, partial cds	8
	AB010154	mRNA for serin/threonine protein kinase expressed in hippocampus, partial cds	9
	S52878	fatty acid-binding protein homolog [rats, ileum, mRNA, 460 nt]	8
	S76404	S76401S2 beta -HKA=H,K-ATPase beta-subunit [rats, Genomic, 8983 nt, segment 2 of 2]	8
	U04934	clone Kc1 Na-Ca exchanger mRNA, partial cds	8
	X62660	X62660mRNA RRGTS8 R.rattus mRNA for glutathione transferase subunit 8	9
	E02468	E02468cds DNA sequence of rat TNF	8
	E01884	E01884cds DNA sequence coding for rat IL-1-beta(interleukin-1 beta)	9
	M27433	M27433 Rattus norvegicus germinal histone H4 gene	8
	X86086	X86086 R.norvegicus RNA for annexin VI	9
	X62528	mRNA for ribonuclease inhibitor	8
	U53859	calpain small subunit (css1) mRNA	8
	AB006137	mRNA for alpha 1,2-fucosyltransferase	8
	D90109	Rat mRNA for long-chain acyl-CoA synthetase (EC 6.2.1.3)	8
	J02646	J02646 Rat translational initiation factor (eIF-2) alpha subunit mRNA	8
	L46791	L46791 Rattus norvegicus cholesterol esterase mRNA, complete cds	10
	X55286	mRNA for HMG-CoA reductase	8
	AF071495	type II pneumocyte CD36-related class B scavenger receptor (SRB1R) mRNA	8

Table 2.4B (continued)

J05156	cholesterol side-chain cleavage enzyme mRNA (P450SCC), complete cds	9

2.4 Discussion

We applied a GA/PLS framework to gene expression and metabolic data to identify a subset of genes whose expression levels could predict metabolic functions. Subsequently, part of the regulatory pathways was reconstructed based upon the functional information of the genes. The selection of relevant genes was obtained using GA guided by a fitness function (equation 2.7) that was defined by the prediction error and the number of latent variables. No *a priori* information was considered in equation 2.7. The gene selection process was based solely upon the gene expression and metabolic data. This provides a framework to uncover information from the data itself.

However, the information extracted by GA/PLS is redundant for both intracellular TG and urea synthesis, i.e., many of the selected genes are not relevant to these cellular functions. It was observed that the genes selected from the supervised group (45 genes added after t-test) were more relevant to these cellular functions than from the unsupervised group (111 genes selected by t-test). To reduce the redundant information, an approach that weighs the genes based on *a priori* knowledge may be better suited. If one knows that some genes are related and others are not related to a metabolic function, this information can be incorporated into the fitness function by adding a factor that awards the inclusion of relevant genes and punishes the inclusion of irrelevant genes. Each gene is assigned a score based upon its relevance to the metabolic function based upon the literature. An overall gene relevance score (*Gs* in equation 2.10) can be obtained

by summing up the score of each individual gene and integrating it into a fitness function as follows:

$$fitness = \frac{1}{\sum (y_i - \hat{y}_i)^2 + LV^{w1}} + Gs^{w2} \qquad (2.10)$$

To illustrate the ability of this fitness function (equation 2.10) to incorporate domain knowledge, we set up a hypothetical example. In this hypothetical example, we included only the genes selected by t-test and assumed we know *a priori* that genes (D90109, J04087, M88592, U37099) are relevant to intracellular TG accumulation and that the other genes are not relevant. Therefore, a score of one was assigned to genes (D90109, J04087, M88592, U37099), and a score of zero was assigned to all the other genes listed in Table 2.1 for intracellular TG. The parameter *w1* and *w2* were assigned values of 0.4 and 0.1, see methods for determination of the *w1 and w2,* so that the prediction error, number of latent variables and gene score are normalized to the same scale. Using this new fitness function in the GA/PLS model with the same gene expression and metabolic data, we evaluated the robustness of the GA/PLS model to the addition of 10% noise in the gene data. The genes selected before and after noise addition are listed in Tables 2.5A and 2.5B, respectively. From the results in Table 2.5, we note that the genes assigned a relevance score of one were selected before and after noise addition and at a high frequency (i.e., selected each time the model was run, regardless of the initial population), which suggests that incorporating *a priori* knowledge of the relevance of the genes into the fitness function (equation 2.10) ensures that their selection is not affected by the noise in the data. In addition to the genes that were defined *a priori* to be relevant, other genes such as P450 genes, Na^+, K^+-ATPase genes, which are not defined *a priori* as relevant,

were also selected by the GA/PLS model with this new fitness function, both before and after noise addition. Therefore, by defining an appropriate fitness function, the GA/PLS framework was able to incorporate available domain knowledge into the model as well as able to recognize genes relevant to a metabolic function without *a prior* knowledge. It is possible that the relevance of a number of genes to a cellular function may currently not be known. Therefore an advantage of not including *a priori* knowledge allows GA/PLS model to be naïve to any biases in the user and in turn may allow the uncovering of genes that may be important and to be investigated further.

Table 2.5A. Genes selected for intracellular TG using fitness function in equation 2.10

Function Category	Accession Number	Gene name	Freq	Score
Lipid metabolism				
	D90109	mRNA for long-chain acyl-CoA synthetase (EC 6.2.1.3)	14	1
	L27061	Phosphodiesterase mRNA	11	0
Transcription factor				
	M88592	peroxisome proliferator activated receptor (PPAR) mRNA	14	1
	J03170	liver specific transcription factor (LF-B1) gene	10	0
	M60617	CCAAT binding transcription factor-B subunit (CBF-A11) mRNA	9	0
	X54250	X54250mRNA RRATBP2 Rat mRNA for zinc finger protein AT-BP2	12	0
Ion transporter				
	J04629	(Na+, K+)-ATPase-beta-2 subunit mRNA	9	0
	M74494	sodium/potassium ATPase alpha-1 subunit truncated isoform mRNA	12	0
Electron transport chain Oxidative phosphoration				
	K00996	cytochrome p-450e (phenobarbital-induced) mRNA	8	0
	K00750	cytochrome c nuclear-encoded mitochondrial gene and flanks	9	0
	J01435	mitochondrial cytochrome oxidase subunits I,II, III genes,	11	0
	U39207	cytochrome P450 4F5 (CYP4F5) mRNA	8	0
	E03344	cDNA sequence of peroxisome forming factor	8	0
Signal pathways				
	U93306	VEGF receptor-2/FLK-1 mRNA	14	0
	M80784	type III TGF-beta receptor mRNA	13	0

Table 2.5A (continued)

	U68544	cyclophilin D mRNA, nuclear gene encoding mitochondrial protein	12	0
	D83538	mRNA for 230kDa phosphatidylinositol 4-kinase	8	0
	J04807	RATINSIIA Rat insulin II gene mRNA	14	1
	U37099	small GTP-binding protein (rab3c) mRNA	14	1
	AF085693	G protein-coupled receptor kinase-associated ADP ribosylation factor GTPase-activating protein (GIT1) mRNA	8	0
	X74833	X mRNA for acetylcholine receptor beta-subunit	12	0
	L22558	adenylyl cyclase-activated serotonin receptor (5-HT7) mRNA	9	0
	M60655	alpha-1B adrenergic receptor mRNA	13	0
	X04979	X04979 Rat gene for apolipoprotein E	9	0
Others				
	Y00156	mRNA for hepatic microsomal UDP-glucuronosyltransferase (UDPGT)	8	0
	J03960	5-lipoxygenase mRNA	11	0
	S59893	La=autoantigen SS-B/La	9	0
	U75923	isoleucyl tRNA synthetase mRNA	9	0
	M64033	Rat secretin gene	13	0
	M27433	Germinal histone H4 gene	10	0
	J05031	Isovaleryl-CoA dehydrogenase (IVD) mRNA	9	0
	X66494	X66494 R.norvegicus CHOT1 mRNA	9	0
	X80477	X80477 R.norvegicus P2X mRNA	9	0
	X86086	X86086 R.norvegicus RNA for annexin VI	12	0
	X91234	X91234 R.norvegicus mRNA for 17-beta hydroxysteroid dehydrogenase type 2	12	0
	U53859	calpain small subunit (css1) mRNA	8	0
	J02646	translational initiation factor (eIF-2) alpha subunit mRNA	9	0

Table 2.5B. Gene selected for intracellular TG with fitness function in equation 2.10 after noise addition.

Function Category	Accession Number	Gene name	Freq	Score
Lipid metabolism				
	D90109	mRNA for long-chain acyl-CoA synthetase (EC 6.2.1.3)	14	1
	L27061	phosphodiesterase mRNA	11	0
Transcription factor				
	M88592	peroxisome proliferator activated receptor (PPAR) mRNA	14	1
	M60617	CCAAT binding transcription factor-B subunit (CBF-A11) mRNA	13	0
	X54250	X54250mRNA RRATBP2 Rat mRNA for zinc finger protein AT-BP2	10	0
Ion transporter				
	M22253	M22253 Rattus norvegicus sodium channel I mRNA	8	0
	M74494	sodium/potassium ATPase alpha-1 subunit truncated isoform mRNA	10	0

Table 2.5B (continued)

Electron transport chain Oxidative phosphoration				
	M20131	cytochrome P450IIE1 gene	13	0
Signal pathways				
	U93306	VEGF receptor-2/FLK-1 mRNA	14	0
	M80784	type III TGF-beta receptor mRNA	9	0
	U68544	cyclophilin D mRNA, nuclear gene encoding mitochondrial protein	14	0
	J04807	RATINSIIA Rat insulin II gene mRNA	14	1
	U37099	small GTP-binding protein (rab3c) mRNA	14	1
	X74833	mRNA for acetylcholine receptor beta-subunit	12	0
	L22558	adenylyl cyclase-activated serotonin receptor (5-HT7) mRNA	9	0
	M60655	alpha-1B adrenergic receptor mRNA	12	0
	X04979	X04979 Rat gene for apolipoprotein E	10	0
	U53859	calpain small subunit (css1) mRNA	10	0
Others				
	Y00156	mRNA for hepatic microsomal UDP-glucuronosyltransferase (UDPGT)	9	0
	J03960	5-lipoxygenase mRNA	11	0
	AB011530	mRNA for MEGF4	12	0
	D10891	mRNA for metabotropic glutamate receptor mGluR5	8	0
	S48325	S48325 diabetes-inducible cytochrome P450RLM6	12	0
	E01884	DNA sequence coding for rat IL-1-beta(interleukin-1 beta)	11	0
	U75923	isoleucyl tRNA synthetase mRNA	10	0
	M64033	Rat secretin gene	10	0
	M27433	Germinal histone H4 gene	10	0
	U61772	merlin (NF2) mRNA	9	0
	U44750	NAD-dependent 15-hydroxyprostaglandin dehydrogenase mRNA	9	0
	J05031	Isovaleryl-CoA dehydrogenase (IVD) mRNA	11	0
	X66494	X66494 R.norvegicus CHOT1 mRNA	8	0
	X80477	X80477 R.norvegicus P2X mRNA	8	0
	X86086	X86086 R.norvegicus RNA for annexin VI	9	0
	X91234	mRNA for 17-beta hydroxysteroid dehydrogenase type 2	13	0
	J02646	translational initiation factor (eIF-2) alpha subunit mRNA	10	0

In this chapter, GA/PLS was used to find a set of possible solutions rather than a single solution. With this method, multiple solutions of different genes give similar prediction accuracy. By selecting genes based upon their frequency of appearance in the multiple runs, we explored the sub-gene-space for high prediction accuracy. The probability of significant features (important genes) appearing in the sub-gene-space was

estimated based upon their frequency. The probabilistic nature of this method improved the robustness of the GA/PLS approach. Increasing the number of runs enhances the effectiveness of the probabilistic nature of gene selection. Although, GA can not guarantee global optimization, it is possible to improve the performance of GA by combining global optimization methods, such as simulated annealing.

GA/PLS can be applied to group genes according to the metabolic function they are involved in. As opposed to gene clustering, GA/PLS used both metabolic profile and gene expression data to functionally group genes according to a metabolic function. In addition, one gene can be assigned to more than one group, which is not permitted with clustering methods. Typically, a gene may be involved in multiple pathways and may regulate different cellular functions. For example, transcription factors could affect more than one gene, which in turn may be involved in a number of different pathways. The grouping of genes according to metabolic functions facilitates the reduction of a complex biological system with thousands of genes to a system of functional subgroups. This could help facilitate the optimization of cellular function.

A subsequent step after identifying the gene subsets is reconstructing the regulatory pathways based upon the gene expression and metabolic profiles. Data driven reverse engineering methods such as Bayesian network analysis, which has been used successfully in reconstructing metabolic sub-networks from flux data, [Li 2004] could be used to infer regulatory networks at both the genetic and metabolic levels as well as the interaction between the two levels. The resulting model could be used to suggest hypothetical pathways involved in regulating metabolic functions, and thus likely pathways for further experimental studies.

The current model does not incorporate translational and post-translational effects in the gene selection process and thus may be a possible reason why some of the important genes disappeared after noise addition. Incorporating other data types, such as protein expression profile, protein-protein and protein-DNA interactions would enhance the ability of modeling frameworks to predict cellular and physiological function more accurately.

In conclusion, GA/PLS was applied for the first time to quantitatively integrate experimental gene expression and metabolic flux profiles for a cellular system. The results indicate that this method is able to select a subset of genes capable of predicting a metabolic function. In addition, using the selected genes, we were able to reconstruct the pathways involved in regulating a metabolic function.

CHAPTER 3 A BAYESIAN FRAMEWORK TO INFER BIOLOGICAL PATHWAYS AND NETWORK

3.1 Introduction

The introduction of array technology has made readily attainable thousands of genes and proteins for a given metabolic or physiological state, providing a wealth of data from which regulatory and functional relationships may be unraveled. A major challenge limiting the application of these data is the ability to integrate this information, from the gene through the metabolic level, in such a way that it reconstructs the biological process. Moreover, environmental factors, in addition to the genetic wiring, contribute to the regulation and control of the connections within the biological system. Thus, new approaches and mathematical tools for analyzing this information are needed. The motivation behind reverse engineering of pathways and networks from experimental data is the belief that the data contain the structure of the biological system or phenotype. Therefore, the goal is to provide a framework to extract the network structure and connections from the experimental data, thus capturing the effects and changes due to external stimuli and its interactions with the genetic wiring. In chapter 2, we introduced GA/PLS to identify important genes relevant to a cellular function. We then used information from the literature to manually reconstruct the pathways for TG and urea functions, which depended upon the known information currently available. In order to discover pathways from the data itself, we introduce a Bayesian framework in this chapter.

As a first level attempt at reconstruction, numerous studies (Friedman et al., 2000; Pe'er et al., 2001) endeavor to infer gene regulatory pathways and networks from micro-array and simulated data from prokaryotes and yeasts. Typically networks are not known a priori, thus the validity and accuracy of these reconstructions cannot be accessed. Instead, the reconstructed networks are used to suggest biological hypotheses. Alternatively, our goal is to develop a mathematical framework that infers biochemical pathways and networks from metabolic data from mammalian cells (systems). The rationale for reconstructing metabolic networks first, prior to applying this framework to gene expression data, is to confirm the ability of the proposed approach to reconstruct known biological networks, such as the TCA and urea cycles. This approach provides a degree of confidence in the novel networks that may be reconstructed from experimental data. Nevertheless, experimental verification of these novel networks is ultimately still required.

Numerous mathematical models, from Boolean or neural network models [Liang et al., 1998] to deterministic (differential equation based) models [Chen et al., 1999] have been applied to represent biological networks. Thus far, the former algorithms have been used predominantly to infer network structures from gene data, both experimental [Hakamada et al., 2001] and simulated [Akutsu et al., 1999; Maki et al., 2001]. The limitation of the Boolean network model is that it captures only the logical relations within the data. With this method, the genes take on binary (1/0) inputs and give binary outputs according to logical or Boolean rules. Boolean methodology provides a context to analyze logical rules such as $A = B$ or C. The information captured by such a model would be, for example, whether gene A can be activated by either gene B or gene C.

Boolean models are limited to time series data and cannot be used to evaluate quantitatively the dependency between pathways. To infer networks and interactions requires knowledge, to a certain degree, of the cause-effect relationships among the variables. Typically, these data-driven approaches are based up correlation, and correlation between two variables does not guarantee causality [Sprites et al., 1993]. Data-driven methods (herein denoted as data-driven reverse engineering methods), infer relations between variables (pathways) from the data itself without any a priori assumptions of the mechanisms involved, precluding statistical hypothesis testing of possible regulatory models [Hartemink et al., 2001]. Similarly, deterministic models require detailed information that is presently unavailable for mammalian systems, certainly not in the detail required to model the system sufficiently to suggest or evaluate hypothetical environmental conditions. Therefore, the complexity of mammalian systems limits the applicability of deterministic models to relatively simple biological systems and not easily to gene expression data.

In contrast, for certain microorganisms, metabolic networks and their control circuits have been reconstructed from currently available gene sequence data combined with a literature search of the known biochemical pathways and their related physiological functions [Schilling et al., 1999]. These networks are built from a knowledge-based methodology and permit hypothesis testing of possible regulatory models (herein denoted as model-driven hypothesis testing methods). These knowledge-based approaches are formulated by predefining or assuming a model and evaluating its likelihood with respect to existing sequence data. Therefore, these models cannot easily assess the effects of changes in the environment on the hypothetical networks, although,

the metabolic regulatory network of the *E. coli* has been redesigned by using metabolic control analysis from experimental data [Farmer and Liao, 2000].

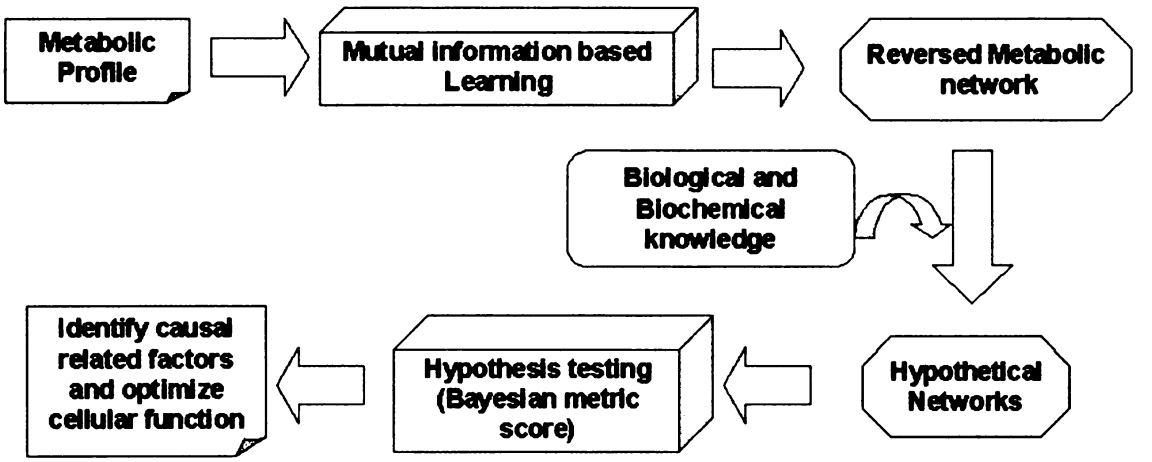


Figure 3.1. Bayesian network based framework. Mutual information based Bayesian network structure learning algorithm was first applied to the metabolic profile to infer the regulatory network. Incorporating known biological knowledge into the formulation of the hypothetical networks further refined the structures. Bayesian metric score identified the mostly likely network structure.

Alternatively, we developed a Bayesian-based framework, schematically shown in Figure 3.1, which combines data-driven reverse engineering and model-driven hypothesis testing methods to re-construct sub-network structures, for example, TCA and urea cycles, from hepatocellular metabolic data. The advantage of the Bayesian framework over other data-driven methods is the ability of the Bayesian approach to perform cause-and-effect analysis, thus providing a basis of identifying causal relationships. This is accomplished without a priori detailed knowledge or assumptions of the biological system and the governing equations but rather is based on the concept of conditional probability. Therefore, similar to model-driven methods, the Bayesian network approach permits the evaluation of several hypothetical networks by using quantitative measures, such as a Bayesian metric score, to access the likelihood of a proposed network structure. In

contrast to other model-driven methods, our framework incorporates experimental data, which allows one to evaluate the affects of environmental changes.

We integrated several Bayesian based techniques into our framework. Data-driven Bayesian network learning method was first applied to learn the raw metabolic regulatory network from the data itself. Many algorithms exist that can infer the Bayesian network structure. Only a few are computationally efficient enough to deal with large datasets. One such method is the information theory-based learning algorithm, which we have applied to the data to infer the underlying metabolic regulatory network. The inferred network was then compared with the known metabolic network to evaluate the ability of our framework to learn for example the TCA and urea cycles. Next, a latent variable detection algorithm was applied to the inferred sub-network with the goal of finding possible latent variables in the network by discovering indirect pathways relevant to our objective function. Methods, such as inductive causation (IC*), were applied to the sub-network composed of the identified pathways with the goal of finding possible latent variables in the network that influence biological (objective) functions, such as triglyceride accumulation and the rate of urea synthesis. Then, based on the pathways learned from the data and currently known biological associations, we postulated several metabolic regulatory network models. The utility and accuracy of these alternative models were evaluated by comparing 1) their Bayesian metric scores, calculated from the experimental data, to assess the model's goodness of fit to the data; and 2) how well the model predicted the objective function(s). Thereupon an optimal model representing the metabolic regulatory network was selected based on these criteria. Finally, a sensitivity analysis was applied to evaluate the stability and robustness of the system or phenotype

with respect to changes, that is, noise, in the system parameters, namely, the environmental or metabolic variables.

The ultimate goal of these models is to aid our understanding of the underlying mechanism that governs the biological systems and the transition of these systems to different states. In the current investigation, our system experiences a transition in its microenvironment going from culture medium to plasma. Although, for the purpose of the current investigation, we evaluated the system's steady-state. The structure of the biological network can be obtained from the steady-state data, which represents the cellular, tissue, or organ phenotype or state. Once the structure is attained, one can evaluate, using transient data, the system dynamics and how it changes from the steady-state. Understanding the system dynamics would permit the management of transitions from a diseased to a more beneficial or normal phenotype.

3.2 MATERIAL AND METHODS

3.2.1 Data Collection

The sources of the dataset used in this study were previously published elsewhere [Chan et al., 2003a,b,c], and what follows is a brief description of the experimental method and the metabolic flux analysis (MFA) used to obtain the dataset that was used in BN analysis. Primary isolated hepatocytes were cultured in a collagen sandwich configuration and incubated in standard hepatocyte culture medium containing either 0.5 U/ml insulin (high insulin) or 50 μU/ml insulin (low insulin). The hepatocytes were cultured in this fashion for at least 6 days prior to plasma exposure. This interval is considered the pre-conditioning period. The six-day-old-sandwiched hepatocyte cultures

were subsequently exposed to unsupplemented, amino acid, hormone, or amino acid plus hormone supplemented plasma solution for an additional seven days (see Chan et al., 2003a,b for details). Therefore, six combinations of pre-conditioning-plasma supplementation were evaluated. They were i) low-insulin, pre-conditioned, and unsupplemented plasma; ii) low-insulin, pre-conditioned, and amino acid-supplemented plasma; iii) high-insulin pre-conditioned and unsupplemented plasma; iv) high-insulin, pre-conditioned, and amino acid-supplemented plasma; v) high-insulin, pre-conditioned, and hormone-supplemented plasma; and vi) high-insulin, pre-conditioned, and amino acid plus hormone-supplemented plasma. A model for hepatocyte metabolism was created based on the known stoichiometry of the hepatic metabolic network. The stoichiometry of each reaction in the MFA model is listed in Chan 2003a. A total of 33 metabolite measurements were coupled to the MFA to estimate the other 43 intracellular fluxes (see Chan et al., 2003a,b for details). The structure of the resulting data, namely, the measured metabolites (in bold) and the estimated intracellular flux values, with their corresponding flux numbers, are illustrated in Table 3.2. To help illustrate the overall model structure see appendix Figure 1, we applied our Bayesian-based methodology to the data in Chan 2003a, to initially infer the known biological networks, such as the TCA and urea cycles. After which, the methodology was applied to infer the network structures relevant to intracellular triglyceride (TG) accumulation.

Table 3.1. Learned TCA cycle relationships

Learned relation	Biological explanation
7 → 9	Synthesis of oxaloacetate (pathway no. 7) controls the synthesis of citrate (pathway no. 9).
9 → 10	Citrate synthesis controls (pathway no. 9) the reaction from citrate to α-ketoglutarate (pathway no. 10).
11 → 12	α-ketodehydrogenase (pathway no. 11) controls the formation of fumarate from succinyl-CoA (pathway no. 12).
13 → 14	Synthesis of malate from fumarate (pathway no. 13) controls the reaction from malate to oxaloacetate (pathway no. 14).
9 ← 7 → 8	Link between the TCA cycle and gluconeogenesis through pyruvate carboxylase (pathway no. 7).

3.2.2 Bayesian networks

Bayesian networks are directed acyclic graphs (DAG) whose nodes correspond to variables and whose arcs represent the dependencies between variables. The dependencies are determined by the conditional probabilities of each node x_i, given its parent node p_a, $\Pr(x_i \mid p_a(x_i))$. A Bayesian network assumes conditional independence, such that each node is independent to its non-descendants, given its parents. In other words, x_i and x_j are conditionally independent to each other given p_a (see Figure 3.2), then

$$Pr(x_i \mid x_j, p_a(x_i)) = Pr(x_i \mid p_a(x_i)) \quad (3.1)$$

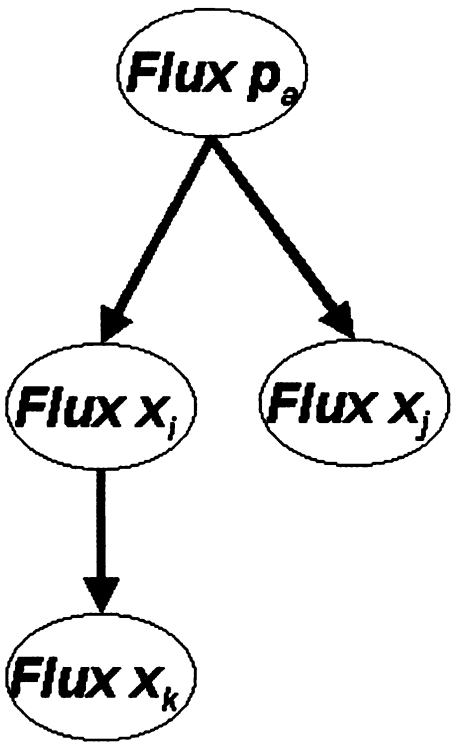


Figure 3.2. An example of a simple Bayesian network. The network consists of three nodes P_a, X_i, X_j and X_k in which P_a is the parent node of X_i and Xj. X_i and X_j are conditionally independent to each other given the parent node P_a. P_a is the cause of or regulates X_i and X_j; X_i and X_j are the effects of P_a.

Also, a Bayesian network consists of the joint distribution defined by a set of variables $\{x_i\}$ as:

$$\Pr(x_1,....,x_n) = \prod_{i=1}^{N} \Pr(x_i \mid p_a(x_i)) \tag{3.2}$$

In the example above:

$$Pr(p_a, x_i, x_j, x_k) = Pr(p_a)\, Pr(x_i|p_a)\, Pr(x_j|p_a)\, Pr(x_k \mid x_i) \tag{3.3}$$

The approaches developed to learn the Bayesian network structure from either experimental or simulated data generally fall into two categories: 1) search and scoring or

2) constraint-based approach (dependency analysis approach). With the search-and-scoring-based approach, the algorithm starts with a graph without arcs, and then new arcs are added and evaluated with a specific scoring function. A scoring function is calculated, and if the score improves, the new arc is kept; otherwise it is eliminated. Thus, the Bayesian network is learned through optimizing this scoring function. Numerous scoring functions have been applied to this type of algorithm, such as Bayesian scoring method [Heckerman et al., 1994] and minimum description length method [Suzuki, 1996]. Given the immense computational time and cost associated with this approach, heuristic methods to reduce the search space, such as the hill climb, greedy searches, and simulated annealing, have been applied. The constraint-based approach applies the conditional independency (CI) test to reconstruct the Bayesian network by uncovering the dependencies within the data. Several algorithms are available for this approach; one of the better-known algorithms is inductive causation (IC) [Sprites et al., 1996]. In the current investigation we combined the constraint-based method with a scoring function to guide the development of the metabolic network.

3.2.3 Inferring the Bayesian Network from information theory

Metabolic flux data are much smaller in scale than micro-array data. Nevertheless, applying Bayesian network analysis to flux data is still computationally prohibitive. Many algorithms exist that can infer the Bayesian network structure, only a few are computationally efficient enough to deal with large datasets. The information theory-based learning algorithm is one such method, which we've applied to flux data to infer the underlying metabolic regulatory network. This algorithm has been applied to real-world data with hundreds of variables and records [Cheng et al., 2002] and is described

briefly below. This constraint-based algorithm is performed in three separate phases: drafting, thickening, and thinning.

Phase 1 computes the mutual information contained in each pair of fluxes as a measure of closeness indicating the correlation between fluxes and creates a draft of the regulatory network based on this information. The mutual information $I(X_i,X_j)$ for each pair of metabolic fluxes (x_i,x_j) is computed as:

$$I(X_i, X_j) = \sum_{x_i, x_j} \Pr(x_i, x_j) \log \frac{\Pr(x_i, x_j)}{\Pr(x_i) P(x_j)} \tag{3.4}$$

Then each pair of metabolic fluxes with mutual information $I(X_i,X_j)$ greater than a threshold value, T, is sorted in a list L from high to low. An arc is drawn for the first two pairs of nodes in L. The pointer is then moved to the next pair of nodes. If no path exists between them, an arc is added. To illustrate this point, we provide a modified example from [Cheng et al., 2002]. Suppose we have five metabolic fluxes A, B, C, D, E, and the mutual information sorted as follows:

$$I(A,B)>I(B,C)>I(B,D)> T > (A,D)> I(A,C)>I(C,D)$$

then

$$L=\{[A,B],[B,C],[B,D]\}$$

and we obtain a draft shown in Figure. 3.3 *A*.

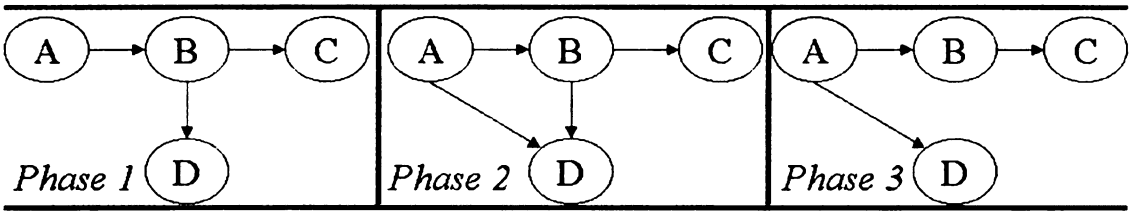


Figure 3.3 An illustration of the three phases of the Bayesian network learning process using mutual information based algorithm. Phase 1: the arcs are added to pairs of nodes whose mutual information are larger than the threshold value E. Phase 2: the arcs are added to pairs of nodes that are not conditionally independent using mutual information based CI test. Phase 3: the arcs are removed between pairs of nodes that are conditionally independent using mutual information based CI test.

In Phase 2, arcs are added when pairs of unconnected metabolic fluxes (i.e., nodes) are not independent as determined by CI test. The CI test is based on conditional mutual information as defined by Eq. 5 below. For $I(Xi,Xj|c)$ less than the specified threshold value T, (Xi,Xj) are said to be independent given c, where c is a set of metabolic fluxes.

$$I(X_i, X_j \mid c) = \sum_{x_i, x_j, c} \Pr(x_i, x_j, c) \log \frac{\Pr(x_i, x_j \mid c)}{\Pr(x_i \mid c) P(x_j \mid c)} \tag{3.5}$$

For example, if the metabolic fluxes (A, D) in Figure. 3.3*A* are not independent based on the CI test, then the arc between (A, D) is added, with the resulting network shown in Figure. 3.3*B*.

In Phase 3, each arc is examined by using the CI test and arcs are removed if the two metabolic fluxes linked by the arc are conditionally independent. For example, if (B, D) are found to be independent given (A) by the CI test, then the arc between (B, D) is removed with the resulting network shown in Figure. 3.3*C*.

3.2.4 Detecting latent variables and comparing alternative models based on a Bayesian scoring metric

Many biological systems are not well defined due to our incomplete understanding of the underlying mechanism involved. As a result, certain relevant factors may be overlooked in the experimental design and measurements. From a mathematical standpoint, these factors would be considered latent variables. To detect the presence of these latent variables, algorithms such as IC* [Pearl, 2000], and the faster FCI [Sprites et al., 1993] have been used.

Previously [Chan et al., 2003c], we found that the intracellular TG accumulation was not captured completely by the partial least-squares (PLS) method. A possible explanation may be due to insufficient details of the metabolic pathways pertinent to intracellular TG. In our current model development, the IC* algorithm was applied to a sub-network of pathways to detect possible latent variables relevant to intracellular TG. The alternative models containing these latent variables were then compared quantitatively by using a Bayesian scoring metric. A model's score, BS(S), is defined as the log of the posterior probability of the model S given the observed data D, p(S|D):

$$BS(S) = \log p(S|D) = \log p(S) + \log p(D|S) - \log p(D) \tag{3.6}$$

where p(D|S) is the likelihood of the observed data D given S. For a given set of data D, all possible structures S are assumed a prior to have equal probability, namely $p(S)$ is constant. $p(D)$ is a normalizing constant, which is the a prior probability of a given dataset that is applied to all structures S being evaluated. $p(D|S)$ is determined by Eq. 7, which was first derived by [Cooper and Herskovits,1992].

$$p(D \mid S) = \prod_{i=1}^{n} \prod_{j=1}^{qi} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{3.7}$$

Because $P(S)$ and $P(D)$ are the same for every network, thus they can be ignored in our application with only $p(D|S)$ remaining in the BS(S). Therefore, the score obtained is a relative posterior probability. A Matlab Bayesian Network Toolbox (BNT) was implemented to detect for latent variables and to evaluate the Bayesian metric score [Murphy, 2001].

3.2.5 Bayesian network predictor and target function prediction

The Bayesian network can be used to predict the value of a particular node as a function of the values of the remaining nodes (e.g., pathways) in the network. The target node being predicted will be denoted here as the class node. The posterior probability that the class node will take on a certain value given the values of the other nodes is determined based on conditional probability. The class node can take on any number of possible values, but the value that gives the highest posterior probability is selected as the "predicted" value for the node.

The predictive capabilities of a Bayesian network model can be improved by incorporating into the Bayesian network structure the current knowledge of the biological system being characterized. The prediction accuracy of the Bayesian network model in turn can be used to assess the correctness of the structure. The model is built as follow. Suppose node p_a in Figure. 3.3 is the class node, and b_1 and c_1 are the known values of nodes x_i and x_j, respectively, we can predict the class value with the highest posterior probability $Pr(pa|\ x_i,\ x_j)$ according to the Bayesian rule:

$$Pr(p_a=a_1|x_i=b_1,x_j=c_1) = Pr(x_i=b_1,x_j=c_1|p_a=a_1)*Pr(p_a=a_1)/Pr(x_i=b_1,x_j=c_1) \qquad (3.8)$$

where a_1 is one of any possible values for node p_a. The dataset of flux values is divided into a test and training set. Because nodes x_i, x_j are conditionally independent to each other given the value of node p_a, $Pr(x_i=b_1,x_j=c_1|pa=a_1)$ can be estimated from the training data as follows:

$$Pr(x_i=b_1,x_j=c_1|p_a=a_1) = Pr(x_i=b_1|p_a=a_1)*Pr(x_j=c_1|p_a=a_1) \quad (3.9)$$

Similarly, $Pr(p_a=a_1)$ can be estimated from the training data, and $Pr(x_i=b_1, x_i=c_1)$ is immaterial since it is the same for each class value of p_a. Therefore, a known Bayesian network structure inferred using information theory-based algorithm can be used to determine the conditional probabilities of a target variable. In this chapter, intracellular triglyceride is the class node predicted based on the values of the other fluxes in the network.

3.2.6 Data discretization

To learn the parameters and structure of the Bayesian network, we must first define a probability table containing all the nodes. This requires that the fluxes be discretized into nominal data. The simplest method to discretize the data is to divide the fluxes into equal-size intervals. This method, however, may be more sensitive to outliers, because extreme values may skew the range. A method that is less sensitive to this type of outlier is to divide the data into equal frequency intervals. In this approach, a continuous variable is divided into k bins where each bin contains equal number of frequency or adjacent values. These two methods are unsupervised and thus do not take into account the culture conditions associated with the metabolic flux data. Alternatively, a supervised algorithm

for discretization that incorporates the culture conditions uses class information entropy to select the boundaries of the bins. Class information entropy is defined as:

$$Ent = \sum_{i=1}^{N} \frac{n_i}{N} H_i \tag{3.10}$$

where n_i is the number of samples in interval i, N is the total number of samples, and H_i is the Shannon entropy in interval i. The Shannon entropy is defined in terms of the probability of observing an event, p_j:

$$H_i = \sum - p_j \log p_j \tag{3.11}$$

where p_j is the probability of culture condition j in interval i. The partition boundaries are determined by minimizing the class information entropy defined in Eq. 10. Each of the aforementioned methods was applied, and the equal frequency method was found to perform the best by comparing the resulting Bayesian networks with the known metabolic network. Therefore, we applied the equal frequency discretization method to our data.

3.2.7 Sensitivity analysis of Bayesian Networks

Bayesian network is a construction of connections between variables where the connections represent their degree of dependency based on conditional probabilities. The conditional probabilities will be denoted here as parameters of the Bayesian network. The importance of a connection to the network can be assessed through a sensitivity analysis. The sensitivity analysis examines the posterior probability of a response variable or network when a parameter is systematically perturbed. Some parameters are likely to

show considerable effects and thus are deemed important to the response variable or network.

Let B be a Bayesian network, X be a parameter, and Y be a response variable, e be the data, the sensitivity of X on Y given e is a partial derivative:

$$S(x \mid y,e) = \frac{\partial p(y \mid e)}{\partial x} \tag{3.12}$$

It was proven by [Castillo et al., 1997] that any posterior probability of a response variable is a fraction of two linear functions of X,

$$p(y \mid e)(x) = \frac{\alpha x + \beta}{\gamma x + 1} \tag{3.13}$$

then $S(x|y,e)$ can be expressed as

$$S(x \mid y,e) = \frac{\partial p(y \mid e)}{\partial x} = \frac{\alpha - \beta\gamma}{(\gamma x + 1)^2} \tag{3.14}$$

To determine the value of α,β,γ, we calculate $p(y|e)(x)$ for three different values of x, for example, if we set x to values of 0, 0.5, 1, then α,β,γ are determined by

$$\begin{aligned} \beta &= p^0 \\ \gamma &= \frac{\beta - p^{0.5}}{p^{0.5} - p^1} - 1 \\ \alpha &= p^1(\gamma + 1) - \beta \end{aligned} \tag{3.15}$$

where p^0, $p^{0.5}$, p^1 denote the corresponding posterior probabilities of $p(y|e)$ given x = 0, 0.5, 1. We applied a sensitivity analysis to our Bayesian network by using a Matlab code

developed by [Wang et al., 2002]. The sensitivity analysis enabled us to identify the important or most sensitive variables (nodes) in the network with regard to the targeted response variable (objective function), that is, intracellular TG or urea. Therefore, carefully controlling or modulating the most sensitive parameters will facilitate the optimization of the response variable or network structure.

3.3 Results

In this chapter, we illustrate the integration for the first time of several Bayesian-based techniques, such as Bayesian network structure inference [Cheng et al., 2002], latent variable detection algorithms [Pearl, 2000], hypothesis testing using Bayesian metric scoring [Hartemink et al., 2001], Bayesian network sensitivity analysis [Wang et al., 2002], and Bayesian network prediction [Friedman and Goldszmidt, 1996], into one united framework. Moreover, we illustrate the utility of this unified framework to biological data, and the first application of this approach to metabolic data for the purpose of inferring metabolic network structure of hepatocellular metabolism from the data. We applied this framework to identify direct and indirect pathways that influence an objective function, such as TG accumulation and the rate of urea synthesis. Previously, we developed a methodology that combined the metabolic flux distribution obtained by MFA with PLS to reveal the underlying structure and correlation within the data. PLS could capture and predict urea synthesis very well, but the intracellular TG levels not as well. Alternatively, the Bayesian-based framework could accurately predict both the level of urea synthesis and TG accumulation for each of the cellular conditions evaluated.

Algorithms to detect latent variables to uncover non-obvious pathways relevant to intracellular TG accumulation were used to help construct the Bayesian network model, thus providing flexibility in the model building process. This contributed to the improved predictive capabilities of the Bayesian network model over the PLS method. To uncover these latent variables, both the measured and estimated intracellular fluxes, the latter obtained from a MFA model, were combined into one data matrix X (36×76). The matrix X was then subjected to the equal frequency discretization to obtain a discretized data matrix D. We applied Bayesian network analysis to the data matrix D to infer sub-networks within the larger metabolic network and to identify possible latent variables not included in D. Based on the latent variables identified, we proposed several alternative metabolic sub-networks and evaluated their "accuracy" with a Bayesian scoring function and its ability to predict the objective function.

3.3.1 Reverse engineering the sub-networks

Using Bayesian network analysis, we reversed engineered parts of the metabolic network from the flux data, namely, the TCA (Figure. 3.4*A*) and urea (Figure. 3.4*C*) cycles. The TCA cycle is important for producing cellular energy (ATP) from the oxidation of fuels and generating intermediates for the gluconeogenic pathway. The analysis enabled us to infer the relationships shown in Figure. 3.4*B* and listed in Table 1. The direct links between the TCA pathways 10–11 and 12–13 were not learned by Bayesian network analysis but rather were identified as being coupled through oxidative phosphorylation (flux nos. 51 to 53). This recognition is rather appropriate because oxidative phosphorylation, whereby ATP is formed by electron transfer from NADH and $FADH_2$ to O_2, is coupled to the TCA cycle. It is noteworthy that, in our model, pathway no. 10–

11 combined two pathways, namely, citrate–isocitrate and isocitrate–α-ketoglutarate, into one, citrate–α-ketoglutarate; likewise two pathways were combined into pathway no. 12–13. If we add the missing links (10 → 11 and 12 → 13) to the inferred network, the Bayesian metric score indeed improved from –198 to –172.

Another important hepatic function is the removal of NH_4^+ derived from protein and amino acid metabolism via its conversion to urea by the liver through the urea cycle (Figure. 3.4*C*). Bayesian network analysis was able to "learn" most of the pathways in the urea cycle (Figure. 3.4*D*, Table 3.2). A comparison of the actual pathways and the inferred network is shown in Figure. 3.4. The link between pathways 16 and 15 was not learned by Bayesian network analysis. Adding the missing relation (15 → 16) improved the Bayesian metric score from –116 to –102. The learned Bayesian-based network shown in Figure. 3.4*D* was evaluated on how well it predicted urea production. We found the Bayesian model could predict urea production as well as the PLS and multi-block PLS models, with ~92% accuracy [Chan et al., 2003c; Hwang et al., 2004].

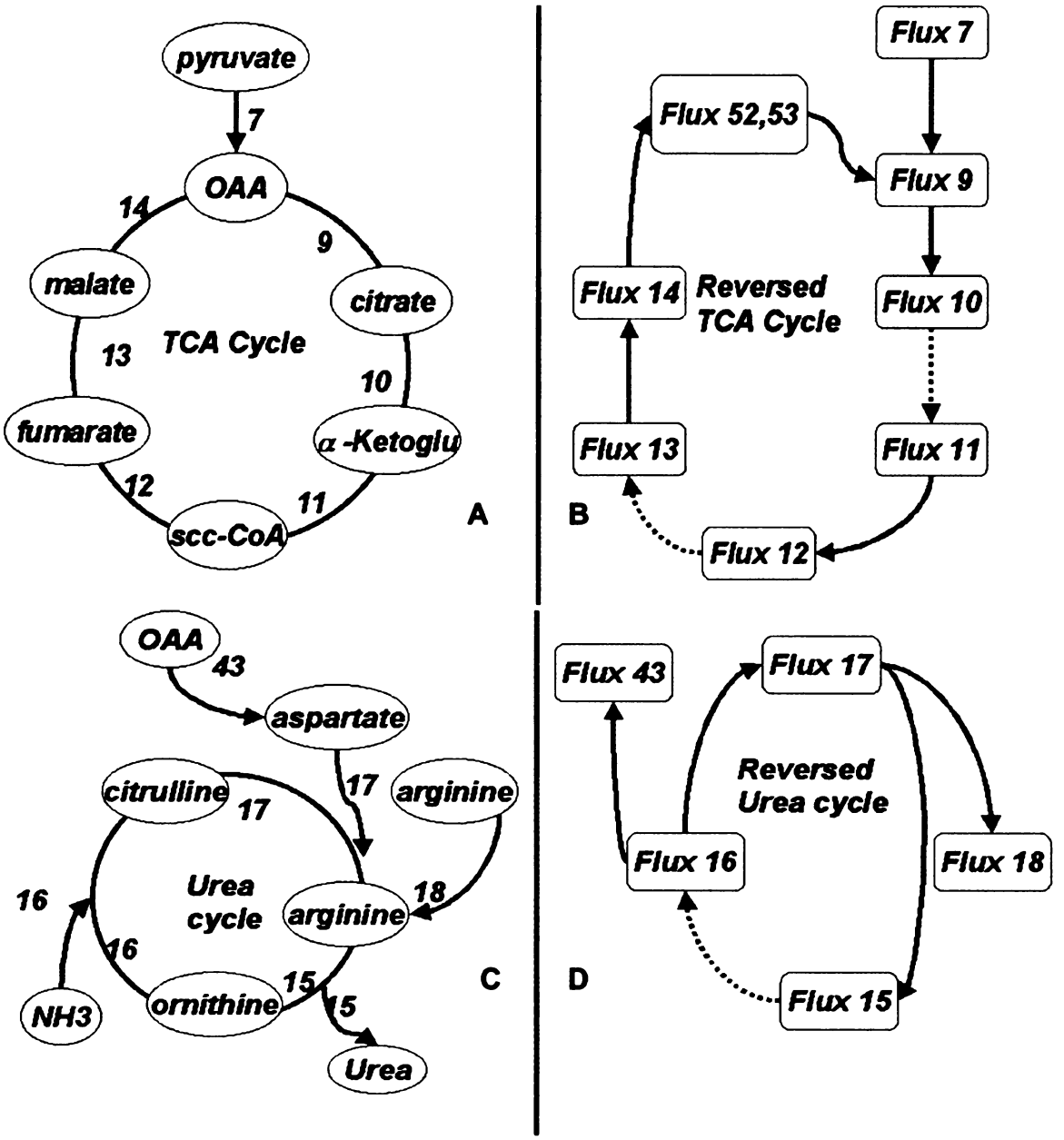


Figure 3.4. Reverse Engineered TCA and urea cycle learned by mutual information based algorithm. A) Actual metabolic network of TCA cycle, each node represents a metabolite and connections (arcs) between nodes represent the metabolic fluxes. B) TCA cycle inferred by Bayesian network analysis, each node represents a flux in Figure 3.4A, each arc between the nodes represent a causal relation between the fluxes, the solid connections were learned and the dashed connections were not learned by Bayesian network analysis. C) Actual metabolic network of urea cycle, each node represents a metabolite and connections between nodes represent the metabolic fluxes. D) Urea cycle inferred by Bayesian network analysis, each arc between the nodes represent a causal relation between the fluxes, the solid connections were learned and the dashed connections were not learned by Bayesian network analysis.

Table 3.2. Learned urea cycle relationships

Learned relation	Biological explanation
16 → 17	Synthesis of arginine (pathway no. 17) is controlled by the formation of citrulline carbamoylate from ornithine (pathway no. 16),
16 → 43	Aspartate aminotranferase (pathway no.43) is controlled by the formation of citrulline carbamoylate from ornithine (pathway no. 16),
17 → 15	Urea production (hydrolysis of arginine) (pathway no.15) is controlled by pathway no. 17
17 → 18	Uptake of arginine (pathway no.18) is controlled by pathway no. 17

3.3.2 Identifying relevant pathways linked to intracellular TG levels

The ability of Bayesian network analysis to infer the known metabolic sub-networks from the flux data lends credence to the latent relationships it uncovers. Despite the tremendous amount of information known about the hepatic metabolic network, our understanding is incomplete. Therefore, due to our imperfect knowledge of the metabolic network, it is plausible that important latent variables that are not directly associated but revealed by the model to be linked to currently unassociated pathways may indeed be consequential, as was noted previously with the uncovering of the link between flux nos. 51–53 and the TCA cycle.

Applying a constraint-based algorithm to infer a sub-network containing pathways linked to TG, we found that intracellular TG (flux no. 76) was influenced by β-hydroxybutyrate (BOH) dehydrogenase (flux no. 50), glutaminase (flux no. 38), and glutamate dehydrogenase I (flux no. 36), whereas extracellular TG (flux no. 70) was

affected by lactate dehydrogenase (LDH; flux no. 8), glyceraldehyde-3-P (G3P; flux no. 5), free fatty acid uptake (flux no. 72), glutamate dehydrogenase I (flux no. 36), and asparaginase (flux no. 45). Latent variable detection (IC*) algorithm was applied to the aforementioned pathways, along with several other pathways (flux nos. 2, 26, 54, 61, 64, 73, 74, 75), a total of 17 pathways, and the results suggest that latent variables may exist that influence intracellular TG and the glutamate related pathways (flux nos. 36 and 38). Those with direct connections learned through IC* are listed in Table 3.3. Although direct connections to flux no. 50 were not found with IC*, a direct connection between flux no. 50 and 76 was indicated by the constraint-based method; therefore, it was included in all subsequent analyses. From the currently known pathways related to TG metabolism and the inferred connections, we postulated several alternative networks, shown in Figure. 3.5. We tested the hypothesis that flux no. 50 may be a latent variable between flux no. 76 and flux no. 36 (Figure. 3.5*A*) and between flux no. 76 and flux no. 38 (Figure. 3.5*C*). Similarly, we tested the hypothesis that flux no. 36 was a latent variable between flux nos. 76 and 38 (Figure. 3.5*B*) and flux no. 38 was a latent variable between flux nos. 76 and 36 (not shown, but will be denoted as 5B alternate). Finally, in Figure. 3.5*D*, we tested the possibility that flux no. 50 was the latent variable influencing flux nos. 76 and 38, in combination with flux no. 38 as the latent variable for flux nos. 76 and 36. We evaluated the likelihood of these models with a Bayesian metric score and prediction accuracy of the inferred model.

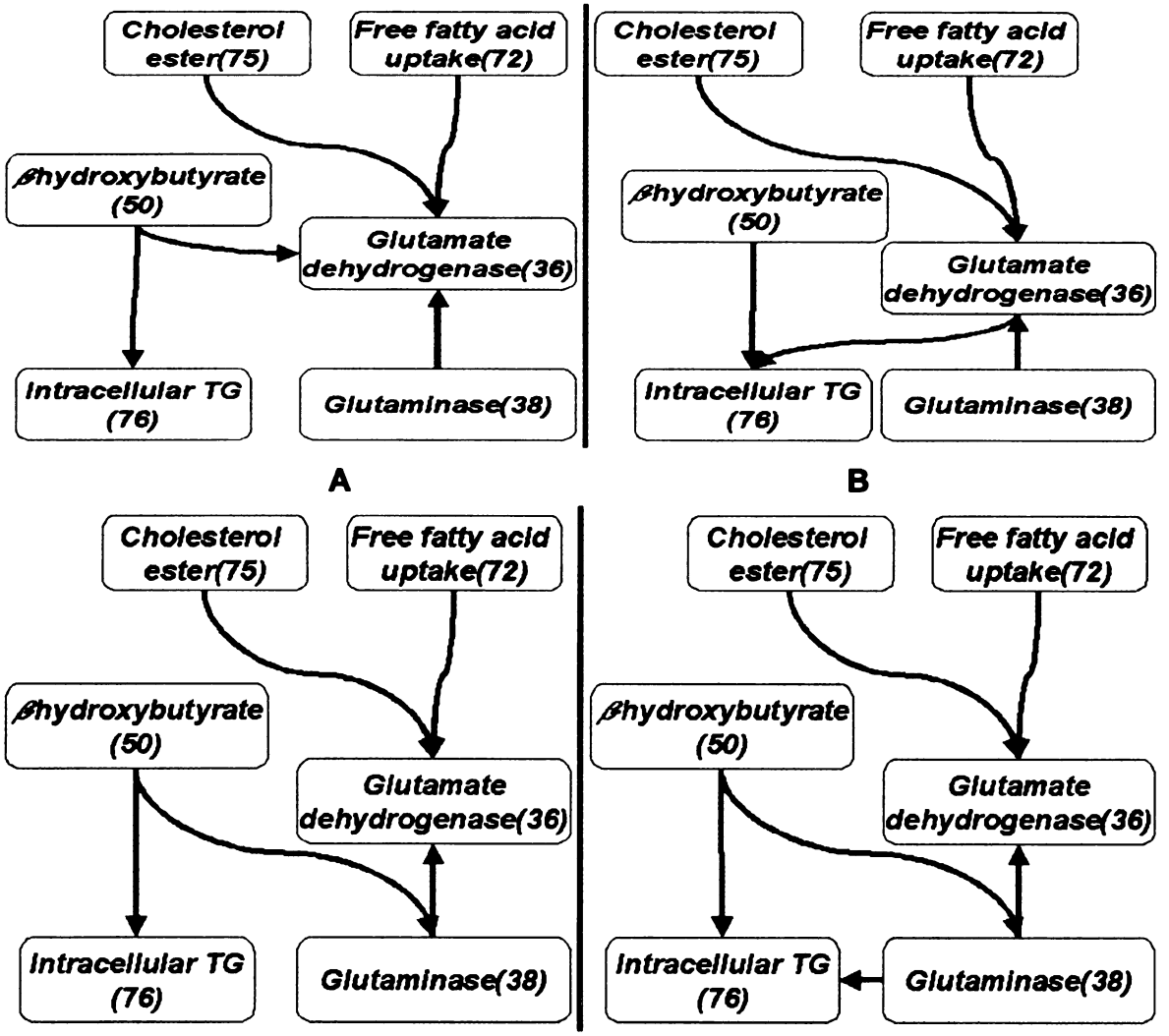


Figure 3.5. Postulated sub-networks for intracellular TG accumulation, learned with IC* algorithm. A) Postulated network that assumes β-hydroxybutyrate as the latent variable regulating intracellular TG accumulation and glutamate dehydrogenase pathways. The network's Bayesian metric score is -168. B) Postulated network that assumes glutamate dehydrogenase pathway as the latent variable between intracellular TG accumulation and glutaminase. The network's Bayesian metric score is -158. C) Postulated network that assumes β-hydroxybutyrate as latent variable regulating intracellular TG accumulation and glutaminase pathways. The network's Bayesian metric score is -155. D) Postulated network that assumes β-hydroxybutyrate as the latent variable regulating intracellular TG accumulation and glutaminase pathways while glutaminase influences intracellular TG accumulation and glutamate dehydrogenase pathways. The network's Bayesian metric score is -152.

Table 3.3. Relations recovered by Bayesian network analysis with IC* algorithm. A value of 0, 1, -1 indicates no direct link, a direct connection, and the existence of a possible latent variable, respectively.

Flux nos.	36	38	50	72	75	76
36	0	1	0	1	1	-1
38	1	0	0	0	0	-1
50	0	0	0	0	0	1
72	1	0	0	0	0	0
75	1	0	0	0	0	0
76	-1	-1	0	0	0	0

3.3.3 Accuracy and sensitivity of the postulated networks

Two methods were used to evaluate the performance of candidate networks, namely, the Bayesian metric score and the prediction accuracy. Linking the known pathways in the urea and TCA cycles that were not inferred by Bayesian network analysis improved the Bayesian metric score of the inferred network. This illustrates that our knowledge of the biological system can be used to help refine the inferred network and the Bayesian metric score can be used to guide the identification of relevant pathways within a network. The metric scores of the models in Figure. 3.5*A*, 3.5*B*, 3.5*B* alternate, 3.5*C*, and 3.5*D* were –168, –158, –153, –155, and –152, respectively. The scores indicate that the model in Figure. 3.5*B* is 22,000 (e^{10}=22,000) times more likely than the model in Figure. 3.5*A* in explaining the experimental data, and the model Figure. 3.5*B* alternate is ~150 times more likely than the model in Figure. 3.5*B*. Similarly, the model in Figure. 3.5*C* is ~440,000 times more likely than the model in Figure. 3.5*A*. Combining Figure. 3.5*C* and 3.5*B* alternate into Figure. 3.5*D* provides a network model that is 20 times more likely than the model in Figure. 3.5*C*, 3 times more likely than the model in Figure. 3.5*B* alternate, and 8.9 million times more likely than the model in Figure. 3.5*A*. The program

used to obtain these scores is available at http://www.ai.mit.edu/~murphyk/Software/bnsoft.html. In the second method, we divided our 36 samples into a training set and a test set consisting of 24 and 12 samples, respectively, to evaluate the hypothetical network structures shown in Figure. 3.5. We trained several Bayesian classifying models with our training dataset to predict intracellular TG. The ability of the postulated networks to predict intracellular TG for the test set are indicated in Table 3.4. The model shown in Figure. 3.5*A* classified the samples with a prediction accuracy of ~67%. However, the model in Figure. 3.5*B*, 3.5*B* alternate, 3.5*C*, and 3.5*D* predicted with an accuracy of 92, 92, 83, and 92, respectively. The Bayesian metric score and the prediction accuracy method both indicated that the model most likely to be correct given our data is the model illustrated in Figure. 3.5*D*. Therefore, it is possible that β-hydroxybutyrate (flux no. 50) may be a latent variable affecting flux nos. 38 and 76, whereas the glutaminase pathway (flux no. 38) may be a latent variable participant in flux nos. 36 and 76 (see Table 3.4). Because glutamine is the amino acid that is supplied to the cells in the highest quantities (4 mM vs. less than 0.8 mM for the other amino acids), a possible role played by the glutaminase (no. 38) is to up-regulate the TCA cycle with the necessary intermediates, such as glutamate and in turn α-ketoglutarate and oxaloacetate. An up-regulated TCA cycle increases the demand on acetyl-CoA, which can be obtained from the breakdown of fatty acids via β-oxidation, with some portion directed to ketone body production, such as β-hydroxybutyrate. Therefore, it is encouraging that the model identified both flux nos. 50 and 38 as relevant influences on intracellular TG accumulation. The program used to determine the

prediction accuracy is Belief Network Power Predictor, available at http://www.cs.ualberta.ca/~jcheng/bnpp.htm.

Table 3.4. Bayesian score and prediction accuracy of the Figure 3.5

Model	Prediction Accuracy	Bayesian Score
3.6 A	67%	-168
3.6 B	92%	-158
3.6 B alternate	92%	-153
3.6 C	83%	-155
3.6 D	92%	-152

A sensitivity analysis was applied to the intracellular TG sub-network shown in Figure. 3.5, with intracellular TG (flux no. 76) as the response variable. The results in Table 3.5 indicate that the intracellular TG level is most sensitive to β-hydroxybutyrate (flux no. 50) and glutaminase (flux no. 38) and least sensitive to glutamate dehydrogenase (flux no. 36) and cholesterol ester uptake (flux no. 75). Although the sensitivity coefficients are all small, the coefficients for β-hydroxybutyrate (flux no. 50) and glutaminase (flux no. 38) are generally an order-of-magnitude higher than the other pathways, indicating that they have more impact on TG storage. Thus, changing flux nos. 50 and 38 would more likely alter the level of intracellular TG accumulation as oppose to changing the other pathways.

Table 3.5 Sensitivity analysis of TG network shown in Figure 3.6

Pathway sensitivity	Cholesterol uptake(75)	FFA uptake (72)	Beta-hydroxy(50)	Glutaminase (38)	Glutamate dehydrogenase(36)
sensitivity5A	0.0001725	0.0014	0.0094	0.0014	0
5B	0.000345	0.0013	0.0115	0.0048	0.0058
5B alternative	0.0021	0.0023	0.0138	0.0134	0
5C	0.000469	0.000525	0.0093	0.0013	0
5D	0.0014	0.0016	0.0093	0.0134	0

3.4 Discussion

We applied the Bayesian-based framework to infer known network structures from metabolic data and evaluated the hypothetical networks using quantitative measures, such as Bayesian metric scores. The framework was applied to hepatocellular systems and successfully inferred the metabolic sub-networks, TCA and urea cycles, from metabolic data. We identified both direct and indirect pathways that influence our objective function, taken to be either intracellular TG accumulation or rate of urea synthesis. It was found in [Chan et al., 2003c] that urea synthesis can be predicted well with a PLS model, but the intracellular TG accumulation was modeled only reasonably well with PLS. Thus one of the objectives in the current study was to develop a model that could capture the causal mechanism or system structure and in turn more accurately predict intracellular TG accumulation.

Compared with PLS analysis [Chan et al., 2003c], Bayesian network analysis offers the advantage/possibility of characterizing the underlying causal structure within the data. PLS is an approach based on correlation analysis; thus it identifies factors that are highly correlated to the target variable or objective function. Thus, PLS cannot distinguish situations in which an identified factor (x_i) and the target variable (x_j) are highly correlated but are caused or regulated by a third variable, the common cause or parent variable (p_a). Under those conditions, to achieve an optimal response in the target or objective function requires modulating the common cause variable as opposed to the variable(s) that are simply highly correlated. For example, in our previous paper [Chan et al., 2003c], flux no. 43 (aspartate aminotransferase) was found to contribute most to optimizing or restoring urea synthesis. However in the sub-network inferred by the Bayesian network analysis, shown in Figure. 3.4*D*, flux nos. 43 and 17 (argininosuccinate synthetase and argininosuccinase) have a common cause, flux no.16 (carbamoyl-P-synthetase and ornithine transcabamylase). It suggested that flux no. 43 is relevant to urea synthesis but it is not the cause of urea synthesis. Thus, to systematically optimize urea production, the variable of common cause (flux no. 16) should be modified rather than the conditionally independent variable, flux no. 43. This suggests that ammonia, regardless of its source, generated in the liver mitochondria and combined with HCO_3^- to form carbamoyl phosphate, is the driving force for urea synthesis and is the source of the up-regulation in pathways 17 and 43, argininosuccinate synthetase, and aspartate aminotransferase, respectively. This is a reasonable finding by the model. The ability of the Bayesian framework to perform cause-and-effect analysis provides an advantage over PLS in optimizing the target variable. However, multivariate analysis, such as PLS, can

be used to pre-select a subset of variables that are highly correlated to the target variable, for example, intracellular TG, thereby decreasing the Bayesian network learning procedure. For example, we selected the variables with absolute regression coefficient value larger than 0.5 and applied Bayesian network analysis on this subset of variables. The results are shown in Figure. 3.6. Although this decreases the complexity of the learning process, it does not preclude the possibility that causal information may be loss by evaluating only the highly correlated variables.

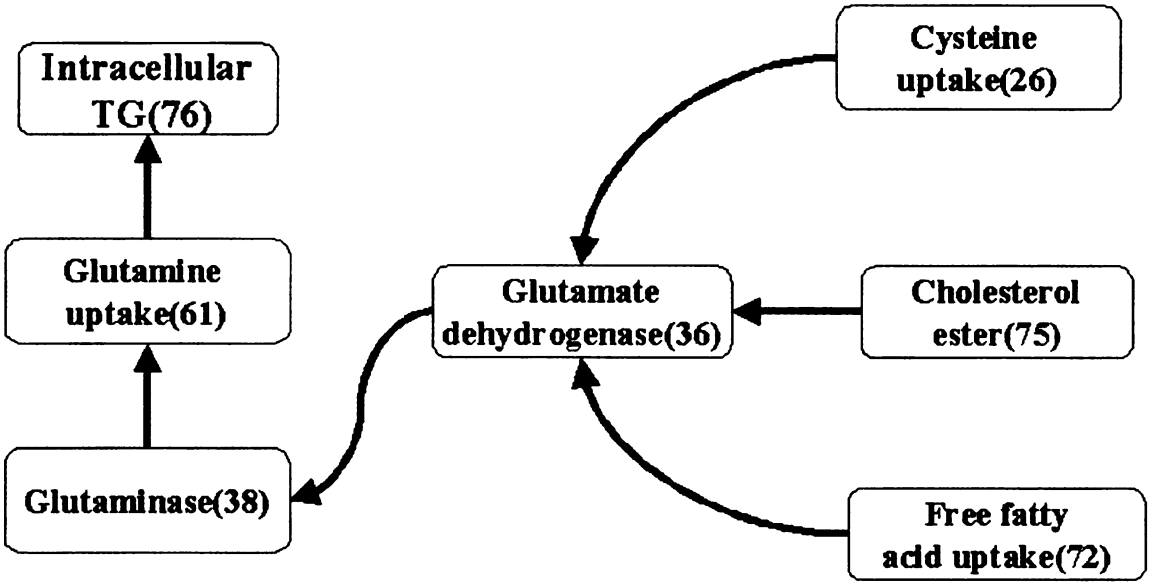


Figure 3.6. PLS flux selection. A subset of variables with absolute PLS regression coefficients larger than 0.5 were selected based on PLS analysis as being important to intracellular TG accumulation were subsequently subjected to the search and score based Bayesian network learning approach.

In the reversed Bayesian network, some relations were missing, for example, pathways 12→13 in Figure. 3.5*B* and 15→16 in Figure. 3.5*D*. Two possible explanations could account for this: 1) noise in the data; and 2) omitted measurements. We tested the first hypothesis by adding noise to our data and examined its effect on the resulting network

structure. Noise in the form of a Gaussian distribution was generated according to the following equation,

$$Noise = randn(n) * std(signal) * ratio \quad (3.16)$$

where *randn*(*n*) is a vector of random numbers of length *n* following a normal distribution, the *std*(*signal*) is the standard deviation of the data, and the ratio is any number between [0,1]. We added three different levels of noise to the data by assigning a value of 0.05, 0.10, or 0.20 to the ratio, representing 5, 10, and 20% noise, respectively. The resulting urea cycles are shown in Figure. 3.7. When 5% noise was added, the pathway 16→17 disappeared but instead pathway 15→16 was learned (Figure. 3.8*B*). This indicated the possibility that the missing pathway 15→16 may have been due to noise in the data. Adding 10% noise to the data resulted in pathway 17→18 being replaced by a new pathway 16→18 (Figure. 3.7*C*), and adding 20% noise eliminated both pathways 15→16 and 17→18 (Figure. 3.7*D*). The results suggested that Bayesian network analysis can tolerate a certain degree (5–10%) of noise within the data to provide a satisfactory level of performance, represented by the number of relations inferred. Although when the noise increased to a higher level, for example, 20%, the performance of Bayesian network approach decreased as reflected in the higher number of missed relations. We tested the second explanation by omitting one measurement from the data and examining the pathways inferred. To illustrate this idea, we removed the lactate measurement from the dataset (see Figure. 3.8*A*), which resulted in the omission of two additional pathways, pathways 8→7 and 7→9, which were no longer inferred by Bayesian network analysis (Figure. 3.8*B*). Therefore, both noise and missing

measurements could contribute to an incomplete or incorrect reversed Bayesian network. Thus care must be taken in collecting the data and, if possible, a comprehensive set of measurements needs to be acquired in order to gain the maximum benefit from Bayesian network analysis. It is appropriate to note that incomplete data would be less of an issue with the availability of high-throughput micro-array technology for collecting gene data.

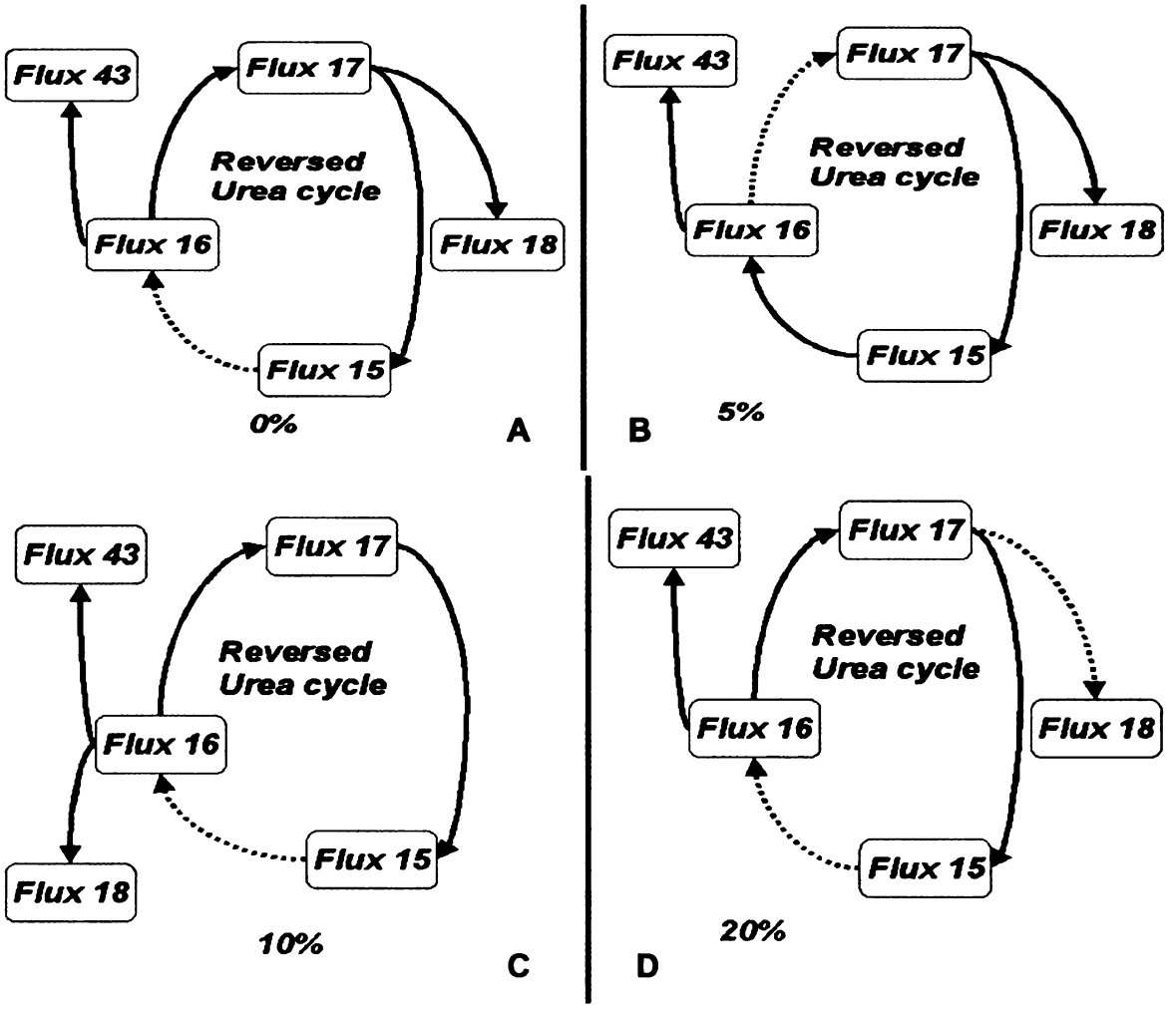


Figure 3.7. The effect of noise in the data on Bayesian network learning of the urea cycle. 5%, 10% and 20% noise were added to experimental data and Bayesian network was applied to each set of data. A) Urea cycle learned from the experimental data without the addition of noise. B) Adding 5% noise resulted in pathway 16→17 not inferred but 15→16 inferred instead, suggesting the possibility that 15→16 were not inferred in Figure 3.7A due to noise in the data. C) Adding 10% noise resulted in pathway 17→18 not inferred but 16→18 inferred instead, which indicates noise could result in incorrect relations being learned. D) Adding 20% noise resulted in pathway 17→18 not inferred as well as pathway 15→16, which indicates noise could cause decrease performance in the Bayesian network learning methods.

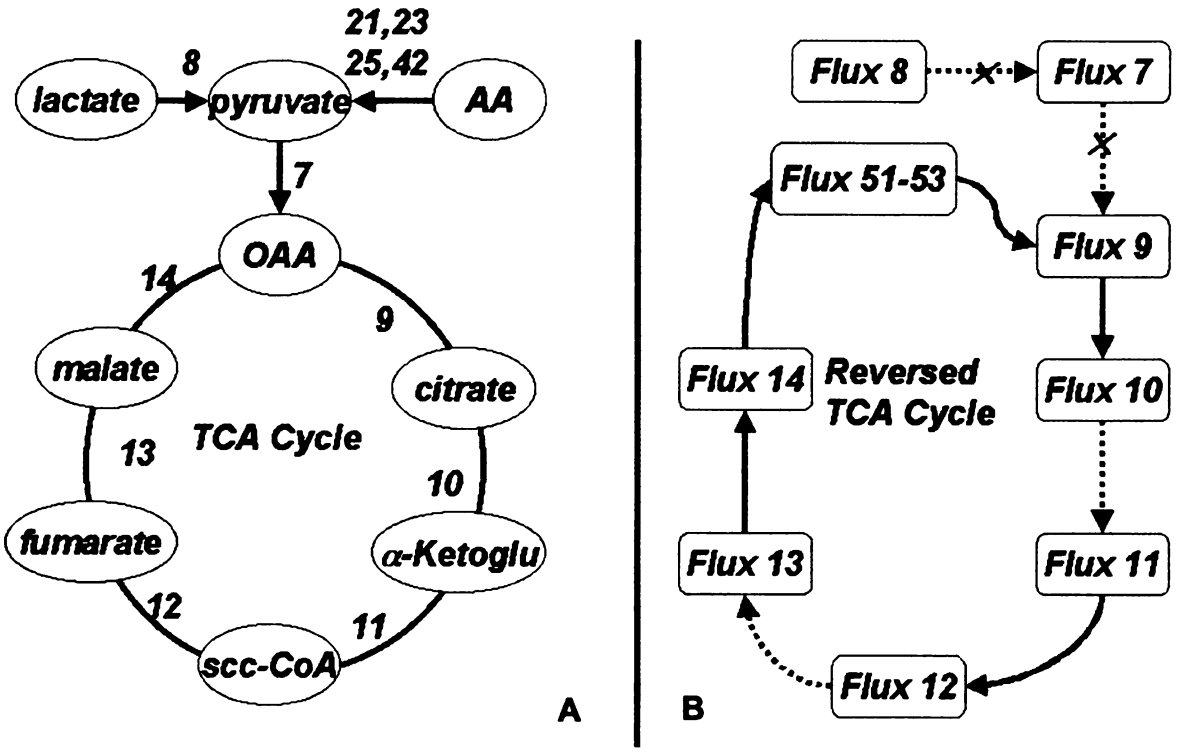


Figure 3.8. The effect of omitted measurements on Bayesian network learning. A) The current network. B) Not including the lactate measurement resulted in the causal relations 8→7 and 7→9 not being inferred, indicating that omitted measurement also could account for the causal relations missed in the networks inferred by the Bayesian network analysis.

According to the methodology developed in the proposed framework, several hypothetical networks describing intracellular TG accumulation were postulated and tested. Data-driven reverse engineering and model-driven hypothesis testing are two essential components of the proposed framework. Of the two, the data-driven portion of the analysis (i.e., the constraint based or mutual information Bayesian learning portion) used to infer the network structure is essentially automatic; thus, this part of the Bayesian network analysis is both fast and efficient. The analysis identifies hypothetical structures containing causal information that can be subsequently refined further by hypothesis testing. This is accomplished by incorporating existing biological knowledge in determining the pathways to include in the hypothetical structures. For instance, in our

intracellular TG example, we manually selected a subset of pathways to perform the IC* algorithm to determine which pathways influence TG accumulation. The method identified the existence of latent variables, which we determined through a process of elimination with the help of the metric score and the model's prediction accuracy to guide us to the "final" model illustrated in Figure. 3.5*D*. As an indirect confirmation of the model in Figure. 3.5*D*, we applied the Markov Chain Monte Carlo (MCMC) algorithm, which is a search-and-score method to the 17 pathways mentioned earlier. Because this method is computationally expensive, it is not feasible to apply this method to large networks. However, we obtained a solution for the sub-network after 500 iterations. This search-and-score method identified the connection between flux nos. 38 and 76 (see Table 3.6), which we postulated to be the latent variable between flux nos. 36 and 76 (Table 3.3). Thus it indirectly lends support to our hypothesis that flux no. 38 may be the latent variable linking flux no. 36 with flux no. 76. However, as illustrated by our example, the hypothesis testing has to be formulated and assessed manually, slowing the entire process. In our example, we manually postulated the four structures shown in Figure. 3.6 for intracellular TG. Therefore, a reverse engineering framework that combines these two steps becomes time-consuming. If only one could automatically incorporate the existing and, as much as possible, comprehensive knowledge of the biological system into the hypothesis formulation of the framework, the effectiveness of this methodology would be greatly enhanced.

Table 3.6. Relations recovered by Bayesian network analysis with MCMC algorithm, a search and score method. A value of 1 indicates a direct connection.

Flux nos.	36	38	50	72	75	76
36	0	1	0	1	1	0
38	1	0	0	0	0	1
50	0	0	0	0	0	1
72	1	0	0	0	0	0
75	1	0	0	0	0	0
76	0	1	0	0	0	0

The data in the current study were based upon a metabolic flux model, which provides a comprehensive overview of the intracellular metabolic state. MFA invokes certain assumptions in the development of the metabolic flux model, and these assumptions influence the results we obtain. Ideally, it would be advantageous to increase the number of measurements and reduce the number of assumptions. This may eventually come to pass as high-throughput technologies facilitate these measurements. Gene chip data have the advantage of more independent measurements. Unfortunately, transcriptional, translational, and post-translational effects currently limit our ability to infer cellular function from gene expression data. The measurements used in this study are at the functional level, with the disadvantage that MFA is needed to augment the data. That said, the Bayesian network analysis is naïve with respect to the assumptions used in MFA. The Bayesian analysis uncovers the dependencies resulting from these assumptions without bias. The ability of the Bayesian-based framework to infer the TCA

and urea cycles, which are well-established cellular networks, provides confidence in this approach to uncover relationships from this, as well as, other types of data.

In addition, the Bayesian methodology allowed us to infer the coupling of the oxidative phosphorylation pathways (flux no. 51–53) to the TCA cycle (flux nos. 9–14). This finding was encouraging, as oxygen uptake is an independent and direct measurement and is not linked directly to the TCA cycle in the flux model. Therefore, the ability of the Bayesian framework to infer this coupling provides a degree of confidence in the ability of this framework to infer, as well, implicit causal relationships from the data.

MFA is performed under the assumption of pseudo-steady-state, which limits the evaluation of the dynamic behavior of the system. In the future, we plan to monitor the temporal metabolic profiles in order to evaluate the dynamic properties of the model. Sensitivity analysis will then be used to identify the most sensitive parameters and hence the most important variables in the network that affect the stability of a metabolic state. Furthermore, this analysis can be used to evaluate how to manage the transition from one steady-state to another and thus guide us in designing experiments to collect the requisite data for these transition studies. Because the costs for collecting experimental data are relatively high in both time and monetary expense, this type of analysis can provide a means to screen and thus optimize the number of experiments necessary to obtain the relevant information.

In conclusion, we have shown that Bayesian network analysis could enable inference of network structures from metabolic data. From the resulting network, we can

identify the relevant variables that most affect a target function(s), thereby providing a framework to optimize the target function(s) and insight into the underlying mechanisms that govern the metabolic state. Finally, the ability of this methodology to infer well-known metabolic structures from experimental data provides confidence in the ability of this methodology to infer other networks, such as, genetic regulatory networks from gene data.

CHAPTER 4 IDENTIFYING PHENOTYPE RELEVANT PAHTWAYS

4.1 Introduction

In chapter 2, GA/PLS was applied to identify important genes relevant to a metabolic function. In chapter 3, Bayesian Network analysis was applied to infer metabolic network from metabolites profiles. In this chapter, approaches, i.e. GA/PLS and Bayesian network analysis are integrated into a unified framework, which we denote as the Three Stage Integrative Pathway Search (*TIPS*©), to identify the genes and pathways that regulate a particular phenotype.

The regulation of cellular function is achieved through the contribution and interactions of genetic, signaling, and metabolic pathways. Consequently, cellular processes may be singularly regulated, i.e., at the gene or transcription level, or controlled by a network of interactions of genes, proteins and metabolites. Therefore, understanding the biological function of the myriad of genes and how these genes and gene products interact and regulate each other to yield a functional cell would help to identify more appropriate pathways that should be targeted or studied for a given disease. An effective drug should be designed to regulate the targeted pathways, while leaving other pathways unaltered. Thus, it is important to identify the pathways in cells perturbed by external stimuli including the drug.

With the advent of high throughput technologies, systems of genes, proteins and metabolites for a given cellular state may be readily obtained, requiring novel analytical frameworks to aid in unraveling the regulatory and functional relationships. Approaches

such as log linear regression [di Bernardo et al. 2005] and mutual information based ARACNe [Basso et al 2005] have been applied to yeast and human B cells, respectively, to reverse engineering genome wide gene regulatory networks. Unfortunately, these approaches require large amount of data. Thus the challenge remains as to how to reconstruct active networks from a small number of observations. Therefore, we developed an approach that identifies the active pathways without requiring interaction measurements or libraries of genetic mutants, and with a limited amount of data, namely, by integrating gene expression and phenotypic profiles. To identify the active pathways, we first developed approaches to identify phenotype relevant (active) genes with GA/PLS [Li and Chan 2004] and constrained independent component analysis (CICA) [Lin et al. 2002, Liebermeister 2002]. BN analysis was then used to reconstruct the active sub-network from the smaller group of genes to reveal which pathways are induced by the external stimuli or environment. BN can detect indirect influences and unmeasured events and is ***not susceptible*** to the existence of unobserved variables. It has been applied to infer gene regulatory network of yeast cell cycle from gene expression data, metabolic subnetworks from metabolic data and protein signaling from protein activity data [Li and Chan 2004a; Friedman 2004; Sachs et al. 2005]. BN is computationally inefficient when applied to large network e.g. genome wide network. However, in our framework BN was applied to a reduced subset of relevant genes, which circumvented this limit. In addition, the reconstructed model can predict the effect of perturbing a gene (or several genes) on the other genes in the network, and thus provide insight into how the genes interact within the network. Mathematical frameworks that can reconstruct the active pathways

will permit better understanding of how the environment interacts with the genes to regulate cellular functions and phenotypes.

As proof-of-concept, the framework was applied to identify pathways that regulate cytotoxicity in human liver cells. Elevated levels of free fatty acids (FFAs) and TNFα have been shown to be involved in the pathogenesis of liver disorders, such as fatty liver disease and steatohepatitis. Saturated FFA, palmitate, was found to induce cytotoxicity and TNFα exacerbated this effect [Srivastava 2006]. Quantification of the genetic responses of hepatocytes to physiologically elevated levels of FFAs and TNFα may provide insight into the physiological actions of these factors and identify the pathways involved in conferring cytotoxicity.

We developed a *TIPS*© framework whereby we assume, as a first approximation, a log linear relationship between gene expression and cytotoxicity. The genes selected by GA/PLS were corroborated with published results to identify known interactions and analyzed by gene ontology to identify known functional associations. Thus, GA/PLS was used to determine the relevance of the genes to LDH release. In order to extract an independent pathway related to a process, such as LDH release, from the gene expression profile, we propose a constrained ICA (CICA) approach. With the relevance as determined by GA/PLS and the LDH profile as constraints, CICA was applied to extract an independent component from the gene expression data. CICA assumes that the expression profile of thousands of genes can be represented by a reduced number of mutually independent processes. Biological meaningful gene groups have been successfully identified by ICA [Liebermeister 2002; Martoglio et al. 2002]. CICA selects a subset of genes whose profiles are statistically independent to the other components and

corresponds closest to the profile of LDH release. This is achieved by minimizing the mutual information between the independent components and maximizing the correlation to the constraints. Finally, Bayesian network (BN) analysis was applied to reconstruct the pathways from the expression profile of the selected genes. The reconstructed network was perturbed to identify i) which genes, when perturbed, have the greatest impact on altering the phenotype (high LDH release) in the palmitate cultures, and ii) how does perturbing one or several genes (nodes) affect the other genes in the network, as well as the phenotype. The simulated perturbations of the reconstructed model were evaluated experimentally to assess the validity of the predictions. The reconstructed network of pathways provided potential explanation(s) on how TNFα and palmitate induce LDH release. The model identified i) the involvement of stearoyl-CoA desaturase (SCD) in the palmitate-induced cytotoxicity, ii) the activation of nuclear factor kappa B (NF-κB) by TNF-α is mediated by protein kinase C (PKC)-δ, , and iii) the role of Bcl-2 and PKR in palmitate induced cytotoxicity.

4.2 Materials and Methods

4.2.1 Three Stage Integrative Pathway Search Framework (*TIPS*©)

We applied a mathematical framework that first integrates genetic algorithm (GA) and partial least squares (PLS) analyses to identify the genes relevant to LDH release, but these genes may be involved in many independent pathways. Therefore, the framework then applies CICA to identify an independent pathway involved in LDH release. Finally the connections between these identified genes are reconstructed using BN analysis to infer how the genes interact with each other in the independent pathways. The

reconstructed network illustrates how the genes interact under the given environmental conditions to regulate LDH release. The framework is shown schematically in Figure 4.1.

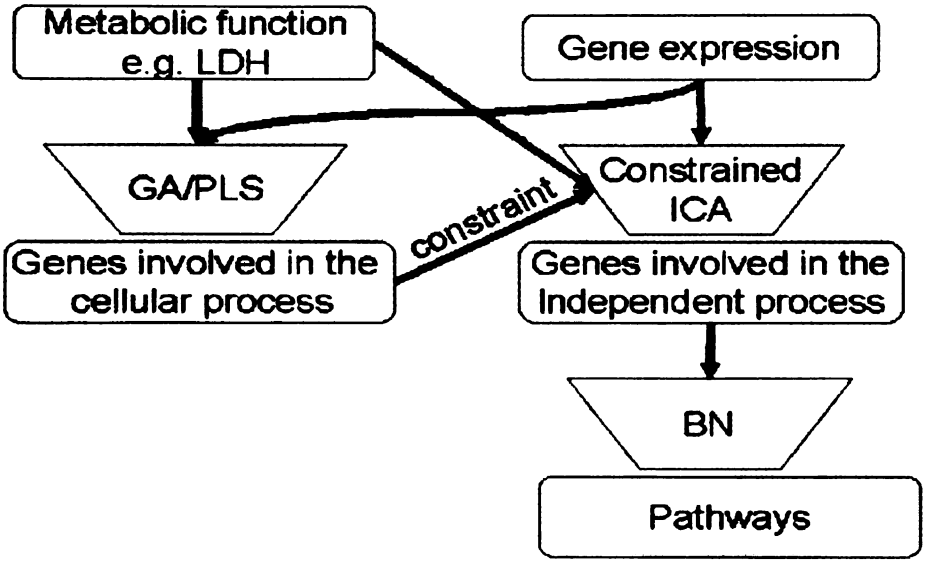


Figure 4.1: Scheme of TIPS©. GA/PLS, ICA and BN are integrated to infer the pathways thtat regulate a cellular function.

The framework being developed involved two central tasks. The first task attempts to identify the factors involved in producing a phenotype and the second task attempts to uncover how these factors associated with each other in a network of pathways. The three stage framework illustrated in Figure 4.1 was developed to integrate gene expression and metabolic profiles to identify the pathways involved in producing a particular phenotype. In the *TIPS*© framework, GA/PLS and CICA were applied to achieve the first task and BN analysis was applied to achieve the second task.

ICA was developed originally for blind source separation (BSS). Original sources S were mixed by matrix M resulting in observation Y. The goal of BSS is then to estimate the original source with the de-mix matrix W resulting in Z. Thus the ICA model can be expressed as

$$Z = WY \tag{4.1}$$

where W is the de-mix matrix and Y is the observations, Z is the estimated source signal. Z and W are simultaneously estimated by minimizing the statistical dependence (e.g. mutual information) between the columns in Z. The problem of ICA is then reduced to:

Minimize (Z) (4.2)

s.t. Z = WY.

In some cases, *a priori* information about the signal source and mix matrix are available. This information can be used to constrain W and Z, and thereby incorporate the *a priori* knowledge. We denote the signal measurements Y to be the gene expression data, S to be the independent components (pathway), and A to be the mixing matrix. The ICA model is then expressed as

$$Y = A S \quad (4.3)$$

The gene expression Y is supplied to the ICA model and the mixing matrix A can be uniquely estimated by assuming that the components in S are statistically independent to each other. *Y(i,t)* represents the expression of gene *i* in experiment t, *A(i,j)* represents weight of gene *i* in independent pathway *j*, *S(j,t)* represents the profile of independent pathway *j* in experiment *t*. Since the objective here is to identify LDH release related independent pathway, A was further constrained with the frequency learned by GA/PLS and S was further constrained with LDH release profile. Let F(i) be the frequency of gene *i* with respect to LDH release and *a(i)* be the weight of gene *i* in the pathway related to LDH release. Thus, *a* is constrained to have a correlation with F by equation (4.4), where $\rho 1$ is a threshold value.

$$Corr1 = a^T * diag(F) * a / (a^T * a) > \rho 1 \quad (4.4)$$

Let s(t) be the profile of LDH related pathway in experiment t and L(t) be the profile of LDH release in experiment *t*. Similarly, s is constrained to have a correlation with L by equation (4.5), where ρ2 is a threshold value.

$$Corr2 = s^T * L * L^T * s/(s * s^T) > \rho 2 \quad (4.5)$$

For further details please refer [Lin et al. 2002].

The advantage of *TIPS*© is that it identifies a subset of genes involved in the active pathways in response to the environmental stimuli and then reconstructs the pathways from this subset of genes using BN analysis. It reduces the number of samples required for pathway reconstruction and provides pathway information specific to a cellular function.

4.2.2 Bayesian Network Inference

Bayesian network inference was used to predict the probabilities of a phenotype, e.g. LDH, or a gene within the network taking on a certain value upon changing a target gene. The posterior probability that the class node will take on a certain value given the values of the other nodes is determined based upon conditional probability. Suppose node A in Figure 4.2 is the target node and b1 and c1 are the known values of evidence nodes B and C, respectively, we can predict the posterior probability Pr(A| xi, xj) according to the Bayesian rule:

$$Pr(A=a_1|B=b_1, C=c_1)=Pr(B=b_1,C=c_1|A=a_1)*Pr(A=a_1)/Pr(B=b_1,C=c_1) \quad (4.6)$$

However, applying this exact inference to a large network is computationally expensive [Cooper 1990]. Therefore, we applied approximate inference algorithm such as logic sampling [Henrion1988] to infer the posterior probabilities. Briefly, logic sampling generates a case by randomly assigning values to each node weighted by the probability

of that value occurring. To estimate the posterior probability Pr(X|E) where X is the target node and E is the evidence node, we compute the ratio of the number of cases where both E and X are true to the number of cases where just E is true, i.e. Pr(X=x|E=e) = Pr(X=x, E=e)/Pr(E=e).

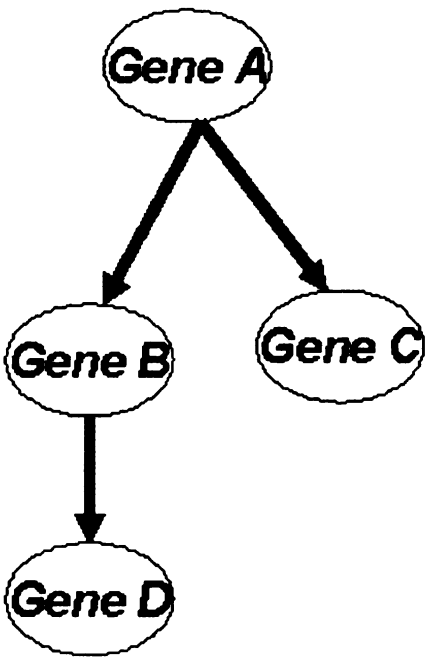


Figure 4.2. Example of a Bayesian network. Gene A is the parent node of Genes B and C, Gene B is the parent node of gene D. Genes B and gene C are conditionally independent given Gene A.

4.3 Results and discussion

4.3.1. Identifying genes relevant to cytotoxicity or LDH release using GA/PLS

We found that exposing human hepatoblastoma cells (HepG2/C3A) to three different types of FFAs (palmitate, oleate, and linoleate at 0.7mM) in the presence and absence of TNFα (0, 20, 100 ng/ml), only palmitate was cytotoxic to the cells and resulted in significantly higher LDH release. TNFα alone was not toxic to the cells. The effect of TNFα on cytotoxicity was observed only in the palmitate-treated cells. To obtain a global

view of the cellular processes, we capitalized upon high-throughput methods to quantify gene expression profiles of the HepG2 cells.

We applied GA/PLS to the gene expression and LDH release data. The GA/PLS algorithm determined the relevancy of each gene to LDH release by counting the frequency with which each gene was selected by GA into a subset of genes used to predict LDH release. The genes identified by GA/PLS to be relevant to palmitate-induced cytotoxicity, as measured by LDH release, included *oxidative stress related genes* (e.g., glutathione peroxidaase (AA664180)), *apoptosis related genes* (e.g., caspase 8 (AA44868), Bcl-2 (AA446839)), *TNF- related genes* (e.g., TNF super-family (AA77863)), *mitochondria membrane permeability related genes* (e.g., translocase of outer membrane (TOM)-34 (AA457118)), *ceramide related genes* (e.g., sphingomyelinase (AA676836), sphingosine kinase (AA630354)), *translational related genes* (e.g., eIF2β (AA027240)), and *signal cascade related genes* (e.g., protein phosphatase 2A (AA490473) and PKCδ (H11054)). The identification by GA/PLS of oxidative stress related genes suggests the involvement of reactive oxygen species (ROS) in the palmitate-induced cytotoxicity. Likewise the model identified ceramide related genes, e.g., sphingomyelinase and sphingosine kinase, are involved. In addition, mitochondria potential and signal transduction are involved in the palmitate-induced cytotoxicity.

4.3.2. Identifying genes involved in an independent pathway related to cytotoxicity using CICA

GA/PLS determined the frequency with which each gene was selected to predict LDH release. The frequency and the profile of LDH release were applied as constraints in

CICA to extract an independent component from the gene expression profile. The independent component in this case identified a subset of genes whose profiles corresponded to the profile of LDH release. CICA determined the weights for each gene by minimizing the mutual information between the independent components and maximizing the correlation to the constraints. Genes that have weights significantly different from zero with a 95% confidence using the Z-test were subjected to BN analysis for pathway reconstruction.

4.3.3. Reconstruct pathways related to cytotoxicity using BN

BN reconstructed how the genes, identified by CICA, are connected in a network and involved in LDH release or cytotoxicity. The resulting network is shown in Figure 4.3. The model revealed possible mechanisms involved in palmitate-induced cytotoxicity. To evaluate these potential mechanisms, we performed perturbation analyses and experimental validation. In the perturbation studies, all connections relevant to the perturbed genes were included.

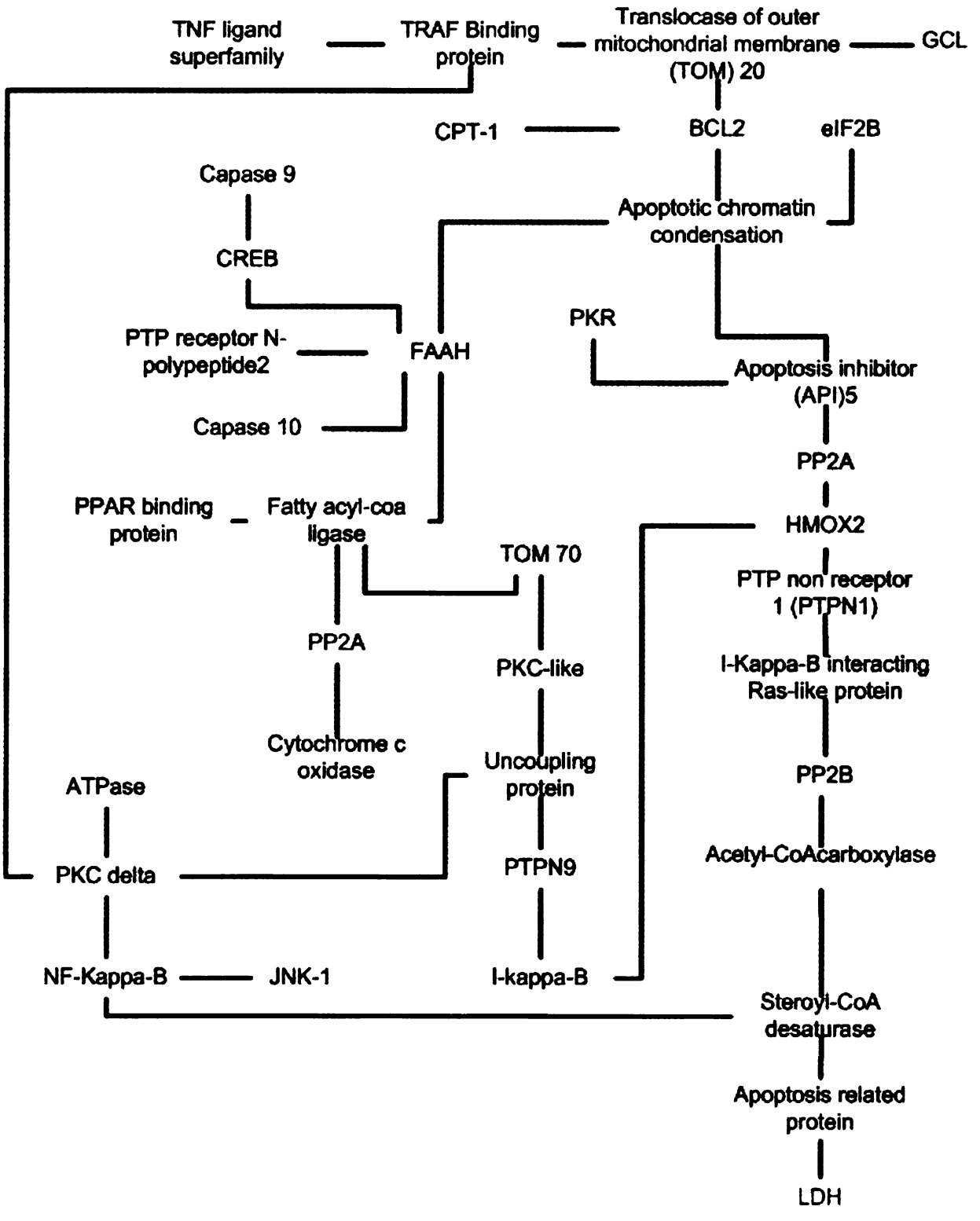


Figure 4.3 A representative sub-network related to cytotoxicity. GA/PLS and ICA select the relevant genes, and BN analysis reconstructs the network using the subset of selected genes. The network provides an overview of the factors and pathways involved in regulating cytotoxicity.

4.3.3.1 Perturbation of the reconstructed network

Based upon the reconstructed network illustrated in Figure 4.3, BN inference was applied to predict the effect of perturbing a single gene/node on i) the other genes within the network and ii) reducing the level of LDH release, in the palmitate and control cultures. The reconstructed network allowed us to artificially perturb the network and identify which nodes were the most sensitive and thus had the most impact on modulating LDH release. An assumption in our model is that perturbations at the gene level will translate correspondingly (qualitatively, but not necessarily quantitatively) to the protein level. Therefore, we altered the activity of the nodes by either down-regulating (up-regulating) the genes or inhibiting (activating) the protein activity, to study how the nodes interact with each other in the network. Genetic perturbations of SCD, PKC-δ , Bcl-2, NF-κβ and double-stranded-RNA-dependent protein kinase (PKR) were simulated with the BN inference to study their effects on the probability of LDH taking on a phenotype either of high or low release (see Table 4.1).

Table 4.1 Simulating genetic perturbation and its effects on LDH release

		Probability of LDH Release			
		Control		Palmitate	
Gene	**Perturbation**	**Low**	**High**	**Low**	**High**
	no perturbation	0.94	0.06	0.02	0.98
SCD	Upregulation	0.94	0.06	0.55	0.45↓
	Downregulation	0.9	0.1	0.02	0.98
PKC-δ	Downregulation	0.9	0.1	0.02	0.98
PKR	Downregulation	0.85	0.15↑	0.08	0.92↓
NF-κβ	Upregulation	0.94	0.06	0.15	0.85↓

Note that the probability of LDH release taking on a high or low level in the control and palmitate cultures. Using a BN inference on the structure shown in Figure 4.4 – simulations of up/down regulation of SCD, down-regulation of PKR, and down-regulation of PKC-δ and upregulation of NF-κβ as compared to the original condition (unperturbed state). "↓" indicates decreased probability and "↑" indicates increased probability.

4.3.3.2 Role of stearoyl-CoA desaturase (SCD) in palmitate induced cytotoxicity

A connection between SCD and acetyl-CoA carboxylase (ACC) was found in the reconstructed network. SCD is the rate-limiting enzyme to produce monounsaturated fatty acids. Its deficiency has been found to increase fatty acid oxidation by activating AMP-activated protein kinase (AMPK) in the liver. AMPK phosphorylates ACC at Ser-79 which inhibits ACC activity and decreases malonyl-CoA concentration [Dobrzyn et al. 2004]. Malonyl-CoA inhibits carnitine palmitoyl-CoA transferase (CPT-1) [Kerner and Hoppel, 2000]. Thus a decreased level in malonyl-CoA activates CPT-1 activity and increases fatty acid beta-oxidation in the mitochondrion [Dobrzyn et al. 2004]. In the palmitate cultures, the protein (Figure 4.4) expression levels of SCD were reduced compared to the control. This may explain in part the preference of hepatocytes to oxidize as oppose to synthesize TG from palmitate, as is the preference with the unsaturated fatty acids. Indeed, co-supplementing palmitate with oleate enhanced the accumulation of TG [Li 2006] and reduced the cytotoxicity of palmitate (Figure 4.4c).

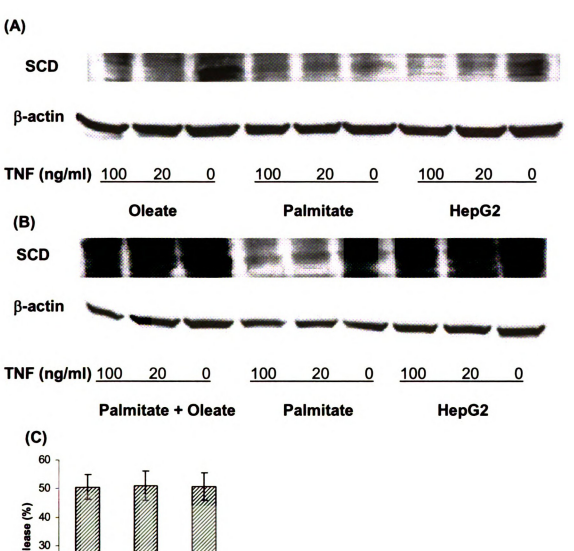


Figure 4.4 Effect of TNFα and palmitate on stearoyl-CoA desaturase (SCD) measured by western blotting. (A) SCD was downregulated in the palmitate (0.7 mM) cultures as compared to the oleate (0.7mM) and control cultures. (B) Co-supplementation of oleate (0.3 mM) with palmitate (0.4mM) prevented the downregulation of SCD. TNFα decreased the expression of SCD in the control, oleate and co-supplementation (oleate + palmitate) cultures. (C) Co-supplementing palmitate (0.4 mM) with oleate (0.3mM) decreased LDH release significantly, $P<0.01$ (t-test). P: treated with 0.7mM palmitate for 48 hours, PO: treated with 0.4 mM palmitate plus 0.3 mM oleate for 48 hours. Number following P or PO represents the concentration of TNFα in ng/ml, e.g. P20 is 0.7 mM palmitate with 20 ng/ml TNFα. Data expressed as average +/- SD from three independent experiments.

To evaluate the role of SCD on LDH release, we simulated up- or down-regulation of SCD in the reconstructed network by setting the SCD gene expression level at either a high or low level, respectively. Up-regulating SCD in the palmitate cultures reduced the probability from ~98% to ~45% that the LDH release would remain high (Table 4.1). While down-regulating SCD in the control cultures suggests that the LDH release would likely remain low (i.e., down-regulating SCD will have no effect on LDH release). The simulation results agreed well with the literature, e.g., over-expressing SCD has been shown to prevent palmitate-induced cytotoxicity [Listenberger et al. 2001]. The induction of SCD1 activity is transcriptionally activated [Ntambi 1995]. In further support of the protective of role of SCD, we supplemented the cultures with 50 μM of clofibrate or ciprofibrate, which are known to increase the activity of SCD through a PPAR independent pathway [Rodriguez et al. 2001; Miller and Ntambi 1996]. The fibrate supplementation significantly decreased LDH release in the palmitate cultures (Figure 4.5). In the control cultures TNFα decreased the level of SCD protein (Figure 4.4 A), but did not affect the LDH release [Srivastava 2006], which agreed with the simulation results shown in Table 6.1.

To examine the connection between TNFα and SCD shown in Figure 4.3, the protein expression level of SCD in control (HepG2 medium), palmitate, oleate, and palmitate + oleate cultures were measured as a function of the level of TNFα by western blotting. The presence of TNFα down-regulated the protein expression level of SCD under all the conditions except palmitate (Figure 4.4A, B). Similarly, palmitate supplementation decreased the protein expression level of SCD as compared to the control and oleate cultures. Co-supplementing the palmitate cultures with oleate restored

the SCD protein expression level (Figure 4.4B) and correspondingly reduced LDH release (Figure 4.4 C). Administrating fibrates increases the liver phospholipid proportion of monounsaturated fatty acids [Miller and Ntambi 1996], which appears to be similar in effect to the addition of oleate to the palmitate cultures.

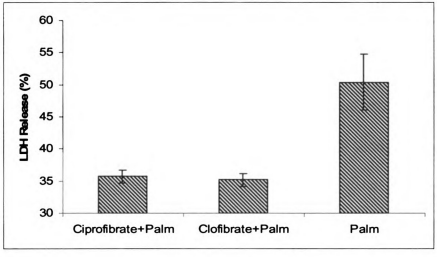


Figure 4.5. Effect of SCD activator, clofibrate and ciprofibrate, on LDH release in the palmitate cultures. Clofibrate and ciprofibrate are known to increase the activity of SCD. Palm: treated with 0.7 mM palmitate for 48 hours, Palm+clofibrate: treated with 50 μM clofibrate and 0.7 mM palmitate, Palm+ciprofibrate: treated with 50 μM ciprofibrate and 0.7mM palmitate. 6 hours pretreatment followed by 48 hours co-supplementation of 50 μM of clofibrate or ciprofibrate significantly decrease LDH release in the palmitate culture, $P<0.01$ (t-test). Data expressed as average +/- SD from three independent experiments.

Oxidative stress appears to be involved in down-regulating the protein level of SCD in the palmitate cultures. Supplementing with 10 mM N-acetyl cysteine (NAC) prevented the down-regulation of SCD (Figure 4.6) and reduced the release of LDH. The involvement of SCD in palmitate-induced cytotoxicity is further supported by a recent study whereby over-expressing SCD was found to protect CHO cells from palmitate induced cytotoxicity [Listenberger et al. 2001].

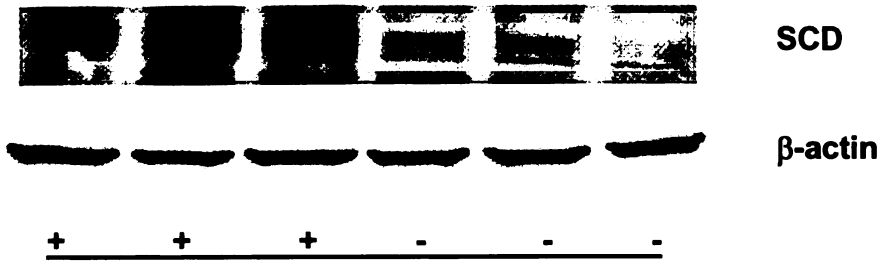


Figure 4.6. Effect of antioxidant N-acetyl-cysteine on SCD measured by western blotting. Down-regulation of SCD in the palmitate culture was prevented by co-supplementing with an antioxidant N-acetyl cysteine (10 mM).

4.3.3.3 Activation of NF-κB by TNF-α is mediated by PKC-δ

In a separate study, we found the phospho-p65 NF-κB levels to be significantly lower in the palmitate cultures than in the control cultures (data not shown). Using the BN inference to simulate an up-regulation of NF-κB in the palmitate cultures predicted that the probability of LDH release being high would decrease (see Table 4.1). NF-κB is an important cytoprotective transcription factor, which can be activated by oxidative stress and cytokines, including TNF-α [Haddad 2002]. From Figure 4.3 we find the connection between TNFα and NF-κB is linked through PKC-δ, suggesting that PKC-δ is an intermediate factor in the activation of NF-κB. PKC-δ has been found to be a redox-sensitive kinase in many cell types [Kanthasamy et al. 2003]. PKC-δ is activated through translocation, tyrosine phosphorylation or proteolysis. During proteolysis, the native PKC-δ (72-74 kDa) is cleaved into two fragments, a catalytically active 41 kDa and a regulatory 38 kDa fragment. The 41 kDa catalytically active fragment plays a key role in promoting apoptotic cell death [Kanthasamy et al. 2003]. In addition, PKC-δ has been found to associate with the P60 TNF receptor during TNFα triggered signaling [Kilpatrick et al. 2002]. As shown in Figure 4.7, TNFα increased the expression of the

PKC-δ 41 kDa catalytic active fragment. This confirmed the connection between TNFα and PKC-δ in the reconstructed network.

Connections between TNFα, PKC-δ and NF-κB have been identified in cells such as neutrophils [Kilpatrick et al. 2002] and pancreatic acinar cells [Satoh et al. 2004].

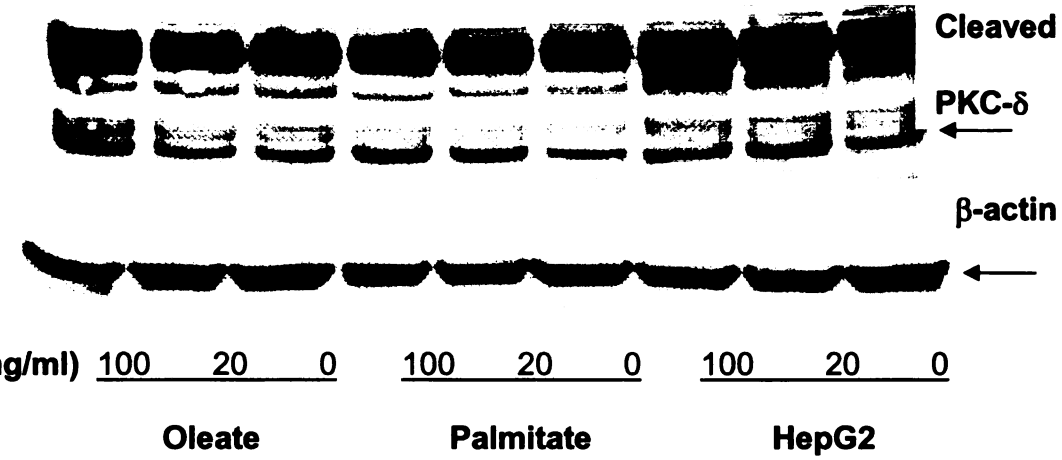


Figure 4.7. Effect of TNFα on PKC-δ expression measured by western blotting. Expression of the catalytically active PKC-δ 41 kDa fragment was increased by TNFα.

Table 4.2 Simulating down-regulation of PKC-δ and its effects on NF-κβ in medium culture (with 20 ng/ml TNF-α).

		Probability of NF-κβ activation	
Gene	**Perturbation**	**Low**	**High**
	no perturbation	0.08	0.92
PKC-δ	downregulation	0.4	0.6 ↓

Note: The probability of NF-κβ taking on a high or low level. The model predicts that simulating a down-regulation in PKC-δ will decrease the probability of NF-κβ taking on a high value. "↓" indicates decreased probability and "↑" indicates increased probability.

Inhibiting PKC-δ has been shown to attenuate TNF-α-mediated activation of the anti-apoptotic transcription factor NF-κB in adherent neutrophils [Kilpatrick et al. 2002]. There has been no study to date suggesting that PKC-δ mediates TNF-α's activation of NF-κB in HepG2 cells. Our model suggests that down-regulating PKC-δ will decrease

the probability of NF-κB taking on a high expression level in the medium (plus TNF-α) cultures (Table 4.2), and maintain a high probability of a high LDH release phenotype in the palmitate cultures (Table 4.1). To determine whether PKC-δ is involved in mediating the activation of NF-κB by TNFα in HepG2 cells, we added rottlerin, an inhibitor of PKC-δ, and measured the activity levels of NF-κB by western blotting, and LDH release. Rotterlin inhibits PKC-δ by inhibiting the tyrosine phosphorylation of PKC-δ. The activity of NF-κB was measured by detecting the levels of phophorylated NF-κB p65 at Ser-536 [Sakurai et al. 2003]. As shown in Figure 4.8, the activation of NF-κB p65 was attenuated by rottlerin. Therefore, PKC-δ was appropriately identified by the model to be an important factor in mediating the TNFα signaling pathway. In addition, the model predicted that inhibiting PKC-δ in the palmitate cultures should maintain a high LDH release phenotype. Indeed, as shown in Figure 4.9, rottlerin supplementation increased LDH release in all the palmitate cultures and the medium cultures with TNFα. Therefore, the cytoprotective role of NF-κB on LDH release in HepG2 cells is mediated in part by PKC-δ.

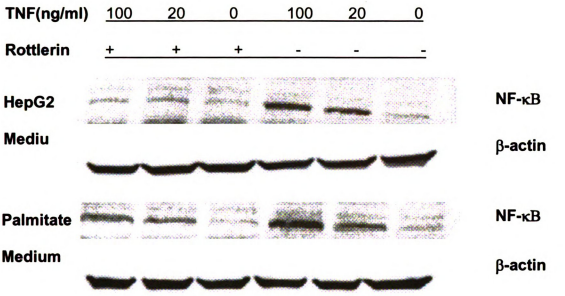


Figure 4.8. Effect of rottlerin on NF-κB measured by western blotting. Expression of phospho-P65 NF-κB in control and palmitate mediums with 0, 20, 100 ng/ml TNFα, with and without rottlerin (5 μM). TNFα activated phospho-P65 NF-κB and this activation was attenuated with the PKC-δ inhibitor, rottlerin.

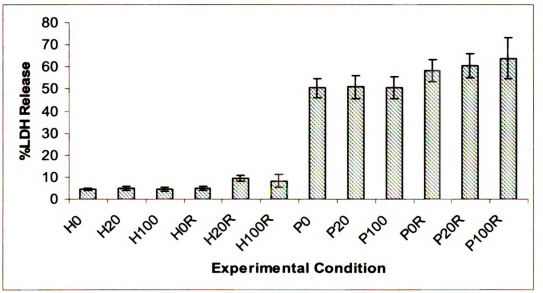


Figure 4.9 Measurement of LDH release in the different culture mediums, with and without rottlerin. Palmitate increased LDH release compared to control cultures. Rottlerin significantly increased LDH release in the palmitate cultures (comparing P0 with P0R, P20 with P20R, P100 with P100R), $P<0.05$ (t-test). H: treated with culture medium (control) for 48 hours, P: treated with 0.7 mM palmitate for 48 hours; R: treated with 5μM rottlerin. Data expressed as average +/- SD from 3 independent experiments.

4.3.3.4 Role of Bcl-2 and PKR in Palmitate induced cytotoxicity

Bcl-2 is a group of proteins including pro-apoptotic members, such as Bax, Bid, Bad, and anti-apoptotic ones such as Bcl-2, Bcl-xl, Bcl-w. Anti-apoptotic Bcl-2 protein inhibits apoptosis by guarding the mitochondrial gate against the release of cytochrome c and the subsequent activation of caspases. Bcl-2 was found to be connected to factors such as TNFα, PKR, TOM20, eIF2B, which in turn influenced LDH release (Figure 4.3). The regulation of Bcl-2 by TNFα is cell dependent [Kim et al. 2002; Tamatani et al. 1999].

The protein expression level of Bcl-2 in cultured HepG2 cells as a function of TNFα concentrations were measured with western blots (Figure 4.10A). In the control cultures, TNFα (20-100 ng/ml) slightly suppressed the protein expression level of Bcl-2 in a dose-dependent manner. Similarly, palmitate significantly decreased Bcl-2 protein expression levels as compared to the control and oleate cultures. The suppression of Bcl-2 may explain in part the palmitate and TNFα induced cytoxicity. In support of this finding, over-expression of Bcl-2 in 2B4.11 T cell hybridoma cell lines have been shown to inhibit palmitate induced cytotoxicity [de Pablo et al. 1999].

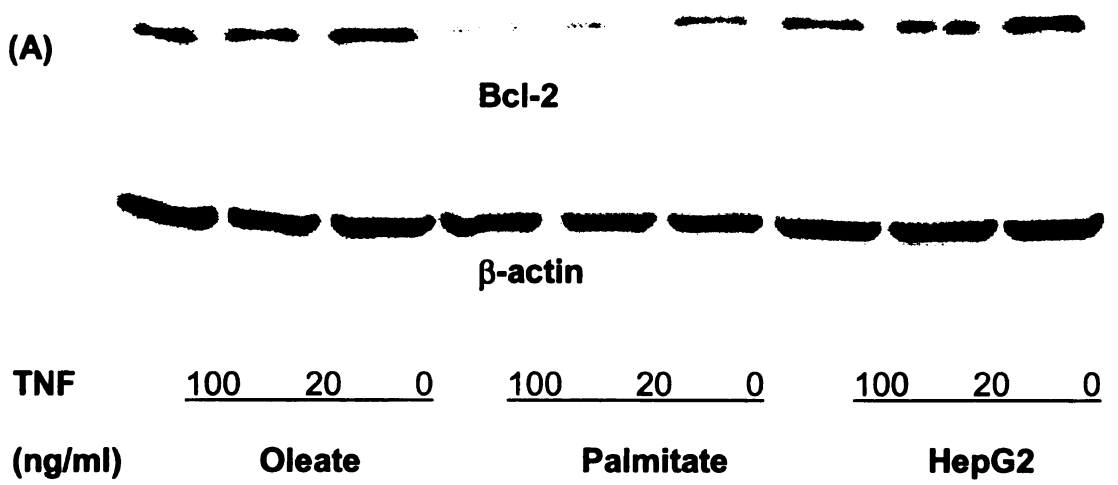


Figure 4.10. Effects of TNFa on Bcl-2. (A) Effect of palmitate and TNFα on Bcl-2 expression measured by western blotting. TNFα supplementation at 20-100 ng/ml downregulated Bcl-2 in the control, palmitate, and oleate cultures. Palmitate downreglated Bcl-2 protein expression level as compared to the control and oleate cultures

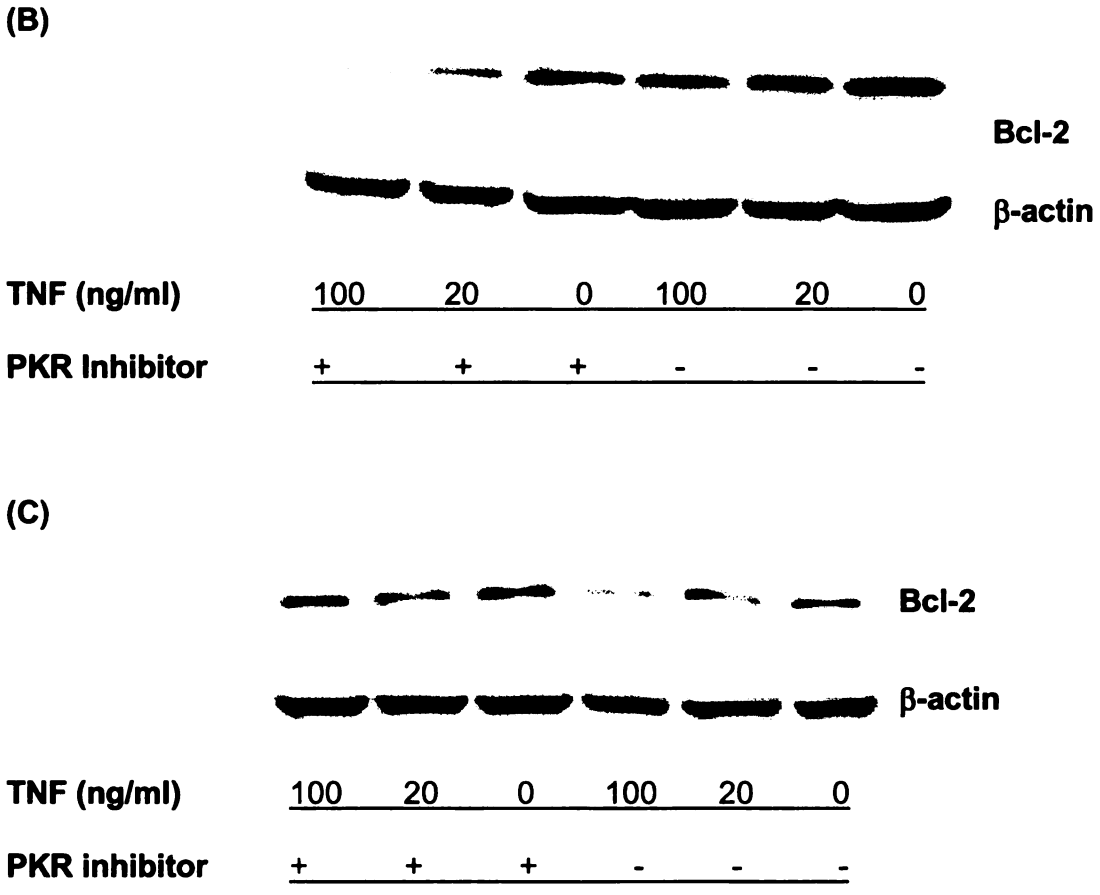


Figure 4.10. Effects of TNFa on Bcl-2. (B) Effect of PKR inhibition on Bcl-2 level in the control cultures. PKR inhibitor (6μM) decreased the expression of Bcl-2. (C) Effect of PKR inhibition on Bcl-2 level in the palmitate cultures. PKR inhibitor (6μM) increased the expression of Bcl-2.

Bcl-2 was connected to the translocase of outer membrane (TOM20), cysteine ligase (GCL) and carnitine palmitoyl transferase (CPT-1) in the reconstructed network. These connections are supported by published results [Motz 2002, Haddad 2004]. Targeting the Bcl-2 protein to the mitochondria is mediated by the interaction between the C terminus of Bcl-2 and TOM20 [Motz et al. 2002], whereas inhibiting GCL in fATIII cells has been found to enhance Bax and p53 expressions over Bcl-2 expression [Haddad 2004]. Finally, direct interaction between Bcl-2 and CPT-1 has been identified with co-immunoprecipitations [Paumen et al. 1997]. These connections predicted by the model

require further analysis into how Bcl-2 may be regulated as well as its role in regulating the mitochondrial membrane in HepG2 cells exposed to palmitate.

PKR was also found in the model to be connected to Bcl-2. Simulating a down-regulation in the PKR node suggests that the LDH release will decrease in the palmitate cultures and increase in the control cultures. Indeed we found that inhibiting PKR (6 μM PKR inhibitor) in the control cultures decreased the Bcl-2 protein expression (Figure 4.10B), correspondingly the LDH release increased from ~3% to ~25%. Inhibiting PKR in the palmitate cultures up-regulated the Bcl-2 protein expression (Figure 4.10C) and decreased the LDH release from ~50% to ~40%. Therefore, the model appropriately identified PKR to be an important factor involved in regulating Bcl-2 protein expression, and in turn LDH release.

In addition, PKR was also found to be connected to eIF2B, PP2A, apoptotic chromatin condensation inducer and apoptosis inhibitor, suggesting that these factors are likely involved in the apoptotic signaling pathway. These connections are supported by published results in the literature. PKR can be cleaved by caspase-3, 7, 8 to liberate the eIF2-α kinase domain, which phosphorylates eIF2-α [Saelens et al. 2001]. Phosphorylation of eIF2-α by PKR will inhibit protein synthesis and lead to apoptosis. Translation of all proteins is dependent upon the binding of the eIF2-GTP-$tRNA_i^{Met}$ ternary complex to the 40S ribosomal subunit, and the amount of eIF2 present in the complex is dependent upon the activity of eIF2B. PKR also can bind to PP2A at the B56 alpha regulatory subunit and increase the phosphatase activity of PP2A [Xu and Williams 2000]. PP2A is a major Ser/Thr phosphatase involved in many signal transduction

pathways. PP2A can dephosphorylate and inactivate the anti-apoptotic Bcl-2 at Ser-70 [Deng et al. 1998].

4.3.4. Model extension

With the availability of high dimensional biological data to characterize a cellular state, one of the challenges is the development of robust methods that can integrate various orthogonal datasets and identify the genes and pathways that induce a phenotype. The significance of the *TIPS*© framework is its ability to extract relevant information, both known and unknown, from the high dimensional data. The phenotypic profile guided the information extraction process. Proteins relay information from the genes to execute biological functions, which define the cellular phenotype, e.g. metabolic functions. Thus, the effects of regulation occurring at the protein level manifest themselves in the phenotype. This chapter illustrates that uncovering this information at the protein (i.e., intermediate) level may be achieved by integrating phenotype and gene expression data.

Metabolite profiles, which characterize the cellular phenotype, may also be used as constraints. Currently, only one phenotypic profile, e.g., LDH release, was used to identify the active network perturbed by TNFα and FFA exposure. The phenotype characterization could be improved by incorporating more metabolic functions. This is discussed in more detail in chapter 6. An approach to obtain more constraints would be to apply ICA or PLS to extract several latent variables from the metabolic profiles [Griffin et al. 2004].

The *TIPS*© framework provides hypotheses of potential mechanisms involved in palmitate and TNFα induced cytoxicity that may be experimentally tested and used to further enhance the model. Experimental approaches, such as ChIP technology, which

measure the physiological interactions between genes and proteins, are becoming more readily available. Integrating interaction data with the gene expression profile using models, i.e. BN, would improve the network inferred from gene expression data alone. One way to accomplish the integration would be to define the *a priori* probabilities of the connections in the gene regulatory network based upon the interaction measurement. Hartemink et al. 2002 provided an example of combining protein-DNA interaction with gene expression data in a BN framework to infer regulatory network underlying pheromone response in yeast.

Due to computational limitation as well as limited availability of data, it is not possible to reconstruct a network with high confidence using all the genes across the genome. GA/PLS and ICA provide an approach to identify relevant genes for further analysis. However, useful information may be missed in the selection process due to low abundance transcripts or the noise in the data. To address the former, methods such as kinetically monitored reverse transcriptase-initiated PCR (kRT-PCR) could be used to measure genes with low abundance transcripts [Holland 2002]. For noise, more replicates of cDNA microarray should be obtained to achieve an accurate estimation of the mRNA expression level. For example, 7 arrrays per sample are required to identify 90% of the truly differentially expressed genes using two-sample t-test [Wang and Chen 2004]. Therefore, another approach to address the latter would be to use a more targeted array with a smaller subset of genes. In conclusion, we have illustrated that *TIPS*© may be applied to reconstruct the active associations, both known and currently unknown, from gene expression and metabolite profiles to help elucidate the pathways involved in regulating palmitate-induced cytotoxicity.

CHAPTER 5 STATE SPACE MODEL TO INFER TRANSCRIPTION FACTOR ACTIVITIES FROM DYNAMIC GENE EXPRESSION DATA

5.1 Introduction

The framework introduced in chapter 4 provided very good predictions of the effect of some, but not all, of the perturbation studies. This may be due in part to Bayesian network analysis' inability to handle transients, such as cycles and feedback loops. Uncovering the additional information from these transients would be valuable in elucidating the mechanism involved in producing a particular phenotype. Therefore, in chapter 5, we developed a dynamic model using time-series gene expression data. To illustrate the ability of the model, we applied the model to *Escherichia coli K12* (*E. coli)* and *Saccharomyces cerevisiae* (yeast), which were more readily amendable to validatation.

Different gene regulatory motifs such as auto-regulatory, single input, multiple input, feed-forward, has been identified from protein-DNA interaction data of *Saccharomyces cerevisiae*[Lee 2002]. These various types of data capture different levels of cellular response to environmental factors and contain within it information about the underlying regulatory network structure. Much effort has been devoted to analyzing these various data sets to reconstruct the regulatory features. A majority of the mathematical approaches, such as clustering, independent component analysis (ICA), principal component analysis (PCA), Boolean network, and Bayesian network have been applied to infer biological information from high dimensional micro-array data. These methods

typically do not incorporate known regulatory information into the structure of the models. Therefore, we developed a state space model (SSM) that integrates gene expression and gene regulatory network structure information to extract underlying information on transcription factor activities.

Cells respond to environmental and physiological changes through an extensive transcriptional regulatory network, which is composed of transcription factors (TF) and genes. These transcription factors bind to the promoter regions of specific genes to either positively or negatively regulate expression. High throughput technologies, such as cDNA microarray, allow the measurement of expression data of the whole genome; however, genome-wide measurement of the regulatory signals, i.e., transcription factor activities (TFA), remains a challenge. Clustering has been applied to gene expression data to identify co-regulated genes [Bar-Joseph et al., 2002; Eisen et al., 1998; Ramoni et al., 2002] and Bayesian network analysis has been applied to infer regulatory networks [Friedman et al., 2000]. The objective of this chapter is to infer TFAs from gene expression data. The advent of the genome-wide binding assay to measure protein-DNA interactions has helped to uncover the network structure describing the connections between TFs and genes in *Escherichia coli K12* (*E. coli)* [Salgado et al., 2000] and *Saccharomyces cerevisiae* (yeast) [Lee et al., 2002]. Given the regulatory network structure and gene expression profiles, the TFAs can be inferred with mathematical modeling.

Several methods have been developed to infer TFAs from gene expression data. A kinetic based approach [Nachman et al., 2004], which modeled mRNA transcription and decay, did not include feedback from genes to TFs. Network Component Analysis (NCA)

[Liao et al., 2003; Kao et al., 2004], which assumed a log-linear relationship between a gene's expression and its regulatory signals, i.e. TFA, modeled the gene regulatory network as multiple-input motifs. Feedback from genes to TFs within network structures, such as in auto-regulation, feed-forward loops, the regulator chain, or the interaction between TFs, is modeled as a "closed-loop" from the TF to the genes, without explicitly modeling the feedback [Tran et al., 2005].

To complement existing approaches, we have developed a state-space model (SSM) with hidden variables that explicitly models feedback in gene regulatory networks to infer the regulatory signals from the gene expression profile. SSM is a subclass of dynamic Bayesian network (DBN). DBN and has been applied to infer the transcriptional regulatory network from gene expression profiles, e.g., T-cell activation [Rangel et al., 2004; Beal et al., 2004]. Other models, such as Hidden Markov model (HMM), the Boolean network, and linear and nonlinear auto-regression models are also subclasses of DBN [Murphy and Mian, 1999]. SSM assumes the existence of state variables that produce observations that are measurable, as well as hidden variables, which are state variables that do not produce an observation. This feature of SSM is attractive for modeling gene regulatory networks. As illustrated in Figure 5.1, a gene regulatory network consisting of TFs and genes can be represented by a SSM. The state of each gene produces observations, such as expression profiles, that can be measured with cDNA microarray. The state of each TF is hidden, and thus, does not produce measurable observations. The structure of the connections can be deduced from measurements of protein-DNA interactions.

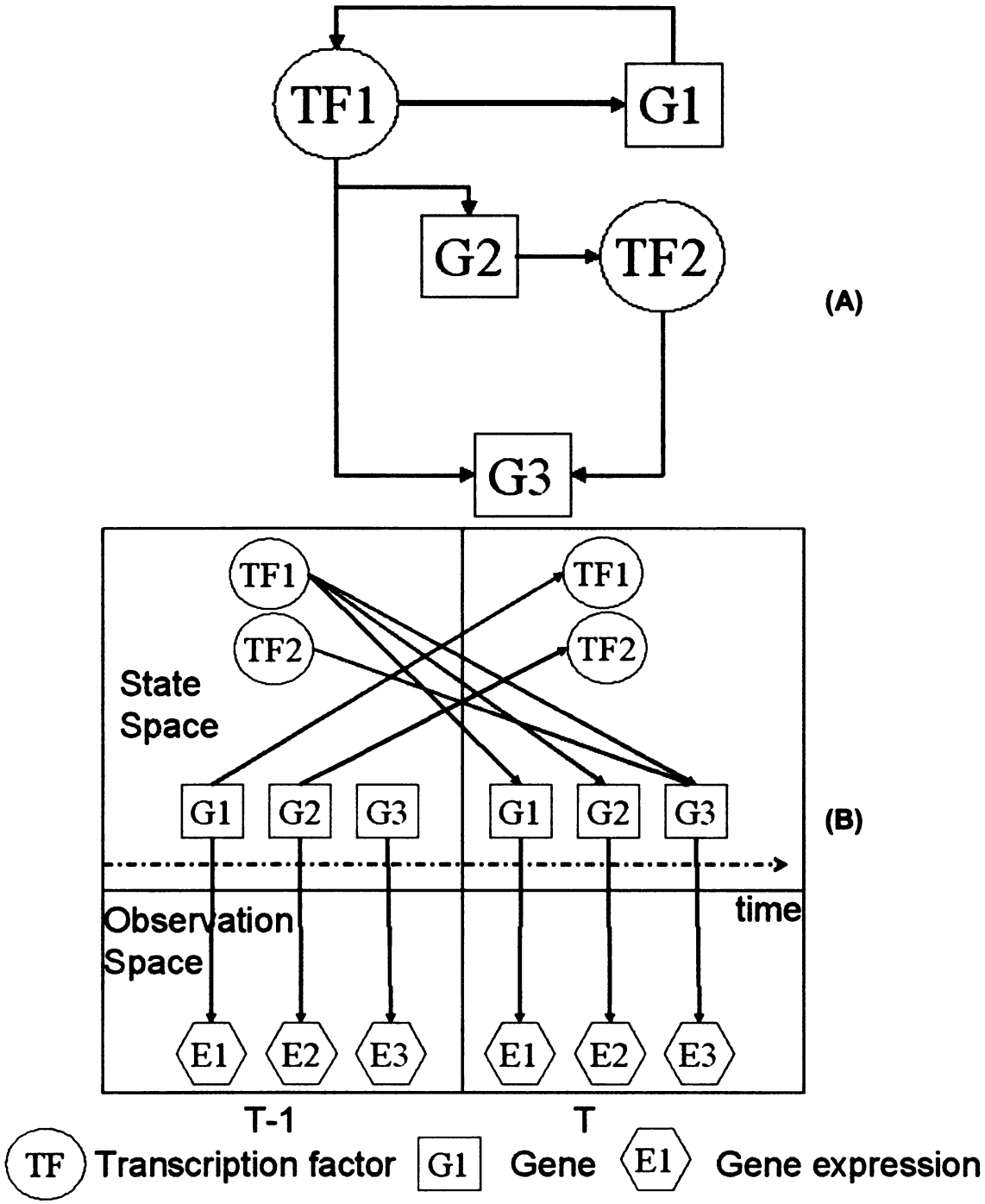


Figure 5.1: SSM representation of gene regulatory network. A) An example of a gene regulatory network with two transcription factors, TF1 and TF2, and three genes, G1, G2, and G3. TF1 regulates G1 and G1 encodes TF1, which formulates auto-regulation between TF1 and G1. TF1 regulates G2, G2 encodes TF2, both TF1 and TF2 regulate G3, which formulates a feed-forward loop between TF1, G2, TF2 and G3. **B) A SSM representation of the gene regulatory network** illustrated in (A). The transcription factors and genes make up the state space, whereas the observation space is comprised of gene expression data.

We demonstrate that SSM can be applied to represent gene regulatory networks of known structures and to infer the TFA from the gene expression profile. We first applied SSM to learn the TFA from data simulated for the gene regulatory network illustrated in Figure 5.1. Then we applied the model to experimental data from *E. coli* transitioning its carbon source from glucose to acetate [Kao et al., 2004] and to cell cycle data from *Saccharomyces cerevisiae* [Spellman et al., 1998]. The inferred activity profile for each TF was validated either by a physical measurement (if available) or activity information from the literature. Finally, further extensions to improve the current SSM model are discussed.

5.2 Materials and methods

State-Space Model

State and Observation: In SSM, a sequence of observations (O_1, O_2,…,O_T) is generated from a sequence of states (S_1, S_2,…,S_T) with the following model:

$$S_t = A\, S_{t-1} + W_t \quad (5.1)$$

$$O_t = B\, S_t + V_t \quad (5.2)$$

where A defines the state transition probability $P(S_t|S_{t-1})$, i.e., how the state at time point T can be determined from the state at time point T-1, and B defines the observation probability $P(O_t|S_t)$, i.e., how the observation at time point at T can be determined from the state at time point T. W~N(0,Q) and V~N(0,R) define the Gaussian noise of the state and observation, respectively. For convenience of notation, parameters A, B, W, V were combined into a single parameter vector θ = (A, B, W, V). In the SSM model, the structure is time invariant and the parameters are also time invariant, i.e., the parameters that determine the transition from T-1 to T are the same as the parameters that govern the

transition from T to T+1. However, the time-scale for each loop is not assumed to be the same. For example, in the yeast dataset [Spellman et al., 1998], the time-scale for each loop is the same (~ 7 minutes). However, in the *E. coli* [Kao et al., 2004] example, the 10 time points were taken at 0, 5, 15, 30, 60, 120, 180, 240, 300, 360 minutes. Therefore, the time-scale for each loop using the first 5 data points is not the same as for the last 5 data points.

Parameter learning: $\theta = (A, B, W, V)$ defines a SSM and is learned from N samples of observation data $O = (O_{1:T}(1),\ldots., O_{1:T}(N))$ by maximizing the likelihood of the observation. An Expectation-Maximization (EM) algorithm is used to learn the parameters. Starting with an initial guess of θ, we perform the E (expectation) step at iteration k to *estimate the value of states S_hat with θ_k and O using inference*; then, we perform the M (maximization) step to *maximize the likelihood of the conditional probability $P(O,S\_hat|\theta)$, such that $\theta_{k+1}=argmax(P(O,S\_hat|\theta))$.* For details of the parameter learning, see [Murphy and Mian, 1999]. We used Bayes Net Toolbox [Murphy, 2001] for the model computation.

State inference: After the parameters are learned with the EM algorithm from the observation data, the value of the state variables, including the hidden variables, can be recursively inferred with the Bayes rule:

$$P(S_t \mid O_{1:t}) = P(O_t \mid S_t) \sum_{O_{t-1}} P(S_t \mid S_{t-1}) P(S_{t-1} \mid O_{1:t-1}) \tag{5.3}$$

Sampling method to generate simulated data

If the structure and parameters of a SSM are defined, data can be generated with sampling methods such as Gibbs sampling. The function of sample_bnet in Bayes Net

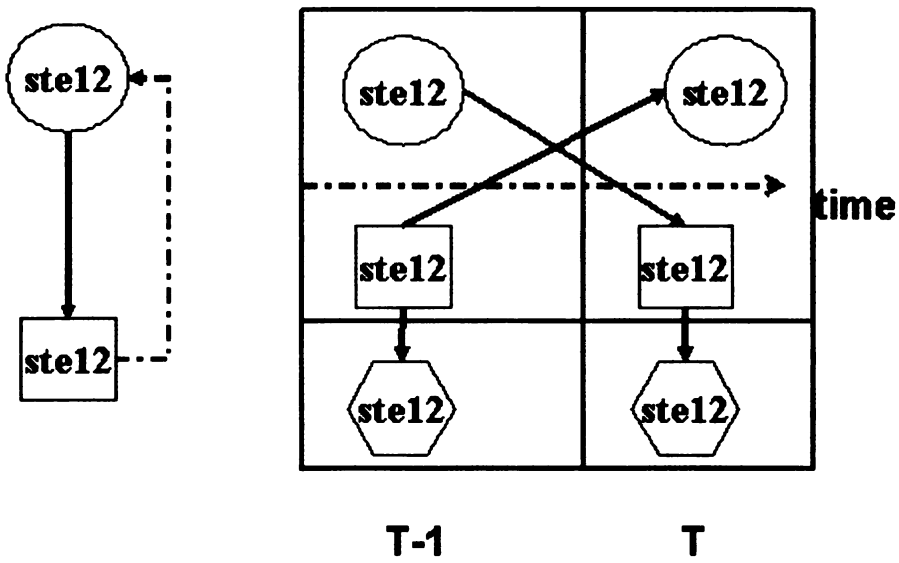
Toolbox [Murphy, 2001] was used to generate the simulated gene expression data and TFA profiles.

Represent gene regulatory motifs within SSM framework

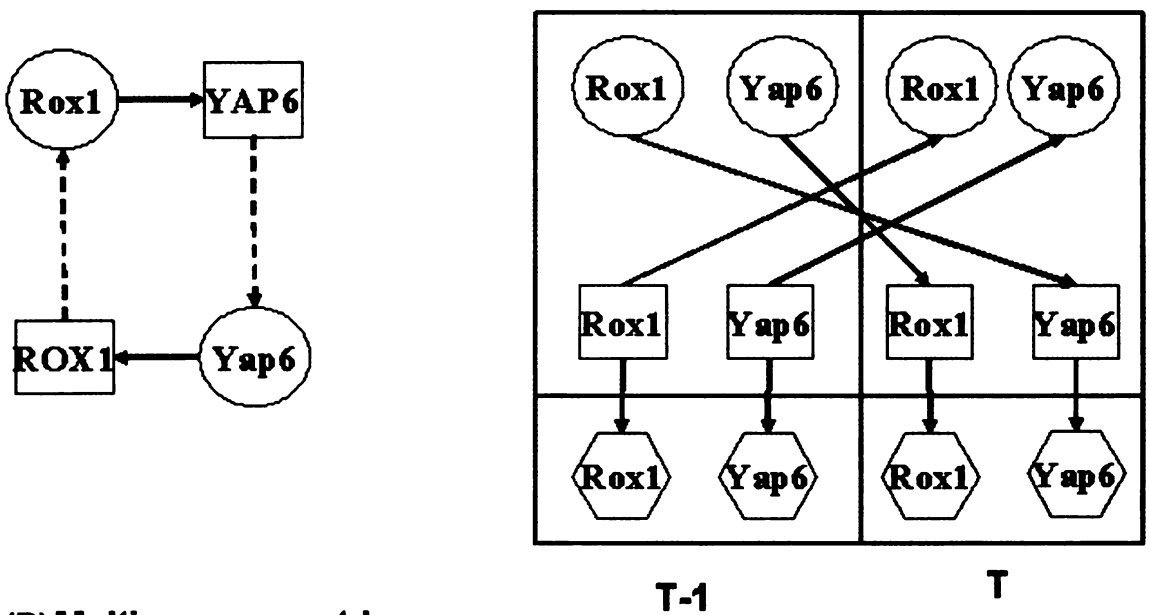
Different motifs, such as auto-regulation, feed-forward loops, multiple-inputs, and single input, have been identified with protein-DNA interaction measurements [Lee et al., 2002]. Protein-DNA interaction measurements identify the DNA sequences of a gene to which a TF will bind. Once a TF binds to a DNA sequence (i.e., binding site) of a gene, the TF regulates the transcription, and in turn the level of expression of that gene. Some TFs bind to similar DNA sequences that may exist on many genes, while other TFs bind to specific DNA sequences present in only one or a few genes. TFA is defined as the concentration of the active conformation of a TF that is capable of DNA binding. In the current SSM representation of a gene regulatory network, each TF has one TFA node and its TFA is the same for all binding sites. Therefore, in the model the level of activity of the transcription factor (i.e., the TFA) indicates only whether a transcription factor is activating (either positively or negatively) its genes or not. The likelihood that a gene is activated by a transcription factor is inferred from the data as conditional probabilities. For example, transcription factor TF1 binds to genes G1 and G2 at the same (or different) binding sites. The activity level of TF1 is assumed to be the same for G1 and G2, however, the probabilities that G1 or G2 is activated by TF1 are different as defined by P(G1|TF1) and P(G2|TF1). In addition, the SSM assumes that there is a time delay between binding and transcription.

Here we demonstrate that SSM is able to model these motifs, i.e., auto-regulation, feed-forward loops, multiple-inputs, and single input. As shown in Figure 5.1, SSM is a

dynamic model composed of two parts: states and observations. State variables generate observations that are measurable, whereas state variables that do not generate observations are called hidden variables. A static gene regulatory motif can be represented dynamically by connecting genes at time point T-1 to the TFs they encode at time point T, and connecting the TFs at time point T-1 to the genes that they regulate at time point T. Each TF has a TFA node. Each gene has a state node and an observation node. Duplicating the state and observable variables of a gene will increase the computational load. However, this limitation has an intrinsic advantage. Having the state variable (gene) and the observation variable (gene expression) as separate entities allows the SSM approach to take into account potential effects, such as post-transcriptional effects and measurement noise. Ideally, if RNA decay and measurement noise did not exist, then a gene that is transcriptionally activated would be measured at the mRNA level to be activated. In other words, Pr (E=1 | G=1) =1. However, mRNA expression is determined by the net effect of mRNA transcription and degradation. Effects, such as RNA decay (post-transcriptional modification) and noise in the microarray measurement, could result in conditional probabilities Pr (E=1|G=1) less than 1. In other words, a gene that is transcriptionally activated may be measured at the mRNA level to be inactivated if mRNA decay dominates over transcription [Wang et al., 2002]. Figure 5.2 illustrates a graphical representation of the auto-regulation, feed-forward loop, multiple-input, and multi-component loop motifs.



(A)autoregulation



(B) Multi-component loop

Figure 5.2: An example and its SSM representation of the gene regulatory motifs of A) auto-regulation. Transcription factor Ste12 binds to the promoter sequence of gene STE12, which encodes transcription factor Ste12. B) multi-component loop. Transcription factor Rox1 binds to the promoter sequence of gene YAP6, which encodes transcription factor Yap6. Transcription factor Yap6 binds to the promoter sequence of gene ROX1, which encodes transcription factor Rox1.

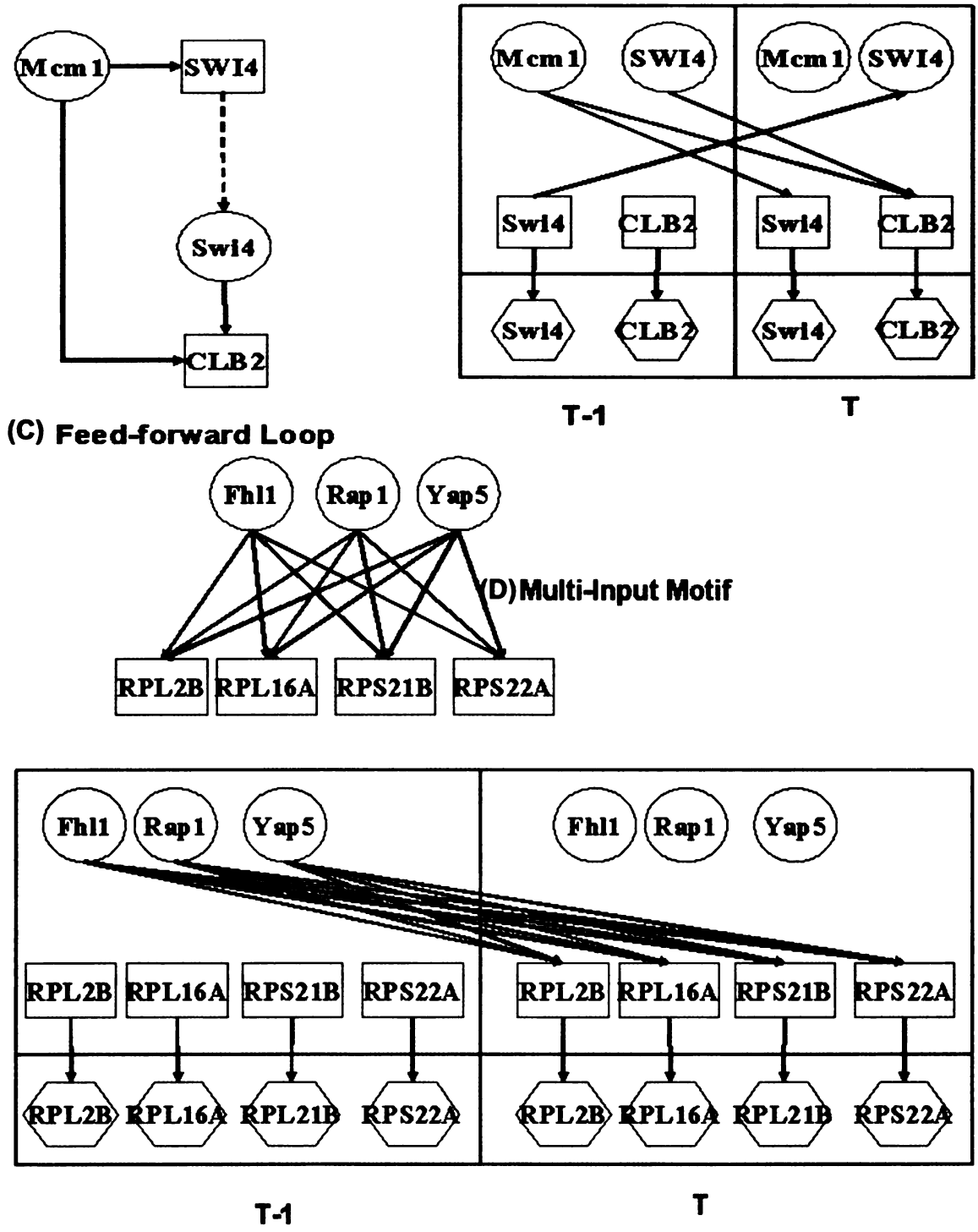


Figure 5.2: An example and its SSM representation of the gene regulatory motifs of C) feed-forward loop. Transcription factor Mcm1 binds to the promoter sequence of gene SWI4, which encodes transcription factor Swi4. Transcription factor Swi4 and Mcm1 bind to promoter sequence of gene CLB2. and D) multi-input motif. Transcription factors Fhl1, Rap2, Yap5 all bind to the promoter sequences of genes RPL2B, RPL16A, RPS21B, RPS22A.

Threshold determination

We used a function with a definable threshold Th to discretize the gene expression data. Any gene that showed a change larger than Th, based upon $\log_2$ ratio, was assigned a discrete value of 1, or otherwise was assigned a value of 0. Thus, a threshold of 1 indicates that a 2-fold (2^1) change in the expression of a gene, relative to its initial state, is significant.

SSM requires an optimal threshold in order to obtain reliable results. To find an optimal threshold, we evaluate the TFA data for each TF over various thresholds between −1 and 1. The optimal threshold, which gives the most appropriate profile for *each* TF, is determined by comparing the TFA results with known (e.g., measured or literature) values. For example, the gene expression data that we used for *Saccharomyces cerevisiae* was taken over 18 time points that spanned two cell cycles [Spellman et al., 1998]. The TFs we studied in *Saccharomyces cerevisiae* are known to have phase-specific activity during the cell cycle [Aerne et al. 1998; Baetz et al., 1999; Kovacech et al., 1996; Lee et al., 2002; MacKay et al., 2001; McInerny et al., 1997; Oehlen et al., 1996; Spellman et al., 1998]. Therefore, we assumed that the activity profile of each TF during the first cell cycle should be repeated in the second cell cycle. We identified the optimal threshold as the one that predicted this cyclic behavior for all the transcription factors.

5.3 Results

5.3.1. Inferring TFA from simulated data

Before testing SSM with experimental data where the transcription factor activities are unknown, we applied SSM to a simulated system. We created a simple regulatory network (see Figure 5.1a) with the feed-forward loop and auto-regulation motifs containing two TFs and three genes. From the known network structure, we constructed a SSM representation of the network (see Figure 5.1b) and pre-defined the θ parameter, which is shown in Table 5.1(a). Using the network reconstructed by the SSM, we simulated the dynamic profiles of TFA and gene expression levels with the sampling method discussed in the Methods section. Given the simulated gene expression profile (shown in table 5.1(c)) and the network structure, SSM learned the parameters θ and the TFA profile. The learned parameters are shown in Table 5.1(b). They match closely to the predefined parameters in Table 5.1(a). The learned or inferred TFA was compared to the simulated TFA. The learned TFA matched 95% of the time points (38 out of 40) of the simulated TFA. The results of this simulation provided confidence that the SSM may be able to analyze a real system. The simulations were performed on Matlab R13, Windows XP, on a PC with Celeron CPU 2.4 GHZ and RAM 512MB. The simulation time was 328.5 seconds.

SSM requires sufficient data to infer the parameters. We evaluated how many data points are needed with the simulation data by varying the number of time points from 6 to 40. Ten time points were needed to correctly infer 80% of the TFA profile and this percentage increased with the number of time points used. The sampling time needs

to be small enough to capture the dynamic profile, which will vary with the biological system being modeled.

Table 5.1 (a) Parameters of the simulated network shown in Figure 5.1

1.Initial probabilities of (TF1, TF2, G1, G2, G3) [P(x=1),P(x=2)] = [0.5, 0.5]
2.Conditional probabilities between gene (G) and its expression (E) [P(E=1|G=1),P(E=2|G=1),P(E=1|G=2),P(E=2|G=2)= [0.9, 0.1, 0.1, 0.9]
3.Conditional probabilities between transcription factors (TF) and genes (G) TF1->G1; TF1->G2 [P(G=1|TF=1),P(G=2|TF=1),P(G=1|TF=2),P(G=2|TF=2)] = [0.9, 0.1, 0.1, 0.9] **(TF1+TF2)->G(3)** P(G=1|TF1=1,TF2=1)=0.9 P(G=1|TF1=1,TF2=2)=0.7 P(G=1|TF=2,TF2=1)=0.7 P(G=1|TF=2,TF2=2)=0.01
4.Conditional probabilities between genes (G) and transcription factors (TF) G1->TF(1);G2->TF2 P(TF=1|G=1)=0.9 P(TF=2|G=1)=0.1 P(TF=1|G=2)=0.1 P(TF=2|G=2)=0.9

Table 5.1 (b) Learned parameters of the simulated network shown in Figure 5.1

1.Initial probabilities of (TF1, TF2, G1, G2, G3)

[P(TF1=1),P(TF1=2)] = [0.25, 0.75]
[P(TF2=1),P(TF2=2)] = [0.55, 0.45]
[P(G1=1),P(G1=2)] = [0.49, 0.51]
[P(G2=1), P(G2=2)] = [0.45, 0.55]
[P(G3=1),P(G3=2)] = [0.44, 0.56]

2.Conditional probabilities between gene (G) and its expression (E)

[P(E1=1|G1=1),P(E1=2|G1=1),P(E1=1|G1=2),P(E1=2|G1=2)= [0.83, 0.17, 0.13, 0.87]
[P(E2=1|G2=1),P(E2=2|G2=1),P(E2=1|G2=2),P(E2=2|G2=2)= [0.95, 0.05, 0.12, 0.88]
[P(E3=1|G3=1),P(E3=2|G3=1),P(E3=1|G3=2),P(E3=2|G3=2)= [0.91, 0.09, 0.15, 0.85]

3.Conditional probabilities between transcription factors (TF) and genes (G)

TF1->G1
[P(G=1|TF=1),P(G=2|TF=1),P(G=1|TF=2),P(G=2|TF=2)]
= [0.94, 0.06, 0.05, 0.95]
TF1->G2
[P(G=1|TF=1),P(G=2|TF=1),P(G=1|TF=2),P(G=2|TF=2)]
= [0.83, 0.17, 0.07, 0.93]

(TF1+TF2)->G(3)
P(G=1|TF1=1,TF2=1)=0.76
P(G=1|TF1=1,TF2=2)=0.86
P(G=1|TF=2,TF2=1)=0.15
P(G=1|TF=2,TF2=2)=0.04

4.Conditional probabilities between genes (G) and transcription factors (TF)

G1->TF(1)
P(TF=1|G=1)=0.92
P(TF=2|G=1)=0.08
P(TF=1|G=2)=0.1
P(TF=2|G=2)=0.9
G2->TF(2)
P(TF=1|G=1)=0.44
P(TF=2|G=1)=0.56
P(TF=1|G=2)=0.34
P(TF=2|G=2)=0.66

Table 5.1 (c) Simulated gene expression data

T	E1	E2	E3
1	1	0	0
2	0	0	0
3	1	1	0
4	0	0	0
5	1	1	1
6	0	0	1
7	1	1	1
8	1	1	0
9	1	0	0
10	0	0	0
11	1	1	0
12	1	1	0
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
17	1	1	1
18	0	1	1
19	1	1	1
20	1	0	1
21	0	0	1
22	1	1	1
23	1	0	1
24	0	1	1
25	0	0	0
26	0	0	1
27	0	0	0
28	0	0	0
29	0	0	0
30	0	0	0
31	0	1	0
32	0	0	1
33	0	0	0
34	0	0	1
35	0	0	1
36	0	0	1
37	0	0	0
38	1	0	1
39	0	0	0
40	1	0	1

Note: T: time point, E1, E2, E3 are expression data of gene 1, 2, 3 respectively. 0: inactive. 1: active.

Table 5.2 Eight genes regulated by ARCA and CRP

Gene	TFs
Acs	CRP(+)
SucD	ArcA(-)CRP(+)
SucC	ArcA(-)CRP(+)
SdhB	ArcA(-)CRP(+)
SdhA	ArcA(-)CRP(+)
SucB	ArcA(-)CRP(+)
mdh	ArcA(-)CRP(+)
gltA	ArcA(-)CRP(+)

Note: (+) positive control, (-) negative control

5.3.2. E. coli

We applied SSM to a model system: *E. coli* transitioning from glucose to acetate as a carbon source. The gene expression data was obtained from [Kao et al., 2004] and the regulatory information is available from the regulonDB database. The SSM model included two TFs, CRP and ArcA, and eight of the genes (shown in Table 5.2) that are regulated by the two TFs. The dynamic profiles of these two TFs were learned from the gene expression levels of the eight genes. In [Kao et al., 2004], cAMP was measured to indirectly indicate the activity of CRP, since the activation of CRP requires the binding of cAMP. From the measurement of cAMP (see Figure 5 in Kao 2004), we can see that the level decreased from its initial high but remained upregulated, around 10 fold above the basal level, from the second time point onward. Indeed, SSM inferred that CRP is active from the second point onwards (see Figure 5.3). For the first time point, however, CRP is predicted to be inactive. The expression level at the first time point is the reference; all subsequent expression levels are measured relative to the expression level at the first time point. SSM also identified that ArcA was inactive for the first four time points and active for the remaining 6 time points (see Figure 5.3). An ArcA measurement is not readily obtainable for comparison.

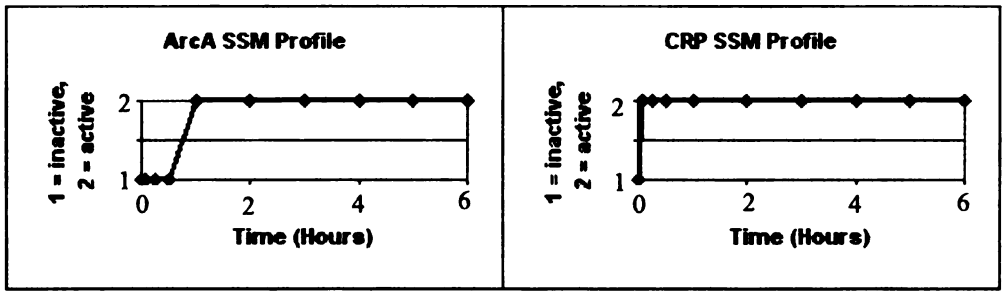


Figure 5.3: The results of using a SSM to analyze an *E. coli* system. SSM predicts that CRP is inactive initially and then active from the second time point onwards, as expected from the cAMP measurement [Kao 2004]. SSM predicts that ArcA is inactive for the first four time points, but is activated for the last six time points.

5.3.3. Saccharomyces Cerevisiae

Next, we applied SSM to model several of the common regulatory motifs in *Saccharomyces cerevisiae*. According to Lee et al. (2002), 39 out of the 106 studied regulators were involved in *feed forward loops* and 10 of the 106 were involved in *auto-regulation*. Lee et al. (2002) also found that in combinations of two or more of the 106 regulators, 295 were involved in *multi-input motifs*. Therefore, we selected TFs with the aforementioned regulatory motifs, and whose activities could be verified by the literature. Namely, MCm1, Swi4, Swi5, and Swi6, which are well studied and well-understood [Aerne et al., 1998; Baetz et al., 1999; Kovacech et al., 1996; Lee et al., 2002; MacKay et al., 2001; McInerny et al., 1997; Oehlen et al., 1996; Spellman et al., 1998] TFs. SSM was applied to analyze the system illustrated in Figure 5.4. The gene regulatory motifs, feed-forward (Mcm1 → SWI5 → Swi5 → YJL160C + PIR1 + PIR3), multiple-input (Mcm1 + Swi4 + Swi6 → CLA4 + SWI4 + LSM4 + PCL1 + SIM1 + GIN4 + YDR509W), and auto-regulation, (Swi5 → SWI5; Swi4 → SWI4) were obtained from [Lee et al. 2002].

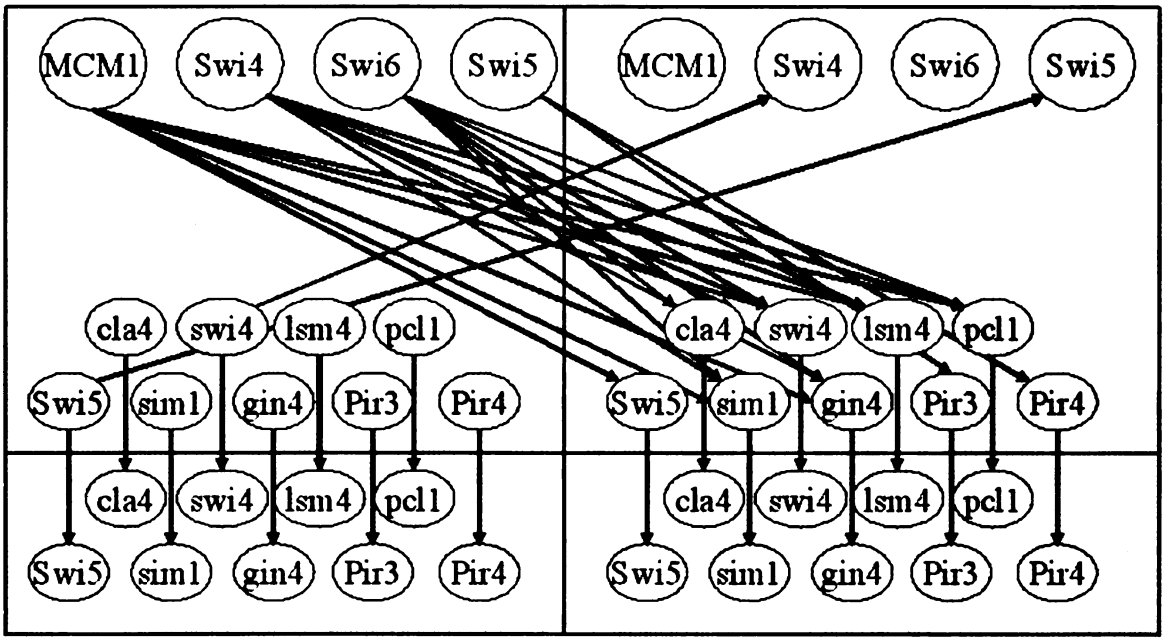


Figure 5.4: A SSM representation of the *Saccharomyces cerevisiae* (yeast) system studied. The gene regulatory motifs were taken from [Lee et al. 2002], including feed-forward (Mcm1->SWI5->Swi5->YJL160C+PIR1+PIR3), multiple-input (Mcm1 + Swi4 + Swi6 -> CLA4 + SWI4+LSM4+PCL1+SIM1+GIN4+YDR509W), auto-regulation (Swi5->SWI5; Swi4->SWI4).

We evaluated the ability of the SSM to infer TFA profiles during the cell cycle of *Saccharomyces cerevisiae*. Lee et al. (2002) performed a genome-wide binding analysis to obtain the connectivity information between the TFs and genes in yeast. We coupled the connectivity information [Lee et al., 2002] with gene expression data taken from yeast cultures synchronized by α-factor arrest [Spellman et al., 1998]. Of the four synchronization methods used by Spellman et al., we chose the data obtained with the α-factor method because it presented the least amount of missing data for the genes studied. With the α-factor arrest method, the data were sampled every 7 minutes, which captured approximately two cell cycles with the 18 time points [Spellman et al., 1998]. This provided 9 time points (~ 63 minutes) for each cell cycle. Each phase (i.e., M, G1, S, and G2) within a cell cycle takes about 15 minutes [Liao et al., 2003]. In other words, one cell cycle is about 60 minutes. Therefore, a 7 minute sampling time is small enough to

capture the phase change profile within a cell cycle. The binding motifs and gene expression data were used by SSM to infer the TFAs.

The SSM predictions are consistent with the literature results. SSM inferred that MCm1 is active during the G2/M/G1 phases, Swi4 and Swi6 are active during the G1 and S phases, and Swi5 is active during the M and G1 phases. We confirmed the predictions (Figure 5.5) made by SSM with the literature. Past studies found that MCm1 induced the expression of many genes during the G2/M/G1 phases. High transcription of both FAR1 and STE2 in the G2/M phases requires MCm1 [Oehlen et al., 1996]. MCm1 is also known to induce the transcription of CLN3, SWI4, and CDC6 at the M/G1 boundary [Spellman et al., 1998]. In contrast, Swi4 induced genes in the G1 and S phases. Swi4 is the DNA binding component of SBF [Baetz et al., 1999]. Baetz et al. (1999) indicated that SBF promotes the induction of gene expression at the G1/S-phase transition of the mitotic cell cycle. MacKay et al. (2000) also showed that a complex containing Swi4 induces CLN1 and CLN2 transcription in the late G1 and drives the transition to S. Similarly, Swi6 induced genes during the G1 and S phases. The activity of Swi6 is very similar to that of Swi4, and these two factors are known to be connected [Baetz et al., 1999; MacKay et al., 2000]. MacKay et al. (2000) showed that a Swi4-Swi6 complex induces CLN1 and CLN2 transcription in late G1 until S. Baetz et al. (1999) also suggested that the DNA binding domain of Swi4 is inaccessible in the full-length protein when not complexed with Swi6. In contrast, Swi5 was activated during the M phase and the M/G1 boundary. Kovacech et al. (1996) found that Swi5 is partially responsible for the peak in EGT2 expression during late M and early G1 phases. Aerne et al. (1998)

found that Swi5 regulates the expression of PCL2, PCL9, and the SIC1 Cdk inhibitor in the late M phase.

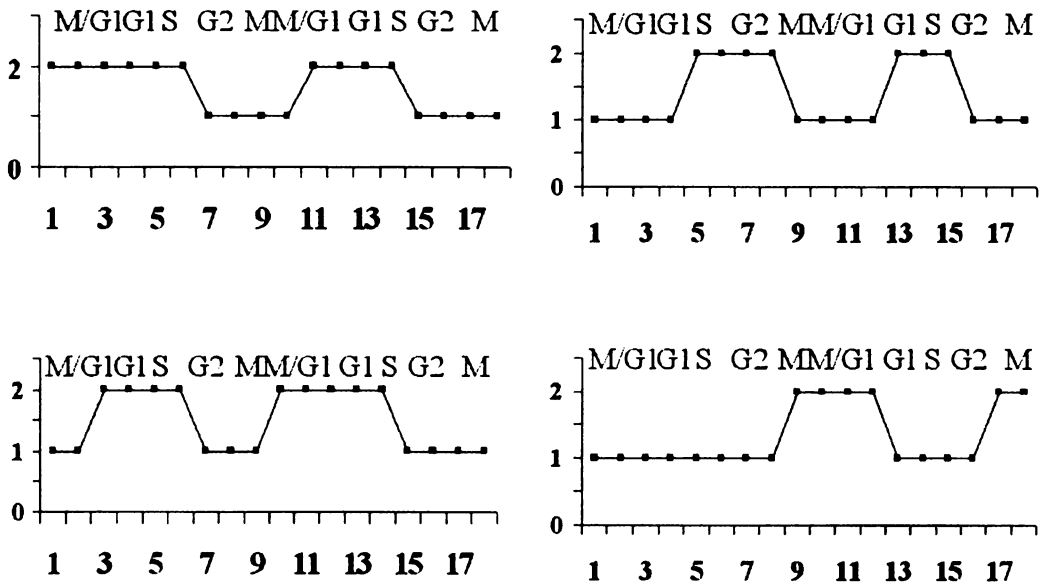


Figure 5.5: The results of using a SSM to analyze a yeast system. The SSM predictions closely followed the experimental trends and phases in which each transcription factor is known to be active or inactive [Aerne et al., 1998; Baetz et al., 1999; Kovacech et al., 1996; Lee et al., 2002; MacKay et al., 2001; McInerny et al., 1997; Oehlen et al., 1996; Spellman et al., 1998]. SSM inferred that MCm1 is active during the G2/M/G1 phases, that Swi4 and Swi6 are active during the G1 and S phases, and that Swi5 is active during the M and G1 phases.

In summary, SSM was applied to three systems, a simulated gene regulatory network and two experimental systems, *E. coli* and *S. cerevisiae.* The simulation study confirmed the ability of SSM to infer network parameters and state values from observational data. Application of SSM to the experimental systems illustrates that TFAs can be inferred from the gene expression data given the regulatory network structure.

5.4 Discussion

SSM is a subclass of DBN. DBN has been applied to infer the structure of regulatory networks from temporal gene expression data [Rangel et al., 2004; Beal et al., 2005; Ong

et al., 2002; Perrin et al., 2003; Nachman et al., 2004]. Ong et al. (2002) explicitly included operons as hidden variables in the model to facilitate the incorporation of *a priori* biological knowledge of the co-expressed genes to improve the quality of the analysis. Here, we explicitly included TFs in the model and included the connections between TFs and genes obtained through binding analysis [Lee et al., 2002]. SSM has the potential to infer unmeasurable, as well as unmeasured, connections and events.

The benefit of the SSM over existing models (e.g., NCA [Liao et al., 2003], kinetic modeling [Nachman et al., 2004) is that all gene regulatory motifs, including feedback from gene to TFs, are explicitly modeled. NCA implicitly models the feedback from gene to TFs [Tran et al., 2005]. In kinetic modeling [Nachman et al., 2004], feedback from gene to TFs is not incorporated. In contrast, SSM can explicitly model the feedback from gene to TFs, such as the auto-regulatory motif. This facilitates the incorporation of domain or experimental knowledge. For example, if a TF is experimentally knocked–out or silenced, the SSM approach could easily incorporate this information for the auto-regulatory motif.

In SSM, the nonlinear relationship between the TFs and genes are quantified with conditional probabilities, i.e. $P(O_t|S_{t-1})$. By using conditional probabilities, SSM does not presuppose a relationship between the TFs and genes in the model, whereas NCA assumes a log-linear relationship and kinetic modeling assumes a form for the rate law, such as Michaelis-Menten kinetics. Conversely, NCA [Liao et al., 2003] and kinetic modeling [Nachman et al., 2004] can infer continuous profiles of TFA. Another advantage of kinetic modeling is the ability to explicitly model both mRNA transcription and decay. In the current application of SSM the gene expression data are discretized,

thus allowing us to infer when the TFAs are active. This was sufficient to allow verification of the model predictions with the literature. For example, the model predicted the phase in the cell cycle in which the genes that are regulated by a TF are activated, which can be compared with the phase of the cell cycle in which the genes are known to be activated in the literature.

The SSM assumes there is a time delay between binding and transcription. By considering the state values at time T as a function of the state values at time T-1, the SSM implicitly assumes a time delay for all TF effects on gene expression. In other words, although the binding of a TF to the DNA sequence of a gene may occur quickly [McAdams and Arkin, 2002], there is a time offset between the binding of the TF to the DNA sequence of a gene and the onset of transcription. This time offset ranges between minutes to hours [Kersberg, 2004]. It has been shown that incorporating a time delay in modeling gene regulatory networks is critical to inferring the oscillatory behavior of NF-kB [Monk, 2003]. We further evaluated this assumption by allowing the genes to be regulated by the current TFA in the yeast dataset. Without the time delay, the cyclic TFA could not be inferred. This, in addition to the previous study [Monk, 2003], suggests that this biologically relevant time delay [Kersberg, 2004] must be incorporated in the model to accurately infer the TFA profiles. In some cases, if the actual time delay is on the order of minutes and the measurements are taken on the order of hours, then considerable error would be introduced. In those cases, it would be more appropriate to incorporate the connection between the TFs and genes in the same time slice.

In the simulation study, the TFA inferred by the SSM matched the simulated TFA well but not exactly. The mismatches may be due in part to the EM algorithm being a

local optimization method [Ong et al., 2002], in other words, the algorithm cannot guarantee a TFA of (global) maximal likelihood. The optimization could be improved by either running EM multiple times from different starting points or using a global search algorithm, such as Markov Chain Monte Carlo (MCMC) [Murphy 2001].

The SSM model determines an optimal threshold value for discretizing the gene expression data based upon *a priori* knowledge of the TFA. If no *a priori* knowledge is available for the TFA dynamics, this can be addressed one of two ways. In one approach, the TFAs could be estimated from other approaches, e.g. NCA, and the estimated TFA could be used to determine the threshold value. In the other approach, the optimal threshold could be determined in the SSM by including the threshold value (Th) as a part of the parameter learning process, i.e., in the parameters θ = (A, B, W, V, Th). How the observation data O(Th) = ($O_{1:T}$ (1),...., $O_{1:T}$ (N)) is discretized depends on the threshold value, e.g., a very high threshold value will set all the genes in the inactive state while a very low threshold value will set all the genes in the active state. The Expectation-Maximization (EM) algorithm could be used to learn the parameters. Starting with an initial guess of θ, we can perform the E (expectation) step at iteration k to *estimate the value of states S_hat with* θ_k *and O(Th) using inference*; then, we can perform the M (maximization) step to *maximize the likelihood of the conditional probability* $P(O(Th),S\_hat|\theta)$, *such that* $\theta_{k+1}=argmax(P(O(Th),S\_hat|\theta))$. Therefore, the parameters, including the threshold value, can be inferred by maximizing the probability of the observation and state values given the inferred parameters (e.g., conditional probabilities).

The current SSM infers the most probable model, given the observed data, by approximating the underlying structure of the noise to be Gaussian [Murphy and Mian, 1999; Perrin et al., 2003]. Gene expression is an inherently stochastic phenomenon [McAdams and Arkin, 2002]. SSM modeled the regulatory networks stochastically using conditional probabilities. This probabilistic SSM may capture some of the stochastic nature of the gene regulatory network, but an accurate representation of the stochasticity requires further understanding of the underlying structure of the noise. Without knowing the structure of the noise, studies have assumed it to be Gaussian [Perrin et al., 2003].

The current SSM model could be extended to incorporate a step to learn the structure before inferring the TFAs by searching for a network that gives the maximal likelihood against the observation. The structural information obtained from the binding analysis could be used to construct the initial network as a starting point for the search [Nachman et al. 2004]. Alternatively, the connections indicated by the interaction data could be used to define a priori probabilities of the connections in the network [Hartemink et al., 2002]. Thus, the connections that are supported by the interaction measurements would have a higher likelihood of being valid connections in the network than the unsupported connections. By including a step to learn the structure, it could help refine the network by inferring the interactions that are unmeasurable or missing due to error (noise) in the measurement. A more accurate network provides more confidence to the inferred TFA profile. In addition, the model size was limited by the computational tool that was used, namely Bayes Net Toolbox [Murphy, 2001], which is on a Matlab platform. We are currently working on developing an executable SSM in a C++ version

of Bayes Net Toolbox, Probabilistic Networks Library (PNL) http://www.intel.com/technology/computing/pnl/ to handle larger model sizes.

CHAPTER 6 EXTENSIONS TO IMPROVE THE *TIPS*© FRAMEWORK

6.1 Introduction

In the previous chapters, we introduced the *TIPS*© framework to identify genes and pathways relevant to a phenotype, e.g. cytotoxicity. This chapter discusses improvements to the current *TIPS*© framework. Improvements include hierarchically integrating phenotypic, metabolic and gene expression profiles with genomic function information, inferring pathways that confer a phenotype involving multiples metabolites and developing a dynamic modeling. Preliminary results are presented in this chapter.

6.2 A Hierarchical Approach Employing Metabolic and Gene Expression Profiles to Identify the Pathways that Confer Cytotoxicity in HepG2 Cells

6.2.1 Introduction

In the *TIPS*© approach introduced in chapter 2 to chapter 4, genes relevant to a phenotype e.g. cytotoxicity were selected with GA/PLS and ICA without incorporating metabolic profiles and functional information of the genes. It is the objective of this section to develop an approach to hierarchically integrate phenotypic, metabolic and gene expression profiles with *a priori* knowledge to select genes relevant to a phenotype. The approach was applied to the cell culture system introduced in chapter 4, namely HepG2 cells exposed to FFAs and TNF-α, to study the genes involved in palmitate induced cytotoxicity.

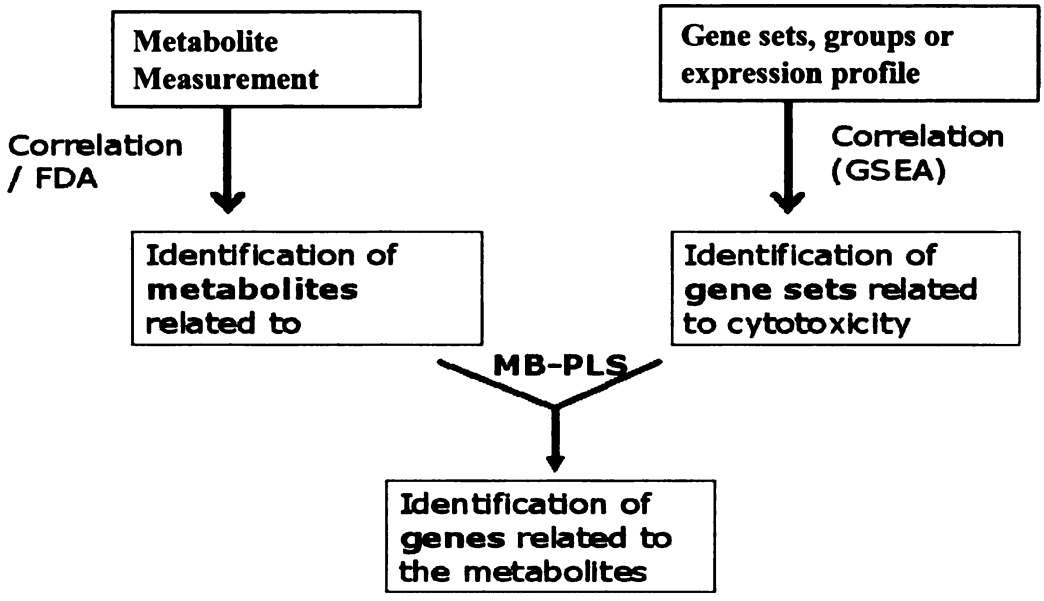


Figure 6.1 An overview of the hierarchical approach. First, important metabolic fluxes relevant to a phenotype were identified with discriminant analysis. Second, to identify the transcriptionally altered pathways, gene set enrichment analysis (GSEA) was applied to the cDNA microarray data. Finally, the expression of the enriched gene sets and the metabolic profiles were integrated with a multi-block partial least squares analysis (MBPLS) regression model to identify the genes and gene sets that regulate the metabolic pathways identified in step 1. This figure was prepared by Shireesh Srivastava.

An overview of the hierarchical framework that was developed is shown in Figure 6.1. The framework consisted of three stages. First, the metabolic fluxes of glucose, fatty acid and amino acid metabolism were measured to identify the metabolic changes responsible for producing the cytotoxic phenotype. Different phenotypes induced by TNF-α and FFA were characterized with Fisher's Discriminant Analysis (FDA) (Chan 2003 c) to identify the metabolic fluxes that contributed most to separating the phenotypes. Ketone body (e.g., beta-hydroxybutyrate (BOH) and acetoacetate) release and TG accumulation were identified to be the most important in separating the phenotypes, suggesting their involvement in inducing the cytotoxicity. However, the FDA of the metabolic changes did not provide information on the changes in the signaling and gene regulatory pathways. To obtain this missing information, the genomic responses were measured using cDNA microarrays, and analyzed with the gene set

enrichment analysis (GSEA) [Subramanian 2005]. GSEA integrates *a priori* knowledge of a gene's functional role with the expression data to detect concerted expression changes in a set of genes responsible for producing a phenotype. GSEA indicated the involvement of mitochondria-related, oxidative stress-related and FFA metabolism pathways in inducing the cytotoxic phenotype. GSEA identifies the possible functional pathways, but does not provide a quantitative model to predict the effects of individual genes on the phenotype. Therefore, in the final step, the gene expression and metabolic flux profiles were integrated with multi-block partial least squares analysis (MBPLS) [Hwang 2004] to identify the genes most relevant to the metabolic fluxes which correlated most to the cytotoxicity. In the MBPLS model, the different gene sets identified to be enriched by GSEA formed the blocks of the MBPLS. This ensured that the blocks were separated according to the functional role of the genes. Block scores were extracted from each block to predict the metabolic fluxes. Some of the identified genes were experimentally perturbed to validate their predicted roles in regulating the cytotoxicity.

6.2.2 Materials and Methods

Fisher's Discriminant Analysis

FDA was applied to identify the measured metabolic fluxes that contributed to the separation of the different phenotypes (cytotoxic versus nontoxic). FDA was applied to 15 samples belonging to 5 groups with 27 metabolic flux measurements for each sample. FDA identifies the projection axes that maximize the ratio of the between-group and the within-group variations. Details on the FDA algorithm can be found in [Chan 2003c].

Two axes T1 and T2 were extracted; each is a linear combination of the 27 metabolic fluxes.

$$T = \sum_i W_i \times Met(i) \tag{6.1}$$

W_i provides the contribution of each metabolic flux in separating the phenotypes in the T1 and T2 directions.

Gene Set Enrichment analysis (GSEA) of the gene data

In order to evaluate the genetic response of the cells to the treatments, cDNA microarray analyses were conducted. The expression levels of 19,522 genes were measured, and the data was analyzed using GSEA. GSEA aims to identify the gene sets whose coordinated change differentiates two phenotypes. The software GSEA-P, from http://www.broad.harvard.edu/gsea/, was used for the GSEA analysis. Thirty seven gene sets from the molecular signature database, MsigDB [Subramanian 2005] functional gene group c2 were selected, as shown in Table 6.1. These sets included 10 metabolic pathways, 26 signal pathways and 1 cellular component. An enrichment score of a gene set S characterizes whether the set of genes randomly distributed across the list or falls mainly at the bottom or top of the list. The null hypothesis that a gene set S randomly distributes across the ranked gene list was tested with Kolmogorov-Smirnov test, with the statistical significance value estimated through 1000 random permutations of the phenotype label. The gene sets with a high significance of enrichment are considered important in separating the distinct phenotypes. Here, the expressions of 19,522 genes were measured and ranked based upon the comparison between the palmitate treatment and the control using t-test.

Integrating the gene expression and metabolic flux profiles

MBPLS is a hierarchical multivariate analysis method [MacGregor 1994; Lopes 2002], where the variables are divided into different blocks based upon *a priori* knowledge, for example according to different stages of a process [MacGregor 1994] or different metabolic pathways in a cell [Hwang 2004]. In this study, we separated the genes into different blocks based upon their functional roles in different pathways. This facilitates the identification of an important block (gene set) to a metabolic flux, and then further identifies the important genes within the block. Gene expression X is divided into K blocks X = [X1,X2,...,Xk], block scores B = [b1, b2, ...bk] are extracted from each block

Table 6.1: Gene sets used in the GSEA analysis

Metabolic pathway	
Glycolysis	fatty acid metabolism
TCA	Oxidative phosphorylation
PPP pathway	beta-oxdiation
Fructose	Sphingoglycolipid
Fatty acid biosynthesis	Glutathione
Singal pathway	
Bcl2family_and_reg_network	Map_kinase
CAPASE	Mapkk
Cell death	Ppara
ceramidePathway	pparG
crebPathway	S1P signaling
NFKB	Programmed cell death
P38	ERK1 Erk2 Mapk pathway
Akt	JNK
EGF	TNFalpha
EIF2	Stress Pathway
EIF4	TNFR1 pathway
Electron Transport Chain	TNFR2 pathways
ERK pathway	
Cellular Component	
Mitochondria	

respectively. A PLS regression model was then constructed to map B to a metabolic flux. For details of the MBPLS algorithm, please refer [Hwang 2004]. Important genes sets

were identified by evaluating the weights of each block and importance of individual genes were then further identified by evaluating the regression coefficients of the genes within the block. The N-way toolbox (http://www.models.kvl.dk/source/nwaytoolbox) [Anderson 2000] was applied to conduct the MBPLS modeling.

6.2.3 Results

Metabolic flux important in separating cytotoxic phenotype

FDA was applied to identify the metabolic fluxes responsible for separating the phenotypic response (cytotoxic versus nontoxic as defined by the level of lactate dehydrogenase (LDH) release). The metabolic flux data were obtained by Shireesh Srivastava. As shown in Figure 6.2 (a), palmitate cultured samples were separated from the rest in a two dimensional space defined by FDA projections T1 and T2. Furthermore, as shown in Figure 6.2 (b) metabolic fluxes of BOH, acetoacetate, intracellular TG accumulation and amino acids such as aspartate, alanine, ornithine were the most significant in separating the cytotoxic (palmitate) from the nontoxic phenotype (control, unsaturated fatty acids, TNF-α and their combinations). Intracellular TG, ornithine, aspartate have the highest coefficients in separating palmitate from the other conditions in the T2 direction, while BOH, acetoacetate and alanine separated the two groups in the T1 direction, which suggested differences in the metabolic state between the phenotypes exist mainly in the aforementioned metabolic fluxes such as TG, BOH and acetoacetate. While FDA can identify the important metabolic fluxes to the phenotype, it does not provide the direction of the relation, i.e., whether the variables are correlated positively or negatively. To obtain the direction of the relationships, we performed correlation analysis. The (Pearson's) correlation coefficients (r) between cytotoxicity and metabolic

fluxes are shown in Table 6.2. BOH, acetoacetate, alanine release had positive correlations. Acetoacetate and BOH are produced from acetyl-CoA, the metabolic intermediate of fatty acid oxidation. Increased fatty acid oxidation is known to increase oxidative stress and cell death [Gill 2006, Sanyal 2002]. TG accumulation, ornithine and aspartate had negative correlations. TG accumulation has been found to protect cells from palmitate induced lipoxicity [Listenberger 2003]. Ornithine and aspartate may affect fatty acid metabolism through the TCA cycle. Thus, both the FDA and correlation analyses suggested that metabolic fluxes relevant to fatty acid oxidation are positively associated with the palmitate-induced cytotoxicity, while metabolic fluxes relevant to the accumulation of TG are negatively correlated to cytotoxicity.

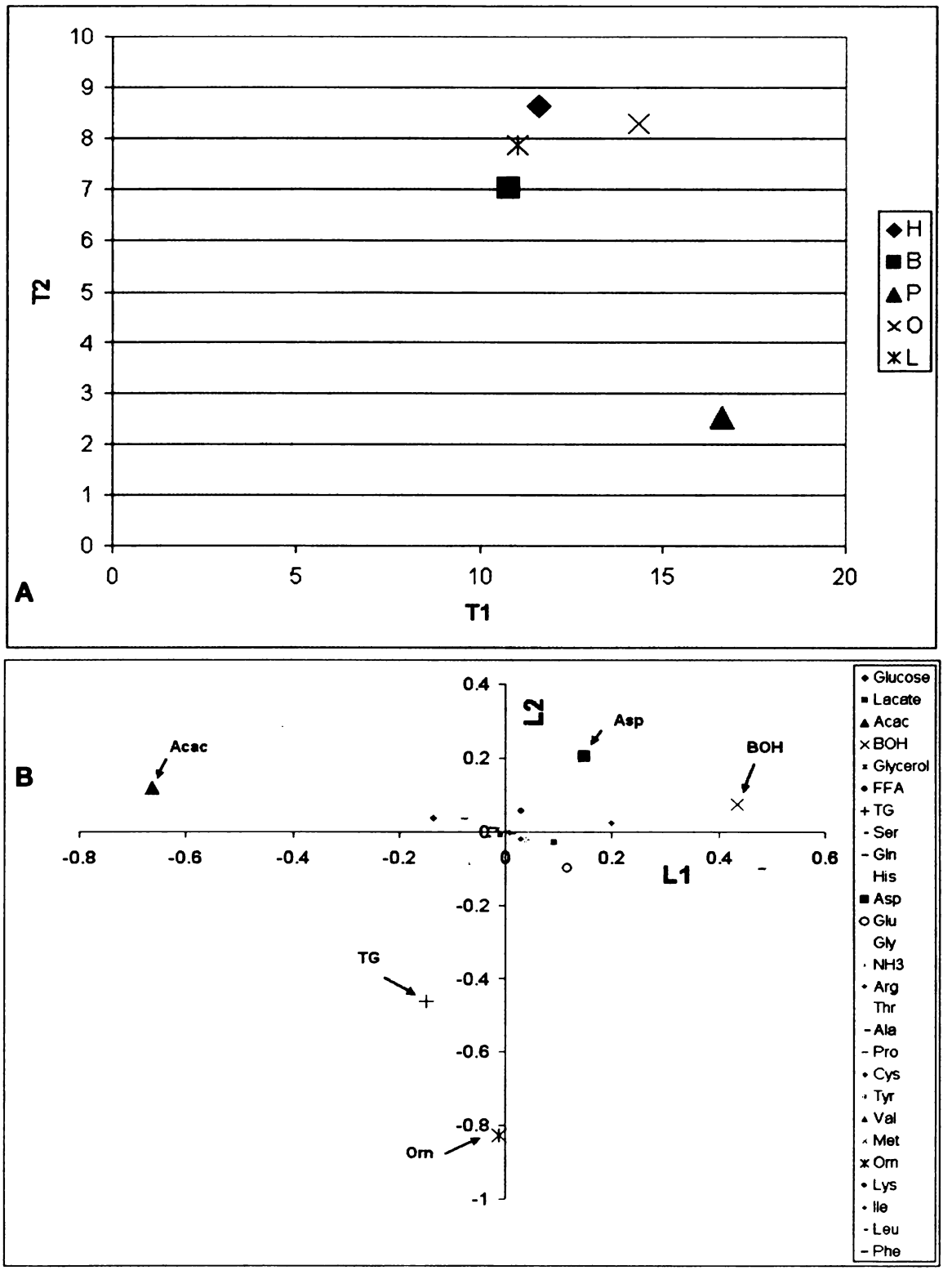


Figure 6.2. Fisher's discriminant analysis of metabolites. (a) Two dimensional score (T1 and T2) plot of the samples, H: control, B: BSA, P: palmitate culture, O: oleate culture, L: linoleate culture. (b) Loading plot of the first two dimensions for the metabolites.

Table 6.2 Correlation between the metabolic fluxes and the LDH release

Metabolite	Correlation
Beta-hydroxybutyrate	0.96
Acetoacetate	0.74
Lactate	0.47
Glutamate	0.37
Alanine	0.27
Fatty acid	0.25
NH4 (in)	0.19
Serine	0.14
Histidine	-0.09
Aspartate	-0.09
Triglyceride	-0.33
Glutamine	-0.35
Arginine	-0.5
Lysine	-0.51
Proline	-0.52
Glycine	-0.55
Tyrosine	-0.55
Isoleucine	-0.56
Leucine	-0.6
Threonine	-0.6
Ornithine	-0.61
Valine	-0.62
Phenylalanine	-0.64
Methionine	-0.69
Glycerol	-0.7
Cysteine	-0.84

Gene Sets Enriched by Palmitate

37 gene sets involved in a variety of cellular processes and organelles, such as fatty acid metabolism, cell death, TNF-α signaling, etc., were evaluated to identify which of these processes were associated with palmitate-induced cytotoxicity (Table 6.1). Of the 37 gene sets evaluated, 14 were found to be significantly enriched with nominal p values less than 0.05, and shown in Table 6.3. The pathways that were enriched included

electron transport chain (ETC), fatty acid metabolism, glycolysis, oxidative phosphorylation, ROS, the pentose phosphate pathway (PPP), cell death, fatty acid beta-oxidation, TCA cycle, fructose metabolism, glutathione, and the ERK 1, ERK 2, MAP kinase pathways. Therefore, it suggested that the altered fatty acid metabolism may be associated with oxidative stress related pathways such as ROS, glutathione, oxidative phosphorylation and ETC to induce cytotoxicity. Interestingly, the gene set belonging to sphingolipid (ceramide) metabolism was not selected, suggesting that ceramide metabolism does not play an important role in the observed toxicity. Notably, all the enriched gene sets belong to the mitochondria, for example fatty acid beta-oxidation, electron transport chain, oxidative phosphorylation, indicating a central role of this organelle in the observed cytotoxicity. To further support this finding, the gene set of mitochondria with 229 genes, defined according to MsigDB, was found to be significantly changed in the palmitate-treated cells, with a nominal p value of 0.00189.

Genes Important to Cytotoxicity

To identify the genes that regulate the metabolic functions most closely associated with cytotoxicity, a MBPLS model was developed to predict the metabolic processes based upon the expression data of the gene sets identified by GSEA. Ketogenesis was identified to be positively related to LDH release by the FDA and correlation analysis. Therefore, a MBPLS model was developed to predict BOH based upon the expression profiles of the 14 enriched gene sets shown in Table 6.3. Cytoxicity-related functional pathways were identified by evaluating the block weights. Based upon the block weigth of the MBPLS

Table 6.3 Enriched Gene sets in GSEA analysis

Gene Sets	SIZE	NES	NOM p-val
Electron Transport Chain	43	-1.998	0
Fatty acid metabolism	33	-1.824	0
Glycolysis	55	-1.765	0
Oxidative Phosphorylation	30	-1.89	0.004
ROS	26	-1.741	0.009
PPP pathway	15	-1.644	0.01
Cell death	15	-1.631	0.01
ERK pathway	31	-1.62	0.016
Fatty acids beta-oxidation	7	-1.58	0.02
Fatty acid biosynthesis	5	-1.574	0.021
TCA	15	-1.631	0.023
Fructose	19	-1.599	0.026
ERK1 Erk2 Mapk pathway	29	-1.542	0.034
Glutathione	23	-1.52	0.036
Mitochondria	229	-1.945	0.002

Note: Size: number of genes in the gene set; NES: normalized enrichment score; NOM p-val: nominal P value.

identified by evaluating the block weights. Based upon the block weight of the MBPLS model, it was identified that functional groups such as glycolysis, oxidative phosphorylation, ETC, ERK, ROS, glutathione, and fatty acid metabolism had relatively higher weights, suggesting the involvement of these pathways in inducing the cytotoxicity. The genes relevant to these functional groups were identified by evaluating the regression coefficients of the individual genes, and are discussed below.

Identification and literature evaluation of the cytotoxic pathways through BOH

BOH and acetoacetate were identified to be important in separating the phenotypes. Increased production of ketone bodies, such as BOH and acetoacetate, were observed in the palmitate treated cells. Increased fatty acid oxidation has been shown to be associated with oxidative stress [Gill 2006, Sanya 2003]. Thus, the genes with large (positive or

negative) regression coefficients to BOH release, for example, could suggest pathways relevant to the cytotoxicity. The regression coefficients of the genes were inspected to identify the individual genes that were important. The genes with the largest regression coefficients are listed in Table 6.4. Genes relevant to ROS production, fatty acid metabolism, and detoxification of lipid peroxidation products were found to have high regression coefficients. Genes involved in ETC and oxidative phosphorylation including NADH dehydrogenase with positive regression coefficients, indicated a positive role in the cytotoxicity. ROS are produced during the electron transfer and oxidative phosphorylation process. NADH dehydrogenase complex, or complex I of the electron transport chain, is a main source of superoxide production [Moller 2001].

Table 6.4 Genes selected by MBPLS model

Num	VAID	GenName
0	933	CASP8 and FADD-like apoptosis regulator
1	1088	glutathione S-transferase M1
2	1095	Glucose-6-phosphatase, catalytic
3	1160	mitogen-activated protein kinase 3
4	1172	dihydrolipoamide dehydrogenase
5	1197	cytochrome P450, family 2, subfamily A, polypeptide 7
6	1222	clusterin (complement lysis inhibitor, SP-40)
7	1247	cytochrome P450, family 2, subfamily E, polypeptide 1
8	1319	aldehyde dehydrogenase 3 family, member B2
9	1322	alcohol dehydrogenase IB (class I), beta polypeptide
10	1332	acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain
11	1341	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C, isoform 1
12	1342	ATPase, Cu++ transporting, alpha polypeptide
13	1554	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)
14	1650	TNF receptor-associated factor 2 [T55353,NM_021138.3,7186]
15	2024	forkhead box M1 [AA136566,NM_202003.1,2305]
16	2056	TNF receptor-associated factor 3 [AA504259,NM_145725.1,7187]
17	2111	brain and reproductive organ-expressed (TNFRSF1A modulator)

Table 6.4 (Continued)

18	2152	peroxiredoxin 2 [H68845,NM_005809.4,7001]
19	2177	HIV-1 Rev binding protein [AA485958,NM_004504.3,3267]
20	2179	Fas (TNFRSF6)-associated via death domain
21	2201	isocitrate dehydrogenase 3 (NAD+) gamma
22	2202	caspase 8, apoptosis-related cysteine protease
23	2244	glutathione S-transferase M4 [AA486570,NM_000850.3,2948]
24	2408	v-raf murine sarcoma 3611 viral oncogene homolog
25	2427	neutrophil cytosolic factor 1 (47kDa)
26	2510	FOS-like antigen 2 [T58873,NM_005253.3,2355]
27	2541	zinc finger protein 11B [N57658,NM_006955.1,7558]
28	3926	(gF) insulin-like growth factor 1 receptor_(AA256532,_,Hs.239176)
29	4111	interleukin 17D [N92873,NM_138284.1,53342]
30	4824	myeloperoxidase [R05801,NM_000250.1,4353]
31	5100	FOS-like antigen 2 [T58873,NM_005253.3,2355]
32	5110	v-raf murine sarcoma 3611 viral oncogene homolog
33	5146	baculoviral IAP repeat-containing 3 [AA002126,NM_001165.3,330]
34	5147	caspase 3, apoptosis-related cysteine protease
35	5148	baculoviral IAP repeat-containing 2 [AA702174,NM_001166.3,329]
36	5186	platelet-derived growth factor receptor, alpha polypeptide
37	5532	alcohol dehydrogenase 4 (class II), pi polypeptide
38	5640	BCL2-associated athanogene 4 [N25897,NM_004874.2,9530]
39	5982	(gC) ESTs, Highly similar to A33983 pyruvate kinase (EC 2.7.1.40)
40	6532	KIAA1117 [AA001870,NM_015018.2,23033]
41	7320	tumor necrosis factor, alpha-induced protein 3
42	7546	aldehyde dehydrogenase 1 family, member A2
43	8188	tumor necrosis factor receptor superfamily, member 1B
44	8331	nerve growth factor receptor (TNFR superfamily, member 16)
45	8512	clusterin-like 1 (retinal) [H91647,NM_199167.1,27098]
46	8692	epithelial membrane protein 2 [T88721,NM_001424.3,2013]
47	8749	Kruppel-like factor 6 [AA055585,NM_001300.4,1316]
48	8959	fucose-1-phosphate guanylyltransferase [R38619,NM_003838.2,8790]
49	9089	glutathione S-transferase M5 [N56898,NM_000851.2,2949]
50	9112	fructose-1,6-bisphosphatase 1 [AA699427,NM_000507.2,2203]
51	9148	mitogen-activated protein kinase kinase 2 [AA425826,NM_030662.2,5605]
52	9160	cytochrome P450, family 2, subfamily J, polypeptide 2
53	9269	aldehyde dehydrogenase 1 family, member A1
54	9282	acyl-Coenzyme A dehydrogenase, C-2 to C-3 short chain
55	9404	ATP citrate lyase [H08548,NM_001096.2,47]
56	9456	acylphosphatase 1, erythrocyte (common) type
57	10037	acyl-CoA synthetase long-chain family member 3
58	10252	mitogen-activated protein kinase kinase kinase 7
59	10378	chemokine (C-C motif) ligand 5 [AA486072,NM_002985.2,6352]
60	10441	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4, 9kDa
61	10486	glutathione peroxidase 3 (plasma) [AA664180,NM_002084.2,2878]
62	11001	ATPase, H+ transporting, lysosomal 13kDa, V1 subunit G isoform 1
63	11072	Phosphoglucomutase 3 [AA189113,NM_015599.1,5238]
64	12701	amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 13

Table 6.4 (Continued)

65	12963	receptor (TNFRSF)-interacting serine-threonine kinase 1
66	13744	interleukin 17B [AA443286,NM_014443.2,27190]
67	13975	oxidative-stress responsive 1 [R95128,NM_005109.1,9943]
68	16995	copper chaperone for superoxide dismutase [N30404,NM_005125.1,9973]
69	17052	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, beta
70	17066	dual specificity phosphatase 2 [AA759046,NM_004418.2,1844]
71	17114	ecotropic viral integration site 1 [AA181023,NM_005241.1,2122]
72	17207	epithelial membrane protein 3 [W73810,NM_001425.1,2014]
73	17299	amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 3
74	17355	TNFRSF1A-associated via death domain [AA916906,NM_003789.2,8717]
75	17461	serine/threonine kinase 25 (STE20 homolog, yeast
76	17484	2,3-bisphosphoglycerate mutase [AA678065,NM_199186.1,669]
77	18054	mitogen-activated protein kinase kinase kinase 3 [
78	18637	NADH dehydrogenase (ubiquinone) flavoprotein 1, 51kDa
79	19717	(gC) alcohol dehydrogenase IB (class I), beta polypeptide
80		ACAC (Metabolic flux)
81		BOH (Metabolic flux)
82		TG (Metabolic flux)
83		LDH (Phenotype)

Experimental Validations

In the hierarchical approach, FDA suggested a protective role of TG accumulation. Indeed, our laboratory had found that supplementing palmitate cultures with oleate increased TG synthesis and reduced the cytotoxicity significantly, to the same levels as the control [Li 2006]. GSEA found ROS to be an important factor associated with the cytotoxicity of saturated FFA and TNF-α. Indeed, our laboratory has found that ROS levels were increased in response to palmitate and that ROS played an important role in the observed cytotoxicity in the palmitate cultures [Srivastava 2006]. NADH dehydrogenase was identified by MBPLS to be one of the genes that is positively related to cytotoxicity. This was experimentally validated for its role in regulating palmitate-induced cytotoxicity [Srivastava 2006]. The levels of ROS and LDH release were

measured in the palmitate cultured cells with NADH dehydrogenase inhibitor and both were significantly reduced [Srivastava 2006, Li 2006], which confirmed the role of NADH dehydrogenase in the observed toxicity as predicted by the model.

6.2.4 Discussion

A hierarchical framework was developed to integrate phenotypic, metabolic and gene expression profiles with functional genomics information to identify the pathways which played an important role in regulating FFA and TNF-α induced cytotoxicity in HepG2/C3A cells. The phenotype was connected to the gene expression profile through the intermediate metabolic level. Metabolic changes that contributed and are relevant to the cytotoxic phenotype were identified by discriminant and correlation analyses. Pathway-based gene expression analysis, namely GSEA, was used to identify the functional pathways that were associated with the cytotoxic phenotype. Metabolic and signaling pathways identified from the gene expression analysis support the results obtained by FDA and correlation analysis of the metabolic analysis and provided additional information on the mechanisms of toxicity. While GSEA was able to identify the gene sets important for the phenotype, it did not provide information on which genes within these sets were responsible for mediating the cytotoxic effect. MBPLS offered the advantage of identifying the individual genes that were directly related to the desired phenotype. We applied this framework to a model system of HepG2/C3A cells and successfully identified the targets to reduce FFA toxicity. For example, GSEA identified that mitochondrial and ROS-related genes, but not ceramide synthesis contribute to the toxicity, which was found experimentally to be indeed the case [Srivastava 2006]. The

MBPLS prediction of the role of NADH dehydrogenase in regulating cytotoxicity was validated with complex I inhibitor studies [Srivastava 2006, Li 2006].

6.2.5 Conclusion

In summary, this section illustrated how phenotypic, metabolic and genetic profiles can be integrated hierarchically to identify phenotype relevant genes and pathways. The metabolites whose levels were most strongly related to the cytotoxicity were identified. The genes associated with these metabolic processes were subsequently identified. Determining the genes which control the levels of these metabolites provided important information on the likely mechanism of toxicity and identified the targets to control the cytotoxicity of palmitate. This approach identified the involvement of ROS generation, altered fatty acid and energy, but not ceramide metabolism in the cytotoxicity. Thus, the integration of metabolic and genetic information can provide a more comprehensive picture of the cellular states and provide potential targets that regulate the cellular responses.

6.3 Inferring active pathways that confer a phenotype involving multiple metabolites

6.3.1 Introduction

In the current application of *TIPS*© to identify cytotoxicity related pathways in HepG2 cells, the profile of LDH release was used to characterize the phenotype. In order to define a phenotype more accurately and robustly, more information, e.g. more metabolites should be incorporated. Metabolites are connected to each other through the metabolic network, which are implicitly captured by the metabolite profiles [Li and Chan

2004b]. Therefore, multiple metabolites can discriminate different phenotypes more readily than a single biomarker.

6.3.2 Materials and Methods

Extracting pathways for multiple metabolites

Let S(t) be the profile of pathways related to a phenotype in experiment t and L1(t) be the profile of metabolite L1 in experiment t, L2(t) be the profile of metabolite L2 in experiment t. S is constrained to have a correlation with L1 and L2 by equation (6.2), where $\rho 2$ is a threshold value.

$$Corr2 = S^T*L1*L1^T*S/(S*S^T) *A+ S^T*L2*L2^T*S/(S*S^T) *B>\rho 2 \quad (6.2)$$

A and B in the above equation defines the weight of each latent variable that contributes to the overall correlation of the pathways related to a phenotype.

6.3.3 Results and discussion

Preliminary study to identify genes relevant to both LDH and TG

We applied the approach described above to identify genes relevant to two cellular processes, TG and LDH, hereby denoted as two factors. The results are presented in Table 6.5. The genes selected using only LDH are listed in Table 6.6 for comparison.

Table 6.5 Important genes selected for two factors TG and LDH.

Access Num	Name
AA018907	protein phosphatase 3 (formerly 2B), regulatory subunit B, 19kDa, alpha isoform
R06605	protein tyrosine phosphatase, non-receptor type 1 (PTPN1)
AA160670	lysophosphatidic acid phosphatase (ACP6)
AA464163	acyl-Coenzyme A dehydrogenase, very long chain (ACADVL)
R39463	aldolase C, fructose-bisphosphate (ALDOC)
AA437389	Lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)
R46512	protein phosphatase 1, regulatory subunit 7

Table 6.5 (Continued)

T55835	histone deacetylase 6 (HDAC6)
AA644448	protein tyrosine phosphatase, receptor type, U (PTPRU)
R46823	N-acetylgalactosaminidase, alpha- (NAGA)
AA457700	stearoyl-CoA desaturase (delta-9-desaturase) (SCD)
T81764	cell division cycle 27 (CDC27)
T40568	phosphatidylinositol-4-phosphate 5-kinase, type II, alpha (PIP5K2A)
T71976	phosphatidic acid phosphatase type 2B (PPAP2B)
AA490466	Gap junction protein, beta 2, 26kDa (connexin 26)_(AA490466,_,Hs.323733)
T77729	Pyruvate carboxylase (PC), nuclear gene encoding mitochondrial protein
AA789328	cyclin-dependent kinase (CDC2-like) 10 (CDK10)
N93686	Aldehyde dehydrogenase 3 family, member B1 (ALDH3B1)
AA598513	protein tyrosine phosphatase, receptor type, F (PTPRF)
N71653	aspartoacylase (aminoacylase 2, Canavan disease) (ASPA)
AA520978	Ubiquitin-conjugating enzyme E2H (UBC8 homolog, yeast) (UBE2H)
N50655	solute carrier family 27 (fatty acid transporter), member 3 (SLC27A3)
AA129171	protein phosphatase 2, regulatory subunit B (B56), beta isoform (PPP2R5B)
AA443899	scavenger receptor class B, member 1 (SCARB1)
AA421269	phosphatidylinositol 4-kinase, catalytic, alpha polypeptide (PIK4CA)
T61256	ketohexokinase (fructokinase) (KHK)
AA487588	ATPase, H+ transporting, lysosomal interacting protein 1 (ATP6IP1)
N53169	apolipoprotein C-III (APOC3)
H21869	COX10 homolog
AA443577	tumor necrosis factor (ligand) superfamily, member 13 (TNFSF13)
H08753	guanine nucleotide binding protein (G protein)
R71691	TNF receptor-associated factor 1 (TRAF1)
R22219	(gF) phosphate cytidylyltransferase 2, ethanolamine (PCYT2)
W92859	protein tyrosine phosphatase
AA186686	prostaglandin E synthase 2 (PTGES2)
R06458	lecithin-cholesterol acyltransferase
R98442	UDP-glucose ceramide glucosyltransferase-like 1
AA434420	protein tyrosine phosphatase, non-receptor type 9 (PTPN9)
AA464957	Pleckstrin homology, Sec7 and coiled/coil domains 2 (cytohesin-2) (PSCD2)
N46830	Cytokine receptor-like factor 3 (CRLF3)
AA010079	protein kinase, interferon-inducible double stranded RNA dependent (PRKR)
N58305	hydroxysteroid (17-beta) dehydrogenase 7
AA279072	inositol polyphosphate phosphatase-like 1 (INPPL1)
AA425450	glycoprotein (transmembrane) nmb (GPNMB)
R95841	ribosomal protein S6 kinase, 90kDa
AA682469	alcohol dehydrogenase IB (class I)
AA022480	CREB binding protein (Rubinstein-Taybi syndrome) (CREBBP)
AA497051	sialyltransferase (STHM)
AA417279	protein tyrosine phosphatase, non-receptor type substrate 1 (PTPNS1)
R53787	protein phosphatase 2, regulatory subunit B (B56), epsilon isoform (PPP2R5E)
AA454819	mitogen-activated protein kinase 3
AA608857	serine/threonine protein kinase SSTK (SSTK)
N62379	I-kappa-B-interacting Ras-like protein 1
R88440	Fas apoptotic inhibitory molecule 2
AA416584	chemokine (C-C motif) ligand 3
H29475	Pyruvate dehydrogenase kinase, isoenzyme 2 (PDK2)

Table 6.5 (Continued)

	cytochrome c oxidase subunit VIII (COX8), nuclear gene encoding mitochondrial
AA862813	protein
AA443121	apoptosis-inducing protein D (APPD)
AA464590	protein tyrosine phosphatase, receptor type, N polypeptide 2 (PTPRN2)
AA236957	Rac/Cdc42 guanine nucleotide exchange factor (GEF) 6 (ARHGEF6)
R83038	low density lipoprotein receptor-related protein 5 (LRP5)
T61271	phospholipase A2, group IIA (platelets, synovial fluid) (PLA2G2A)
AA446839	BCL2/adenovirus E1B 19kDa interacting protein 3 (BNIP3)
N34876	TRAF-binding protein domain
R62612	fibronectin 1 (FN1)
AA676836	acid sphingomyelinase-like phosphodiesterase (ASM3A)
AA464421	zinc finger protein 144 (Mel-18) (ZNF144)
H75862	Apoptotic chromatin condensation inducer in the nucleus (ACINUS)
AA452802	solute carrier family 4, sodium bicarbonate cotransporter,
AA098980	protein kinase C-like 2 (PRKCL2)
AA485397	Ubiquitin-like 4
AA490501	UV radiation resistance associated gene (UVRAG)
N80129	metallothionein 1L (MT1L)
	fatty acid binding protein 3, muscle and heart (mammary-derived growth inhibitor)
AA148548	(FABP3)
N59764	guanine monphosphate synthetase (GMPS)
AA460646	peroxisomal biogenesis factor 6 (PEX6)
AA28062	diacylglycerol kinase, delta 130kDa (DGKD)
AA455102	heat shock 70kDa protein 2
W80489	acylphosphatase 1, erythrocyte (common) type (ACYP1)
AA644550	translocase of outer mitochondrial membrane 20 (yeast) homolog (KIAA0016)
H98694	PI-3-kinase-related kinase SMG-1 (SMG1)
AA451886	cytochrome P450, subfamily I (dioxin-inducible)
N48178	phosphoinositide-binding protein PIP3-E
AA405901	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 3, 9kDa (NDUFA3)
AA678095	G protein-coupled receptor 48 (GPR48)
AA464580	acetyl-Coenzyme A carboxylase alpha
AA001614	insulin receptor_
AA485347	protein phosphatase 1, regulatory (inhibitor) subunit 11 (PPP1R11)
H03436	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 3 (B3GNT3)
AA488373	Phosphoglucomutase 1 (PGM1)
AA156571	alanyl-tRNA synthetase (AARS)
AA459265	(gF) C/EBP-induced protein (LOC81558)
H92232	guanine nucleotide binding protein (G protein), alpha 11 (Gq class) (GNA11)
AA429946	peroxisomal short-chain alcohol dehydrogenase (humNRDR)
T62040	electron-transfer-flavoprotein, beta polypeptide (ETFB)
AA663439	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator)
H53340	metallothionein 1G (MT1G)
AA401111	glucose phosphate isomerase (GPI)
AA42934	Rho GTPase activating protein 1 (ARHGAP1)
T81033	zinc finger protein 148 (pHZ-52)

Table 6.6 Important genes selected for LDH alone

Access Num	Gene Name
AA626370	(gM) heme oxygenase (decycling) 2 (HMOX2)
AA862813	(gC) cytochrome c oxidase subunit VIII (COX8)
AA464590	(gC) protein tyrosine phosphatase, receptor type, N polypeptide 2 (PTPRN2)
AA453404	(gM) PPAR binding protein
AA400973	(g) lipocalin 2 (oncogene 24p3) (LCN2)
AA495981	(gC) Rho GTPase activating protein 6 (ARHGAP6)
T67006	(gF) glucokinase (hexokinase 4) regulatory protein (GCKR)
T81033	(gC) zinc finger protein 148 (pHZ-52)
AA430654	(gN) ATPase, H+ transporting, lysosomal V0 subunit a isoform 1 (ATP6V0A1)
T73556	(gM) fatty-acid-Coenzyme A ligase, long-chain 2 (FACL2)
AA485898	(gC) apoptosis related protein APR-3 (APR-3)
W92859	(gC) protein tyrosine phosphatase, non-receptor type 1
H75862	(gC) apoptotic chromatin condensation inducer in the nucleus (ACINUS)
H57850	(gC) protein phosphatase 2 (formerly 2A), regulatory subunit A (PR 65), beta isoform (PPP2R1B)
T47788	(gC) apoptosis-inducing factor (AIF)-homologous mitochondrion-associated inducer of death (AMID)
W85710	(gC) carnitine palmitoyltransferase I, muscle (CPT1B)
H11054	(gC) protein kinase C, delta
AA872420	(gN) collagen, type VIII, alpha 1 (COL8A1)
N72115	(gC) cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4) (CDKN2C)
R22219	(gF) phosphate cytidylyltransferase 2, ethanolamine (PCYT2)
AA644550	(gM) translocase of outer mitochondrial membrane 20 (yeast) homolog (KIAA0016)
H77542	(gC) hyperpolarization activated cyclic nucleotide-gated potassium channel 3
H92556	(gC) eukaryotic translation initiation factor 2B, subunit 2 beta, 39kDa (EIF2B2)
R31562	(gC) CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 1 (CDS1)
R50407	(gM) chemokine (C-X-C motif) ligand 2 (CXCL2)
AA676836	(gC) acid sphingomyelinase-like phosphodiesterase (ASM3A)
AA482070	(gC) Rho guanine exchange factor (GEF) 16 (ARHGEF16)
R95841	(gC) ribosomal protein S6 kinase, 90kDa, polypeptide 3
T64223	(gM) carboxypeptidase A3 (mast cell) (CPA3)
AA481464	(gC) peptidylprolyl isomerase B (cyclophilin B) (PPIB)
H09819	(gC) phosphomevalonate kinase (PMVK)
AA169176	(gC) glycerol-3-phosphate dehydrogenase 2 (mitochondrial)
AA411656	(gC) chemokine (C-X-C motif) ligand 16 (CXCL16)
R36587	(gM) phosphatidylinositol (4,5) bisphosphate 5-phosphatase homolog
H98694	(gC) PI-3-kinase-related kinase SMG-1 (SMG1)
AA160670	(gC) lysophosphatidic acid phosphatase (ACP6)
H95038	(gI) uncoupling protein 4
AA458563	(gN) ribulose-5-phosphate-3-epimerase (RPE)
AA776294	(gC) Rab geranylgeranyltransferase, alpha subunit (RABGGTA)
H66617	(gC) golgi apparatus protein 1 (GLG1)
AA451935	(gC) apoptosis inhibitor 5 (API5)
AA098980	(gC) protein kinase C-like 2 (PRKCL2)
N34876	(gC) TRAF-binding protein domain
AA281945	(gC) MAP-kinase activating death domain (MADD)

Table 6.6 (Continued)

AA454810	(g) tumor-associated calcium signal transducer 2 (TACSTD2)
T65790	(gC) farnesyl diphosphate synthase (farnesyl pyrophosphate synthetase
R53787	(gC) protein phosphatase 2, regulatory subunit B (B56), epsilon isoform (PPP2R5E)
AA022480	(gC) CREB binding protein (Rubinstein-Taybi syndrome) (CREBBP)
AA129171	(gC) protein phosphatase 2, regulatory subunit B (B56)
AA018907	(gC) protein phosphatase 3 (formerly 2B), regulatory subunit B, 19kDa
AA279072	(gK) inositol polyphosphate phosphatase-like 1 (INPPL1)
H17513	(gF) heat shock 70kDa protein 1-like_(H17513,_,Hs.80288)
AA463931	(gC) inositol 1,3,4-triphosphate 5/6 kinase (ITPK1)
AA600173	(gC) ubiquitin-conjugating enzyme E2A (RAD6 homolog) (UBE2A)
AA609421	(gC) membrane protein, palmitoylated 6 (MAGUK p55 subfamily member 6) (MPP6)
R94191	(gC) translocase of outer mitochondrial membrane 70 homolog A (yeast) (TOMM70A)
AA459265	(gF) C/EBP-induced protein (LOC81558), mRNA. (AA459265,NM_030802,Hs.9851)
AA431988	(gC) fatty acid amide hydrolase (FAAH), mRNA. (AA431988,NM_001441,Hs.288828)
W65461	(gC) dual specificity phosphatase 5 (DUSP5), mRNA. (W65461,NM_004419,Hs.2128)
AA437389	(g) lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)_(AA437389,_,Hs.93199)
AA487586	(gC) inorganic pyrophosphatase (SID6-306), mRNA. (AA487586,NM_006903,Hs.5123)
AA434420	(gC) protein tyrosine phosphatase, non-receptor type 9 (PTPN9)
AA487588	(gC) ATPase, H+ transporting, lysosomal interacting protein 1 (ATP6IP1)
AA464580	(gC) acetyl-Coenzyme A carboxylase alpha_(AA464580,_,Hs.7232)
N71497	(gC) low density lipoprotein receptor-related protein 6 (LRP6)
R98442	(gC) UDP-glucose ceramide glucosyltransferase-like 1_(R98442,_,Hs.105794)
H04769	(gC) protein kinase (cAMP-dependent, catalytic) inhibitor beta_(H04769,_,Hs.106106)
AA490494	(gC) tumor necrosis factor receptor superfamily, member 21 (TNFRSF21)
T61256	(gM) ketohexokinase (fructokinase) (KHK)
AA486407	(gN) phosphodiesterase 4D interacting protein (myomegalin)
T55835	(gC) histone deacetylase 6 (HDAC6)
R39463	(gC) aldolase C, fructose-bisphosphate (ALDOC)
AA490501	(gC) UV radiation resistance associated gene (UVRAG)
AA451886	(gC) cytochrome P450, subfamily I (dioxin-inducible)
AA280677	(gC) zinc finger protein 258 (ZNF258)
AA457700	(gC) stearoyl-CoA desaturase (delta-9-desaturase) (SCD)
R83038	(gC) low density lipoprotein receptor-related protein 5 (LRP5)
N62847	(g) lysosomal-associated membrane protein 2 (LAMP2)
T73468	(gC) glutathione S-transferase A2 (GSTA2)
AA443577	(gC) tumor necrosis factor (ligand) superfamily, member 13 (TNFSF13)
AA682469	(gC) alcohol dehydrogenase IB (class I), beta polypeptide
T77729	(gC) pyruvate carboxylase (PC)
AA644448	(gC) protein tyrosine phosphatase, receptor type
W20487	(gC) BCL2-related protein A1 (BCL2A1)
N91135	(gN) chloride intracellular channel 3 (CLIC3)

Table 6.6 (Continued)

AA425450	(gC) glycoprotein (transmembrane) nmb (GPNMB)
R71691	(gC) TNF receptor-associated factor 1 (TRAF1)
AA186686	(gC) prostaglandin E synthase 2 (PTGES2)
AA454588	(gC) BCL2-like 2 (BCL2L2)
T71976	(gC) phosphatidic acid phosphatase type 2B (PPAP2B)
N46830	(gC) cytokine receptor-like factor 3 (CRLF3)
T81764	(gC) cell division cycle 27 (CDC27)
AA434130	(gC) thioredoxin reductase 2 (TXNRD2)
AA520978	(gC) ubiquitin-conjugating enzyme E2H (UBC8 homolog, yeast)
AA010079	(gC) protein kinase, interferon-inducible double stranded RNA dependent (PRKR)
AA446839	(gC) BCL2/adenovirus E1B 19kDa interacting protein 3 (BNIP3)
AA418907	(gC) cytochrome P450, subfamily I (aromatic compound-inducible), polypeptide 1 (CYP1A1)
R06605	(gC) protein tyrosine phosphatase, non-receptor type 1 (PTPN1)
N62379	(gC) I-kappa-B-interacting Ras-like protein 1
R46823	(g) N-acetylgalactosaminidase, alpha- (NAGA)

Comparing Tables 6.5 with 6.6, we can see that many important genes such as SCD, ACC, TNF super family, TRAF1, Bcl-2 were selected in both cases. However, the two factor case selected more lipid metabolism related genes, e.g. acyl-CoA dehydrogenase and diacylglycerol kinase. In addition, in the two factors case selected NADH dehydrogenase, which is involved in complex I of the electron transport chain. It is a major site of ROS production and has been shown to play an important role in the palmitate-induced toxicity [Srivastava 2006, Li 2006]. Therefore, more useful information was extracted with multiple factors. The next paragraph describes how the *TIPS*© framework would be extended to incorporate more than two metabolites.

To incorporate more than two metabolites or cellular processes such as the entire metabolome profile into the *TIPS*© framework, latent variables can be extracted from the metabolome data with methods such as PCA, PLS [Griffin 2004], and FDA. L1 and L2 in Equation 6.2 would be substituted with the extracted latent variables that represent the metabolome data. PCA and PLS extracts latent variables that account for most of the

variance in the data while FDA extracts latent variables that have the largest discriminating capability. To incorporate nonlinearity effects, kernels may be applied to the aforementioned methods. The number of latent variables may be determined based upon the variance captured by the latent variables. For example, to capture 90% of the variance in the raw data, we included enough latent variables to extract 90% (or more) of the variance. In the network reconstruction stage by BN, the latent variables (rather than simply LDH and TG) may be used as pseudo nodes to represent pathways, the biological meaning of the pathways are then determined from the genes in the latent variables [Griffin 2004]. An alternative method is to select the metabolites with high weights to be the latent variables that would then be used for the final network reconstruction in the BN step. This method is not limited to two factors as illustrated in equation 6.2, but may be extended to multiple factors, L1, L2, ..., Ln.

6.4 Using a Dynamic Bayesian Network (DBN) to identify active pathways from the dynamic profiles

6.4.1 Introduction

The current *TIPS*© model handles static data, i.e. data at a single time point. However, cells continuously reprogram gene regulatory networks when they sense the changes in the environmental conditions. In order to understand how cells are regulated in response to environmental changes, time series data (dynamic data) are required. In our system, gene expression data as well as metabolites over three consecutive days (24, 48, 72 hours after FFA and TNFα exposure) have been profiled. It has been observed that cytotoxicity was regulated by FFA and TNFα differently as a function of time [Srivastava 2006]. For example, TNFα increased cytoxicity on day 1 and had no effects on day 2. However it

decreased cytotoxicity on day 3. This suggests that the gene regulatory network changes over time. Uncovering this information would be valuable in helping to elucidate the mechanism of toxicity and cell death. Models, such as DBN, can be applied to model time series data. To apply DBN to identify the pathways that regulate cytotoxicity in HepG2 cells, we first applied a hierarchical framework introduced in section 6.2 to select the genes relevant to cytotoxicity.

6.4.2 Materials and Methods

Dynamic Bayesian network analysis

We applied Banjo (http://www.cs.duke.edu/~amink/software/banjo/) from Alexander Hartemink's group to learn the DBN. Every node in the Bayesian network represents a gene, a metabolic flux or a phenotypic response. A first order Markov DBN was used, i.e. a node at a given time point can be influenced by itself or its parents in the immediate previous time point. The search and score approach was used to find a network structure with the maximum score. Bayesian Dirichlet equivalence (BDe) was used to score a network. Greedy search algorithm was used to search for the optimal network.

6.4.3 Results and Discussion

Network Reconstruction

Table 6.7: Experimental conditions for DBN learning

TNF-(*ng/ml*) / FFA	0	20	100
HepG2 medium (H)	H-0	H-20	H-100
BSA(B)	B0	B20	B100
Palmitate(P)	P0	P20	P100
Oleate(O)	O0	O20	O100
Linoleate(L)	L0	L20	L100

Based upon the hierarchial approach in section 6.2, we selected 80 genes, 3 metabolites (BOH, ACAC, TG) and one phenotype (LDH) for DBN reconstruction. The genes, metabolites and phenotypic profiles are listed in Table 6.4. The network was learned for each treatment, i.e. HepG2 medium (H, control), Bovine Serum Albumin (B), Palmitate (P), Oleate (O), Linoleate (L) and their combinations with TNF-α, separately, based upon the dynamic profile of the genes measured at 24, 48 and 72 hours after the treatments. As summarized in Table 6.7, we learned the network for 15 conditions with two replicates for each condition. The experiments were performed by Shireesh Srivastava and the gene expression data were obtained at the Van Andel Institute. All connections that appeared more than twice were combined together for network reconstruction. The network is shown in Figure 6.3, with the numbers in the network corresponding to the Nums in Table 6.4.

From the network shown in Figure 6.4 we can see that the phenotype is mainly regulated by two categories of factors, oxidative stress related factors and TNF-α signal pathway related factors. The oxidative stress related factors include NADH

dehydrogenase, aldehyde dehydrogenase (ALDH), glutathione-S transferase (GST), copper chaperone for superoxide dismutase (CCS) and CYP2E1. TNF-α related factors include NF-κβ, IKB, TRAF, TNF superfamily, capase 8, inhibitor of apoptosis protein (IAP). These two categories interact with each other in the network. LDH was found connected directly to genes such as copper chaperone for superoxide dismutase (CCS) and biophosphoglycerate mutase (BPGM). CCS was connected to ALDH1A2, NADH dehydrogenase, CYP2E1, and GSTM5. All these factors are relevant to cellular ROS and lipid peroxidation and thus play a role in the palmitate-induced cytotoxicity. Excessive accumulation of ROS leads to oxidative damage, e.g. peroxidizing membrane to produce deleterious aldehydes such as malonedialdehyde and 4-hydroxy-2,3-(E)-nonenal (4-HNE). ROS are produced during the electron transfer and oxidative phosphorylation process. NADH dehydrogenase complex, or complex I of the electron transport chain, is a main source of superoxide production [Moller 2001]. Superoxide dismutase (SOD) protects cells from free radical damage by reducing superoxide anion radical into hydrogen peroxide. 4-HNE has been found to elicit cytotoxic effects in diverse cells [Hartley 1995]. The detoxification pathways of 4-HNE include oxidation of its carbonyl group into carboxylic acid by ALDH or conjugation with glutathione by GSTs. Cytochrome P-450 e.g. CYP2E1 sequentially transfers two electrons from NADPH to molecular oxygen. In the course of electron transfer some of the activated oxygen is released as superoxides or H_2O_2 [Moller 2001]. Excessive accumulation of ROS leads to oxidative damage. Therefore, the connections among CCS, ALDH, NADH dehydrogenase, GST, CYP2E1 suggest the involvement of ROS in the observed cytotoxicity and possible detoxification mechanisms. BPGM is an enzyme involved in

glycolysis to convert 3-phospho-D-glyceroyl phosphate to 2,3-bisphospho-D-glycerate and it is connected to two apoptosis inhibitor genes of IAP2 and IAP3. The connections identified between LDH and IAP and BPGM suggested the involvement of glycolysis and apoptosis in the palmitate induced cytotoxicity. High glucose level has been reported to synergize with saturated fatty acids to cause increased cell death through apoptosis in pancreatic beta cells because elevated glucose level has inhibitory effects on the lipid detoxification via beta-oxidation [El-Assaad 2003]. In addition, high glucose metabolism may increase superoxide production, which also induces apoptosis [Okuyama 2003]. Indeed the glucose level in the HepG2 culture medium is very high. Apoptosis was initiated in the palmitate cultures as was evident by a significantly higher caspase activity detected in these cultures as compared to the control and unsaturated fatty acid cultures [Srivastava 2006].

The activation of NF-κβ by the TNF-α signaling pathway was uncovered by DBN. TNFRSF1 connects to TRAF3, which then connects to IKB. IKB is connected to NF- κβ through GSTM4. TNF-α is a cytokine that can be detected by the TNF receptor which activates TRAF and IKK. Activation of IKK will phosphorate and degrade IKB to release NF-κβ. It is encouraging that the DBN model was able to reconstruct this signaling pathway. The model also indicated that the NF-κβ activation may be mediated by genes such as GST and ALDH.

Discriminant and correlation analyses found the metabolic flux, BOH, to be the most relevant to the palmitate-induced cytotoxicity. BOH was found to be connected to NADH dehydrogenase and SOD. BOH are produced from acetyl-CoA, which is a metabolic intermediate of fatty acid oxidation. Fatty acid oxidation induces ROS

production [Gill 2006, Sanya 2003]. NADH dehydrogenase complex, or complex I of the electron transport chain, is a main site of superoxide production. Thus, the connection between NADH dehydrogenase, SOD and BOH indicates the involvement of ROS production in the palmitate-induced cytotoxicity. Indeed, experimentally inhibiting NADH dehydrogenase in the palmitate cultures significantly reduced the LDH release [Srivastava 2006, Li 2006].

Therefore based upon the important genes identified by the hierarchical approach (described in section 6.2), DBN was able to reconstruct a regulatory network from the dynamic gene expression and metabolic profiles. The resulting network suggested possible interactions between the identified genes and metabolic fluxes, which indicated the involvement of some pathways such as apoptosis, NF-κβ and ROS in the palmitate induced cytotoxicity. The inferred network structure may be used to construct an *in silico* regulatory network to simulate how BOH or LDH release is dynamically regulated. Thus, potential hypothesis and mechanisms involved in inducing cytoxicity may be identified based upon the reconstructed network.

6.5 Future work

We discussed in this chapter the improvements to *TIPS*© including hierarchically integrating phenotypic metabolic, and gene expression profiles with genomic function information, inferring pathways that confer a phenotype using multiples metabolites and dynamic modeling. Future work would be to integrate all these features into a dynamic *TIPS*© framework. Briefly, metabolites important to a phenotype can be identified with the hierarchical approach using FDA. This method would be able to identify multiple metabolites that may be important in characterizing the phenotype. GSEA can be used to

define the functional gene groups and identify significantly enriched ones, which can be integrated with the multiple metabolites, identified in FDA, by GA/MB-PLS to select important functional gene sets for the several metabolites that are most relevant to a phenotype. Finally, important gene sets may be subjected to DBN analysis to learn the regulatory network and reconstruct the pathways from time series data.

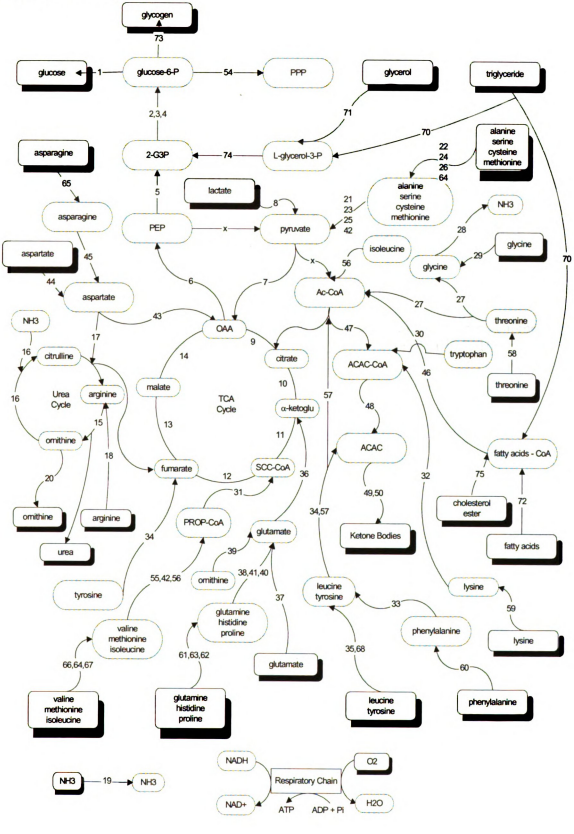
Figure 6.3 DBN reconstrunction of the network. The nubers in the nodes correspond to the nums in Table 6.4.

CHAPTER 7 CONCLUSIONS

It is the objective of this research to develop approaches that can identify the pathways involved in conferring a phenotype or cellular state. Once these pathways are elucidated, targets may be identified and engineered to more accurately modulate or optimize the cellular state. Microarray gene expression and metabolic flux data were obtained to characterize cellular states. In addition to the measurements, *a priori* knowledge about the genes, metabolites and pathways available in literature and public databases were also used. Systems approaches were developed in this thesis to integrate these multiple sources of data to identify the genes and active pathways relevant to a particular phenotype. *TIPS*© was developed in chapter 2 to chapter 4 to identify the active pathways that regulate a phenotype. *TIPS*© applied GA/PLS and CICA to identify a subset of relevant genes, and this subset of genes was subsequently subjected to BN analysis to reconstruct active pathways. *TIPS*© was successfully applied to identify active pathways relevant to palmitate and TNF-α induced cytotoxicity in HepG2 cells. A dynamic SSM was developed in chapter 5 to integrate gene expression and *a priori* knowledge of the gene regulatory network structure to infer underlying transcription factor activity profiles. SSM was successfully applied to infer transcription factor activities of E. Coli and yeast, which were validated by the literature. Extensions to *TIPS*© were discussed in chapter 6 to incorporate functional information in the gene selection process, infer pathways relevant to multiple metabolites, and from time series data using DBN analysis. In summary, systems approaches were developed to integrate gene expression, metabolic and phenotypic profiles, along with *a priori* information, to elucidate mechanisms and

pathways involved in conferring a particular phenotype. As a proof of concept, the approaches developed were applied to HepG2 cells exposed to FFAs and TNF-α to identify genes and pathways that confer a cytotoxic vs. a cytoprotective phenotype.

Appendix Figure 1



BIBLIOGRAPHY

Abreo K., Sella M., Alvarez-Hernandez X., Jain S. (2004) "Antioxidants prevent aluminum-induced toxicity in cultured hepatocytes" J Inorg Biochem. 98(6):1129-34.

Aerne B.L., Johnson A.L., Toyn J.H., Johnston L.H. (1998). "Swi5 Controls a Novel Wave of Cyclin Synthesis in Late Mitosis" Mol. Biol. Cell **9**, 945-956.

Akutsu T., Miyano S., Kuhara S. (1999) "Identification of Genetic Networks from a Small Number of Gene Expression Patterns Under the Boolean Network Model" Pacific Symposium on Biocomputing, 4:17-28.

Al-Shahrour F., Minguez P., Vaquerizas J.M., Conde L., Dopazo J.(2005) "Babelomics: a suite of web-tools for functional annotation and analysis of group of genes in high-throughput experiments" Nucleic Acids Research, 33(Web Server issue), W460-W464.

Alter O., Brown P.O., Bostein D. (2000) "Singular value decomposition for genome-wide expression data processing and modeling" PNAS, **97**(18):10101-10106.

Andersson C. A. and Bro R. (2000) "The N-way Toolbox for MATLAB", Chemometrics & Intelligent Laboratory Systems. 52 (1):1-4

Basso K., Margolin A.A., Stolovitzky G., Klein U., Dalla-Favera R., Califano A.(2005) "Reverse engineering of regulatory networks in human B cells" Nat Genet. 37(4):382-90.

Bangalore A.S., Shaffer R. E., Small G.W. (1996) "Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy." Anal. Chem., **68**: 4200-4212.

Baetz K., Andrews B. (1999). "Regulation of Cell Cycle Transcription Factor Swi4 through Auto-Inhibition of DNA Binding." Mol. Cell Biol. **19**, 6729-6741.

Bar-Joseph Z., Gerber G., Gifford D., Jaakkola T., Simon I. (2002). "A New Approach to Analyzing Gene Expression Time Series Data." 6th Annual International Conference on Research in Computational Molecular Biology. pp. 39-48

Beal M.J., Falciani F., Ghahramani Z., Rangel C., Wild, D.L., (2005) "A Bayesian approach to reconstructing genetic regulatory networks with hidden factors." Bioinformatics **21**(3): 349-356.

Bligh E.G., Dyer W.J., (1959) "A rapid method of total lipid extraction and purification", Can J Biochem Physiol. 37(8):911-7.

Bolon C., Gauthier C., Simonnet H. (1997) "Glycolysis inhibition by palmitate in renal cells cultured in a two-chamber system." Am J Physiol. 273(5 Pt 1):C1732-8.

Brenner D. A. (1998)"Signal transduction during liver regeneration", J Gastroenterol. Hepatol. 13 Suppl: S93-95.

Brookes P. S. (2005)"Mitochondrial H(+) leak and ROS generation: an odd couple." Free Radic Biol Med. 38(1):12-23.

Castillo E. F., Gutiérrez J. M., and Hadi A. S. (1997) "Sensitivity analysis in discrete Bayesian networks", IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 27:412-423.

Chan C., Berthiaume F., Washuzu J., Toner M., and Yarmush M. L. (2002) "Metabolic Pre-Conditioning of Cultured Cells in Physiological Levels of Insulin: Generating Resistance to the Lipid Accumulating Effects of Plasma in Hepatocytes" Biotechnology and Bioengineering, **78**: 753-760.

Chan C., Berthiaume F., Lee K., and Yarmush M. L. (2003a) "Metabolic Flux Analysis of Cultured Hepatocytes Exposed to Plasma", Biotechnology and Bioengineering, 81: 33-49.

Chan C., Berthiaume F., Lee K., and Yarmush M. L. (2003b) "Effect of Hormones on Liver-specific Function and Lipid Accumulation of Hepatocytes during Plasma Exposure Analyzed by Metabolic Flux Analysis", Metabolic Engineering, 5: 1-15.

Chan C., Hwang D. H., Stephanopoulos G. N., Yarmush M. L., and Stephanopoulos G. (2003c) "Application of Multivariate Analysis to Optimize Function of Cultured Hepatocytes", Biotechnology Progress, 19: 580-598.

Chen J.L., Peacock E., Samady W., Turner S.M., Neese R.A., Hellerstein M.K., Murphy E.J.(2005) "Physiologic and pharmacologic factors influencing glyceroneogenic contribution to triacylglyceride glycerol measured by mass isotopomer distribution analysis", J Biol Chem. 280(27):25396-402.

Chen T., He H. L., Church G.M. (1999) "Modeling Gene Expression with Differential Equations", Pacific Symposium on Biocomputing, 4:29-40

Cheng J., Kelly J., Bell D.A. and Liu W. (2002) "Learning belief networks from data: An information theory based approach", Artificial Intelligence Journal, 137: 43-90.

Ciccoli L., Ferrali M., Casini A. F., Comporti M. (1981) "Effect of reduced glutathione on carbon tetrachloride induced fatty liver: various considerations" Boll Soc Ital Biol Sper. 57(13):1463-9.

Cooper G.F. (1990) "The computational complexity of probabilistic inference using Bayesian belief networks". Artificial Intelligence, 42(2-3):393-405

Cooper G. and Herskovits E. (1992) "A Bayesian method for the induction of probabilistic networks from data", Machine Learning, **9**:309-347.

de Pablo M.A. (1999) "Palmitate induces apoptosis via a direct effect on mitochondria." Apoptosis **4**, 81 (1999).

Deng X., Ito T., Carr B., Mumby M., May Jr, W.S. (1998) "Reversible Phosphorylation of Bcl2 following Interleukin 3 or Bryostatin 1 Is Mediated by Direct Interaction with Protein Phosphatase 2A". J. Biol. Chem. **273**, 34157-34163.

Dentin R., Benhamed F., Pegorier J. P., Foufelle F., Viollet B., Vaulont S., Girard J., Postic C. (2005) "Polyunsaturated fatty acids suppress glycolytic and lipogenic genes through the inhibition of ChREBP nuclear protein translocation." J Clin Invest. 115(10):2843-54.

di Bernardo D., Thompson M. J., Gardner T. S., Chobot S. E., Eastwood E. L., Wojtovich A. P., Elliott S. J., Schaus S. E., and Collins J.J. (2005) "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks." Nature Biotechnology **23**(3):377-383.

Di Giovanni S., Mirabella M., Papacci M., Odoardi F., Silvestri G., Servidei S. (2001) "Apoptosis and ROS detoxification enzymes correlate with cytochrome c oxidase deficiency in mitochondrial encephalomyopathies", Mol Cell Neurosci. 17(4):696-705.

Ding W. X., Yin X. M.(2004) "Dissection of the multiple mechanisms of TNF-alpha-induced apoptosis in liver injury", J Cell Mol Med, 8(4): 445-54.

Dobrzyn P., Dobrzyn A., Miyazaki M., Cohen P., Asilmaz E., Hardie D.G., Friedman J. M., Ntambi J. M. (2004) "Stearoyl-CoA desaturase 1 deficiency increases fatty acid oxidation by activating AMP-activated protein kinase in liver." PNAS, 101(17):6409-6414.

Eisen M.B., Spellman P.T., Brown P.O. and Botstein D. (1998) "Cluster analysis and display of genome-wide expression patterns." Proc. Natl. Acad. Sci. USA, **95**(25): 14863-14868.

El-Assaad W., Buteau J., Peyot M. L., Nolan C., Roduit R., Hardy S., Joly E., Dbaibo G., Rosenberg L., Prentki M. (2003) "Saturated fatty acids synergize with elevated glucose to cause pancreatic beta-cell death." Endocrinology. 144(9):4154-63.

Farmer W.R., and Liao J.C. (2000) "Improving lycopene production in Escherichia coli by engineering metabolic control", Nature biotechnology, 18:533:537.

Felber J. P. and Golay A. (2002) "Pathways from obesity to diabetes." International Journal of Obesity **26**(Suppl. 2): S39-S45.

Fellenberg K., Hauser N.C., Brors B., Neutzner A., Hoheisel J.D., Vingron M. (2001) "Correspondence analysis applied to microarray data." Proc. Natl. Acad. Sci. USA, **98**(19):10781-10786.

Fleury C. (1997) "Uncoupling protein-2: a novel gene linked to obesity and hyperinsulinemia." Nature Genetics **15**, 269-272.

Friedman N., Linial M., Nachman I., and Pe' er D. (2000) "Using Bayesian networks to analyze expression data" J. Comp. Bio., 7:601-620.

Friedman N. (2004) "Inferring Cellular Networks Using Probabilistic Graphical Models." Science, 303(5659):799-805.

Geladi P., and Kowalski B. P. (1986) "Partial least-squares regression: a tutorial." Anal. Chim. Acta., 185, 1.

Gill H., Wu G. (2006) "Non-alcoholic fatty liver disease and the metabolic syndrome: Effects of weight loss and a review of popular diets. Are low carbohydrate diets the answer?" World J Gastroenterol 12(3):345-353

Goldberg D.E, Deb K. (1989) "Genetic algorithms in search, optimization, and machine learning" Addison-Wesley, Reading, MA.

Gomez E.O., Mendoza-Milla C., Ibarra-Sanchez M.J., Ventura-Gallegos J.L. (1996). "Ceramide reproduces late appearance of oxidative stress during TNF-mediated cell death in L929 cells", Biochem Biophys Res Commun. 228(2): 505-9.

Griffin J.L. (2004) "An integrated reverse functional genomic and metabolic approach to understanding orotic acid-induced fatty liver." Physiol. Genomics 17, 140-149.

Haddad J. (2002) "Oxygen-sensing mechanisms and the regulation of redox-responsive transcription factors in development and pathophysiology" Respiratory Research 3, 26.

Haddad J.J. (2004) "On the antioxidant mechanisms of Bcl-2: a retrospective of NF-[kappa]B signaling and oxidative stress." Biochemical and Biophysical Research Communications 322, 355.

Hakamada K., Hanai T., Honda H., and Kobayashi T., (2001) "Identifying genetic network using experimental time series data by Boolean algorithm", Genome Informatics, 12:272-273.

Henrion M. (1988) "Propagating uncertainty in Bayesian networks by probabilistic logic sampling". In Uncertainty in Artificial Intellgience 2, pages 149-163, New York, N. Y., Elsevier Science Publishing Company, Inc.

Hartley D. P., Ruth J. A., Petersen D. R. (1995) "The hepatocellular metabolism of 4-hydroxynonenal by alcohol dehydrogenase, aldehyde dehydrogenase, and glutathione S-transferase"Arch Biochem Biophys. 316(1):197-205.

Hartemink A.J., Gifford D. K., Jaakkola T. S., and Young R.A. (2001) "Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks" Pacific Symposium on Biocomputing, 6:422-433.

Hayes J. D., Pulford D. J. (1995) "The glutathione S-transferase supergene family: regulation of GST and the contribution of the isoenzymes to cancer chemoprotection and drug resistance." Crit Rev Biochem Mol Biol. 30(6):445-600.

Heckerman D., Geiger D. and Chickering D.M. (1995) "Learning Bayesian networks: the combination of knowledge and statistical data" Machine Learning Journal, 20:197-243.

Heyninck K., Wullaert A., Beyaert R. (2003) "Nuclear factor-kappa B plays a central role in tumour necrosis factor-mediated liver disease", Biochem. Pharmacol. 66(8):1409-15.

Holland J. (1975) "Adaptation in natural and artificial system." The university of Michigan press, Ann Arbor.

Holland M. (2002) "Transcript abundance in yeast varies over six orders of magnitude." J. Biol. Chem. **277**, 14363-14366.

Houck C.R., Joines J.A., Kay M.G. (1995), "A genetic algorithm for function optimization: a Matlab implementation," NCSU-IETR 95-09.

Hwang D. H., Stephanopoulos G., Chan C. (2004) "Inverse modeling using multi-block PLS to determine the environmental conditions that provide optimal cellular function," Bioinformatics 20, 487-499.

Ideker T., Ozier O., Schwikowski B., Siegel Andrew F. (2002) "Discovering regulatory and signalling circuits in molecular interaction networks." Bioinformatics 18 Suppl 1:S233-240.

Idris I., Gray S., Donnelly R. (2001) "Protein kinase C activation: isozyme-specific effects on metabolism and cardiovascular complications in diabetes" Diabetologia, 44: 659-673.

Ikuko M., Hiroto M., Yuki K., Tetsuya K., Sachiko S., Shinya T. (2003) "Acetaldehyde Induces Granulocyte Macrophage Colony-Stimulating Factor Production in Human Bronchi through Activation of Nuclear Factor-κB" Allergy and Asthma Proceedings, 24 (5): 367-371

Izpisua J.C., Barber T., Cabo J., Hrelia S., Rossi C.A., Parenti C. G., Lerker G., Biagi P.L., Bordoni A., Lenaz G. (1989) "Lipid composition, fluidity and enzymatic activities of rat liver plasma and mitochondiral membranes in dietary obese rats." Int J Obes., 13(4): 531-542.

Janicke R., Droge W. (1985) "Effect of L-ornithine on proliferative and cytotoxic T-cell responses in allogeneic and syngeneic mixed leukocyte cultures" Cell Immunol. 92(2):359-65.

Jin R., Si L., Srivastava S., Li Z., Chan, C. (2006) "A Knowledge Driven Regression Model for Gene Expression and Microarray Analysis", EMBC, accepted.

Jetten A. M., Suter U. (2000) "The peripheral myelin protein 22 and epithelial membrane protein family", Prog Nucleic Acid Res Mol Biol. 64:97-129.

Ji J., Zhang L., Wang P., Mu Y. M. (2006) "Saturated free fatty acid, palmitic acid, induces apoptosis in fetal hepatocytes in culture", Exp Toxicol Pathol. 56(6):369-76.

Jia Z., Xu S. (2005) "Clustering expressed genes on the basis of their association with a quantitative phenotype." Genet Res. 86(3):193-207.

Jump D.B. (2004) "Fatty acid regulation of gene transcription", Crit Rev Clin Lab Sci. 41(1):41-78.

Kao K.C., Yang Y.L., Boscolo R., Sabatti C., Roychowdhury V.P., Liao J.C., (2004) "Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis." PNAS, **101**(2): 641-646.

Kao K.C., Tran L.M., Liao J.C., (2005) "A global regulatory role of gluconeogenic genes in Escherichia coli revealed by transcriptome network analysis," The Journal of Biological Chemistry, Vol 280 (43):36079-36087

Kanthasamy A.G., Kitazawa M., Kanthasamy A., Anantharam V. (2003) "Role of Proteolytic Activation of Protein Kinase C in Oxidative Stress-Induced Apoptosis", Antioxid Redox Signal. 5(5):609-620.

Kerszberg M. (2004) "Noise, delays, robustness, canalization and all that" Current Opinion in Genetics and Development 14: 440-445

Kerner J., Hoppel C. (2000) "Fatty acid import into mitochondria." Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids **1486**, 1).

Kilpatrick L. E., Lee J. Y., Haines K.M., Campbell D.E., Sullivan K.E., Korchak H.M. (2002) "A role for PKC-delta and PI 3-kinase in TNF-alpha -mediated antiapoptotic signaling in the human neutrophil." Am J Physiol Cell Physiol 283(1):C48-57.

Kim B.C. (2002) "Tumor Necrosis Factor Induces Apoptosis in Hepatoma Cells by Increasing Ca2+ Release from the Endoplasmic Reticulum and Suppressing Bcl-2 Expression." J. Biol. Chem. **277**, 31381-31389.

Kim J. S. He L., and Lemasters J. J. (2003) "Mitochondrial permeability transition: a common pathway to necrosis and apoptosis", Biochem Biophys Res Commun 304(3):463-70.

Kobayashi M. (1998) "Molecular mechanism of insulin resistance." Saishin Igaku 53(6): 1210-1216.

Kovacech B., Nasmyth K., Schuster T. (1996). "EGT2 Gene Transcription is Induced Predominantly by Swi5 in Early G1." Mol. Cell. Biol. **16**, 3264–3274.

Leardi R. (2000) "Application of genetic algorithm-PLS for feature selection in spectral data sets." J. Chemometrics, 14: 643-655.

Lee T. I., Rinaldi N. J., Robert F., Odom D. T., Bar-Joseph Z., Gerber G. K., Hannett N. M., Harbison C. T., Thompson C. M., Simon I., Zeitlinger J., Jennings E. G., Murray H. L., Gordon D. B., Ren B., Wyrick J. J., Tagne J.-B., Volkert T. L., Fraenkel E., Gifford D. K., and Young R. A. (2002) "Transcriptional Regulatory Networks in Saccharomyces cerevisiae". Science, **298**(5594): 799-804.

Lewandowski E. D., Kudej R. K., White L. T., O'Donnell J. M., Vatner S. F. (2002) "Mitochondrial preference for short chain fatty acid oxidation during coronary artery constriction." Circulation. 105(3):367-72.

Liang S., Fuhrman S. and Somogyi R. (1998) "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures." Pacific Symposium on Biocomputing, **3**:18 -29.

Liao, J.C., Boscolo, R., Yang, Y.L., Tran, L.M., Sabatti, C., Roychowdhury, V.P., (2003), "Network component analysis: Reconstruction of regulatory signals in biological systems." PNAS, **100**(26):15522-15527.

Liebermeister W. (2002) "Linear modes of gene expression determined by independent component analysis." Bioinformatics **18**, 51-60 (2002).

Listenberger L. L., Han X., Lewis S. E., Cases S., Farese R.V., Jr. Ory D. S., Schaffer J.E. (2003) "Triglyceride accumulation protects against fatty acid-induced lipotoxicity." Proceedings of the National Academy of Sciences of the United States of America 100(6):3077-3082.

Listenberger L.L., Ory D.S., Schaffer J.E. (2001) "Palmitate-induced apoptosis can occur through a ceramide-independent pathway." Journal of Biological Chemistry 276, 14890-14895.

Li Z., Chan C. (2004a) "Integrating Gene Expression and Metabolic Profiles." J Biol Chem 279(26):27124-27137.

Li Z., Chan C. (2004b) "Inferring pathways and networks with a Bayesian framework" FASEB J 18(6):746-748.

Li Z., Srivastava S., Mittal S., Norton P., Resau J., Haab B., Chan C. (2006) "A Hierarchical Approach to Identify Pathways that Confer Cytotoxicity in HepG2 Cells from Metabolic and Gene Expression Profiles", in review.

Lin S. M., Liao X., McConnell P., Vata K., Carin L., Goldschmidt P. (2001) "Using functional genomic units to corroborate user experiments with the Rosetta compendium." *Methods of Microarray Data Analysis II, Papers from CAMDA '01, Durham, NC, United States, Oct 15-16, 2001* 2002:123-137.

Loew L.M., Schaff J.C. (2001) "The virtual cell: a software environment for computational cell biology." Trends in Biotechnology, **19**(10): 401-406.

Lopes J.A., Menezes J.C., Westerhuis J.A., Smilde A.K.. (2002) "Multiblock PLS analysis of an industrial pharmaceutical process" Biotechnol Bioeng. 80(4):419-27

Lu Z.H., Mu Y.M., Wang B.A., Li X.L.(2003) "Saturated free fatty acids, palmitic acid and stearic acid, induce apoptosis by stimulation of ceramide generation in rat testicular Leydig cell", Biochemical and Biophysical Research Communications 303(4): 1002-1007.

MacKay V.L., Mai B., Waters L., Breeden L.L. (2001). "Early Cell Cycle Box-Mediated Transcription of CLN3 and SWI4 Contributes to the Proper Timing of the G1-to-S Transition in Budding Yeast." Mol. Cell Biol. **21**, 4140-4148.

McAdams H. and Arkin A. (1997). "Stochastic mechanisms in gene expression." PNAS **94**(3): 814-819.

McInerny C.J., Partridge J.F., Mikesell G.E., Greemer D.P., Breeden L.L. (1997). "A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G1-specific transcription." Genes & Dev. **11**, 1277–1288.

Monk N. A.M. (2003) "Oscillatory expression of Hes1, p53 and NF-kB driven by transcriptional time delays," Current Biology 13: 1409-1413

Murphy K. and Mian S. (1999). "Modelling gene expression data using Dynamic Bayesian Networks." Technical report, University of California, Berkeley.

Murphy K. (2001) "The Bayes net toolbox for matlab." Computing Science and Statistics: Proceedings of Interface. 33

MacGregor J. F., Jaeckle C., Kiparissides C., Koutoudi, M., (1994) "Process monitoring and diagnosis by multiblock PLS methods" AIChE J., 40, 826-838.

Maki Y., Tominaga D., Okamoto M., Watanabe S., Eguchi Y. (2001) "Development of a System for the Inference of Large Scale Genetic Networks" Pacific Symposium on Biocomputing, 6:446-458.

Magana M.M., Koo S.H., Towle H.C., Osborne T.F. (2000) "Different sterol regulatory element-binding protein-1 isoforms utilize distinct co-regulatory factors to activate the promoter for fatty acid synthase." The journal of biological chemistry, 275(7): 4726-4733.

Martens G.A., Cai Y., Hinke S., Stange G., Van de Casteele M., Pipeleers D. (2005) "Glucose suppresses superoxide generation in metabolically responsive pancreatic beta cells." J Biol Chem. 280(21):20389-96.

Martin-Requero A., Cipres G., Rodriguez A., Ayuso M. S., and Parrilla R. (1992) "On the mechanism of stimulation of ureagenesis by gluconeogenic substrates: role of pyruvate carboxylase." American Journal of Physiology 263: E493-E499

Martoglio A.M., Miskin J.W., Smith S.K., MacKay D.J.C. (2002) "A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer." Bioinformatics **18**, 1617-1624.

Meijer A. J., Lamers W. H., Chamuleau R. A. F. M. (1990) "Nitrogen metabolism and ornithine cycle function." Physiol. Rev., **70**: 701-748.

Miller C.W., Ntambi J.M. (1996) "Peroxisome proliferators induce mouse liver stearoyl-CoA desaturase 1 gene expression." *PNAS* **93**, 9443-9448.

Moller I. M. (2001) "Plant mitochondria and oxidative stress: Electron Transport, NADPH Turnover, and Metabolism of Reactive Oxygen Species" Annu Rev Plant Physiol Plant Mol Biol. 52:561-591.

Mootha V. K., Lindgren C. M., Eriksson K. F., Subramanian A., Sihag S., Lehar J., Puigserver P., Carlsson E., Ridderstrale M., Laurila E., Houstis N., Daly M. J.,

Patterson N., Mesirov J.P., Golub T.R., Tamayo P., Spiegelman B., Lander E.S., Hirschhorn J.N., Altshuler D., Groop L.C. (2003) "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes" Nat Genet. 34(3):267-73.

Morris S. M., Jr. (1992) "Regulation of enzymes of urea and arginine synthesis." Annu. Rev. Nutr., **12**: 81-101.

Morris S. M., Jr. (2002) "Regulation of enzymes of the urea cycle and arginine metabolism." Annu. Rev. Nutr., **22**: 87-105.

Motz C., Martin H., Krimmer T., Rassow J. (2002) "Bcl-2 and Porin Follow Different Pathways of TOM-dependent Insertion into the Mitochondrial Outer Membrane." Journal of Molecular Biology **323**, 729-738.

Murphy K. and Mian S. (1999). "Modelling gene expression data using Dynamic Bayesian Networks." Technical report, University of California, Berkeley.

Murphy K.P. (2001) "The Bayesian Net Toolbox for Matlab", *Computing Science and Statistics*, 33. Available at http://www.cs.berkeley.edu/~murphy/Bayes/usage.html.

Nachman I., Regev A. (2004). "Inferring quantitative models of regulatory networks from expression data." Bioinformatics **20**(suppl_1): i248-256

Nagai H., Matsumaru K., Feng G., Kaplowitz N. (2002) "Reduced glutathione depletion causes necrosis and sensitization to tumor necrosis factor-alpha-induced apoptosis in cultured mouse hepatocytes". Heptaol. 36(1): 55-64.

Nakao M., Bono H., Kawashima S., Kamiya T., Sato K., Goto S., Kanehisa M. (1999) "Genome-scale Gene Expression Analysis and Pathway Reconstruction in KEGG." Genome Inform Ser Workshop Genome Inform, **10**:94-103

Ni T. C. and. Savageau M. A (1996). "Application of biochemical systems theory to metabolism in human red blood cells. Signal propagation and accuracy of representation." Journal of biological chemistry 271(14): 7927-41.

Ni T. C. and Savageau M. A. (1996). "Model assessment and refinement using strategies from biochemical systems theory: application to metabolism in human red blood cells." Journal of theoretical biology 179(4): 329-68.

Ntambi J.M. (1995) "The regulation of stearoyl-CoA desaturase (SCD)." Progress in Lipid Research **34**, 139.

Nystrom T., Larsson C., Gustafsson L., (1996) "Bacterial defense againt aging: role of the Escherichia coli ArcA regulator in gene expression readjusted energy flux and survival during stasis", The EMBO Journal 15(13): 3219-3228

Oehlen L.J., McKinney J.D., Cross F.R. (1996). "Ste12 and Mcm1 Regulate Cell Cycle-Dependent Transcription of FAR1." Mol. Cell Biol. **16**, 2830–2837.

Okuyama R., Fujiwara T., Ohsumi J. (2003) "High glucose potentiates palmitate-induced NO-mediated cytotoxicity through generation of superoxide in clonal beta-cell HIT-T15." FEBS Lett. 545(2-3):219-23.

Olsen C., (1971) "An enzymatic fluorimetric micromethod for the determination of acetoacetate, beta-hydroxybutyrate, pyruvate and lactate", Clin Chim Acta. 33(2):293-300.

Ong I., Glasner J. and Page D. (2002) "Modelling regulatory pathways in E.coli from time series expression profiles." Bioinformatics **18**, S241–S248.

Ontko J. A. (1972) "Metabolism of free fatty acids in isolated liver cells. Factors affecting the partition between esterification and oxidation" J Biol Chem. 247(6):1788-800.

Parke D.V. (1982) "Mechanisms of chemical toxicity--a unifying hypothesis." Regul Toxicol Pharmacol. 2(4):267-86.

Paumen M.B. (1997) "Direct Interaction of the Mitochondrial Membrane Protein Carnitine Palmitoyltransferase I with Bcl-2." Biochemical and Biophysical Research Communications **231**, 523.

Pearl J. (2000) Causality: models, reasoning and inference, Cambridge University Press, New York.

Pe'er D., Regev A., Elidan G. and Friedman N. (2001) "Inferring subnetworks from perturbed expression profiles", Bioinformatics, 17, Sup l.1: s215-s224.

Perrin B.E., Ralaivola L., Mazurie A., Bottani S., Mallet J., and d'Alche Buc F. (2003). "Gene networks inference using dynamic Bayesian networks." Bioinformatics **19**, S138–S148.

Ramoni M.F., Sebastiani P., Kohane I.S. (2002). "Cluster analysis of gene expression dynamics." Proc. Natl. Acad. Sci. USA **99**, 9121-9126.

Randle P. J., Garland P. B., Newsholme E. A., Hales C. N. (1965) "The glucose fatty acid cycle in obesity and maturity onset diabetes mellitus", Ann N Y Acad Sci. 131(1):324-33.

Rangel C., Angus J., Ghahramani Z., Lioumi M., Sotheran E., Gaiba A., Wild D. L., and Falciani F. (2004). "Modelling T cell activation using gene expression profiling and state space models." Bioinformatics **20**, 1361–1372.

Re D.B., Nafia I., Melon C., Shimamoto K., Goff L.K., Had-Aissouni L. (2006) "Glutamate leakage from a compartmentalized intracellular metabolic pool and activation of the lipoxygenase pathway mediate oxidative astrocyte death by reversed glutamate transport", Glia. 54(1):47-57.

Rego A.C., Santos M.S., Oliveira C.R. (1996) "Oxidative stress, hypoxia, and ischemia-like conditions increase the release of endogenous amino acids by distinct mechanisms in cultured retinal cells" J Neurochem. 66(6):2506-16.

Roder K., Wolf S.S., Sickinger S., and Schweizer M. (1999) "FIRE3 in the promoter of the rat fatty acid synthase (FAS) gene binds the ubiquitous transcription factors CBF and USF but does not mediate an insulin response in a rat hepatoma cell line." Eur J. Biochem, 260: 743-751.

Rodriguez C.(2001) "Differential induction of stearoyl-CoA desaturase and acyl-CoA oxidase genes by fibrates in HepG2 cells." Biochemical Pharmacology **61**, 357.

Rousset S., Bringuier A., Lardeux B., Feldmann G. (2003) "Apoptosis induced by tumor necrosis factor a in human hepatoma cell lines", Falk Symposium, 113: 303-13.

Sachs K., Perez O., Pe'er D., Lauffenburger D.A., Nolan G.P. (2005) "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data." Science **308**, 523-529.

Satoh A. (2004) "PKC-{delta} and -{epsilon} regulate NF-{kappa}B activation induced by cholecystokinin and TNF-{alpha} in pancreatic acinar cells." Am J Physiol Gastrointest Liver Physiol **287**, G582-591.

Sakurai H. (2003) "Tumor Necrosis Factor-{alpha}-induced IKK Phosphorylation of NF-{kappa}B p65 on Serine 536 Is Mediated through the TRAF2, TRAF5, and TAK1 Signaling Pathway." J. Biol. Chem. **278**, 36916-36923.

Saelens X., Kalai M., Vandenabeele P. (2001) „Translation Inhibition in Apoptosis. CASPASE-DEPENDENT PKR ACTIVATION AND eIF2-alpha PHOSPHORYLATION." J Biol Chem 276(45):41620-41628.

Salgado H., Santos-Zavaleta A., Gama-Castro S., Millan-Zarate D., Diaz-Paredo E., Sanchez-Solano F., Perez-Rueda E., Bonavides-Martinez C., Collado-Vides J. (2001) "RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12", Nucleic Acid Research 29: 72-74

Sanyal A. J., Campbell-Sargent C., Mirshahi F., Rizzo W. B., Contos M. J., Sterling R. K., Luketic V. A., Shiffman M. L. and Clore J. N. (2001) "Nonalcoholic steatohepatitis: association of insulin resistance and mitochondrial abnormalities", Gastroenterology. 120(5):1183-92.

Schroder J.M., Weber R., Weyhenmeyer S., Lammers-Reissing A., Meurers B., Reichmann H. (1991) "Adult onset lipid storage in gastric mucosa and skeletal muscle fibers associated with gastric pain, progressive muscle weakness and partial deficiency of cytochrome C oxidase." Pathol Res Pract. 187(1):85-95

Schilling C.H., Edwards J.S., and Palsson B.O., (1999) "Toward metabolic phenomics: analysis of genomic data using flux balances", Biotechnol. Prog. 15:288-295.

Segal E., Shapira M., Regev A., Pe'er D., Botstein D., Koller D., Friedman N. (2003) "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nature Genetics 34(2):166-176

Smith S.A. (2002) "Peroxisome proliferator-activated receptors and the regulation of mammalian lipid metabolism". Biochemical society transactions, 30: 1086-1090.

Soltoff S.P. (2001) "Rottlerin Is a Mitochondrial Uncoupler That Decreases Cellular ATP Levels and Indirectly Blocks Protein Kinase Cdelta Tyrosine Phosphorylation." J. Biol. Chem. **276**, 37986-37992.

Spellman P. T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Botstein D., Futcher B.(1998). "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization." Mol. Biol. Cell **9**, 3273-3297.

Spirtes P., Glymour C. and Scheines R. (1993) Causation, Prediction, and Search, Springer-Verlag, New York.

Srivastava S., Chan C. (2006) "Hydrogen peroxide and hydroxyl radical mediate palmitate-induced cytotoxicity to hepatoma cells: relation to MPT", accepted.

Staal F. J., Roederer M., Herzenberg L.A., Herzenberg L.A. (1990) "Intracellular thiols regulate activation of nuclear factor kappa B and transcription of human immunodeficiency virus." Proc Natl Acad Sci U S A. 87(24):9943–9947

Steffen M., Petti A., Aach J., D'Haeseleer P., Church G. (2002) "Automated modelling of signal transduction networks." BMC Bioinformatics, 3:34-44.

Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., Mesirov J.P. (2005) "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles" Proc Natl Acad Sci U S A.102(43):15545-50.

Suzuki J. (1996) "Learning Bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique", Proceedings of the International Conference on Machine Learning, Bari, Italy.

Takahashi Y., Campbell E.A., Hirata Y., Takayama T., Listowsky I.(1993) "A basis for differentiating among the multiple human Mu-glutathione S-transferases and molecular cloning of brain GSTM5", J Biol Chem. 268(12):8893-8.

Taylor V., Suter U. (1996) "Epithelial membrane protein-2 and epithelial membrane protein-3: two novel members of the peripheral myelin protein 22 gene family", Gene, 175(1-2): 115-20.

Tamatani M. (1999) "Tumor Necrosis Factor Induces Bcl-2 and Bcl-x Expression through NFkappa B Activation in Primary Hippocampal Neurons." J. Biol. Chem. 274, 8531-8538.

Tanay A., Sharan R., Kupiec M., Shamir R. (2004) "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data." Proceedings of the National Academy of Sciences of the United States of America 101(9):2981-2986.

Tilg H., Diehl A.M. (2000) "Cytokines in alcoholic and nonalcoholic steatohepatitis", New Eng J Med. 343(20): 1467-1476.

Tilg H. (2001) "Cytokines and liver diseases." Can J Gastroenterol **15**(10): 661-8.

Townsend A.J., Leone-Kabler S., Haynes R.L., Wu Y., Szweda L., Bunting K.D. (2001) "Selective protection by stably transfected human ALDH3A1 (but not human ALDH1A1) against toxicity of aliphatic aldehydes in V79 cells", Chem Biol Interact.130-132(1-3):261-73.

Trach V., Buschmans-Denkel E., Schaper W. (1986) "Relation between lipolysis and glycolysis during ischemia in the isolated rat heart", Basic Res Cardiol. 81(5):454-64.

Tran L.M., Brynildsen M.P., Kao K.C., Suen J.K., Liao J.C. (2005). "gNCA: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation." Met Eng. 7, 128-141.

Valle Blazquez M., Luque I., Collantes E., Aranda E., Solana R., Pena J., Munoz E. (1997) "Cellular redox status influences both cytotoxic and NF-kappa B activation in natural killer cells. Immunology." 90(3):455-60.

Van Helden J., Naim A., Mancuso R., Eldridge M., Wernisch L., Gilbert D., Wodak S. J. (2000) "Representing and analyzing molecular and cellular function using the computer." Biol. Chem. **381**(9-10): 921-935.

Wang H., Rish I., and Ma S. (2002) "Using Sensitivity Analysis for Selective Parameter Update in Bayesian Network Learning," 2002 AAAI Spring Symposium Proceedings on Information Refinement and Revision for Decision Making: odeling for Diagnostics, Prognostics, and Prediction, Stanford, Palo Alto, March 25-27, 2002.

Wang S.J., Chen J. J. (2004) "Sample Size for Identifying Differentially Expressed Genes in Microarray Experiments." Journal of Computational Biology Vol. 11, 714-726.

Wang Y., Liu C. L., Storey J.D., Tibshirani R.J., Hershlag D., Brown P.O. (2002). "Precision and functional specificity in mRNA decay" PNAS 99(9): 5860-5865.

Watada H. and Kawamori R.(2003) "Insulin resistance and NASH." BIO Clinica, 18(10): 874-879.

Wei Y., Wang D., Topczewski F., Pagliassotti M.J. (2006) "Saturated fatty acids induce endoplasmic reticulum stress and apoptosis independently of ceramide in liver cells", Am J Physiol Endocrinol Metab.291(2):E275-81

Wold S., Ruhe A., Dunn W.J. (1984) "The ollinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses", SIAM Journal on Scientific and Statistical Computing, 5, 735-743

Xie, Z., and Askari, A. (2002) "Na^+,K^+-ATPase as a signal transducer." Eur. J. Biochem, 269: 2434-2439.

Xu Z., Williams B.R. (2000) "The B56alpha regulatory subunit of protein phosphatase 2A is a target for regulation by double-stranded RNA-dependent protein kinase PKR." Molecular and cellular biology 20(14):5285-5299.

Yamaguchi M., Miyashita Y., Kumagai Y., Kojo S. (2004) "Change in liver and plasma ceramides during D-galactosamine-induced acute hepatic injury by LC-MS/MS", Bioorg Med Chem Lett. 14(15): 4061-64.

Zhang B., Schmoyer D.(2004). "GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies." BMC Bioinformatics **5**(1): 16.