PATHWAY MODELING: FROM GENE EXPRESSION TO PATHWAY DYNAMICS

by

AYSE MERIC OVACIK

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Chemical and Biochemical Engineering

Written under the direction of

Ioannis P. Androulakis, PhD

and approved by

________________________

________________________

________________________

________________________

________________________

New Brunswick, New Jersey

January, 2011
ABSTRACT OF THE DISSERTATION

Pathway Modeling: From Gene Expression to Pathway Dynamics

By AYSE MERIC OVACIK

Dissertation Director:

Ioannis P. Androulakis, PhD

Biological pathways represent a critical level of biological organization and understanding of biochemical pathways is identified as key to future advances in biological sciences (Schaefer, 2004). The overall goal of this thesis is to develop a pathway based approach that integrates different aspects of biological pathways, specifically the structure and the dynamics of a pathway in order to characterize cells’ behavior. Our objectives are to assess structural and functional cross-species comparison of pathways (Chapter 2), to formulate a reliable pathway activity metric based on gene expression data (Chapter 3), to demonstrate that our pathway activity formulation can predict the underlying dynamics (Chapter 4) and finally to demonstrate that the pathway activity formulation can identify cell’s response to a stimulant (Chapter 5). Chapter 3-5 present how a significant pathway can be identified. Then, cross-species comparison of pathways (Chapter 2) can be used. Note that we could have Chapter 2 and Chapter 5 swapped for a more fluent flow. Nevertheless, we present the chapters in this order for a better read.

In Chapter 2, we propose an improvement of the reaction alignment method, emerged as the most successful pathway comparison method, by accounting for sequence similarity in addition to reaction alignment method. Using nine species, including human and some model organisms and test species, we evaluate the standard and improved
comparison methods by analyzing glycolysis and citrate cycle pathways conservation. In addition, we demonstrate how organism comparison can be conducted by accounting for the cumulative information retrieved from nine pathways in central metabolism as well as a more complete study involving 36 pathways common in all nine species. In Chapter 3, we explore an extension of the pathway activity methodology which entails singular value decomposition (SVD) of the expression data of the genes constituting a given pathway. We show that pathway analysis enhances our ability to detect relevant changes in pathway activity using synthetic data. In addition, we illustrate that pathway activity formulation should be coupled with a significance analysis to distinguish significant information from random deviations. In Chapter 4, we perform an unsupervised pathway level analysis, based on the formulation presented in Chapter 3, on a rich time series of transcriptional profiling in rat liver. The over-represented five specific patterns of pathway activity levels, which cannot be explained by random events, exhibit circadian rhythms. The identification of the circadian signatures at the pathway level identify pathways related to energy metabolism, amino acid metabolism, lipid metabolism and DNA replication and protein synthesis, which are biologically relevant in rat liver. In Chapter 5, we demonstrate that our pathway activity formulation enables us to detect relevant changes in pathways due to in utero di-butyl-phthalate (DBP) exposure. Our findings suggest that the pathways that produce precursors to cholesterol synthesis exhibit more significant change compared to the rest of the affected pathways. In addition, pathway activity levels of certain biological functions accompany testosterone decrease, which is the critical event for male reproductive developmental effects of DBP, such as steroid hormone metabolism and biosynthesis of steroids.
ACKNOWLEDGEMENTS

I have been blessed to have Prof. Androulakis as my thesis advisor. I cannot even find the words to explain my gratitude for his guidance and support. I know one thing for sure; this thesis would not have come out without him. His passion for science, honest words, and constant help carried me through my PhD. I just hope that I would hold on what he thought me and not stray.

I would like to acknowledge my committee members – Prof. Gallo, Prof. Georgopoulos and Prof. Ierapetritou for their valuable comments on my proposal. Their suggestions definitely helped me to finish this dissertation. I would like to thank all of them for their time and insights overall.

I would like to acknowledge the funding sources which supported me throughout my graduate study. Support for this work has been partially provided by USEPA-funded Environmental Bioinformatics and Computational Toxicology Center (ebCTC), under grant number GAD R 832721-010. Further, I would like to acknowledge the following students at Rutgers University: Eric Yang (Rutgers Alumni), Panagiota Foteniou, Jeremy Scheff, Nguyen Tung, Kaiyuan He, Qian Yang and Pantelis Mavroudis.

Finally, I would like to thank my family and close friends. Being away from them is hard, yet I always feel their love and support even we are many miles apart. I would never been able to finish my PhD without them. I feel lucky to be surrounded with such a nice crowd.
DEDICATION

to my mother, nergiz
# Contents

ABSTRACT .......................................................................................................................... ii  

ACKNOWLEDGEMENTS .................................................................................................... iv  

DEDICATION ...................................................................................................................... v  

List of Tables .................................................................................................................... x  

List of Figures ................................................................................................................... xi  

Chapter 1 ............................................................................................................................ 1  

Biological Pathways ......................................................................................................... 1  
1.1 Definition of biological pathways ........................................................................ 1  
1.1.1 Metabolic pathways ......................................................................................... 2  
1.1.2 Signaling pathways ......................................................................................... 3  
1.2 Pathway databases .............................................................................................. 5  
1.3 Pathway quantification ....................................................................................... 8  
1.3.1 Kinetic modeling ............................................................................................ 9  
1.3.2 FBA analysis ................................................................................................. 9  
1.4 Pathway based research ................................................................................... 10  
1.5 The goal and the objectives of the thesis ..................................................... 11  

Chapter 2 .......................................................................................................................... 15  

Cross-species Comparison of Metabolic Pathways .................................................. 15  
2.1 Cross-species comparison studies of metabolic pathways ....................... 17
Chapter 3 ................................................................................................................................. 44

Representing the Activity of a Pathway Based On Gene Expression ................. 44

3.1 Pathway activity level analysis ................................................................................. 47
3.2 Significance analysis of pathway activity level ..................................................... 48
3.3 Identifying underlying dynamics within a pathway .............................................. 49
  3.3.1 Comparing two experimental conditions ..................................................... 49
  3.3.2 Pathway activity of time course data .......................................................... 53
3.4 Results and discussion ......................................................................................... 55
3.5 Summary and conclusions ..................................................................................... 59

Chapter 4 ................................................................................................................................. 61

Circadian Signatures in Rat Liver: From Gene Expression to Pathways ........... 61
4.1 Circadian signatures of pathways in rat liver ................................................ 64
4.2 Materials and methods .................................................................................. 64
   4.2.1 Experimental data ..................................................................................... 64
   4.2.2 Pathway database ...................................................................................... 65
   4.2.3 Clustering analysis of pathway activity levels .......................................... 67
4.3 Results and discussion .................................................................................. 68
4.2 Summary and conclusions ............................................................................ 83

Chapter 5 ................................................................................................................. 86

Pathway Changes in the Testes following in utero Exposure to Di-butyl-phthalate (DBP) ................................................................................................................. 86

5.1 Testicular dysgenesis due to in utero DBP exposure: implications of exposure .................................................................................................................... 91
   5.1.1 Materials and methods .............................................................................. 91
   5.1.2 Results and discussion ............................................................................ 94
5.2 Testicular dysgenesis due to in utero DBP exposure: implications of duration ..................................................................................................................... 101
   5.2.1 Materials and methods ............................................................................ 102
   5.2.3 Results and discussion ............................................................................ 103
5.3 Summary and conclusions ......................................................................... 107

Chapter 6 ................................................................................................................. 109

Conclusions and Future Perspectives .................................................................... 109

6.1 Summary ..................................................................................................... 109
6.2 Future directions .................................................................................................................... 111

Bibliography .................................................................................................................................. 113

Appendix: Transcriptional and Metabolic Flux Profiling of Triadimefon Effects on Cultured Hepatocytes .................................................................................................................. 130

A.1 Materials and Methods........................................................................................................... 133
A.1.1 Hepatocyte isolation and culture ..................................................................................... 133
A.1.2 RNA Extraction and Subsequent Microarray Experiments ........................................... 134
A.1.3 Microarray Data Analysis: ............................................................................................. 134
A.1.4 Clustering Analysis ......................................................................................................... 135
A.1.5 Functional analyses of microarray data ........................................................................... 138
A.1.6 Computational Identification of Transcriptional Regulators ....................................... 138
A.1.7 Extracellular metabolite measurements ......................................................................... 139
A.1.8 Metabolic network flexibility analysis (MNFA) ............................................................. 140
A.2 Results ..................................................................................................................................... 147
A.2.1 Clustering and Functional Analysis of Microarray Data ................................................. 150
A.2.2 Metabolite measurements ............................................................................................. 157
A.2.3 Metabolic network flexibility analysis (MNFA) ............................................................. 159
A.3 Discussion ............................................................................................................................... 164

Curriculum Vitae ........................................................................................................................... 169
List of Tables

Table 2-1 Dissimilarity values between reconstructed phylogenetic tree and NCBI taxonomy retrieved from TOPD/FMTS software. + ................................................................. 40

Table 2-2 Dissimilarity values between reconstructed phylogenetic tree and NCBI taxonomy retrieved from TOPD/FMTS software. + ................................................................. 41

Table 4-1 Online databases from which MSigDB (Subramanian et al., 2005) is curated. ...................................................................................................................................................................................... 65

Table 4-2 Circadian pathways and associated cluster numbers........................................ 77

Table 4-3 Enriched pathways by circadian genes The circadian genes were mapped to canonical pathways provided by http://www.broadinstitute.org/gsea/msigdb/ . p-values indicate the significance of the overlap of the circadian genes within a pathway............. 79

Table 5-1. Summary of the identified pathways ............................................................ 96
List of Figures

Figure 1-1: Metabolic pathways: series of enzymatic reactions. A) In enzymatic reactions, substrates, the molecules at the beginning of the process, are converted to the products by enzymatic activities. Enzymes are selective for their substrates. Therefore, set of enzymes in a cell determines the metabolic activity. The figure is adapted from http://simpert.com/resources B) Series of enzymatic reactions form metabolic pathways that do not exist in isolation. An end product of a metabolic pathway is a substrate for another pathway. The figure is adapted from http://student.biology.arizona.edu/. ........... 3

Figure 1-2: A significant example of signal transduction even in a cell: Apoptosis. Apoptosis is signaling transduction event characterized as a cellular suicide mechanism. Specific set of components is activated by receptors and the signal transduction continues until the intended function is carried out. The figure is adapted from www.cellsignal.com........................................................................................................... 4

Figure 1-3: The system flow of FMM is displayed. FMM is a web server established to reconstruct metabolic pathways. FMM uses information from multiple resources. The figure is adapted from (Chou et al., 2009). ................................................................. 7

Figure 1-4: Metabolic network representing hepatic metabolism and triadimefon detoxification pathways in primary rat hepatocytes. It includes important pathways such as glycolysis/gluconeogenesis, pentose phosphate pathway, lipid metabolism, urea production, TCA cycle, triadimefon oxidation and subsequent detoxification and reactions involving amino acid metabolism. Metabolites in gray ovals signify extracellular metabolites. For the sake of clarity, dashed lines link metabolites that are
connected and occur in different parts of the pathway. The figure is adapted from (Iyer et al., 2010a).

**Figure 2-1: Modeling metabolic pathways and evaluating pathway similarity:** 1) A metabolic pathway consists of series of enzymatic reactions to fulfill a particular biological function. 2) The pair wise relations are translated to an incidence matrix. This model describes metabolic pathways as a graph G (V,E) where V is the set of nodes (vertices) and E is the set of edges connecting pairs of nodes (Acton et al., 1996; Deville et al., 2003). 3) Clement et al. (2005) proposed an algorithm that considers the robustness of a pathway by searching the alternative routes in a pathway. Alignment of the metabolic pathways occurs by mapping each non-common compound, enzyme, and reaction in one metabolic pathway to the most similar compound, enzyme, and reaction, respectively, in the other metabolic pathway.

**Figure 2-2: Simple example of the reaction alignment method.** Organism I has one reaction, R1. Organism II has two reactions, R2 and R3. The similarity between each reaction in Organism I and Organism II is evaluated (Eq. (2.1) or Eq. (2.4)). If \( \text{sim}(R1,R2) > \text{sim}(R1,R3) \), then R1 is aligned with R2 (Eq.(2.2)). i) The constituents of reactions in two substitute pathways; reactions include metabolites and enzymes. ii) The similarity between reactions is evaluated by calculating similarity between sets of compounds and the similarity between enzymes.

**Figure 2-3: Phylogenetic tree reconstruction based on a distance matrix.**

**Figure 2-4: Comparison of reaction alignment algorithm and reaction alignment algorithm with enzyme sequence similarity employing glycolysis pathway.** A) The reconstructed phylogenetic tree based on reaction alignment and B) reaction alignment
with enzyme sequences of glycolysis pathway. Homo sapiens (hsa), Mus musculus (mmu), Rattus norvegicus (rno), Danio rerio (dre), Drosophila melanogaster (dre), Caenorhabditis elegans (cel), Arabidopsis thaliana (ath), Saccharomyces cerevisiae (sce) and Escherichia coli (eco).

Figure 2-5: Comparison of reaction alignment algorithm and reaction alignment algorithm with enzyme sequence similarity employing citrate cycle. A) The reconstructed phylogenetic tree based on reaction alignment and B) reaction alignment with enzyme sequences of citrate cycle. Homo sapiens (hsa), Mus musculus (mmu), Rattus norvegicus (rno), Danio rerio (dre), Drosophila melanogaster (dre), Caenorhabditis elegans (cel), Arabidopsis thaliana (ath), Saccharomyces cerevisiae (sce) and Escherichia coli (eco).

Figure 2-6: The phylogenetic tree based on the information retrieved from NCBI (Benson et al., 2009) for selected organisms: Homo sapiens (hsa), Mus musculus (mmu), Rattus norvegicus (rno), Danio rerio (dre), Drosophila melanogaster (dre), Caenorhabditis elegans (cel), Arabidopsis thaliana (ath), Saccharomyces cerevisiae (sce) and Escherichia coli (eco).

Figure 2-7: Comparison of reaction alignment algorithm and reaction alignment algorithm with enzyme sequence similarity employing 9 pathways belong to the central metabolism. A) The reconstructed phylogenetic tree based on reaction alignment and B) reaction alignment with enzyme sequences of glycolysis pathway. Homo sapiens (hsa), Mus musculus (mmu), Rattus norvegicus (rno), Danio rerio (dre), Drosophila melanogaster (dre), Caenorhabditis elegans (cel), Arabidopsis thaliana (ath), Saccharomyces cerevisiae (sce) and Escherichia coli (eco).
Figure 2-8: Comparison of reaction alignment algorithm and reaction alignment algorithm with enzyme sequence similarity employing all 36 common pathways in 9 selected species. A) The reconstructed phylogenetic tree based on reaction alignment and B) reaction alignment with enzyme sequences of glycolysis pathway. Homo sapiens (hsa), Mus musculus (mmu), Rattus norvegicus (rno), Danio rerio (dre), Drosophila melanogaster (dre), Caenorhabditis elegans (cel), Arabidopsis thaliana (ath), Saccharomyces cerevisiae (sce) and Escherichia coli (eco). ............................................ 41

Figure 3-1: The outline comparing two experimental conditions within a pathway. Pathway activity levels are calculated using SVD and the significance of pathway activity levels are evaluated. Subsequently, the identification of the differentiation between two treatment groups of pathway activity levels is evaluated. ............................................ 51

Figure 3-2: The outline for identifying significant time-course data pattern within pathways. Pathway activity levels are calculated using SVD and the significance of pathway activity levels are evaluated. Then, the parameters of the sinusoidal model $A \cdot \sin(B \cdot t + C)$ that would best fit the pathway activity levels are characterized. Finally, the correlation between fitted sinusoidal model and the pathway activity levels is evaluated. ....................................................................................................................................... 55

Figure 3-3: Effect of noise level on the differential expression of the synthetic pathway. Gene expression values within the synthetic pathway are generated representing two different treatment groups (given in Eq.(3.1)). The significance of $PAL_p$ and $PA_p$ (denotes the difference between different treatment groups) of the synthetic pathway and the events that 10 %, 50 % and 90 % of the genes within the synthetic pathway are DEGs are illustrated at different noise levels. .............................................. 57
Figure 3-4: A) Effect of noise level on the sinusoidal dynamics of the synthetic pathway. As the noise level is increased, the significance (1- \( p\)-value) of the event that pathway activity level has sinusoidal pattern and the events that 10 \%, 50 \% and 90 \% of the genes within the synthetic pathway have sinusoidal pattern are illustrated. B) Effect of noise level on the significance of PAL. As the noise level is increased, the significance (1- \( p\)-value) of the event that pathway activity level of the synthetic pathway has sinusoidal pattern and the significance of PAL are illustrated.

Figure 4-1: The suprachiasmatic nucleus (SCN) if located in the hypothalamus in mammals. SCN is synchronized to daylight and communicates the timing information to the body via direct and indirect signals (feeding time, body temperature and circulating hormones)(Ben-Shlomo and Kyriacou, 2002). A group of genes in peripheral tissues bypass the circadian signaling responding light stimulation (Ishida et al., 2005). The figure is retrieved from (Ben-Shlomo and Kyriacou, 2002).

Figure 4-2: The outline for clustering analysis of pathway activity levels. Pathway activity analysis begins with mapping gene expression onto known pre-defined groups of genes, pathways. Subsequently, the pathway activity levels are calculated using SVD and the significance of pathway activity levels are evaluated. Pathways are filtered based on the significance of the PALs. Following, the over-populated patterns are identified by using a consensus clustering approach proposed in (Nguyen et al., 2009). Then, the parameters of the sinusoidal model \( A \cdot \sin(B \cdot t + C) \) that would best fit the centroids of the pathway activity levels (in each cluster) are characterized. Finally, the correlation between fitted sinusoidal model and the centroids of the pathway activity levels in each cluster is evaluated.
Figure 4-3: The five significant clusters identified by a consensus clustering analysis (Nguyen et al., 2009) using $\delta = 0.65$. The pathway activity level (PAL) of pathways represents the presented curves and the exact reverse curves; PAL = (-) PAL. The signs of PAL are chosen so that PAL has the similar patterns for a better representation and clustering purposes. The centroids of each cluster is shown with the red error bars, the fitted sinusoidal model to the centroids of each cluster is depicted in white.

Figure 4-4: A) The comparison of the fp to the permuted fp for MAPK Pathway, nicotinate and nicotinamide metabolism and glycine, serine and threonine metabolism pathway. The mean and the standard deviation interval of permuted fp are given. The same value of fp can be obtained by randomly permuted data in MAPK Pathway and nicotinate and nicotinamide metabolism, whereas the fp captured by randomly permuted data is much lower compared to fp in glycine, serine and threonine metabolism pathway.

B) Pathway activity levels and fitted sinusoidal models for the pathways. The mean and the standard deviation interval of the pathway activity levels are given. The correlation between pathways activity level and fitted sinusoidal model is presented for each pathway on top of each graph.

Figure 4-5: The relationships between weight and the correlation of the genes within glycine, serine and threonine metabolism. The correlation is between gene expressions and the fitted sinusoidal models and is set to identify circadian genes. The threshold for circadian genes is correlation > 0.8. The weights are evaluated from the SVD analysis. The absolute value of the weights represents the contribution of the individual genes to the pathway activity level. The genes that have higher correlation values have relatively higher absolute weights.
Figure 4-6: Selected gene expressions within glycine, serine and threonine metabolism. The correlation between the gene expression levels and the fitted sinusoidal models and the weights, which are evaluated via SVD analysis, of the genes are given on top of each graph. The signs (positive and/or negative) of weights indicate opposite direction in the gene expression................................. 74

Figure 4-7: The relative values of the associated eigenvalues for glycine, serine and threonine metabolism. The bars indicate the variation in the data captured by each individual eigenvector for glycine, serine and threonine metabolism pathways. The solid line represents the data variability captured by the corresponding eigenvectors when randomly generated data (of the same dimension) were used. No apparent distinction between the actual data and randomly generated data was identified after the first eigenvalue, as quantified by the calculated p-values. ....................................................... 75

Figure 4-8: The most significant data patterns in Glycine, serine and threonine. The first 4 rows of $V_p(t,1)$ that are retrieved from SVD calculations of Glycine, serine and threonine metabolism the elements of $S_p(k,1)$ are sorted from the highest to the lowest. 1) $V_p(t,1)$, 2) $V_p(t,2)$, 3) $V_p(t,3)$, 4) $V_p(t,1)$ ....................................................... 76

Figure 4-9: Fitted sinusoidal models of pathway activity levels for cholesterol biosynthesis and bile acid biosynthesis. ................................................................. 81

Figure 4-10: Individual gene expressions in cholesterol biosynthesis. Associated weights and correlations with the fitted sinusoidal model were given on top of each panel. ........................................................................................................... 82

Figure 4-11: Individual gene expressions in bile acid biosynthesis. Associated weights and correlations with the fitted sinusoidal model were given on top of each panel. .......... 83
Figure 5-1: Exposure response array for candidate endpoints and PODs for RfD derivation presented in the external review draft IRIS Tox Review for DBP (U.S.EPA, 2006) The studies are arrayed by toxicological endpoint category. All studies were performed in the rat using the oral route of exposure (although method of oral dosing varied). The study DBP exposure interval and reference are shown on the X axis. The dose (Y axis) is shown in a logarithmic scale. See the articles for additional study details. The vertical line indicates the dose range tested. The lowest and highest doses tested in the study are indicated by the horizontal lines. The green circle indicates the study NOAEL (no observed-adverse-effect level) and the blue triangle indicates the study LOAEL (lowest-observed-adverse-effect level). Neuro, neurological; GD, gestation day; PND, postnatal day. The figure is adapted from (U.S.EPA, 2009) ................................. 88

Figure 5-2: The proposed DBP mechanism of action for the male reproductive developmental effects. The mechanism of action is defined as all of the steps between chemical exposure at the target tissue to expression of outcome. The steps shown are based on male reproductive developmental toxicity and toxicogenomics studies. Some of the affected pathways and individual genes whose expression was significantly affected after DBP exposure in multiple studies are included. By contrast, MOAs are shown in purple letters including the well-characterized MOAa and one example of an unidentified MOA. The figure is adapted from (Shultz et al., 2001; Barlow and Foster, 2003; Thompson et al., 2004; Wilson et al., 2004; Liu et al., 2005) and published in (U.S.EPA, 2009) ................................................................................................................. 90

Figure 5-3: Metabolic sub-network of active pathways. Active pathways are connected to each other via metabolites and illustrated from the most active pathway
towards to least active pathway. Data source is (Liu et al., 2005). The connections of the active pathways were retrieved from KEGG (Kanehisa and Goto, 2000), December 2008. The active pathways also have connections to non-active pathways; but only the active pathways are included in this metabolic sub-network.

Figure 5-4: The relationship between differential expression of individual genes and PA using the Liu et al. (2005) DBP data. The $PA_P$ of a given pathway is first evaluated using the gene that has the highest SNR. Subsequently, the genes are added in the order of their SNR, from highest to lowest. Pathways identified as active for DBP, such as biosynthesis of steroids (A) and butanoate metabolism (B), maintain high $PA_P$ values even when all genes in the pathway are added. Alternatively, pathways not identified as active for DBP such as pentose and glucuronate interconversions (C) and ether lipid metabolism (D), exhibit a decrease in $PA_P$ as the less discriminating genes (i.e., those with a lower SNR value) are added. The Y axis is the $PA_P$.

Figure 5-5: A gene network created using ingenuity pathway analysis (IPA). The figure illustrated that the interactions among informative genes following in utero DBP exposure in the rat testis from (Liu et al., 2005). Interactions among genes are derived from the IPA database. Genes or gene products are represented as nodes. Diamonds, enzymes; horizontal ovals, transcription regulators; Squares, cytokines; rectangles, nuclear receptors. Solid lines represents direct relationships between nodes (i.e. molecules that physical contact with each other such as binding or phosphorylation). Dashed lines represent indirect interactions (i.e. not requiring physical contact between the two molecules such as signaling events.).
Figure 5-6: Comparison of control and DBP-treated pathway activity profiles over time. The control (dashed line) and DBP (solid line) treated profiles of active pathways. The dots are the replicate information for the associated experimental condition and time point. Pathway activity levels of associated pathways do not indicate up-regulation or down regulation, however they represent relative change across different experimental conditions. Most of the pathways exhibit largest differentiation between control and DBP treated profiles at the 6th and 18th hour of exposure, whereas few pathways exhibit differentiation at the 3rd hour of DBP exposure such as glycosaminoglycan degradation and glutathione metabolism.

Figure 5-7: A temporal gene network model over time created by IPA from the informative gene list based on time-course data after in utero DBP exposure (Thompson et al., 2005). The informative genes were evaluated at each time point and mapped onto a global molecular network using the Ingenuity Pathways Knowledge Base. Diamonds, enzymes; Horizontal ovals, transcription regulators; Squares, cytokines; Rectangles, nuclear receptors. Solid lines represent direct relationships (also called edges) between nodes (i.e., molecules that make physical contact with each other, such as binding or phosphorylation). Dashed lines represent indirect interactions (i.e., not requiring physical contact between the two molecules, such as signaling events). CP, Canonical pathway. Low (green), down regulated expression with respect to control. High (red), unregulated expression with respect to control.
Chapter 1

Biological Pathways

1.1 Definition of biological pathways

A living cell is composed of many types of molecules such as proteins, DNA, RNA and small molecules (metabolites). The activity of the cell depends on the interaction between these molecules. The interactions and the collective behavior of a subset of these molecules form a functional unit that reveals certain tasks in a cell, such as protein synthesis or TCA cycle. These functional units are identified as biological pathways. By definition, a biological pathway is a series of reactions whose function is separable from those of others (Alberts et al., 2002; BioPAX, 2010). Biological pathways need not to be rigid fixed structures; such as a given component may belong to different pathways. Some examples of biological pathways, such as protein synthesis, DNA replication and glycolysis have been successfully reconstituted in vitro (Hartwell et al., 1999).

There are two main types of biological pathways depending on the interactions between the pathway components; metabolic pathways and signaling pathways. Metabolic pathways consist of enzymatic reactions, whereas signaling pathways involve signal transduction mostly via protein-protein interactions.
1.1.1 Metabolic pathways

The biochemical reactions would normally occur at higher temperatures at which cells can function. For this reason, each reaction needs a specific catalyzing event by the cell. This requirement is crucial, because it allows each reaction to be controlled by the cell (Alberts et al., 2002).

Enzymes are the specialized proteins that catalyze one of the many possible reactions in the cell (Figure 1-1A). Enzyme-catalyzed reactions are usually linked in series, so that the product of one reaction becomes the substrate, for the next. These long reaction series, called metabolic pathways, are connected to one another (Figure 1-1B) and defined as series of enzymatic reactions that converts a specific substrate to another specific product.

Two opposing streams of chemical reactions occur in cells: (1) the catabolic pathways and (2) anabolic pathways. Catabolic pathways break down complex molecules into smaller molecules to generate energy and building blocks of cells, whereas anabolic pathways use the generated energy by catabolic pathways to synthesize other molecules that cells need (Alberts et al., 2002).
Figure 1-1: Metabolic pathways: series of enzymatic reactions. A) In enzymatic reactions, substrates, the molecules at the beginning of the process, are converted to the products by enzymatic activities. Enzymes are selective for their substrates. Therefore, a set of enzymes in a cell determines the metabolic activity. The figure is adapted from http://simpert.com/resources B) Series of enzymatic reactions form metabolic pathways that do not exist in isolation. An end product of a metabolic pathway is a substrate for another pathway. The figure is adapted from http://student.biology.arizona.edu/.

1.1.2 Signaling pathways

Signaling pathways are described as a group of molecules that work together to control one or more cell functions. Cells receive signals via specialized proteins called receptors which are very specific to the signals they receive. The signal can be chemical or electrical signals, which include all the hormones and neurotransmitters secreted within
the body. After interacting with a receptor, the signal travels through the cell where its message is transmitted by specialized proteins to fulfill a specific action in the cell. Such as after the first molecule in a pathway receives a signal, it activates another molecule. This process is repeated until the last molecule is activated and the cell function involved is carried out (Alberts et al., 2002).

Figure 1-2: A significant example of signal transduction even in a cell: Apoptosis. Apoptosis is a signaling transduction event characterized as a cellular suicide mechanism. Specific set of components is activated by receptors and the signal transduction continues until the intended function is carried out. The figure is adapted from www.cellsignal.com.
1.2 **Pathway databases**

New technologies and advanced research enable us to study the life in great detail. The chemical reactions and significant biological processes that make every species alive can now be examined intensively. The amount of information available about biological systems is overwhelming. It is estimated that information about biological systems is growing at an exponential rate (Olken and Jagadish, 2004). Especially, with the advent of high-throughput experimentation, our ability to identify and better characterize biological information is promoted. To make the emergent findings about biological processes readily accessible, the systematic collection of pathway information is deposited in the form of pathway databases.

A pathway database is a bioinformatics tool that describes the biological pathways, their components and the interactions (Schaefer, 2004; Bauer-Mehren *et al.*, 2009). A significant issue with pathway databases is that they need special requirements and challenges of data management for cellular and molecular biology (Raschid, 2004). In most of the pathway databases (e.g.: BioCARTA (www.biocarta.com)), the information is not in computer-readable form needed for automatic retrieval and processing. Though there are standard formats for data mining of pathway information such as BIOPAX (BioPAX, 2010) and SBML (*Gauges et al.*, 2006; *Bornstein et al.*, 2008), only few of the pathway repositories follow these standards.

Generally, expert curators have encoded their knowledge of particular biological processes, backed by small-scale experiments. In some cases, the detailed representations are supplemented with new descriptions synthesized by the curator (*Joshi-Tope et al.*, 2003). In other cases, literature references may be added to the database at the level of the
pathway or the individual interaction (Bader et al., 2003). There have been attempts to construct databases by automatically mining the literature (Thomas and Ganji, 2006).

In addition, biological pathway data may include a significant amount of noisy data. Experiments performed to generate biological pathways information generate huge amounts of information, often on the order of terabytes for even extremely short-duration experiments. Because of the large amount of data, statistical methods are needed to sufficiently analyze the data (van Helden et al., 2000; Berg et al., 2004). However, these analyses cannot essentially differentiate between results produced by the actual function and chance occurrence of the statistical analysis (Berg et al., 2004).

Some databases rely on assumed conservation of pathways across organisms to fill in missing data such as KEGG (Kyoto Encyclopedia of Genes and Genomes, (Kanehisa et al., 2006)), the BioCyc database family, (Karp et al., 2005) and ERGO (Overbeek et al., 2003). To annotate pathways across species, these databases analyze well studied organisms and map the enzymes in well-studied organisms onto other organisms via the enzyme commission numbers (Pellegrini et al., 1999) and orthology relationships (Tatusov et al., 2001).
Figure 1-3: The system flow of FMM (From Metabolite to Metabolite) is displayed. FMM is a web server established to reconstruct metabolic pathways. FMM uses information from multiple resources. The figure is adapted from (Chou et al., 2009).

The resulting pathways often include gaps such as missing steps in a chain of biochemical reactions. There is a strong need of tools for the automatic integration of different pathway databases. The missing information in pathways is addressed by gap-filling algorithms (Green and Karp, 2004) and via comparative genomics in SYSTOMONAS (Systems Biology of Pseudomonas, (Choi et al., 2007)) and FFM (From metabolite to metabolite, (Chou et al., 2009). Figure 1-3) where large data integration is performed for the well established databases BRENDA (Chang et al., 2009)) KEGG and PRODORIC (Prokaryotic Database of Gene Regulation, (Munch et al., 2003)).

Pathguide (www.pathguide.org) presents the available online pathway databases and is updated regularly (Bader et al., 2006). Pathguide contains information about 325
biological pathway related resources. Among the various pathway databases, KEGG is the most widely used (Bader et al., 2006) and most comprehensively compiled web source for metabolic and signaling pathways. KEGG database provides a standard for other pathway databases. As such, other important metabolic pathways databases (e.g. Ingenuity, www.ingenuity.com and GeneGO, www.genego.com) are generally dependent on the information in the KEGG database. As a result, the KEGG database has been widely used in studying mammalian systems (Arum et al.,; Liu et al.,; Kitami and Nadeau, 2002; Bono and Okazaki, 2005; Guo et al., 2006; Kovacs et al., 2007; Singh et al., 2007; Zhou et al., 2007; Fardet et al., 2008), even though KEGG is largely based on simpler organisms. KEGG contains the largest amount of species with relevant information such as orthology, ligands and sequence information of associated genes and enzymes in a given pathway. As more specific pathway databases such as tissue specific pathway databases or cellular compartment specific pathway databases are created and the pathway knowledge databases are improved, the power of pathway based approaches will increase.

1.3 Pathway quantification

The quantification of pathways is defined to characterize the activity in a pathway such as the rates at which metabolites and substrates are produced and degraded in cells. Genome sequencing projects and further systematic analyses of “omics” studies are producing an unprecedented mass of molecular information. This provides us with detailed account of the cell with which scientists begin to build models for simulating intracellular molecular processes to predict the dynamic behavior of pathways.
1.3.1 Kinetic modeling

Kinetic modeling of biochemical pathways has been one of the most active areas of cellular simulation. Many attempts have been made to simulate biological pathways; both metabolic (Mendes, 1993; Mendes, 1997; Tomita, 2001; Takahashi et al., 2003) and signaling (de Jong, 2002; Calder et al., 2010) in using numerical integration of rate equations. Kinetic modeling, however, requires very detailed information on cells. Besides concentration of each component, kinetic information of proteins, cell volume, pH and temperature conditions should be provided for kinetic modeling.

1.3.2 FBA analysis

Flux balance analysis (FBA) involves determination of intracellular fluxes in a biological system, using constraints based on reaction stoichiometry, and extracellular measurements in an optimization framework (Varma and Palsson, 1994; Stephanopoulos et al., 1998). Assuming pseudo-steady state conditions, balances around each intracellular metabolite are written and the equations are solved for the intracellular fluxes. There are different modifications of FBA such as metabolic network flexibility analysis (MNFA)(Wiback et al., 2004; Llaneras and Picó, 2007b) and thermodynamics based metabolic flux analysis (TMFA) (Henry et al., 2007). This procedure involves solving, separately, minimization and maximization linear programming problems for each of the unmeasured fluxes, thereby calculating a flux range for each unmeasured flux, subject to system constraints.
Figure 1-4: Metabolic network representing hepatic metabolism and triadimefon detoxification pathways in primary rat hepatocytes. It includes important pathways such as glycolysis/gluconeogenesis, pentose phosphate pathway, lipid metabolism, urea production, TCA cycle, triadimefon oxidation and subsequent detoxification and reactions involving amino acid metabolism. Metabolites in gray ovals signify extracellular metabolites. For the sake of clarity, dashed lines link metabolites that are connected and occur in different parts of the pathway. The figure is adapted from (Iyer et al., 2010a).

1.4 Pathway based research

The recognition of biological pathway is a critical level of biological organization. Biological pathways capture many of the essential characteristics of the biological system and elaborating biological pathways is advantageous over using single genes, proteins or metabolites for studying living cell because they represent a more robust and complete response (Stelling et al., 2004). High throughput methods for sequence identification, mRNA expression, proteins and metabolites (-omics data) are making it possible to define molecular profiles of diseases and to identify molecular targets for drugs (Brown and Superti-Furga, 2003). However, in many cases the real target is not an individual
protein but a biological process such as uncontrolled cell proliferation (Hopkins, 2007; Hopkins, 2008; Berger and Iyengar, 2009).

Biological pathways have become the hallmarks of both the understanding cancer (Graeber and Sawyers, 2005; Schadt et al., 2009) and development of cancer therapies (Tan and White, 2008; Crawford, 2009; Mastellos, 2009; Neergheen et al., 2009; She and Chen, 2009; Zhu et al., 2009). National Academy of Sciences stated that that human health risk assessment for chemicals is moving toward a pathway based approach (NRC, 2007). As such, pathway analysis can be used to identify key regulatory points of xenobiotic exposure (Iyer et al., 2010a). The details of this study are given in the Appendix section, Appendix A.

Numerous approaches and methods based on biological pathways have been developed in elucidating the mechanisms of the cell. However, we still need techniques that could address biological pathways in more informative ways. One of the main challenges is to develop reliable representations of biological pathways that could help us to understand the dynamics and regulation of biochemical pathways.

1.5 The goal and the objectives of the thesis

“Systems biology is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different. It means changing our philosophy, in the full sense of the term” (Nobel, 2006).

The developments in high throughput studies, the technology and the computational techniques have provided scientists massive information regarding gene function, gene
expression, protein sequence, metabolic and signaling pathways. This information flow definitely advances our understanding of cells’ activities. However, to fully elucidate the biological phenomena, it is necessary to take functional information of all levels of information as well as their interactions. Therefore, integrating different types of information is recognized as a challenge that needs to be addressed to successfully understand the cell’s mechanism (Joyce and Palsson, 2006). Our overall goal is to develop a pathway based approach that integrates different aspects of biological pathways, specifically the structure and the dynamics of a pathway in order to characterize cells’ behavior.

Pathway based research has become an important source of information for understanding the effect of a chemical or pharmaceutical (NRC, 2007) as the information on biological pathways has accumulated. Methods of cross-species extrapolation could aid in establishing whether one biological pathway in a species can be used to build a predictive model for another species (Benson and Di Giulio, 2004). However, cross-species comparison of metabolic pathways is rather complex as there are multiple features of a pathway that can be modeled and compared. We need, therefore, to develop appropriate computational methodologies that will allow to assess structural and functional cross-species comparison of pathways and are presented in Chapter 2.

Transcriptional profiling is an important tool in systems biology that can help us to understand the mechanism of cells (Clauset et al., 2008). Numerous methods and algorithms have been developed to identify significant pathways within gene expression data. One of the significant issues is to be able to separate signal from the noise. Thus, our next objective is to develop more informative methods in order to characterize
pathways’ behavior. In Chapter 3, we formulate a reliable pathway activity metric based on gene expression data so that underlying dynamics within a pathway can be identified and represented.

Previous studies have shown that circadian rhythms are observed at all levels of organization in liver from molecular to the cellular level such as enzyme activity, gene expression, metabolite concentration, DNA synthesis and morphological changes (Davidson et al., 2004). An appreciation of the circadian characteristics of the biological pathways in liver is essential for understanding both the normal physiological and pathophysiological activity of liver. Our next objective is to identify circadian patterns at the pathway level in rat liver. We demonstrate in Chapter 4 that our pathway activity metric, formulated in Chapter 3, can identify the underlying circadian dynamics in biological pathways based on the transcriptional profiling in rat liver.

DBP is one of the most commonly used plasticizers, and the exposure routes include crop cultivation, packaging materials, air, and drinking water (NRC, 2008). In rats, in utero DBP exposure results in developmental toxicity in a number of male reproductive organs, including testes (Mylchreest et al., 1998; Mylchreest et al., 1999; Mylchreest and Foster, 2000; Mylchreest et al., 2002; Barlow and Foster, 2003; Barlow et al., 2004; Lehmann et al., 2004; Kleynenova et al., 2005; Ferrara et al., 2006). Previous analyses of DBP microarray data in the testes (Shultz et al., 2001; Liu et al., 2005; Thompson et al., 2005; Plummer et al., 2007) focused on either individual gene expression changes or changes in the expression of specific genes that are hypothesized, or known, to be important in testicular development and testosterone synthesis. Our final objective is to demonstrate that pathway level analysis enhances our ability to characterize cells’
response to in utero DBP exposure. In Chapter 5, we demonstrate that our pathway activity metric, formulated in Chapter 3, can characterize biologically relevant changes in rat testes after in utero DBP exposure.
Chapter 2

Cross-species Comparison of Metabolic Pathways

Scientists use model organisms to understand underlying biological phenomena. In many cases, approaches to extrapolate results from model organisms and test species to humans are needed. One of the key issues associated with interspecies differences relates to the mechanism of action for a chemical or a drug and determining whether there is evidence that the mechanism(s) is similar or different between the test species and humans. Methods of cross-species extrapolation could aid in establishing whether the potential mechanisms in a species can be used to build a predictive model for another species. With the availability of omics data, increasing systems biology studies and approaches, it is now possible to identify affected pathways and hypothesize possible mechanisms of action from these data (Benson and Di Giulio, 2004).

Phylogenomics analysis, defined here as the comparison of a common feature across a number of species and the reconstruction of a phylogenetic tree, is a bioinformatics approach that enables the identification of evolutionary relationships among different organisms. Commonly, this type of analysis is also used to improve the functional prediction and classification of genes and proteins through the detection of evolutionarily constrained processes, such as conserved, active and functional sites of homologous proteins (Casari et al., 1995), and through the identification of key regulatory elements in non-coding genomic regions (Hardison et al., 1997). Evolutionary and organizational relationships among species may be investigated through multiple sequence alignment of
certain characteristic sequences: sequence of a single protein, single rRNA (Field et al., 1988) or single gene from each organism. In addition, phylogenomics analysis can reveal significant functional information which is not anticipated. For example, a reconstruction of the phylogenetic tree based on the protein sequences of the CYP2A family of cytochrome P450 demonstrated that the rat liver isoform Cyp2a1 has diverged significantly from the human CYP2A6 and mouse Cyp2a4, having a distinct branch of the tree rooted outside of the rest of the family (Searls, 2003). This considerable deviation is associated with a well-known functional shift in which the rat enzyme Cyp2a1 causes coumarin to be metabolized to a hepatotoxic epoxide. On the other hand, the human CYP2A6 and mouse Cyp2a4 act on the same substrate by a harmless hydroxylation (Lake, 1999).

Phylogenomics analyses are not limited to primary sequence data. The same principles can be extended to the whole genome (Fitz-Gibbon and House, 1999; Li et al., 2001). In addition, different levels of biological organizations such as metabolic networks (complete set of metabolic processes of an organism) (Yamada et al., 2004; Parter et al., 2007; Mazurie et al., 2008) and protein interaction networks (Sharan et al., 2005; Liang et al., 2006) and metabolic pathways (specific set of reactions to fulfill a particular function) (Forst and Schulten, 1999; Heymans and Singh, 2003; Clemente et al., 2005; Forst et al., 2006) have also been compared using phylogenomics analysis.

Cross-species comparison of metabolic pathways is not trivial as there are multiple features of a biological pathway that contribute to the function of a metabolic pathway such as pathway structure that incorporates the relations between metabolites and the enzymes, and the sequence information of the enzymes catalyzing the reactions within a
metabolic pathway. Most of the pathway cross-species comparison studies include prokaryotic organisms. This is due to the fact that the available data is far more complete on prokaryotes than higher organisms. However, the amount of information available on metabolic pathways for different organisms is increasing remarkably via both experiments and computational studies. This expansion offers the possibility of performing various analyses on the pathways across different organisms.

In this Chapter, we propose an improvement over the reaction alignment algorithm (Clemente et al., 2007), which is emerged as the most successful pathway comparison method, by including sequence similarity of enzymes.

2.1 Cross-species comparison studies of metabolic pathways

The glycolysis pathway is one of the most studied metabolic pathways in the literature. It is highly conserved evolutionarily and is found in all free living organisms (Dandekar et al., 1999). Five studies (Forst and Schulten, 1999; Forst, 2002; Heymans and Singh, 2003; Clemente et al., 2005; Tun et al., 2006) have used the glycolysis pathway to test their pathway similarity metrics. The main assumption of these studies is that the similarity among different organisms can be studied by analyzing the similarity of their respective glycolysis pathways. Similarly, the patterns of evolutionary changes in amino acid biosynthesis was investigated to derive an empirical estimate of evolution of selected organisms (Rutter and Zufall, 2004). The motivation is that amino acid biosynthesis pathways are highly informative, well characterized both biochemically and genetically, and utilized for similar functions across organisms (Rutter and Zufall, 2004). A broader species comparison study was performed by Oh et al (2006) where they used nine
pathways that belong to the central metabolism pathway (glycolysis/gluconeogenesis, citrate cycle (TCA cycle), pentose phosphate pathway, fructose and mannose metabolism, galactose metabolism, pyruvate metabolism, glyoxylate and dicarboxylate metabolism, propanoate metabolism, butanoate metabolism). These nine pathways are known to generate energy and metabolites that are essential in almost all living cells.

Despite the missing information in metabolic pathway databases, several studies utilize large scale-metabolic data and analyze all the common pathways across different species (Zhang et al., 2006; Clemente et al., 2007). Such an analysis requires the identification of pathway presence in a species. One approach is to use the information retrieved from pathway databases. If there are annotated reactions in a pathway of a species, that pathway is considered present in the species of interest (Clemente et al., 2007). However, the enzymes may be present in a species, yet they may not be part of a successive reaction series. Zhang et al. (2006) addressed this issue by testing the connectivity or completeness in a pathway. First, they calculate the average path length and the diameter of a pathway. The average length of a pathway is calculated as the average number of enzymes in the shortest paths between any two metabolites (Watts and Strogatz, 1998) and the diameter of the graph is calculated as the maximum shortest length in the pathway (Watts and Strogatz, 1998). Then, the diameter of an organism-specific pathway is compared to the average path length of the reference pathway. Both organism-specific pathway and reference pathway are defined by KEGG database and the constituents of an organism specific pathway compose a subset of constituents of the reference pathway. If the diameter of an organism-specific pathway is larger than the average path length of the reference pathway, the pathway is considered as present in the
organism, if shorter; the pathway is considered as absent. If a pathway is defined as present in an organism, the pathway graph in that specific organism must contain at least one series of continuous chemical reactions with relative large path compared to the size of reference graph. On the other hand, if a pathway is defined as absent in an organism, it would at least be very incomplete or unconnected. This analysis, however, is solely based on the available data.

### 2.1.1 Mathematical representation of metabolic pathways

Each model of a metabolic pathway provides its own rendition of reality and takes into account particular aspects while neglecting others (Deville et al., 2003). Models that are used in pathway comparison studies differ in the kinds of interactions that they can represent. Every model presented in this section has its own advantages and enables specific type of analysis.
Figure 2-1: Modeling metabolic pathways and evaluating pathway similarity: 1) A metabolic pathway consists of series of enzymatic reactions to fulfill a particular biological function. 2) The pair wise relations are translated to an incidence matrix. This model describes metabolic pathways as a graph $G(V,E)$ where $V$ is the set of nodes (vertices) and $E$ is the set of edges connecting pairs of nodes (Acton et al., 1996; Deville et al., 2003). 3) Clement et al. (2005) proposed an algorithm that considers the robustness of a pathway by searching the alternative routes in a pathway. Alignment of the metabolic pathways occurs by mapping each non-common compound, enzyme, and reaction in one metabolic pathway to the most similar compound, enzyme, and reaction, respectively, in the other metabolic pathway.
The simplest model of a metabolic pathway embraces the presence and absence of enzymes in a given pathway. In this representation, a vector containing binary information of enzyme presence (1 for presence 0 for absence) is created for each pathway (Forst and Schulten, 1999; Forst, 2002). In this approach, the structure of the pathway is considered implicitly by the presence of an enzyme, but the directionality of the reactions or the relations between reactions is not considered.

A similar approach, the length of a biochemical pathway, has also been used to compare pathways (Rutter and Zufall, 2004). The longer pathways are defined as those that require more enzymes or reactions to convert an initial substrate into the required product than shorter pathways. This characteristic has also been called “pathways distance” or “metabolic distance” (Kolesov et al., 2001). As with the presence/absence approach, the directionality of the reactions is not explicitly considered; however, the sequential nature of the pathway is implicit in the model.

In order to account for the sequential nature of metabolic pathways, pair-wise relations can be translated into an incidence matrix (Deville et al., 2003). This type of model describes metabolic pathways as a graph $G (V, E)$, where $V$ is the set of nodes (metabolites) and $E$ is the set of edges (enzymes) connecting pairs of nodes. The relations between enzymes may be identified in one of two approaches. A relation between enzymes $e_1$ and $e_2$ exists if $e_1$ activates reaction $A \rightarrow B$ (with substrate $A$ and product $B$) and $e_2$ activates reaction $B \rightarrow C$ or $e_2$ activates reaction $A \rightarrow D$ (with substrate $A$ and product $D$). Two enzymes are considered to be related if they activate reactions which share at least one substrate. According to the second approach, a directed relation from enzyme $e_1$ to enzyme $e_2$ exists only if $e_1$ activates reaction $A \rightarrow B$ and $e_2$ activates
reaction \( B \rightarrow C \). Thus, the two enzymes are considered to be related if they activate sequential reactions.

### 2.1.2 Evaluating pathway similarity

Forst et al. (1999) evaluated the similarity between two pathways by incorporating common enzyme sequences and their associated orthology and paralogy information. Orthologous enzymes are encoded by orthologous genes that evolved from a common ancestral gene through speciation; whereas paralogous enzymes are encoded by paralogous genes that evolved by duplication within the genome of a species. The distinction between orthologous and paralogous enzymes is included to account for the fact that paralogs are more likely to have diverged in function than orthologs (Tatusov et al., 1997). However, orthology and the paralogy information are not available for every enzyme in every organism. To evaluate whether two enzymes to be compared are orthologous or paralogous, Forst and Schulten (1999) rely on the sequence similarity. If the sequence similarity between two enzymes in two substitute pathways is defined with an E-value, the expected fraction of false positives, smaller than \( 10^{-4} \) then the enzymes are considered as potential orthologs.

Oh et al. (2006) used a kernel-based algorithm that calculates pathway similarity based on the pathway structure. Their algorithm incorporates directionality into the analysis and searches for possible paths in a given pathway. The possible series of reactions are then used to compare the pathway structure across different species. Zhang et al. (2006) used a similar approach, with the additional incorporation of the number of
shared enzymes and the number the neighbors of each shared enzyme, to evaluate pathway similarity across different species.

Heymans and Singh (2003) also incorporated the directionality of the metabolic reactions to evaluate the similarity between pathways. Their algorithm uses the similarities obtained between every pair of nodes (enzymes) and the relations between neighbor enzymes of every pair of nodes to find the best matching between the two pathways. Subsequently, a similarity score was evaluated by summing the similarity values over the matched nodes.

Alternatively, Clement et al. (2005) proposed an algorithm, the reaction alignment method, which considers the robustness of a pathway by searching for alternative reactions. Their algorithm considers all compounds, enzymes, and reactions present in the metabolic pathways. In this model, alignment of metabolic pathways in different species is achieved by mapping each non-common compound, enzyme, and reaction in one species’ pathway to the most similar compound, enzyme, and reaction, respectively, in the other species’ pathway. The rationale behind this analysis is the following; if one substrate is not available in the environment, then the metabolic pathway may evolve in a way that the required product is produced from another substrate. It has been shown that this reaction alignment algorithm can outperform previous approaches when using NCBI taxonomy as the basis of performance (Clemente et al., 2005; Clemente et al., 2007).

The performance of a pathway similarity algorithm is commonly validated by comparing the reconstructed phylogenetic trees (reconstructed based on the pathway similarity across species) against a tree representing the current understanding of the evolutionary relationships among those species, as reflected in the, NCBI taxonomy.
(Benson et al., 2009; Sayers et al., 2009). The NCBI taxonomy refers to the relationships between species of interest retrieved from (http://www.ncbi.nlm.nih.gov/Taxonomy/). The NCBI taxonomy database is not a primary source for taxonomic or phylogenetic information. Furthermore, the database does not follow a single taxonomic treatise but rather attempts to incorporate phylogenetic and taxonomic knowledge from a variety of sources, including the published literature, web-based databases, and the advice of sequence submitters and outside taxonomy experts. Comparing a reconstructed phylogenetic tree to NCBI taxonomy gives an indication of tree reliability that is conditional on the data and the method (Delsuc et al., 2005). In other words, the reconstructed tree depends on the data and the method. The reconstructed tree can only represent a part of the “true tree.” Reproducing NCBI taxonomy is the most commonly used method to validate a pathway similarity measure in the absence of any functional measurements. However, comparing a reconstructed phylogenetic tree to NCBI taxonomy cannot be used while comparing pathways which are not evolutionarily conserved. Therefore, glycolysis and TCA cycle pathways were used in the literature to assess the performance of pathway similarity measures.

2.1.3 What can be learned by comparing pathways across species

Two broad questions can be addressed by comparing pathways across species. For a known taxonomy one can compare pathways in order to understand the evolution of metabolic pathways. Alternatively, pathway comparison can provide basis for inferring differences in pathways across species.
As an example of the first application, Zhang et al. (2006) calculated different topological parameters such as average pathway length, diameter, and total enzyme number over a number of species. They did not find any obvious correlation between the topological parameters they measured and phylogenetic conservation (based on NCBI taxonomy) except total enzyme number. Furthermore, potential topological variations were not observed particularly in larger pathways. Similar results were observed in (Rutter and Zufall, 2004), where they examined the relationship between pathway length and evolutionary lability of selected organisms. Longer amino acid biosynthesis remained evolutionarily static in their structure, whereas shorter pathways demonstrated greater evolutionary lability. Overall, the conclusion of these studies was that more complex and larger pathways may show slower rates of evolutionary adaptation.

Another approach that is similar in spirit analyzes the phylogenetic extent and conservation of enzymes and pathways across a known taxonomy. (Peregrin-Alvarez et al., 2003) and (Freilich et al., 2008) show that the conserved enzymes participate in the pathways that are related to central metabolism such as sugars, amino acid and nucleotides biosynthesis/metabolism.

Comparing pathways across species is also used to inform the critical steps or key components of a given pathway. Tun et al. (2006) use their pathway comparison method to identify crucial components of the glycolysis pathway. Since their reconstructed phylogenetic tree reflected the evolutionary relationships among 25 species, they concluded that the similarity matrix based on the structure of glycolysis can be used for further applications (Tun et al., 2006). As such, they performed multiple alignments of 11 enzymes of glycolysis that are common in 25 species and constructed a similarity matrix.
for each enzyme. They applied principal component analysis (PCA) on the similarity matrices for both glycolysis pathway and individual enzyme alignments across 25 species. They observed that there was a perfect correlation between the first principal component of the similarity matrices for the glycolysis pathway and the multiple alignments of phosphopyruvate hydratase. The sequence similarity in phosphopyruvate hydratase better predicts pathway similarity compared to sequence similarity in other ten enzymes. Phosphopyruvate hydratase is located at the interface, as a node between the glycolysis and gluconeogenesis pathways and represents the link of the pathway with other biochemical processes such as photosynthesis, amino acid biosynthesis and TCA cycle pathways. It is hypothesized that phosphopyruvate hydratase represents a crucial part of the pathway by integrating the pathway functioning with the global metabolism of the organism. Further, similar relations were discovered in yeast metabolome by the same group (Palumbo et al., 2005). Identification of critical steps of metabolic pathways such as this may allow for characterizing modes of action of potential pharmaceutical or environmental chemical targets (e.g., through the deactivation or activation of particular pathways).

Another application of pathway comparison is to analyze differences in metabolic reactions across species (Forst, 2006; Forst et al., 2006). Forst et al (2006) reconstructed a phylogenetic tree based on the metabolic reactions of 18 microbial pathogens and then analyzed the unique metabolic reactions in one of the sub-trees. They observed that the three species (Pyrococcus horikoshii, Pyrococcus abyssi, Pyrococcus furiosus) represented in the same sub-tree contain species-specific reactions that are absent in other species of the phylogenic tree. The specific set of reactions can be marked as the potential
biomarkers for pathogen identification. Such an analysis can be applied to a variety of problems, for example host-pathogen systems to estimate the determination of organisms as potential pathogens based on their metabolic repertoire. For example, if one species is a known biological threat for a host, then another species with a similar network structure may be considered as a potential threat for the same host. This type of analyses can also be used for xenobiotic threats. When cross-species comparisons include toxicity test species, this information could be applied to human health risk assessment.

2.2 Materials and methods

2.2.1 Reaction alignment algorithm

The reaction alignment of metabolic pathways method maps each non-common compound, enzyme, and reaction in one metabolic pathway to the most similar compound, enzyme, and reaction, respectively, in the other metabolic pathway (Clemente et al., 2005; Clemente et al., 2007). This method calculates the similarity of between two sets of objects $A$ and $B$, in this case enzymes and metabolites, by mapping elements belonging only to $A$ to elements of $B$ ($A \setminus B \rightarrow B \setminus A$), mapping elements belonging only to $B$ to elements of $A$ ($B \setminus A \rightarrow A \setminus B$) and the mapping of common elements to common elements ($A \cap B \rightarrow B \cap A$). Each reaction in the metabolic pathway is aligned with the most similar reaction ($\text{max sim}$) in the same pathway for a different organism. Figure 2-2 presents a simple example how reactions are compared across two organisms.

The similarity of two enzymatic reactions $R = (C, E)$ and $S = (D, F)$, $\text{sim}(R, S)$, where $C, D$ are sets of compounds and $E, F$ are sets of enzymes is given by Eq. (2.1).
\[ \text{sim}(R,S) = \frac{\alpha}{|C \cup D|} \left( \left| C \cap D \right| + \sum_{C \in C \cap D} \max_{D \in D} \text{sim}(C,D) + \sum_{D \in D \setminus C} \max_{D \in D} \text{sim}(C,D) \right) \]
\[ + \frac{1 - \alpha}{|E \cap F|} \left( \left| E \cap F \right| + \sum_{E \in E \setminus F} \max_{F \in F} \text{sim}(E,F) + \sum_{F \in F \setminus E} \max_{F \in F} \text{sim}(E,F) \right) \]

(2.1)

**Figure 2-2: Simple example of the reaction alignment method.** Organism I has one reaction, R1. Organism II has two reactions, R2 and R3. The similarity between each reaction in Organism I and Organism II is evaluated (Eq. (2.1) or Eq. (2.4)). If sim (R1,R2) > sim (R1,R3), then R1 is aligned with R2 (Eq.(2.2)). i) The constituents of reactions in two substitute pathways; reactions include metabolites and enzymes. ii) The similarity between reactions is evaluated by calculating similarity between sets of compounds and the similarity between enzymes.

The similarity between compounds was evaluated by \( \text{sim}(C,D) \) and the similarity between enzymes was evaluated by \( \text{sim}(E,F) \) when two enzymatic reactions were compared, R and S. Each of the compounds and enzymes in reaction R is compared to each of the compounds and enzymes in reaction S.

**Current metabolites**, defined as cofactors in many reactions, namely: H2O, adenosine-5’-triphosphate (ATP), nicotinamide adenine dinucleotide (NAD+), reduced form of nicotinamide adenine dinucleotide (NADH), nicotinamide adenine dinucleotide
phosphate (NADP+), reduced form of nicotinamide adenine dinucleotide (NADPH), oxygen O2, adenosine diphosphate (ADP), orthophosphate, co-enzyme A, carbondioxide (CO2), pyrophosphate, ammonia (NH3), and uridine diphosphate (UDP) are discarded from the analysis as suggested in (Ma and Zeng, 2003; Clemente et al., 2005). Similarity of compounds (metabolites) is 1 for identical compounds and 0 for distinct compounds.

Enzyme similarities, however, are assessed using the hierarchical similarity (Tohsato et al., 2000), which is defined as the number of common most significant Enzyme Commission (E.C.) numbers. The E.C. number is a numerical classification scheme for enzymes based on the reactions the enzyme catalyzes. Every enzyme is characterized by four numbers. Each of the four numbers represents a progressively finer classification of the enzyme. The five possible values of hierarchical similarity are thus: 0, for two dissimilar enzymes (with their first digit different); 0,50, 0.75 and 1, for two identical enzymes (with all four digits identical). The rationale behind hierarchical similarity is to measure similarity of two enzymes in the enzyme hierarchy.

After all the reactions are considered between two species, each reaction in one pathway is aligned with the most similar reaction (max sim) in the other pathway of a different organism. The similarity of two metabolic pathways \( P=(R) \) and \( Q=(S) \), where \( R, S \) are sets of enzymatic reactions, is given by Eq.(2.2).

\[
\text{sim}(R,S) = \frac{\alpha}{|R \cup S|} \left( |R \cap S| + \sum_{R \in S \setminus R} \max_{S \in S} \text{sim}(R,S) + \sum_{S \in S \setminus R} \max_{R \in R} \text{sim}(R,S) \right) \tag{2.2}
\]

where \( \alpha \) is a weight parameter. The \( \alpha \) parameter establishes the relative weight of compound similarity to enzyme similarity in the assessment of enzymatic reaction
similarity. $\alpha = 0.5$ is used to assess the same weight for reaction and enzyme similarity (Clemente et al., 2007).

### 2.2.2 Reaction alignment algorithm including enzyme sequence similarity

The reaction alignment algorithm is reported to be the leading contender to compare metabolic pathways (Clemente et al., 2005). However, this algorithm cannot inform interspecies differences if the reactions are fully conserved across species. To address this issue, we propose to include an index of enzyme sequence similarity in the reaction alignment of metabolic pathways. First, we evaluate the similarity between two protein sequences by using a commonly used alignment and alignment scoring algorithm, “bl2seq”. The bl2seq software, provided by (Altschul et al., 1990; Tatusova and Madden, 1999), is part of the BLAST (Basic Local Alignment Search Tool), and includes the BLOSUM 62 (Blocks of amino acid substitution matrix (Eddy, 2004). bl2seq performs the direct alignment and comparison of two protein sequences, identifies, and scores the differences present between the two sequences. There might be multiple proteins for the same enzyme number (E.C.).

The sequences of enzymes in two substitute pathways of two species, $i$ and $j$, are defined as $PS_{k,i}$ and $PS_{l,j}$, where $k$ and $l$ are the number of sequences associated with the same enzyme number. $\Delta X_{ij}^{ES}$ is the enzyme sequence similarity of enzyme ES between organism $i$ and $j$ in pathway $P$ (given in Eq. (2.3)). The output of bl2seq function is the sequence similarity between two protein sequences. The similarity between same sequence is evaluated as 100, therefore the output of the bl2seq is normalized with 100.
\[ \Delta X_{ij}^{ES} = \max \left( \sum_{i} \sum_{j} bl2seq(PS_i, PS_j) \right) \] (2.3)

Following, the similarity of two enzymatic reactions \( R = (C, E) \), \( S = (D, F) \), where \( C \), \( D \) are sets of compounds and \( E, F \) are sets of enzymes is given by Eq. (2.4).

\[
sim(R, S) = \frac{\alpha}{|C \cup D|} \left[ |C \cap D| + \sum_{C \in C \cap D} \max(E \cap F)_{E,F} \right] + \sum_{E \in E \cap F} \max \left( \sum_{E \in E \cap F} \sum_{F \in F} \max(E \cap F)_{E,F} \right) \]
\[
+ \frac{1 - \alpha}{|E \cup F|} \left[ |E \cap F| + \sum_{E,F \in E \cap F} \max(E \cap F)_{E,F} \right] \] (2.4)

\( \alpha = 0.5 \) is used to assess the same weight for reaction and enzyme similarity. Different than Eq. (2.1), Eq. (2.4) evaluates the enzyme similarity between identical reactions (first term of the enzyme similarity part, \( |E \cap F| \)). Subsequently, Eq. (2.4) is substituted in Eq. (2.2) to evaluate pathway similarity.

### 2.2.3 Phylogenetic reconstruction and evaluation of similarity to NCBI taxonomy

The phylogenetic tree reconstruction begins by evaluating a distance matrix based on a similarity measure. Here, the similarities between species are calculated based on reaction alignment algorithm and reaction alignment algorithm with enzyme sequence similarity. The similarity between species is defined between 0 and 1 (1 being identical, 0 being completely dissimilar), the distances are evaluated by subtracting the similarity from unity (distance = 1 - similarity). A distance matrix is simply a symmetric two-dimensional array containing the pair wise distances for a set of organisms (Figure 2-3). The term ‘distance’ is used as a subjective measure of dissimilarity between organisms. Phylogenetic tree reconstruction is performed by the hierarchical clustering of the entries.
in the distance matrix. As clustering progresses, rows and columns of the distance matrix are merged as the distances in the matrix update with new clusters. We used the MATLAB® seqlinkage function for tree reconstruction employing the complete linkage method.

![Phylogenetic tree reconstruction based on a distance matrix.](http://lectures.molgen.mpg.de/Phylogeny)

**Figure 2-3: Phylogenetic tree reconstruction based on a distance matrix.**

Adapted from: http://lectures.molgen.mpg.de/Phylogeny

### 2.2.4 Pathway information

We use pathway information from nine species: human (*Homo sapiens (hsa)*), mouse (*Mus musculus (mmu)*), rat (*Rattus norvegicus (rno)*), zebra fish (*Danio rerio (dre)*), fruit fly (*Drosophila melanogaster (dre)*), worm (*Caenorhabditis elegans (cel)*), thale cress (*Arabidopsis thaliana (ath)*), yeast (*Saccharomyces cerevisiae (sce)*) and the common bacteria, *Escherichia coli (eco)*.

If there are annotated reactions in a pathway of a species, that pathway is considered present in the species of interest (Clemente *et al.*, 2007). We downloaded the necessary
information from the KEGG database in April 2010 (ftp://ftp.genome.jp/pub/kegg/). The pathway information such as the enzymes, metabolites and enzymes sequences were obtained from KEGG. The data mining and the calculations were performed in Python®.

2.3 Results and discussion

We use similarity based on enzyme sequences along with reaction alignment instead of similarity based on enzyme hierarchy (Tohsato et al., 2000). To test whether this approach improves the method, we compare the glycolysis and citrate cycle pathways across nine selected species. First, we reconstruct the phylogenetic trees based on the reaction alignment algorithm as well as the proposed reaction alignment algorithm with enzyme sequence similarity. Then, we compare the reconstructed tree from reaction alignment algorithm and the proposed reaction alignment algorithm with enzyme sequence similarity to the tree retrieved from the NCBI taxonomy (Benson et al., 2009).

The reconstructed phylogenetic trees for glycolysis pathway retrieved from reaction alignment algorithm and reaction alignment algorithm with enzyme sequence similarity are depicted in Figure 2-4 A and B, respectively. The reconstructed phylogenetic tree of the glycolysis pathway retrieved from the reaction alignment algorithm (Figure 2-4 A) did not inform species differences across human, mouse and rat. Since the glycolysis pathway is fully conserved across human, mouse and rat in terms of the reactions and the enzymes that catalyze the reactions, the similarity retrieved from reaction alignment algorithm does not identify a difference glycolysis pathway in these organisms. Furthermore, the difference between zebra fish and human, mouse, and rat in glycolysis
pathway are the enzymes that catalyze the conversion of alpha-glucose to alpha-glucose-6-phosphate and the reaction that converts beta-glucose to beta-glucose-6-phosphate (KEGG reaction id: R01786). In human, mouse and rat R01786 is catalyzed by E.C: 2.7.1.1 and E.C: 2.7.1.2 whereas in zebra fish R01786 is catalyzed only by E.C: 2.7.1.1. This slight difference is reflected in the reconstructed phylogenetic tree (Figure 2-4 A). When the enzyme sequence similarity is incorporated, the reconstructed phylogenetic tree (Figure 2-4 B) can reflect a difference across human, mouse and rat.

Figure 2-4: Comparison of reaction alignment algorithm and reaction alignment algorithm with enzyme sequence similarity employing glycolysis pathway. A) The reconstructed phylogenetic tree based on reaction alignment and B) reaction alignment with enzyme sequences of glycolysis pathway. Homo sapiens (hsa), Mus musculus (mmu), Rattus norvegicus (rno), Danio rerio (dre), Drosophila melanogaster (dre), Caenorhabditis elegans (cel), Arabidopsis thaliana (ath), Saccharomyces cerevisiae (sce) and Escherichia coli (eco).

Incorporating enzyme sequence similarities into the reaction alignment algorithm may change the reaction alignment between two metabolic pathways. For example, the reaction that converts alpha-D-glucose-6-phosphate to alpha-D-glucose 6-phosphate (R01788) and 2,3-bisphospho-D-glycerate to 3-phospho-D-glycerate (R01516) is not present in thale cress. When the glycolysis pathway is compared between human and
thale cress, R01788 and R01516 in human cannot be aligned with any reaction in thale cress (when Eq. (2.1) is used). This is because the reactions present in human but not present in thale cress, and the reactions present in thale cress but not present in human, do not have any common compounds or enzymes that share any hierarchical similarity. However, when the enzyme sequence similarity is used instead of hierarchical enzyme similarity, R01516 is aligned with the reaction that converts oxaloacetate to phosphopheonol-pyruvate (R00341), and R01788 is aligned with the reaction 2-hydroxyethyl-ThPP to acetaldehyde (R00755). The similarities between the reaction pairs are 0.33 and 0.28, respectively.

The reconstructed phylogenetic trees for the citrate cycle pathway retrieved based on the reaction alignment algorithm and the reaction alignment with enzyme sequence algorithm are depicted in Figure 2-5 A and B, respectively. The citrate cycle pathway is fully conserved across human, mouse, rat, zebra fish, fruit fly, and worm in terms of the reactions and the enzymes that catalyze the reactions. Because of this, the similarity retrieved from the reaction alignment algorithm does not differentiate citrate cycle pathway in these organisms as it can be seen Figure 2-5 A. On the other hand, the reconstructed phylogenetic tree of the citrate cycle pathway retrieved from the reaction alignment algorithm with enzyme sequence (Figure 2-5 B) can inform the species differences across human, mouse, rat, zebra fish, fruit fly, and worm.

Sequence similarity between two proteins may provide an indication of their function similarity (Schlick, 2010). The sequence of a protein is one of the features that determine the tertiary structure of the protein and thus the function of the protein (Eisenhaber et al., 1995; Ruan et al., 2005; Schlick, 2010). Therefore, we propose to use protein sequence
similarity. If a reaction and the enzymes that catalyze that reaction are conserved between two species and if this condition is valid for all the reactions in a pathway, reaction alignment algorithm will decide that there is no difference between two species (Figure 2-4 A and Figure 2-5 A). For the same reaction that is conserved across species of interest, Eq. (2.1) results unity, whereas Eq. (2.4) evaluates sequence similarity between the associated enzymes that is less than unity. The difference of the results between Eq. (2.1) and Eq. (2.4) is reflected in the similarity of two metabolic pathways (evaluated by Eq. (2.2)), and finally in the reconstructed phylogenetic trees of glycolysis pathway and citrate cycle pathway (Figure 2-4 B and Figure 2-5 B).

Figure 2-5: Comparison of reaction alignment algorithm and reaction alignment algorithm with enzyme sequence similarity employing citrate cycle. A) The reconstructed phylogenetic tree based on reaction alignment and B) reaction alignment with enzyme sequences of citrate cycle. Homo sapiens (hsa), Mus musculus (mmu), Rattus norvegicus (rno), Danio rerio (dre), Drosophila melanogaster (dre), Caenorhabditis elegans (cel), Arabidopsis thaliana (ath), Saccharomyces cerevisiae (sce) and Escherichia coli (eco).
To test which of the two methods can better represent the differences in conserved pathways across multiple organisms, we compare the reconstructed phylogenetic trees for the glycolysis and citrate cycle pathways with the phylogenetic tree retrieved from NCBI taxonomy (Figure 2-6). The distances between the reconstructed trees are calculated by TOPD/FMTS software (http://genomes.urv.es/topd/) using the nodal method (see the methods section). The distance is 0 for identical trees, and its value increases as the two trees become more different. For the glycolysis and citrate cycle pathways, the distances between the reconstructed tree retrieved from the reaction alignment algorithm with enzyme sequence similarity and the NCBI tree are evaluated as 0.62 for both, whereas, the distance values between the reconstructed trees retrieved from the reaction alignment algorithm and the NCBI tree are 1.51 and 2.19, respectively.

Figure 2-6: The phylogenetic tree based on the information retrieved from NCBI (Benson et al., 2009) for selected organisms: Homo sapiens (hsa), Mus musculus (mmu), Rattus norvegicus (rno), Danio rerio (dre), Drosophila melanogaster (dre), Caenorhabditis elegans (cel), Arabidopsis thaliana (ath), Saccharomyces cerevisiae (sce) and Escherichia coli (eco).
Additionally, we reconstruct the phylogenetic trees for two additional examples. In the first example, we use cumulative information from nine metabolic pathways (glycolysis, citrate cycle, pentose phosphate pathway, fructose and mannose metabolism, galactose metabolism, pyruvate metabolism, glyoxylate and dicarboxylate metabolism, propanoate metabolism and butanoate metabolism). These nine pathways are part of central metabolism which generates energy and forms the metabolic precursors essential for almost all living cells (Oh et al., 2006). In the second example, we use the cumulative information from 36 pathways (Table 2-1) that are all of the common metabolic pathways in the nine species of this study defined in KEGG. After reconstructing the phylogenetic trees for both examples with each method (Figure 2-7 and Figure 2-8), we compare the reconstructed trees to the tree from NCBI taxonomy. For the 9 pathway and 36 pathway examples, the distances between the reconstructed tree retrieved from the reaction alignment algorithm with enzyme sequence similarity and the NCBI tree are 0.88 and 0.62 respectively; whereas, the distances between the reconstructed tree retrieved from the reaction alignment algorithm and the NCBI tree are 1.20 and 1.15, respectively. Not all the reconstructed trees of 36 pathways exhibit strong similarity to NCBI taxonomy (Table 2-1). Therefore, the cumulative information obtained from 36 pathways may exhibit higher distance to NCBI taxonomy compared to glycolysis and citrate cycle pathways. Overall, reaction alignment with enzyme sequence similarity results in better representation of NCBI taxonomy compared to reaction alignment algorithm (Table 2-2).
The reaction alignment method with enzyme sequence similarity outperforms the original reaction alignment method, because enzyme sequence similarity is a more informative similarity measure compared to hierarchical similarity. For example, if two enzymes only differ by the fourth E.C. number, the hierarchical similarity is represented by 0.75, and this value is the same for every organism pair. However, enzyme sequence similarity implicitly includes species differences. Therefore, the addition of enzyme sequence similarity directly affects the reaction similarity when reactions are compared in two species. Consequently, the pathway similarities are evaluated differently.

Figure 2-7: Comparison of reaction alignment algorithm and reaction alignment algorithm with enzyme sequence similarity employing 9 pathways belong to the central metabolism. A) The reconstructed phylogenetic tree based on reaction alignment and B) reaction alignment with enzyme sequences of glycolysis pathway. Homo sapiens (hsa), Mus musculus (mmu), Rattus norvegicus (rno), Danio rerio (dre), Drosophila melanogaster (dre), Caenorhabditis elegans (cel), Arabidopsis thaliana (ath), Saccharomyces cerevisiae (sce) and Escherichia coli (eco).
Table 2-1 Dissimilarity values between reconstructed phylogenetic tree and NCBI taxonomy retrieved from TOPD/FMTS software. *

<table>
<thead>
<tr>
<th>Pathway name</th>
<th>Reaction Alignment</th>
<th>Reaction Alignment with Sequence Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine, aspartate and glutamate metabolism</td>
<td>1.65</td>
<td>1.18</td>
</tr>
<tr>
<td>Amino sugar and nucleotide sugar metabolism</td>
<td>1.45</td>
<td>0.71</td>
</tr>
<tr>
<td>Arginine and proline metabolism</td>
<td>0.97</td>
<td>0.62</td>
</tr>
<tr>
<td>beta-Alanine metabolism</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Butanoate metabolism</td>
<td>1.51</td>
<td>1.41</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle)</td>
<td>2.19</td>
<td>0.62</td>
</tr>
<tr>
<td>Cysteine and methionine metabolism</td>
<td>0.97</td>
<td>0.75</td>
</tr>
<tr>
<td>Fatty acid biosynthesis</td>
<td>1.78</td>
<td>2.15</td>
</tr>
<tr>
<td>Fatty acid metabolism</td>
<td>1.47</td>
<td>1.27</td>
</tr>
<tr>
<td>Folate biosynthesis</td>
<td>1.86</td>
<td>0.85</td>
</tr>
<tr>
<td>Fructose and mannose metabolism</td>
<td>1.84</td>
<td>1.15</td>
</tr>
<tr>
<td>Galactose metabolism</td>
<td>1.68</td>
<td>1.37</td>
</tr>
<tr>
<td>Glutathione metabolism</td>
<td>1.43</td>
<td>1.43</td>
</tr>
<tr>
<td>Glycerolipid metabolism</td>
<td>1.55</td>
<td>1.03</td>
</tr>
<tr>
<td>Glycine, serine and threonine metabolism</td>
<td>0.97</td>
<td>0.75</td>
</tr>
<tr>
<td>Glycolysis / Gluconeogenesis</td>
<td>1.55</td>
<td>0.62</td>
</tr>
<tr>
<td>Glyoxylate and dicarboxylate metabolism</td>
<td>1.60</td>
<td>1.33</td>
</tr>
<tr>
<td>Histidine metabolism</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Lysine degradation</td>
<td>1.76</td>
<td>1.80</td>
</tr>
<tr>
<td>Nitrogen metabolism</td>
<td>1.91</td>
<td>1.03</td>
</tr>
<tr>
<td>One carbon pool by folate</td>
<td>1.87</td>
<td>1.67</td>
</tr>
<tr>
<td>Pentose and glucuronate interconversions</td>
<td>1.33</td>
<td>1.20</td>
</tr>
<tr>
<td>Pentose phosphate pathway</td>
<td>1.55</td>
<td>1.15</td>
</tr>
<tr>
<td>Phenylalanine metabolism</td>
<td>1.51</td>
<td>1.76</td>
</tr>
<tr>
<td>Phenylalanine, tyrosine and tryptophan biosynthesis</td>
<td>1.51</td>
<td>1.15</td>
</tr>
<tr>
<td>Propanoate metabolism</td>
<td>1.45</td>
<td>1.35</td>
</tr>
<tr>
<td>Purine metabolism</td>
<td>1.22</td>
<td>1.22</td>
</tr>
<tr>
<td>Pyrimidine metabolism</td>
<td>0.97</td>
<td>0.62</td>
</tr>
<tr>
<td>Pyruvate metabolism</td>
<td>0.88</td>
<td>1.20</td>
</tr>
<tr>
<td>Riboflavin metabolism</td>
<td>1.31</td>
<td>0.97</td>
</tr>
<tr>
<td>Sulfur metabolism</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Tryptophan metabolism</td>
<td>1.63</td>
<td>1.73</td>
</tr>
<tr>
<td>Tyrosine metabolism</td>
<td>1.67</td>
<td>1.39</td>
</tr>
<tr>
<td>Ubiquinone and other terpenoid-quinone biosynthesis</td>
<td>2.03</td>
<td>0.62</td>
</tr>
<tr>
<td>Valine, leucine and isoleucine biosynthesis</td>
<td>1.45</td>
<td>1.45</td>
</tr>
<tr>
<td>Valine, leucine and isoleucine degradation</td>
<td>1.84</td>
<td>0.97</td>
</tr>
</tbody>
</table>

*If the reconstructed tree and the tree retrieved from NCBI is the same, the dissimilarity value is 0. As the two trees become more different, the dissimilarity value increases.
Figure 2-8: Comparison of reaction alignment algorithm and reaction alignment algorithm with enzyme sequence similarity employing all 36 common pathways in 9 selected species. A) The reconstructed phylogenetic tree based on reaction alignment and B) reaction alignment with enzyme sequences of glycolysis pathway. Homo sapiens (hsa), Mus musculus (mmu), Rattus norvegicus (rno), Danio rerio (dre), Drosophila melanogaster (dre), Caenorhabditis elegans (cel), Arabidopsis thaliana (ath), Saccharomyces cerevisiae (sce) and Escherichia coli (eco).

Table 2-2 Dissimilarity values between reconstructed phylogenetic tree and NCBI taxonomy retrieved from TOPD/FMTS software.*

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Reaction Alignment With Sequence Information</th>
<th>Reaction Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycolysis</td>
<td>0.62</td>
<td>1.51</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle)</td>
<td>0.62</td>
<td>2.19</td>
</tr>
<tr>
<td>Nine Pathways in Central Metabolism</td>
<td>0.88</td>
<td>1.20</td>
</tr>
<tr>
<td>36 Common Pathways</td>
<td>0.62</td>
<td>1.15</td>
</tr>
</tbody>
</table>

*If the reconstructed tree and the tree retrieved from NCBI is the same, the dissimilarity value is 0. As the two trees become more different, the dissimilarity value increases.
2.4 Summary and conclusions

We propose an improvement over the reaction alignment method (Clemente et al., 2007) in Chapter 2, which is emerged as the most successful to be the best method for pathway comparison, by including sequence similarity along with the reaction alignment. We perform cross-species comparisons for two highly evolutionarily conserved pathways (glycolysis and citrate cycle) using either reaction alignment or reaction alignment with enzyme sequence similarity. Furthermore, we test reaction alignment with enzyme sequence similarity combining the nine pathways that compose the essential central metabolism and all pathways common, 36 pathways, in the nine species. We demonstrate that our reaction alignment method with enzyme sequence similarity results in a better reconstruction of the known evolutionary relationships among these organisms, based on NCBI taxonomy, compared to the standard reaction alignment.

Further improvements can be made by integrating additional features of a metabolic pathway. As such, there is additional information that contributes to an enzymatic reaction such as non-coding region (promoter region) of the gene that code for the enzyme and rate-limiting step of a metabolic pathway and reaction kinetics. Incorporating other properties of metabolic reactions such as stoichiometry, rate and dynamics can potentially yield more comprehensive analysis (Rokhlenko et al., 2007).

Depending on the aim of the study, cross-species comparison analyses of individual pathways, such as the method we describe here, can be used to select the most relevant test species. For example, when a specific metabolic pathway is known to be affected by a xenobiotic or a drug in human or when a biological phenomenon is analyzed, appropriate pathway comparison analysis can be used to decide on a surrogate species for
further testing and analysis. Further, Human health risk assessment for chemicals is moving toward a toxicity pathway based approach that has been proposed by the National Academy of Sciences (NRC, 2007). In this approach, toxicity pathway perturbations would be used as indicators of various types of toxicity. As pathway perturbation assays are typically performed in test organisms, extrapolation of these toxicity pathways to humans is a necessary step. Thus, the methods described here could be applied to a toxicity pathway-based risk assessment paradigm.
Chapter 3

Representing the Activity of a Pathway Based On Gene Expression

High throughput microarray data provide a snapshot of the expression levels of thousands of genes within a cell or tissue sample. Nearly every toxic response is accompanied by changes in gene expression, thus, it is widely accepted that gene expression changes may reflect a toxic response or compensatory mechanisms (Daston, 2008). Despite the increasing number of microarray studies that have been published, interpreting genome-wide transcriptional data remains a major challenge.

DeRisi et al. (1997) showed that the behavior of large groups of functionally related genes can provide broader view of the systematic way in which the yeast cell adapts to a changing environment such as several classes of genes were coordinately induced and decreased by glucose exhaustion.

A widely accepted group of methods for interpreting transcriptional data is driven by the over-representation analysis (ORA) (Khatri and Draghici, 2005). The ORA method first identifies differentially expressed genes (DEGs) by using a statistical filter (Ideker et al., 2000; Tusher et al., 2001; Bar-Joseph, 2004; Storey et al., 2005) and then maps these genes into associated pathways. The ORA of pathways can be performed with a number of statistical models including hypergeometric (Cho et al., 2001) binomial, $x^2$ (chi-square) (van Belle et al., 2004) and Fisher’s exact test (Man et al., 2000). In most cases, the differences between the models are not remarkable. Over-represented pathways are
assumed to be part of the underlying biological response. However, when changes in gene expression are more subtle (observed but not statistically significant), it becomes difficult to differentiate gene expression change from noise due to the large number of genes analyzed, variability between individuals, and limited sample sizes (Tomfohr et al., 2005).

Alternative approaches have been introduced focusing on biochemical pathways (Rison et al., 2002; Allocco et al., 2004) that “act in an orchestrated fashion to mediate the response of a cell toward internal and external signals” (Soyer and Bonhoeffer, 2006). The starting point of such an analysis is that moderate but steady changes in the gene expression levels within a pathway could be missed if relatively few individual genes appear significantly changed. The effectiveness of such an approach was illustrated in a study comparing gene expression profiles in muscle of type 2 diabetics (DM2) relative to non-diabetics by Mootha et al. (2003). No single gene appeared as differentially expressed between samples when the data was analyzed using the ORA approach. Yet, gene-set enrichment analysis (GSEA) revealed a subset of genes involved in oxidative phosphorylation as being differentially expressed. The relationship between oxidative phosphorylation and DM2 is supported by a number of published studies (Mootha et al., 2003). In GSEA, genes in a given pathway are ranked according to the difference in expression using either the t-test or signal-to-noise-ratio). Then, the significance of the enrichment of top-ranked genes is evaluated using maximization of a Kolmogorov-Smirnov running sum statistic.

Another set of methods focusing on pathways (Zien et al., 2000; Hanisch et al., 2002; Rahnenfuhrer et al., 2004; Sohler et al., 2004), considers all gene members using pair-
wise correlation of gene expressions within a pathway and assign a significance value to a given pathway. These methods implicitly use permutation based analysis and apply random gene label assignments as null hypothesis. However, the outcome of these analyses can only determine whether a significant gene expression pattern exists within a pathway, and cannot describe the significant pattern itself.

A further set of studies incorporate the topology of the pathways into aforementioned correlation analysis (Kurhekar et al., 2002; Vert and Kanehisa, 2003; Draghici et al., 2007). The output is an ordered list of significant patterns with corresponding genes. This type of analysis can potentially find new subsets rather than scoring an existing pathway.

Alternatively, Tomfohr et al. (2005) presented a pathway based approach that translates gene expression levels within a pathway into a reduced form, called pathway activity levels, derived from singular value decomposition (SVD). A given pathway represented by pathway activity levels can then be analyzed using the same approaches used for analyzing gene expression levels (Tomfohr et al., 2005). The pathway activity level method can be applied to time-course data, which also includes two different sample groups across different time points.

In Chapter 3, we address two important questions to extend the current state of pathway activity analysis for microarray data. First, can we separate signal from noise; i.e., can we identify pathways that are altered as a result of treatment versus altered due to random variation? Second, can we extend the pathway activity level analysis to time-course data?
3.1 Pathway activity level analysis

In pathway activity analysis, gene expression levels within a given pathway are reduced to pathway activity levels using singular value decomposition (SVD). SVD is a linear transformation procedure to reduce the number of variables that represents the data. The data is linearly transformed, without losing its essence to represent the maximum variability in the data. For example, a point is represented by three coordinates x, y, z. This point can be projected on two of the three coordinates provided that the variability information is kept at maximum. This analysis is suitable for representing gene expression data of a pathway. Instead of monitoring individual gene expressions within a pathway, pathway activity levels capture significant data pattern in the pathway and represents pathway behavior in a single response.

The singular value decomposition (SVD) of $\Xi_P(k,t)$ is given as in Eq.(3.1) where $\Xi_P(k,t)$ is the gene expression data associated with a given pathway $P$ composed of $k$ genes measured at $t$ different conditions (time, samples, etc.), and normalized to have a mean of 0 and a standard deviation of 1.

$$\Xi_P(k,t) = U_P(k,k) \times S_P(k,t) \times V'_P(t,t)$$

(3.1)

The columns of the matrix $U_P(k,k)$ are the orthonormal eigenvectors of $\Xi_P(k,t)$. $S_P(k,t)$ is a diagonal matrix containing the associated eigenvalues, and the columns of the matrix $V'_P(t,t)$ are projections of the associated eigenvectors of $\Xi_P(k,t)$. As the elements of $S_P(k,t)$ are sorted from the highest to the lowest, the first row of $V'_P(t,t)$, represents the most significant pattern within a pathway across different samples. Hence,
Pathway activity levels (PALs) are defined as the first eigenvector $V'_P(t,1)$ of the given in Eq.(3.2).

$$PAL_p(t) = V'_P(t,1)$$ (3.2)

The first column of $U_P(k,k)$ is a vector of weights, one weight for each gene within the pathway. The weights can be positive or negative values indicating the direction of the expression levels with respect to the pathway activity levels. A higher absolute weight of a gene specifies a higher contribution to $PAL_p(t)$.

### 3.2 Significance analysis of pathway activity level

Pathway activity level represents the most significant data pattern within a pathway. However, the pathway activity levels, represented by the largest deviation in the overall gene expressions within a pathway, may be attributed to random deviations in the data (Alter et al., 2000; Cangelosi and Goriely, 2007). To assure that pathway activity levels characterize significant data pattern within a pathway, we evaluate the significance of the data captured by pathway activity levels.

The fraction of the overall gene expression that is captured by $PAL_p(t)$ is evaluated through Eq.(3.3).

$$f_p = \frac{S_P(1,1)^2}{\sum_{g=1}^{g} S_P(g,g)^2}$$ (3.3)

We perform an additional analysis, referred as the significance analysis of $PAL_p$, to evaluate whether $PAL_p$ can represent significant information about the effects of a chemical exposure on a pathway. The significance analysis of $PAL_p$ indicates whether there is significant expression pattern shared by individual genes within a pathway.
(Levine et al., 2006). To evaluate the significance of the $f_P$ value, 10,000 random gene sets of the same size of each pathway are generated from the microarray by permutating the microarray data. Next, the $f_P$ values for the random data sets are evaluated and compared to the actual $f_P$ value. The $p$-value of $f_P$ is computed as the fraction of the $f_P$ of the randomly generated matrices that exceeded the actual $f_P$. If the $f_P$ of the randomly generated matrices exceeds the actual $f_P$ by more than 5%, then the actual $f_P$ is attributed to a random variation in the microarray data ($p$-value < 0.05). Finally, the pathways are filtered based on the associated $p$-value of their $f_P$ value.

3.3 Identifying underlying dynamics within a pathway

We have used two different synthetic data sets to show that

i) pathway activity levels can identify the underlying dynamics within a pathway

ii) significance analysis is necessary for a reliable pathway activity measure

Synthetic data represents a case study where we know and control the underlying data structure so that we can assess the performance of pathway activity level analysis. Our first synthetic data set mimics two different experimental conditions. The purpose is to identify the significant change within a pathway. Our second synthetic data set mimics a time course experiment. The purpose of the second synthetic data set is to identify the underlying dynamics within a pathway. All the computations and the analysis in this chapter are performed in MATLAB.

3.3.1 Comparing two experimental conditions

A hypothetical pathway consisting of 45 genes across $t = 6$ points is constructed following previously described methods (Troyanskaya et al., 2002; Yeung et al., 2003;
Yao et al., 2008). Gene expression values within the synthetic pathway, $g_i$, are generated representing two different treatment groups, as shown in Eq. (3.4):

$$g_i = [(-1 + \alpha \chi_i), (-1 + \alpha \chi_i), (-1 + \alpha \chi_i), (1 + \alpha \chi_i), (1 + \alpha \chi_i), (1 + \alpha \chi_i)]$$

(3.4)

where $\alpha$ is the level of noise and $\chi_i$ is a random variable drawn from a standard normal distribution to create the variability in the data. When the noise level is zero, all 45 genes have the same pattern across the samples. When the noise level is increased, the expression profiles of the individual genes deviate from their initial pattern and converge to random variation.

To compare the outcomes of the pathway activity formulation, differentially expressed genes (DEGs) are calculated by signal to noise (SNR) analysis (Golub et al., 1999). SNR is performed to identify differences in gene expression levels between two different experimental conditions and is defined as the ratio of the mean to standard deviation of gene expression measurements. The difference between the mean values of two set of samples is compared to the summation of the standard deviation of the two set of samples. A high SNR represents a better differentiation between gene expression levels of different treatment groups compared to a low SNR. SNR analysis has been extensively used for the identification of differentially expressed genes (Tu et al., 2002) and feature selection in classification studies (Golub et al., 1999; Shipp et al., 2002; Liang et al., 2004; Goh et al., 2007).
Figure 3-1: The outline comparing two experimental conditions within a pathway. Pathway activity levels are calculated using SVD and the significance of pathway activity levels are evaluated. Subsequently, the identification of the differentiation between two treatment groups of pathway activity levels is evaluated.

To increase our confidence level that a given percentage of the genes within the synthetic pathway are DEGs, the *p-value* is assessed at different noise levels as described earlier by a randomization process. To quantify the effect of the noise level on the individual genes within the synthetic pathway, 1000 replicates of the synthetic pathway are generated at different noise levels. For each generated replicate, the fraction of the DEGs within the synthetic pathway is evaluated and then compared to a given percentage value, i.e. 50%. If the actual the fraction of the DEGs within the synthetic pathway is smaller than the 0.5, the event that 50% of the genes within the synthetic pathway are DEGs is attributed to a random variable. The ratio of the total number of the event that 50% of the genes within the synthetic pathway are DEGs to 1000 identifies the *p-value*. In addition to *p-value* for the event that 50% of the genes within the synthetic pathway
are DEGs, *p-values* for the event that 10% and 90% of the genes within the synthetic pathway are DEGs are evaluated at different noise level.

As the noise level increases, pathway activity levels of the synthetic pathway are evaluated by Eq. (3.1) and (3.2). Furthermore, the significance of PAL of synthetic pathway Eq. (3.4) is evaluated at different noise levels. For each generated replicate, the $f_p$ of the synthetic pathway is compared to the $f_p$ of a random data that is created from shuffling the synthetic pathway. Finally, the *p-value* of the $f_p$ of synthetic pathway is calculated as the fraction of the randomized $f_p$ that exceeded the $f_p$ of the synthetic data.

The differentiation between PALS of different samples is denoted as pathway activity ($PA_p$) and is analogous to the signal to noise ratio (SNR). If $n_1$ samples are associated with a specific treatment and $n_2$ samples with another treatment, then the activity levels associated with each treatment group are given by Eq.(3.5) and Eq.(3.6), respectively.

\[
P_{\text{PAL}_p}^{\text{group1}} = V_p^1 (1 : n_1, 1) \tag{3.5}
\]

\[
P_{\text{PAL}_p}^{\text{group2}} = V_p^2 (n_1 : n_1 + n_2, 1) \tag{3.6}
\]

The evaluation of pathway activity ($PA_p$) is through Eq.(3.7) where $\mu$ and $\sigma$ represent the mean and standard deviation, respectively.

\[
PA_p = \frac{\mu(P_{\text{PAL}_p}^{\text{group1}}) - \mu(P_{\text{PAL}_p}^{\text{group2}})}{\sigma(P_{\text{PAL}_p}^{\text{group1}}) + \sigma(P_{\text{PAL}_p}^{\text{group2}})} \tag{3.7}
\]

Similar to SNR, a high $PA_p$ represents a greater distinction between pathway activity levels between two or among multiple treatment groups. The *p-value* of the event that the synthetic pathway exhibit significant change between different experimental conditions is calculated as following; for each generated replicate, the $PA_p$ of the synthetic pathway is
compared to the $PA_p$ of a random data. Subsequently, the fraction of the randomized $PA_p$ that exceeded the $PA_p$ of the synthetic data determines the $p$-value.

### 3.3.2 Pathway activity of time course data

A hypothetical pathway that consists of 45 gene expressions across $T = 54$ samples (3 replicates at 18 time points) is constructed following previously described methods. The gene expression values within the synthetic pathway, $g_i$, are generated based on a widely accepted model of periodic gene expression (Wichert et al., 2004; Ptitsyn et al., 2006) given in Eq. (3.8)

$$g_i = \beta \cdot \cos(\omega \cdot t + \phi) + \epsilon_i$$  \hspace{1cm} (3.8)

where $\beta$ is a positive constant, $\omega \in (0, \pi)$, $\phi$ uniformly distributed in $(-\pi, \pi]$ where $\epsilon_i$ is a sequence of uncorrelated random variables with mean 0 and variance $\sigma^2$, independent of $\phi$. We assume $\phi = 0$ for all simulated profiles. In order to simulate different signal to noise ratios we also assume the amplitude for baseline variation constant, but add different noise component $\epsilon$ for individual profiles. The $\epsilon$ value for each fraction was taken as a random number $\epsilon_i \in [0.50 \cdot i], i = 0, 1, 2, ..., 100$. When the noise level, $i$, is zero, all 45 genes have the same circadian pattern. As we increase the noise level, the profiles of the individual gene expressions deviate from the circadian pattern and converge to random variation.

To compare the outcomes of the pathway activity formulation, the sinusoidal pattern of a gene expression is approximated using the sinusoidal model $A \cdot \sin(B \cdot t + C)$ (Almon et al., 2008). The coefficients are amplitude (A), frequency (B), and phase (C) of
the model. A non-linear curve fitting algorithm is used to define the parameters of the sinusoidal model that would fit best to the gene expression levels over time. Once a model is built for a given gene expression level, the correlation between the data and the model is the criterion to define the sinusoidal pattern. Genes are characterized as exhibiting sinusoidal pattern if the correlation between the gene expression and the fitted sinusoidal model is equal or greater than 0.8. Similar to the previous case, \textit{p-values} for the events that 10 \%, 50 \% and 90 \% of the genes within the synthetic pathway have sinusoidal pattern are evaluated at different noise levels.

\( \text{PAL}_p \) (Eq. (3.2)) is then applied to describe the pathway activity levels over time. Each entry of \( \text{PAL}_p \) represents the pathway activity level of corresponding experimental condition (\( \Xi_p(k,t) \) and includes replicate measurements at each time point). However, \( \text{PAL}_p \) do not indicate any up- or down-regulation in pathway behavior, instead \( \text{PAL}_p \) evaluates the relative change across different experimental conditions. Additionally, the \textit{p-value} of the event that \( \text{PAL}_p \) (Eq. (3.2)) of the synthetic pathway have sinusoidal pattern and significance of \( \text{PAL}_p \), are evaluated. For each generated replicate, the correlation between \( \text{PAL}_p \), and the fitted sinusoidal model is calculated and compared to 0.8. If the correlation between \( \text{PAL}_p \) and the fitted sinusoidal model of the synthetic pathways is lower than 0.8, the event that the synthetics pathway has sinusoidal pattern is attributed to a random variable. The \textit{p-value} of the event that the synthetic pathway has sinusoidal pattern is computed as the ratio of the total number of the correlation between \( \text{PAL}_p \) and the fitted sinusoidal model is lower than 0.8 to 1000. Furthermore, the significance of \( \text{PAL}_p \) of synthetic pathway is evaluated at different noise levels as described earlier.
Figure 3-2: The outline for identifying significant time-course data pattern within pathways. Pathway activity levels are calculated using SVD and the significance of pathway activity levels are evaluated. Then, the parameters of the sinusoidal model \( A \cdot \sin(B \cdot t + C) \) that would best fit the pathway activity levels are characterized. Finally, the correlation between fitted sinusoidal model and the pathway activity levels is evaluated.

3.4 Results and discussion

To test the hypothesis that pathway activity analysis can identify changes that emerge at the pathway level that cannot be identified at the individual gene expression level, we construct two different synthetic pathways consisting of 45 genes. The genes in the first pathway are representative of two treatment groups, and are generated at different noise levels. Subsequently, we compare the significance of the event when 90 %, 50 % and 10 % of the genes within the synthetic pathway are DEGs. These results are compared with the significance (i.e., the difference in activity between control and treated samples) of \( PA_p \) of the synthetic pathway. A significance value close to unity indicates that the event is highly likely. A typical threshold used to consider the significance of an event is 0.95.
The purpose of this analysis is to evaluate the effect of the noise level on the number of DEGs within the pathway. In addition, the $p$-value of the $PAL_p$, and $PA_p$, of the synthetic pathway is computed at different noise levels (Figure 3-3).

An identical analysis is performed for the second synthetic pathway where the gene expressions are representative of a sinusoidal pattern. This time, we compare the significance of the event when 90 %, 50 % and 10 % of the genes within the synthetic pathway have sinusoidal pattern with the significance of $PAL_p$ having sinusoidal pattern (Figure 3-4A).

From Figure 3-3, we observe that at low noise levels ($0 < \alpha < 0.4$) there is a greater confidence that at least 90 % of the genes within the synthetic pathway are DEGs. However, the confidence level of detecting 90 % of the genes is DEGs decreases sharply as the noise level increases further. At a noise level of 0.7, we can only confidently conclude that even 50 % of the genes are differentially expressed between the two treatments. At higher noise levels (e.g., $\alpha = 1.8$) we cannot even conclude that 10 % of the genes are differentially expressed. Thus, differential expression alone will not be able to provide information about the differences between the two groups at this noise level. However, $PA_p$ predicts with a high confidence level ($p$-value $< 0.001$) that a distinction between the two treatment groups is apparent.
Figure 3-3: Effect of noise level on the differential expression of the synthetic pathway. Gene expression values within the synthetic pathway are generated representing two different treatment groups (given in Eq.(3.1)). The significance of $PAL_p$ and $PA_p$ (denotes the difference between different treatment groups) of the synthetic pathway and the events that 10 %, 50 % and 90 % of the genes within the synthetic pathway are DEGs are illustrated at different noise levels.

Similarly, we observe from Figure 3-4A that at low noise levels ($0 < i < 6$) we are confident that at least 90% of the genes within the synthetic pathway have sinusoidal pattern. At this noise level, the underlying sinusoidal pattern can be identified via both evaluating the individual genes and pathway activity levels. At higher noise levels, i.e. $i = 30$, the significance of the event that 10% of the genes have sinusoidal pattern is low ($p$-value > 0.05). However, pathway activity analysis predicts with high confidence level ($p$-value < 0.0001) that there is an underlying sinusoidal pattern within the synthetic pathway at this noise level ($i = 30$). Therefore, pathway activity levels are more robust.
than the gene expression levels in identifying underlying expression pattern within a
pathway.

A critical issue arises when we consider whether the variation captured by \( PAL_p \) can
represent the overall gene expression within a pathway. We cannot be confident that the
identified dynamics of a pathway is based on biological significance or due to random
variations. To address this issue of random noise in the data vs. real gene expression
changes, we evaluated the significance of the \( PAL_p \) at different noise levels (Figure 3-3
and Figure 3-4B).

From Figure 3-3, we observe that even though \( PAp \) might predict a significant event,
that event could be the result of random variability in the data, as quantified by the
significance of \( PAL_p \). For example, at \( \alpha = 2.5 \), the significance of \( PAp \) is high; however,
the significance of \( PAL_p \) is considerably lower. Also, from Figure 3-4B we observe that
even though \( PAL_p \) might predict confidently a sinusoidal pattern, that event could be the
results of random variability in the data, as quantified by the significance of \( PAL_p \). For
example, at \( i = 10 \), the significance of the synthetic pathway being circadian is high;
however, the significance of \( PAL_p \) is considerably lower. Taken together, these results
indicate that the determining the significance level of \( PAL_p \) is necessary for a reliable
representation of differentially expressed pathways. If the significance of \( PAL_p \) exhibit
high values (\( p\)-value < 0.05) the observed patterns by \( PAL_p \) cannot be solely attributed to
the underlying structure of the data.
3.5 Summary and conclusions

In Chapter 3, we show that pathway analysis enhances our ability to detect relevant changes in pathway activity using two sets of synthetic data. First data set simulates two experimental conditions where we want to identify the significant changes within a pathway. The second data set simulates a time course study where we want to identify the underlying dynamic change within a pathway. We demonstrate that pathway activity formulation can reveal the underlying dynamics within a pathway; however an additional significance analysis is required to represent the changes. Our improved formulation of pathway activity method is a reliable representation of biological pathways. Nevertheless, synthetic data represents a biased approach, where we priori know the underlying
structure. Therefore, the biological relevance of our pathway activity formulation is an important issue to be addressed.
Chapter 4

Circadian Signatures in Rat Liver: From Gene Expression to Pathways

Biological rhythms are important phenomena observed in almost in all living organisms. These rhythms can be categorized under circadian, ultradian and infradian based on the period of the rhythm. Circadian rhythms are 24 hour oscillations in many behavioural, physiological, cellular and molecular processes that are controlled by an endogenous clock which is entrained to environmental factors including light, food and stress (Panda et al., 2002). These oscillations synchronize biological processes with changes in environmental factors thus allowing the organism to adapt, anticipate, and respond to changes effectively (Figure 4-1). Some examples of the biological processes and parameters that show circadian oscillations include body temperature, sleep-wake cycles, endocrine functions, hepatic metabolism and cell cycle progression (Sukumaran et al., 2010). Furthermore, disruption of circadian oscillations is linked to many diseases and disorders including cancer, metabolic syndrome, obesity, diabetes, and cardiovascular diseases.

Circadian oscillators that are present in other parts of the brain and in other organs are referred to as “peripheral clocks” and are controlled by the central master clock. At the molecular level the clock mechanism involves a transcriptional and post-transcriptional auto-regulatory negative feedback loop consisting of BMAL1 and CLOCK transcription factors which form the positive arm and the PERIOD and CRYPTOCHROME
transcription factors which form the negative arm of the feedback loop (Dunlap, 1999; Mirsky et al., 2009). BMAL1 heterodimerizes with CLOCK and binds to E-box elements, thereby transactivating the expression of Per (Per1, Per2 and Per3) and Cry (Cry1 and Cry2) genes. After reaching a critical concentration the proteins encoded by these genes form heterotypic complexes and repress the transcriptional activity of the BMAL1:CLOCK complex. In addition to these core transcription factors, many other transcription factors which are directly regulated by the core factors including REV-ERBs, RORs and PAR-bZip transcription factors are also involved in the regulation of the circadian expression of the transcriptome which in turn regulates various biological processes (Preitner et al., 2002; Gachon, 2007; Jetten, 2009).

Figure 4-1: The suprachiasmatic nucleus (SCN) if located in the hypothalamus in mammals. SCN is synchronized to daylight and communicates the timing information to the body via direct and indirect signals (feeding time, body temperature and circulating hormones)(Ben-Shlomo and Kyriacou, 2002). A group of genes in peripheral tissues bypass the circadian signaling responding light stimulation (Ishida et al., 2005). The figure is retrieved from (Ben-Shlomo and Kyriacou, 2002).
Liver is an important organ that is involved in carrying out a wide variety of critical processes including systemic energy regulation processes, metabolism and detoxification of both endogenous and exogenous compounds and hormonal production (Almon et al., 2008). Liver is the only tissue that stores glucose in the form of glycogen that can be released in response to glucagon or epinephrine to maintain systemic concentrations (Tirone and Brunicardi, 2001). In addition to glucose storage and release, liver can also synthesize glucose de novo through the process of gluconeogenesis. In addition to carbohydrate metabolism, the liver is central to whole body lipid metabolism. About one-half of the cholesterol in the body is produced in the liver, much of which is used for bile acid synthesis (Russell, 1992). Furthermore, liver is the most important organ that is involved in the metabolism of many drugs and hence contributes to the disposition of these compounds from the body (Sukumaran et al., 2010). Proper timing of these processes is of utmost importance for the maintenance of the homeostasis in the system. Previous studies have shown that circadian rhythms are observed at all levels of organization in liver from molecular to the cellular level such as enzyme activity, gene expression, metabolite concentration, DNA synthesis and morphological changes (Davidson et al., 2004). One of the important levels of organization in the cell is biochemical pathways. An appreciation of the circadian characteristics of the biological pathways in liver is essential for understanding both the normal physiological and pathophysiological functioning of liver.

Transcriptional analyses of circadian patterns (Harmer et al., 2000; Panda et al., 2002; Almon et al., 2008) demonstrate that genes showing circadian rhythms are part of a wide variety of biological pathways. The expression of several circadian rhythms in a
single pathway may ensure a tighter circadian regulation of a pathway or be parts of the circadian clock taking place in other biological functions. The issue of this type of analysis, however, is that moderate but steady changes in the gene expression levels within a pathway could be missed if relatively few individual genes appear significant. Consequently, the identification of biological pathways related to circadian phenomenon could be missed. Therefore, we propose to analyze the gene expression data at the pathway level.

4.1 Circadian signatures of pathways in rat liver

To identify the underlying dynamics of microarray data, we follow an unsupervised approach. First, the microarray data is mapped onto a comprehensive database. Next, the pathway activity levels and the significance of the pathway activity levels are evaluated as detailly explained in Chapter 3 through Eq.(3.1)-Eq.(3.3). Finally, significant pathway activity level patterns are retrieved using an unsupervised clustering analysis. The outline of this analysis is depicted in Figure 4-2.

4.2 Materials and methods

4.2.1 Experimental data

Fifty-four male normal Wistar rats (250-350 g body weight) were housed in a stress free environment with light: dark cycles of 12 hr: 12hr. Three animals were sacrificed at each of 18 selected time points within the 24 hour cycle. The time points were 0.25, 1, 2, 4, 6, 8, 10, 11, 11.75 hr after lights on to capture light period and 12.25, 13, 14, 16, 18, 20, 22, 23, 23.75 h after lights on to capture the dark period. To obtain a clear picture of an entire
cycle, two 24 hour periods were concatenated to obtain a 48 hour period. Our research protocol adheres to the ‘Principles of Laboratory Animal Care’ (NIH publication 85-23, revised in 1985) and was approved by the University at Buffalo Institutional Animal Care and Use Committee. The details of the experiment can be found in (Almon et al., 2008). The data is available under the accession number GSE8988 (http://www.ncbi.nlm.nih.gov/geo/).

4.2.2 Pathway database

Pathway annotations of gene expressions are retrieved from the publicly available database; The Molecular Signatures Database (MSigDB) (Subramanian et al., 2005) (Table 4-1). The microarray data is mapped onto 638 defined pathways in MSigDB which is a collection of online databases (given in table), biomedical literature and gene sets compiled from published mammalian microarray studies (Newman and Weiner, 2005).

Table 4-1 Online databases from which MSigDB (Subramanian et al., 2005) is curated.

<table>
<thead>
<tr>
<th>Pathway database name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioCarta</td>
<td><a href="http://www.biocarta.com">http://www.biocarta.com</a></td>
</tr>
<tr>
<td>Signaling pathway database</td>
<td><a href="http://www.grt.kyushu-u.ac.jp/spad/menu.html">http://www.grt.kyushu-u.ac.jp/spad/menu.html</a></td>
</tr>
<tr>
<td>Signaling gateway</td>
<td><a href="http://www.signaling-gateway.org/">http://www.signaling-gateway.org/</a></td>
</tr>
<tr>
<td>Signal transduction knowledge environment</td>
<td><a href="http://stk.scicenmag.org/">http://stk.scicenmag.org/</a></td>
</tr>
<tr>
<td>Human protein reference database</td>
<td><a href="http://www.hprd.org/">http://www.hprd.org/</a></td>
</tr>
<tr>
<td>GenMAPP</td>
<td><a href="http://www.genmapp.org/">http://www.genmapp.org/</a></td>
</tr>
<tr>
<td>KEGG</td>
<td><a href="http://www.genome.jp/kegg/">http://www.genome.jp/kegg/</a></td>
</tr>
<tr>
<td>Gene ontology</td>
<td><a href="http://www.geneontology.org/">http://www.geneontology.org/</a></td>
</tr>
<tr>
<td>Sigma-Aldrich pathways</td>
<td><a href="http://www.sigmaaldrich.com/Area_of_Interest/Biochemicals">http://www.sigmaaldrich.com/Area_of_Interest/Biochemicals</a></td>
</tr>
<tr>
<td>Gene arrays, BioScience Corp</td>
<td><a href="http://www.superarray.com/">http://www.superarray.com/</a></td>
</tr>
<tr>
<td>Human cancer genome anatomy consortium</td>
<td><a href="http://cgap.nci.nih.gov/">http://cgap.nci.nih.gov/</a></td>
</tr>
<tr>
<td>NetAffx</td>
<td><a href="http://www.affymetrix.com/index.affx">http://www.affymetrix.com/index.affx</a></td>
</tr>
</tbody>
</table>
Figure 4-2: The outline for clustering analysis of pathway activity levels. Pathway activity analysis begins with mapping gene expression onto known pre-defined groups of genes, pathways. Subsequently, the pathway activity levels are calculated using SVD and the significance of pathway activity levels are evaluated. Pathways are filtered based on the significance of the PALs. Following, the over-populated patterns are identified by using a consensus clustering approach proposed in (Nguyen et al., 2009). Then, the parameters of the sinusoidal model \( A \cdot \sin(B \cdot t + C) \) that would best fit the centroids of the pathway activity levels (in each cluster) are characterized. Finally, the correlation between fitted sinusoidal model and the centroids of the pathway activity levels in each cluster is evaluated.
4.2.3 Clustering analysis of pathway activity levels

To identify emergent pathway activity level patterns, we apply an unsupervised clustering approach proposed by Nguyen et al. (Nguyen et al., 2009). This approach was applied to detect the significant clusters of co-expressed genes. In this study, we use pathway activity levels instead of gene expression levels.

First, we apply ANOVA (p < 0.01) to remove the pathway activity levels that are not statistically changing across time points. There are three replicates of each measurements; ANOVA analysis ensures that the changes in pathway activity levels occur over time since repeated measurements are averaged for clustering (Yeung et al., 2003). Subsequently, the optimum number of clusters are decided after considering several clustering methods (hclust, diana, kmeans, pam, som, mclust), metrics (Euclidian, Pearson correlation, and Manhattan) and an agreement matrix that quantifies the frequency which two pathways belong to the same cluster based on the pathway activity levels. Then a subset of pathways is selected to ensure that no pathway is present with an ambiguous cluster assignment with any other pathway in the analysis with a confidence level $\delta$. The $\delta$ is the threshold to say whether the agreement level of two pathways belong to one ($\delta$) or two clusters ($1 - \delta$) is consistent or not. The last step is dividing the selected subset into a number of patterns based on the agreement matrix. The details of the algorithm can be found in (Nguyen et al., 2009). In this analysis we use $\delta = 0.65$. 
4.3 Results and discussion

As a result of the significance analysis of PAL, 486 of the 638 defined pathways in MSigDB are considered for further analysis. Having eliminated the pathway activity levels that do not exhibit a significant change over time (ANOVA, \( p\)-value < 0.01), the clustering analysis yielded five significant patterns of pathway activity levels (Figure 4-3).

![Cluster analysis](image)

Figure 4-3: The five significant clusters identified by a consensus clustering analysis (Nguyen et al., 2009) using \( \delta = 0.65 \). The pathway activity level (PAL) of pathways represents the presented curves and the exact reverse curves; PAL= (-) PAL. The signs of PAL are chosen so that PAL has the similar patterns for a better representation and clustering purposes. The centroids of each cluster is shown with the red error bars, the fitted sinusoidal model to the centroids of each cluster is depicted in white.
The emergent pathway activity level patterns appear to have sinusoidal circadian patterns. The significant clusters represent the most populated pathway activity levels patterns within the data, whereas the rest of the data can be associated with random deviations. It can be seen from the Figure 4-3 that the five clusters exhibit circadian patterns. To ascertain that the five clusters actually represent circadian patterns, we use a non-linear curve fitting algorithm to identify the parameters of a sinusoidal model \[ A \cdot \sin(Bt + C). \] The coefficients are amplitude (A), frequency (B), and phase (C) of the model. The frequency of the sinusoidal model identifies the essence of the circadian behavior, which is characterized by one full period in 24 hour. The multiplication of total time (t, 24 hr) and frequency (B) should be equal to \( 2\pi \) in order to characterize one full period (circadian) by the sinusoidal model. Therefore, the fitted models that have the coefficient B between 0.24 and 0.28 are kept for further analysis to assure the circadian dynamics. Once a model is built for a given pathway activity level, the correlation between the pathway activity level and the model is the criterion to define the circadian signature. The correlation between the centroid of each cluster and the associated fitted sinusoidal model exhibits high correlation (correlation => 0.96, given on top of each graph in Figure 4-3).

Evaluating pathway activity levels resulted cases where two pathways have similar fraction of overall gene expression captured by \( PAL_p, f_p \) values, however the associated \( p\)-values, vary significantly. In example, \( f_p \) MAPK Pathway, Nicotinate and nicotinamide metabolism and glycine, serine and threonine metabolism pathway are 0.23, 0.21 and 0.22 respectively (top panel of Figure 4-4). On the other hand, their associated \( p\)-values are rather different; 0.66, 0.12 and 0, respectively (top panel of Figure 4-4).
Figure 4-4: A) The comparison of the fp to the permuted fp for MAPK Pathway, nicotinate and nicotinamide metabolism and glycine, serine and threonine metabolism pathway. The mean and the standard deviation interval of permuted fp are given. The same value of fp can be obtained by randomly permuted data in MAPK Pathway and nicotinate and nicotinamide metabolism, whereas the fp captured by randomly permuted data is much lower compared to fp in glycine, serine and threonine metabolism pathway B) Pathway activity levels and fitted sinusoidal models for the pathways. The mean and the standard deviation interval of the pathway activity levels are given. The correlation between pathways activity level and fitted sinusoidal model is presented for each pathway on top of each graph.
Depending on the size of the pathways, which is number of the genes within a pathway, \( f_p \) value can be obtained from random variations. Therefore, \( f_p \) value itself is not an objective feature to identify whether the information captured overall gene expression by \( PAL_p \) is significant. The significance analysis of \( PAL_p \) enables us to filter out pathways that exhibit circadian rhythms by chance. For example, MAPK pathway and Nicotinate and nicotinamide metabolism may be identified as exhibiting circadian pattern without the significance analysis of \( PAL_p \) because \( PAL_p \) of MAPK Pathway and Nicotinate and nicotinamide metabolism exhibit high correlation with the fitted sinusoidal model (bottom left and bottom middle panels in Figure 4-4).

Glycine, serine and threonine metabolism exhibit both significant \( PAL_p \) and high correlation with the fitted sinusoidal model (top right and bottom right panels in Figure 4-4). To study the effect of individual gene expression on the pathway activity level, we depict the relationship between the weights and the correlation of the individual genes (the correlation between gene expression levels and the fitted sinusoidal model that represent the circadian pattern) in glycine, serine and threonine metabolism pathway Figure 4-5. The weight of a gene characterizes its contribution to the pathway activity level compared to the rest of the genes in the pathway. The weights of the genes are retrieved from the first column of \( U_p(k,k) \) in Eq.(3.1).
Figure 4-5: The relationships between weight and the correlation of the genes within glycine, serine and threonine metabolism. The correlation is between gene expressions and the fitted sinusoidal models and is set to identify circadian genes. The threshold for circadian genes is correlation > 0.8. The weights are evaluated from the SVD analysis. The absolute value of the weights represents the contribution of the individual genes to the pathway activity level. The genes that have higher correlation values have relatively higher absolute weights.

It can be seen from Figure 4-5, that Gldc, Cth, Chka, Chkb, Cbs, Bhmt and Shtm1 exhibit circadian patterns (correlation > 0.8) and also their weights are among the highest (weight > | -0.25|). In addition, the genes, which correlation is slightly under the threshold (correlation ~> 0.7) such as Gatm, Shtm2 and Alas1, have comparably higher absolute weights (weight ~> | -0.25|). The positive and negative values of weights indicate the direction of the gene expression when compared to the pathway activity level. In example, the genes that have negative weights have their peak in the early light period and their nadir in the early dark period (e.g. Chka, Cth), whereas the genes that have...
positive values have their nadir in the early light period and peak in the early dark period (e.g. Shmt1) (Figure 4-6). The pathway activity levels of glycine, serine and threonine metabolism (bottom right panel in Figure 4-4) follow the genes that have the positive weight value (e.g. Chka, Cth) and have its turning point in the early light period. The sign (positive or negative) of the weights can be chosen to represent pathway activity level as pathway activity levels indicate the overall orchestrated significant change in the gene expression within a pathway. Furthermore, we observe that there are genes, which correlation is slightly under the threshold (correlation ~> 0.7) but they have low absolute weights (weight ~< 0) such as Atp6voc and Sardh. The expression pattern of these genes, (as an example we depicted the expression pattern of Atp6voc in Figure 4-6) does not coincide with the rest of the genes that have higher absolute weights, therefore do not contribute to the pathway activity level as much and has low weights.
Figure 4-6: Selected gene expressions within glycine, serine and threonine metabolism. The correlation between the gene expression levels and the fitted sinusoidal models and the weights, which are evaluated via SVD analysis, of the genes are given on top of each graph. The signs (positive and/or negative) of weights indicate opposite direction in the gene expression.

By applying SVD, a number of possible correlated variables (gene expressions) are mapped onto a smaller number of uncorrelated variables (the rows of $V_P(t,t)$ in Eq. (3.1)). Pathway activity is denoted as the most significant data pattern which corresponds to the first row of $V_P(t,t)$ (Eq.(3.2))as the elements of $S_P(k,t)$ are sorted from the highest to the lowest (Figure 4-7). The latter rows correspond to the other patterns which significances are determined with the associated eigenvalues. The matrix $V_P(t,t)$ is orthonormal matrix; therefore the rows represent different data patterns. The two sets of circadian patterns in glycine, serine and threonine metabolism (Figure 4-6) are retrieved
via the first two rows of $V'_{p}(t,t)$. $V'_{p}(t,1)$ and $V'_{p}(t,2)$ have high correlation with fitted sinusoidal model (Figure 4-8). The $p$-value of $V'_{p}(t,1)$ is statistically significant whereas the $p$-value of $V'_{p}(t,2)$ is not statistically significant.

Figure 4-7: The relative values of the associated eigenvalues for glycine, serine and threonine metabolism. The bars indicate the variation in the data captured by each individual eigenvector for glycine, serine and threonine metabolism pathways. The solid line represents the data variability captured by the corresponding eigenvectors when randomly generated data (of the same dimension) were used. No apparent distinction between the actual data and randomly generated data was identified after the first eigenvalue, as quantified by the calculated $p$-values.
Figure 4-8: The most significant data patterns in Glycine, serine and threonine. The first 4 rows of $V_p(t,t)$ that are retrieved from SVD calculations of Glycine, serine and threonine metabolism the elements of $S_p(k,t)$ are sorted from the highest to the lowest. 1) $V_p(t,1)$, 2) $V_p(t,2)$, 3) $V_p(t,3)$, 4) $V_p(t,4)$

Table 4-2 provides the detailed list of identified pathways in each cluster. In total, there are 78 pathways in five clusters. The identification of the circadian signatures at the pathway level identified biologically relevant processes. As such, gene expression, metabolite concentration and enzyme activity in energy metabolism (e.g. glycolysis and gluconeogenesis), amino acid metabolism (e.g. lysine degradation, urea cycle) (Robinson et al., 1981; Froy, 2007), lipid metabolism (e.g. fatty acid biosynthesis) (Akhtar et al., 2002) and DNA replication and protein synthesis (e.g. DNA replication reactome, Purine metabolism) (Schibler, 2003) exhibited having circadian dynamics in mammals liver.
Moreover, we observe series of pathways related to protein synthesis and degradation having circadian patterns. Studies examining the gene expression and enzyme activities related to amino acid metabolism showed persistent circadian rhythms (Davidson et al., 2004). These studies indicate that amino acid metabolism components tend to correlate with food intake. Though no conclusive evidence is available, transport and metabolic substrates of amino acids have shown clock-regulated changes.

Table 4-2 Circadian pathways and associated cluster numbers.

<table>
<thead>
<tr>
<th>Pathway name</th>
<th>Cluster ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCORBATE AND ALDARATE METABOLISM</td>
<td>1</td>
</tr>
<tr>
<td>BUTANOATE METABOLISM</td>
<td>1</td>
</tr>
<tr>
<td>PURINE METABOLISM</td>
<td>1</td>
</tr>
<tr>
<td>LIMONENE AND PINENE DEGRADATION</td>
<td>1</td>
</tr>
<tr>
<td>DNA POLYMERASE</td>
<td>1</td>
</tr>
<tr>
<td>ATP SYNTHESIS</td>
<td>1</td>
</tr>
<tr>
<td>DNA REPLICATION REACTOME</td>
<td>1</td>
</tr>
<tr>
<td>LYSINE DEGRADATION</td>
<td>1</td>
</tr>
<tr>
<td>HISTIDINE METABOLISM</td>
<td>1</td>
</tr>
<tr>
<td>PHENYLALANINE METABOLISM</td>
<td>1</td>
</tr>
<tr>
<td>3 CHLOROACRYLIC ACID DEGRADATION</td>
<td>1</td>
</tr>
<tr>
<td>G1 TO S CELL CYCLE REACTOME</td>
<td>2</td>
</tr>
<tr>
<td>FATTY ACID METABOLISM</td>
<td>2</td>
</tr>
<tr>
<td>BILE ACID BIOSYNTHESIS</td>
<td>2</td>
</tr>
<tr>
<td>UREA CYCLE AND METABOLISM OF AMINO GROUPS</td>
<td>2</td>
</tr>
<tr>
<td>VALINE LEUCINE AND ISOLEUCINE DEGRADATION</td>
<td>2</td>
</tr>
<tr>
<td>TRYPTOPHAN METABOLISM</td>
<td>2</td>
</tr>
<tr>
<td>P53 SIGNALING PATHWAY</td>
<td>2</td>
</tr>
<tr>
<td>CELL CYCLE KEGG</td>
<td>2</td>
</tr>
<tr>
<td>G2 PATHWAY</td>
<td>2</td>
</tr>
<tr>
<td>ARGININE AND PROLINE METABOLISM</td>
<td>2</td>
</tr>
<tr>
<td>RNA POLYMERASE</td>
<td>2</td>
</tr>
<tr>
<td>IFNA PATHWAY</td>
<td>2</td>
</tr>
<tr>
<td>ST TYPE I INTERFERON PATHWAY</td>
<td>2</td>
</tr>
<tr>
<td>POLYUNSATURATED FATTY ACID BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>CELL COMMUNICATION</td>
<td>3</td>
</tr>
<tr>
<td>ANTIGEN PROCESSING AND PRESENTATION</td>
<td>3</td>
</tr>
<tr>
<td>MRP PATHWAY</td>
<td>3</td>
</tr>
<tr>
<td>FRUCTOSE AND MANNOSE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>TYROSINE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>ETC PATHWAY</td>
<td>4</td>
</tr>
<tr>
<td>TYROSINE METABOLISM</td>
<td>4</td>
</tr>
<tr>
<td>MALATEX PATHWAY</td>
<td>4</td>
</tr>
<tr>
<td>PROTEASOME PATHWAY</td>
<td>4</td>
</tr>
<tr>
<td>ALANINE AND ASPARTATE METABOLISM</td>
<td>4</td>
</tr>
<tr>
<td>GLYCOLYSIS AND GLUCONEOGENESIS</td>
<td>4</td>
</tr>
<tr>
<td>SA CASPASE CASCADE</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 4-2 continued Circadian pathways, associated cluster numbers

<table>
<thead>
<tr>
<th>Pathway name</th>
<th>Cluster ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHOLESTEROL BIOSYNTHESIS</td>
<td>5</td>
</tr>
<tr>
<td>GLYCEROPHOSPHOLIPID METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>TERPENOID BIOSYNTHESIS</td>
<td>5</td>
</tr>
<tr>
<td>RNA TRANSCRIPTION REACTOME</td>
<td>5</td>
</tr>
<tr>
<td>BIOSYNTHESIS OF STEROIDS</td>
<td>5</td>
</tr>
<tr>
<td>CIRCADIAN EXERCISE</td>
<td>5</td>
</tr>
<tr>
<td>CYANOAMINO ACID METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>FEEDER PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>GLYCEROLIPID METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>GLYCINE SERINE AND THREONINE METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>METHIONINE METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>LYSINE BIOSYNTHESIS</td>
<td>5</td>
</tr>
<tr>
<td>NUCLEOTIDE SUGARS METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>ETHER LIPID METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>SPHINGOLIPID METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>ONE CARBON POOL BY FOLATE</td>
<td>5</td>
</tr>
<tr>
<td>BASAL TRANSCRIPTION FACTORS</td>
<td>5</td>
</tr>
<tr>
<td>CIRCADIAN RHYTHM</td>
<td>5</td>
</tr>
<tr>
<td>LYSINE BIOSYNTHESIS</td>
<td>5</td>
</tr>
<tr>
<td>LYSINE DEGRADATION</td>
<td>5</td>
</tr>
<tr>
<td>MEF2D PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>METHANE METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>METHIONINE METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>METHIONINE PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>ONE CARBON POOL BY FOLATE</td>
<td>5</td>
</tr>
<tr>
<td>SA G1 AND S PHASES</td>
<td>5</td>
</tr>
<tr>
<td>SELENOAMINO ACID METABOLISM</td>
<td>5</td>
</tr>
<tr>
<td>TID PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>TOLL PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>APOPTOSIS</td>
<td>5</td>
</tr>
<tr>
<td>APOPTOSIS GENMAPP</td>
<td>5</td>
</tr>
<tr>
<td>CARMER PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>EPONFKB PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>FXR PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>G1 PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>GSK3 PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>LEPTIN PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>PS3 PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>RACCCYCD PATHWAY</td>
<td>5</td>
</tr>
<tr>
<td>SA REG CASCADE OF CYCLIN EXPR</td>
<td>5</td>
</tr>
<tr>
<td>TALL1 PATHWAY</td>
<td>5</td>
</tr>
</tbody>
</table>

*) Since gene products can function in multiple pathways, some pathways that may not be active in liver can be identified as circadian. For example small cell lung cancer, SNARE interactions in vesicular transport, prion disease are not defined in liver tissue. For the statistical analysis, we are not biased by the tissue specific pathways; however an additional filtering is performed for the biologically relevant pathways.
In addition, we evaluated the enrichment of the pathways with the genes that exhibited circadian patterns in (Almon et al., 2008). MSigDB database (Subramanian et al., 2005) offers an annotation tool that explore gene set annotations to gain further insight into the biology behind a gene set in question). The end result is a *p*-value indicating the significance of the overlap of the genes with a pathway. The genes that exhibit circadian dynamics in (Almon et al., 2008) have been mapped to 34 pathways nine of which have significant *p*-value < 0.05. Our unsupervised approach identified the entire 34 mapped pathway, whereas nine of mapped pathway exhibited statistically significant enrichment (Table 4-3). Additional biologically relevant pathways were identified by pathway activity level analysis such as pathways related to cell cycle, DNA replication and apoptosis exhibited having circadian dynamics in mammals (Schibler, 2003; Levi and Schibler, 2007). These results indicate that similar to synthetic data in Chapter 3, analysis of biological data emphasizes studying at the individual gene expression levels could miss changes at the pathway level.

**Table 4-3 Enriched pathways by circadian genes** The circadian genes were mapped to canonical pathways provided by http://www.broadinstitute.org/gsea/msigdb/. *p-values* indicate the significance of the overlap of the circadian genes within a pathway.

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Enrichment p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOSYNTHESIS OF STEROIDS</td>
<td>0.0004</td>
</tr>
<tr>
<td>CHOLESTEROL BIOSYNTHESIS</td>
<td>0.0009</td>
</tr>
<tr>
<td>TERPENOID BIOSYNTHESIS</td>
<td>0.0069</td>
</tr>
<tr>
<td>CIRCADIAN EXERCISE</td>
<td>0.0072</td>
</tr>
<tr>
<td>GLYCINE SERINE AND THREONINE METABOLISM</td>
<td>0.0148</td>
</tr>
<tr>
<td>GLYCOlysISPATHWAY</td>
<td>0.0271</td>
</tr>
<tr>
<td>SARSPATHWAY</td>
<td>0.0271</td>
</tr>
<tr>
<td>SELENOAMINO ACID METABOLISM</td>
<td>0.0376</td>
</tr>
<tr>
<td>METHIONINE METABOLISM</td>
<td>0.0432</td>
</tr>
</tbody>
</table>
To further explain the biological significance of the pathway activity level analysis, we studied the coordination between different pathways that is another level of organization in cellular processes, especially in cases where the product of one pathway is the substrate of another pathway. One classic example is the production of bile acids and it needs cholesterol as its starting material. Previous studies have shown that the pathways for steroid and bile acid biosynthesis are coordinated and coupled with cholesterol biosynthesis pathway for maximizing the efficiency of these processes. It has been established that bile acid levels are tightly controlled to ensure appropriate cholesterol catabolism, and promote optimal solubilization and absorption of fat and other essential nutrients (Akhtar et al., 2002; Akhtar et al., 2005). Figure 4-9 shows the fitted sinusoidal models of $PAL_p$ curves for cholesterol and bile acids biosynthesis.

From the Figure 4-9, we could see that both pathways shows circadian rhythmicity with the phase of oscillations for cholesterol biosynthesis with a peak reaching at 15 hours after lights on, but the bile acid biosynthesis pathway shows a slight time lag in its oscillation with the peak occurring at 17 hours after lights on. In the figure, the $PAL_p$ curves reach its peak during the mid-dark period and nadir during the mid-light period. As mentioned previously, the peak and nadir of $PAL_p$ curves represent the maximum variation in the temporal gene expression in the pathway and the exact reverse of the $PAL_p$ curve is mathematically same as the $PAL_p$ curve itself ($PAL_p \equiv -PAL_p$). But from the literature, we know that these pathways peak during the dark period when the animals are actively feeding. Furthermore, the circadian oscillations in expression of many of the genes involved in the pathway (including the rate limiting genes like HMGCR for cholesterol biosynthesis (Russell, 1992) and CYP7A1 for bile acid biosynthesis (Russell
and Setchell, 1992) peaks during the dark/active period in the 24 hours light/dark cycle. **Figure 4-10** and **Figure 4-11** provides the expression of individual genes in these pathways. So to deduce the biological significance of the $PAL_p$ curve, along with the $PAL_p$ curve pattern one should take into account of the oscillation patterns of the individual gene expression (including the rate limiting genes) along with any existing knowledge about the biological function and regulation of a given pathway.

![Figure 4-9: Fitted sinusoidal models of pathway activity levels for cholesterol biosynthesis and bile acid biosynthesis.](image-url)
Figure 4-10: Individual gene expressions in cholesterol biosynthesis. Associated weights and correlations with the fitted sinusoidal model were given on top of each panel.
4.2 Summary and conclusions

In this chapter, pathway activity level formulation enabled us to identify circadian signatures of pathways by reducing the overall gene expression level to a single response. We improved the former formulation of the pathway activity level analysis with an additional significance analysis that enhanced our ability to detect relevant circadian changes and reduce the false positives.
Characterizing the circadian regulation at the pathway level is an important piece of information that helps reveal the complex relationship between environment and circadian clock. The identification of circadian pathways can be used to develop therapeutic strategies. In example, the proper timing of administration of anticancer agents improved the survival of rodents between 2- to 8-fold (Levi, 2002). This phenomenon is a result of the fact that circadian rhythms can modulate cytotoxicity in the cell leading to cell dysfunction or cell death (Levi et al., 2007). Understanding the interplay of circadian behavior between biological pathways that affect the cytotoxicity of anticancer agents in cells such as pathways related to cell cycle, apoptosis and DNA synthesis as well as drug metabolism and detoxification would help to further optimize the timing of anticancer drug administration.

We believe that our analysis of circadian pathways based on transcriptional profiling can contribute to filling the gaps between circadian regulation and biochemical activity. While transcriptional profiling is a valuable tool for unrevealing potential connections between the circadian clock and biochemical activity (Rutter et al., 2002), complementing the transcriptional studies with proteomic and metabolomics analyses will provide new insights to the circadian phenomenon.

This current analysis is limited, as any pathway method, by currently available pathway knowledge. For example, there are two genes, SHMT1 and SHMT2, which have exactly opposite circadian oscillations in gene expression and hence opposite weights. SHMT1 is a cytosolic enzyme and SHMT2 is a mitochondrial enzyme. Though they catalyze the same reaction, the cellular purposes of these enzymes are different. In addition, several genes not linked to known pathways are not considered in pathway
analysis. As more specific pathway databases such as tissue specific pathway databases or cellular compartment specific pathway databases are created and the pathway knowledge databases are improved, the power of this pathway analysis method will increase.
Chapter 5

Pathway Changes in the Testes following in utero Exposure to Di-butyl-phthalate (DBP)

Di-n-butyl phthalate (DBP) is one of the most commonly used plasticizers to introduce flexibility to products containing nitrocellulose, polyvinyl acetate and polyvinyl chloride such as food wraps (NTP, 2003). DBP is also used in cosmetics as a solvent and fixative for perfumes a suspension agent for solids a lubricant for aerosol valves, an anti-foamer, a skin emollient and a plasticizer in nail polish and hair spray (NRC, 2008). The estimates to DBP exposure range from 0.84 to 113 µg/kg/day as determined from urine samples (Blount et al., 2000; Kohn et al., 2000). Women of reproductive age (20-40 years) were found to have significantly higher levels of DBP than other age/gender groups with significantly raising the concern that women of childbearing age may have higher exposure to DBP than the rest of the general population (Blount et al., 2000). NRC states that human exposure to phthalates, including DBP, has considerably increased (NRC, 2008).

There are extensive studies documenting developmental toxicity of DBP and its primary metabolite, monobutyl phthalate, in rodents (Ema et al., 1993; Ema et al., 1994; Ema et al., 1995; Ema et al., 1996; Ema et al., 1997; Ema et al., 1998; Mylchreest et al., 1998; Mylchreest et al., 1999; Ema et al., 2000b; Ema et al., 2000a; Mylchreest and Foster, 2000; Ema and Miyawaki, 2001a; Ema and Miyawaki, 2001b; Mylchreest et al., 2002; Barlow and Foster, 2003; Barlow et al., 2004). DBP exposure to the developing
male rat fetus during a critical window of development in late gestation causes a variety of structural malformations of the reproductive tract (e.g., hypospadias); a decrease in anogenital distance (AGD); delayed preputial separation (PPS); agenesis of the prostate, epididymis, and vas deferens; degeneration of the seminiferous epithelium; interstitial cell hyperplasia of the testis; and retention of thoracic areolas and/or nipples (Ema et al., 1994; Ema et al., 1997; Ema et al., 1998; Mylchreest et al., 1998; Saillenfait et al., 1998; Mylchreest et al., 1999; Ema et al., 2000a; Mylchreest and Foster, 2000; Mylchreest et al., 2002; Barlow and Foster, 2003; Fisher et al., 2003; Bowman et al., 2005; Kleymenova et al., 2005).

**Figure 5-1** shows the studies that were candidates for the development of the reference dose (RfD) presented in the 2006 external review draft IRIS Tox Review for DBP (U.S.EPA, 2006). The point of departure (POD) selected for derivation of the RfD for all exposure durations (acute, short-term, subchronic, and chronic) was the no-observed-adverse-effect level (NOAEL) of 30 mg/kg-d for reduced fetal testicular Testosterone (Lehmann et al., 2004). In this study, a statistically significant decrease in Testosterone concentration in the fetal testis was detected at 50 mg/kg-d.
The reduction in fetal testicular Testosterone is one of the well-characterized MOAs for DBP that occurs after in utero DBP exposure (during the critical window), initiating the cascade of events for a number of malformations in the developing male reproductive tract. Studies using radioimmunoassay of Testosterone levels in fetal testes and studies using RT-PCR, microarrays, and/or immunohistochemical staining found a decrease in the expression of protein and mRNA for several enzymes in the biochemical pathways for cholesterol metabolism, cholesterol transport, and Testosterone biosynthesis (also called
steroidogenesis more generally) in the fetus (Thompson et al., 2004; Thompson et al., 2005; Plummer et al., 2007) (Shultz et al., 2001; Barlow and Foster, 2003; Fisher et al., 2003). Collectively, these studies document that exposure to DBP disrupts Testosterone synthesis in the fetal testis. Thompson et al. (2004) established that following in utero exposure to 500 mg/kg-d, the Testosterone levels in the testes return to normal after the metabolites of DBP are cleared from the circulation. However, the malformations induced by 500 mg/kg-d exposure persist into adulthood (Barlow et al., 2004; Barlow and Foster, 2003). Thus, although the inhibition of Testosterone synthesis can be reversed, the biological effects resulting from the decrease in Testosterone during the critical developmental window are irreversible. Figure 5-2 demonstrates, based on reproducibility of microarray and RT-PCR studies about, the proposed mechanism of DBP exposure in rat testes.
Figure 5-2: The proposed DBP mechanism of action for the male reproductive developmental effects.

The mechanism of action is defined as all of the steps between chemical exposure at the target tissue to expression of outcome. The steps shown are based on male reproductive developmental toxicity and toxicogenomics studies. Some of the affected pathways and individual genes whose expression was significantly affected after DBP exposure in multiple studies are included. By contrast, MOAs are shown in purple letters including the well-characterized MOAa and one example of an unidentified MOA. The figure is adapted from (Shultz et al., 2001; Barlow and Foster, 2003; Thompson et al., 2004; Wilson et al., 2004; Liu et al., 2005) and published in (U.S.EPA, 2009).

Previous analyses of DBP microarray data in the testes (Shultz et al., 2001; Liu et al., 2005; Thompson et al., 2005; Plummer et al., 2007) focused on either individual gene expression changes or changes in the expression of specific genes that are hypothesized, or known, to be important in testicular development and testosterone synthesis. Moreover, not all pathways for the identified significant genes were discussed or presented in detail in the published studies because of this focus. Therefore, we perform pathway analysis of genome-wide transcriptional responses from data provided from two published DBP studies (Liu et al., 2005; Thompson et al., 2005).
We perform pathway analysis of genome-wide transcriptional responses from data provided from two published DBP studies (Liu et al., 2005; Thompson et al., 2005). (Liu et al., 2005) has a comprehensive exposure scenario that covered the critical window for developmental exposure to DBP (GD 12-19), whereas (Thompson et al., 2005) is the time-course study where rats were exposed to DBP for different time intervals before sacrificing on GD 19. Both of these studies exposed rats during the critical window for male reproductive development at the dose of 500 mg/kg-d, a dose which all animals exhibiting one or more phthalate syndrome effects. Thus, the development of the testis of each animal would be expected to be disrupted by treated.

5.1 Testicular dysgenesis due to in utero DBP exposure: implications of exposure

Using data from Liu et al. (2005), we use three different computational methods, enrichment of pathways by DEG, the pathway activity formulation and Gene set enrichment analysis (GSEA), to identify significantly altered pathways relative to control. We use KEGG database to annotate gene expression data. The KEGG database is widely used for studying mammalian gene pathways (Liu et al., ; Kitami and Nadeau, 2002; Bono and Okazaki, 2005; Guo et al., 2006; Kovacs et al., 2007; Singh et al., 2007; Zhou et al., 2007; Fardet et al., 2008; Arum et al.).

5.1.1 Materials and methods

Experimental Data

Pregnant rats were dosed by gavages with either corn oil (1ml/kg/day) alone as a vehicle control or one of 7 phthalates (500mg/kg/day) in corn oil from gestation day (GD) 12 to
In this case study, only vehicle and DBP data were analyzed. The dose level was based on previous studies showing that DBP at 500 mg/kg per day produced significant changes in gene expression in the male offspring without maternal toxicity or fetal death (Shultz et al., 2001; Barlow et al., 2003; Lehmann et al., 2004). All dams were killed on GD 19, the fetuses were removed, and the testes removed from male fetuses. Total RNA was isolated from the testes of 3 pups from different litters. Each RNA sample was labeled and hybridized to Affymetrix GeneChips A and B.

**Computational Methods**

**Enrichment Analysis:**

To compare the outcomes of the pathway activity formulation, DEGs and their associated over-represented pathways are determined. SNR analysis (Golub et al., 1999) is performed to identify differences in gene expression levels between the control and treated samples. *p-values* are evaluated, as described earlier, to assess the statistical significance of the results. For pathways that contain at least 3 genes, the enrichment is evaluated by the 1-tailed Fisher test as stated in (Bluthgen et al., 2005). The formulation of the *p-value* of pathway P is given in Eq.(5.1).

\[
p(P) = \sum_{c=1}^{n11} \frac{R_1! R_2! C_1! C_2!}{N! c!(n_{12} - c)! n_{21}! n_{22}}
\] (5.1)

where

R_1 = Number of total amount of pathways present in the analysis. There are 205 pathways identified for *Rattus Norvegicus* in KEGG.

R_2 = Total number of the pathways of interest

C_1 = Total number of genes that mapped onto the pathways in KEGG.
C_2 = Total number of DEGs
N = R_1 + R_2 = C_1 + C_2
n_{11} = Total number of DEGs in the pathway P
n_{12} = Total number of DEGs
n_{21} = Total number of genes in the pathway P
n_{22} = Total number of genes in the dataset besides the pathway P

The computations are performed in perl.

**Gene set enrichment analysis (GSEA)**

The goal of GSEA is to determine whether any pathway are enriched at the top of a list of genes ordered on the basis of expression difference between two treatment conditions (Mootha et al., 2003; Subramanian et al., 2005) (for example DBP treated and control samples). Genes are ordered on the basis of expression difference using SNR. To determine whether the members of a pathway are enriched at the top of this list, a Kolmogorov-Smirnov (K-S) running sum statistic is computed: beginning with the top-ranking gene, the running sum increases when a gene annotated to be a member of the pathway is encountered and decreases otherwise. The enrichment score (ES) for a single gene set is defined as the greatest positive deviation of the running sum across all the genes present in the analysis. When many genes within the pathway appear at the top of the list, ES is high. The ES is computed for every pathway using actual data, and the maximum ES (MES) achieved is recorded. To determine whether one or more of the pathways are enriched in one treatment condition relative to the other class, the entire procedure is repeated 1,000 times, using permuted data and building a histogram of the maximum ES achieved by any pathway in a given permutation. The MES achieved using
the actual data is then compared to this histogram, providing us with a global $p$-value for assessing whether any gene set is associated with the treatment condition.

We use KEGG database with the implementation tool of GSEA (Subramanian et al., 2005) provided by Broad institute (http://www.broadinstitute.org/gsea/msigdb/downloads.jsp). The threshold for the significance is used as $p$-value $< 0.05$.

**Pathway activity analysis**

Gene expression levels within every pathway are reduced to the pathway activity levels using singular value decomposition (SVD). It is considered that pathway activity levels express the underlying dynamics of a pathway. In Chapter 3, we showed that a significance analysis is necessary for a reliable representation of pathway activity level. We apply this modified pathway activity level analysis to the microarray data as details explained in Chapter 3 through Eq.(3.1)- Eq.(3.3) and Eq.(3.5)- Eq.(3.7). The pathways with statistically significant $P_{A_p}$ (Eq.(3.7)), evaluated through a randomization process and assessing a corresponding $p$-value as described earlier, are referred to as active pathways. The term active pathway indicates an overall difference between treatment groups and does not indicate an increased or reduced gene expression activity in a particular pathway. Pathway activity level computations are performed in MATLAB®.

**5.1.2 Results and discussion**

The outcomes of the three methods, enrichment of DEGs, GSEA and modified pathway activity analysis, are summarized in Table 5-1. Fourteen active pathways were identified by pathway analyses, whereas eleven and fourteen pathways were identified based on the DEG method and GSEA, respectively. The common pathways include previously
identified pathways involved in biosynthesis of steroids (Shultz et al., 2001; Barlow et al., 2003). In addition, (Plummer et al., 2007) reported genes affected by DBP that are related to cellular metabolism (e.g., TCA cycle, glycolysis, pentose phosphate pathway and butanoate metabolism). Furthermore, VLI degradation was identified by the three methods.

Besides the commonly identified pathways, each method identified unique pathways. For example, enrichment of DEGs identified the glyoxylate and dicarboxylate metabolism, whereas fatty acid metabolism and biosynthesis of unsaturated fatty acids were identified only by the pathway activity method. In addition, GSEA identified cell cycle, folate biosynthesis, p53 signaling pathway and urea cycle and metabolism of amino groups. DBP has previously been shown to alter expression of genes related to the PPAR signaling pathway (Thompson et al., 2005) and the fatty acid metabolism pathway (Liu et al., 2005), although the biological relevance of the other identified pathways following DBP exposure remains to be determined. Furthermore, the pyruvate metabolism and butanoate metabolism pathways include genes shared by glycolysis and TCA cycle pathways. Namely, 25 of 39 genes identified in pyruvate metabolism are identified also in glycolysis and TCA cycle pathways. Similarly, 8 of 32 genes identified in Butanoate metabolism are identified in glycolysis and TCA cycle pathways.

The identified active pathways can be ranked based on p-values of identified change. A similar analysis was performed by (Pang et al., 2006). In their study, gene expressions data were collected from left extramural coronary arteries from two groups of dogs, untreated controls and treated with adenosine receptor agonist CI-947, to identify the molecular pathogenesis of drug-induced vascular injury in coronary arteries. The
highest three ranking pathways, LDL pathway during atherogenesis, Msp-Ron receptor signaling pathway, and hypoxia and p53, were related to vascular injury, whereas the significant pathways in the next ranking group were less informative. Their analysis indicated that the highest ranking pathways were the most informative in distinguishing between the healthy and vascular injury groups.

Table 5-1. Summary of the identified pathways

<table>
<thead>
<tr>
<th>Pathway name</th>
<th>PA(p-val)</th>
<th>PAL(p-val)</th>
<th>Enrichment of DEG (p-val)</th>
<th>Enrichment Calculated by GSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reductive carboxylate cycle (CO2 fixation)</td>
<td>0.002</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Valine, leucine and isoleucine degradation</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.002</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle)</td>
<td>&lt;0.001</td>
<td>0.002</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Glutathione metabolism</td>
<td>0.006</td>
<td>0.002</td>
<td>0.000</td>
<td>0.496</td>
</tr>
<tr>
<td>Pentose phosphate pathway</td>
<td>&lt;0.001</td>
<td>0.002</td>
<td>0.008</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Biosynthesis of steroids</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Glycolysis / Gluconeogenesis</td>
<td>&lt;0.001</td>
<td>0.003</td>
<td>0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Butanoate metabolism</td>
<td>0.006</td>
<td>0.004</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Pyruvate metabolism</td>
<td>&lt;0.001</td>
<td>0.004</td>
<td>N/A</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Glyoxylate and dicarboxylate metabolism</td>
<td>0.480</td>
<td>0.012</td>
<td>0.016</td>
<td>N/A</td>
</tr>
<tr>
<td>Biosynthesis of unsaturated fatty acids</td>
<td>0.048</td>
<td>0.012</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Fatty acid metabolism</td>
<td>0.030</td>
<td>0.020</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>C21Steroid hormone metabolism</td>
<td>0.048</td>
<td>0.006</td>
<td>0.012</td>
<td>0.119</td>
</tr>
<tr>
<td>Nicotinate and nicotinamide metabolism</td>
<td>0.068</td>
<td>0.028</td>
<td>0.012</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Propanoate metabolism</td>
<td>0.018</td>
<td>0.030</td>
<td>&lt;0.001</td>
<td>0.490</td>
</tr>
<tr>
<td>PPAR signaling pathway</td>
<td>&lt;0.001</td>
<td>0.042</td>
<td>N/A</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Folate biosynthesis</td>
<td>0.552</td>
<td>0.644</td>
<td>0.195</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>&lt;0.001</td>
<td>1.000</td>
<td>N/A</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ribosome</td>
<td>&lt;0.001</td>
<td>0.868</td>
<td>0.715</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>p53 signaling pathway</td>
<td>0.024</td>
<td>0.866</td>
<td>0.703</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Urea cycle and metabolism of amino groups</td>
<td>0.190</td>
<td>0.016</td>
<td>N/A</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

PA, pathway activity; PAL, pathway activity level; DEG, differentially expressed gene; p-val, p-value.
As it can be seen from Table 5-1, valine, leucine and isoleucine (VLI) degradation and biosynthesis of steroids exhibit the most significant change due to DBP exposure (Table 1). In utero DBP exposure is known to affect the steroid biosynthesis pathway in the fetal testes (Shultz et al., 2001; Barlow et al., 2003; Lehmann et al., 2004; Thompson et al., 2004; Thompson et al., 2005; Plummer et al., 2007) and the VLI degradation pathway produces precursors for biosynthesis of steroids. The pathways related to cellular metabolism and energy production (TCA cycle, pentose phosphate pathway, and glycolysis) exhibit a slightly less significant PA. Pathways associated with fatty acid metabolism and PPAR signaling exhibit the least significant p-values indicating that these pathways are less informative for distinguishing between control and DBP dose groups (Thompson et al., 2005).

Metabolic pathways do not function in isolation and are affected by changes in other pathways and similarly the changes in a pathway affect other pathways. The identification of gene expression changes at the pathway level, thus, enables us to integrate gene expression data and relations between metabolic pathways (Ghazalpour et al., 2005). In the present study, we integrate the statistical outcome of the pathway activity analysis with the known relationships between pathways. We start with the pathway that exhibit the most significant change towards the pathways that exhibit less significant changes querying KEGG database. The metabolic sub-network of the active pathways and their connections via metabolites are depicted in Figure 5-3. Though, we cannot comment on the timing of the critical events, our findings suggest that the pathways that produce precursors to cholesterol synthesis exhibit more significant change compared to the rest of the active pathways.
Figure 5-3: Metabolic sub-network of active pathways. Active pathways are connected to each other via metabolites and illustrated from the most active pathway towards to least active pathway. Data source is (Liu et al., 2005). The connections of the active pathways were retrieved from KEGG (Kanehisa and Goto, 2000), December 2008. The active pathways also have connections to non-active pathways; but only the active pathways are included in this metabolic sub-network.
We explored the contribution of DEGs to the $PA_P$ for a given pathway (Figure 5-4A, B, C, and D). The $PA_P$ of each pathway is calculated by adding genes one-by-one starting with the gene with the highest SNR and adding genes sequentially in the order of their SNR until all genes in the pathway have been added. Figure 5-4 A and B illustrate examples of active pathways, whereas Figure 5-4 C and D are examples of pathways that were not identified as active in our analysis. For pathways that were identified as active or not active, the cumulative $PA_P$ value undergoes a decrease as genes of lower SNRs are added. Yet for the active pathways, the cumulative $PA_P$ remains high enough to be statistically significant. For pathways identified as not active, the cumulative $PA_P$ reaches a low level when all of the genes are added. Accordingly, their $PA_P$ value is not statistically significant. The four pathways are composed of a similar number of genes; therefore, the number of genes in the pathway is not an issue in this comparison. We hypothesize that there is a subset of genes that maintain the $PA_P$ value high enough within active pathways, even when all genes are added. The cumulative behavior of this subset enables us to differentiate the active and non-active pathways. Differentially expressed genes in active pathways are defined as “informative genes”. We identified a relatively small number of genes as informative, and these may represent genes that DBP has most greatly affected.
Figure 5-4: The relationship between differential expression of individual genes and PA using the Liu et al. (2005) DBP data. The $PA_P$ of a given pathway is first evaluated using the gene that has the highest SNR. Subsequently, the genes are added in the order of their SNR, from highest to lowest. Pathways identified as active for DBP, such as biosynthesis of steroids (A) and butanoate metabolism (B), maintain high $PA_P$ values even when all genes in the pathway are added. Alternatively, pathways not identified as active for DBP such as pentose and glucuronate interconversions (C) and ether lipid metabolism (D), exhibit a decrease in $PA_P$ as the less discriminating genes (i.e., those with a lower SNR value) are added. The Y axis is the $PA_P$. 
Figure 5-5: A gene network created using ingenuity pathway analysis (IPA). The figure illustrated that the interactions among informative genes following in utero DBP exposure in the rat testis from (Liu et al., 2005). Interactons among genes are derived from the IPA database. Genes or gene products are represented as nodes. Diomands, enzymes; horizontal ovals, transcription regulators; Squares, cytokines; rectangles, nuclear receptors. Solid lines represents direct relationships between nodes (i.e. molecules that physical contact with each other such as binding or phosphorylation).Dashed lines represent indirect interactions (i.e. not requiring physical contact between the two molecules such as signaling events.)

5.2 Testicular dysgenesis due to in utero DBP exposure: implications of duration

Using the time-course microarray data from Thompson et al. (2005), we addressed the question of adapting the pathway activity level method to data with multiple conditions.
5.2.1 Materials and methods

*Experimental Data*

Pregnant rats were dosed with vehicle (corn oil 1ml/kg) or DBP (500mg/kg) at either 1, 3, 6 or 18 h prior to sacrifice on GD 19 for microarray experiments. Thus, the shortest duration exposure of 1 h began on GD 18; the longest duration exposure began on GD 18 and continued until GD 19. Then the fetuses were removed and the testes from male fetuses were removed. Total RNA was isolated from individual testes and each RNA sample was labeled and hybridized to Affymetrix GeneChips A.

*Pathway activity levels*

Modified pathway activity analysis is applied to describe the pathway activity levels over time as dentally explained through Eq.(3.1)- Eq.(3.3) and Eq.(3.5)- Eq.(3.6). Each entry of the $\text{PAL}_p$ represents the pathway activity levels of the corresponding experimental condition ($\overline{\Xi}_p(k,t)$ (Eq.(3.1)) includes both control and treated samples). However, the absolute values of the $\text{PAL}_p$ do not signify the direction of effect on a pathway (i.e., negative or positive values of $\text{PAL}_p$ do not indicate any up or down regulation for a given pathway). Instead, $\text{PAL}_p$ provides a measure of the relative difference across samples exposed for different durations.

To detect the statistical difference among pathway activity profiles of different treatment groups over time, EDGE (http://www.genomine.org/edge/), an open source software package based on a method specifically designed for time course experiments (Storey *et al.*, 2005; Leek *et al.*, 2006), is used. EDGE evaluates a *p*-value that provides a
5.2.3 Results and discussion

To identify significant pathways, we first calculated pathway activities for both control and DBP treated samples (Eq.(3.1)- Eq.(3.3) and Eq.(3.5)- Eq.(3.6)) and then evaluated the statistical significance of the difference between control and DBP treated activity profiles. The pathways that exhibit significant differentiation between control and DBP treated activity profiles, which are illustrated in Figure 5-6, are referred as active pathways.

This analysis identified pathways known to be biologically relevant to DBP exposure and the fetal testes, including steroid metabolism and biosynthesis of steroids and pathways related to energy metabolism (e.g., TCA cycle, glycolysis, pyruvate metabolism, butanoate metabolism). In addition, Metabolism of xenobiotic by cytochrome P450 and glutathione metabolism were identified. Furthermore, five pathways involved in biosynthesis of secondary metabolites and vitamins (terpenoid biosynthesis, glycan structure degradation, coenzyme-A biosynthesis, synthesis and degradation of ketone bodies and nicotine and nicotinamide metabolism) were also identified. The terpenoid biosynthesis and synthesis and degradation of ketone bodies pathways, contain a relatively small number of genes and are in fact subparts of other biologically relevant pathways. For example, 5 of the 6 genes in the terpenoid biosynthesis pathway are also found in the biosynthesis of steroids pathway and 5 of 7
genes in the synthesis and degradation of ketone bodies pathway are also found in the fatty acid metabolism and VLI degradation pathways.

**Figure 5-6: Comparison of control and DBP-treated pathway activity profiles over time.** The control (dashed line) and DBP (solid line) treated profiles of active pathways. The dots are the replicate information for the associated experimental condition and time point. Pathway activity levels of associated pathways do not indicate up-regulation or down regulation, however they represent relative change across different experimental conditions. Most of the pathways exhibit largest differentiation between control and DBP treated profiles at the 6th and 18th hour of exposure, whereas few pathways exhibit differentiation at the 3rd hour of DBP exposure such as glycosaminoglycan degradation and glutathione metabolism.
Second, this analysis allowed for a comparison of a pathway activity levels across time (Figure 5-6). Most of the pathways exhibit largest differentiation between control and DBP treated profiles at the 6\textsuperscript{th} and 18\textsuperscript{th} hour of exposure, whereas few pathways exhibit differentiation at the 3\textsuperscript{rd} hour of DBP exposure such as glycosaminoglycan degradation and glutathione metabolism. In addition, glycolysis pathway, pyruvate metabolism pathway, tryptophan metabolism pathway and biosynthesis of steroids pathway showed a greater affect after 18 hours of DBP exposure compared with 6 hours of DBP exposure. Thus this analysis provides information about effects on individual and groups of pathways changes over time.

However, we did not observe any significant differences between control and treated pathway activity profiles from 1 and 3 hours of DBP exposure. At the individual gene level, the expressions of 6 genes after 1 hour of e DBP exposure and 50 genes after 3 hours of DBP exposure were identified as significantly altered using DEG (Thompson et al., 2005). By contrast, the analysis described herein is a pathway level effect and not individual gens. Thus, it is possible that the affected genes at 1 and 3 hours did not constitute a comprehensive biological response such as metabolic and signaling pathways that are defined in KEGG.

We used the PAL method described earlier to identify biologically active pathways at each time point. We evaluated the informative genes at each time point and the resulting preliminary temporal gene network, based on the Thompson et al. (2005) data, is shown in Figure 5-7. The analysis showed a preponderance of signaling pathways such as JAK/STAT, PPAR, and MAPK perturbed at the earlier exposure durations. After the longest DBP exposures (18 hours), the metabolic pathways, including amino acid
metabolism, lipid metabolism, and carbohydrate metabolism, were affected. Thompson et al. (2005) hypothesized that the decrease in Testosterone level after a short duration of DBP exposure might be due to cholesterol unavailability and their findings support this hypothesis. To have a complete understanding of the temporal sequence of gene expression and pathway affect events after in utero DBP exposure, data from an exposure-duration series across the entire critical window of exposure are needed.

Figure 5-7: A temporal gene network model over time created by IPA from the informative gene list based on time-course data after in utero DBP exposure (Thompson et al., 2005). The informative genes were evaluated at each time point and mapped onto a global molecular network using the Ingenuity Pathways Knowledge Base. Diamonds, enzymes; Horizontal ovals, transcription regulators; Squares, cytokines; Rectangles, nuclear receptors. Solid lines represent direct relationships (also called edges) between nodes (i.e., molecules that make physical contact with each other, such as binding or phosphorylation). Dashed lines represent indirect interactions (i.e., not requiring physical contact between the two molecules, such as signaling events). CP, Canonical pathway. Low (green), down regulated expression with respect to control. High (red), unregulated expression with respect to control.
5.3 Summary and conclusions

There is concern that the traditional pathway analysis method, of first identifying differentially expressed genes and then as a second step performing pathway mapping, might result in a loss of information. In addition, identifying changes in a pathway based on a number of DEGs, which is the basic premise of ORA, can generate both false positives and negatives and therefore can potentially misrepresent the effect of a xenobiotic. These results are in agreement with the simulations performed in Chapter 3.

For use in risk assessment and in any study, it is important to be able to separate signal from the noise. Thus, the pathway activity level method considers all of genes in a pathway and provides more confidence when evaluating microarray data. The modified pathway activity method enabled us to detect relevant changes in pathways compared to ORA such as fatty acid metabolism and biosynthesis of unsaturated fatty acids were identified only by the pathway activity method. Both adapted pathway activity level method and GSEA identified biosynthesis of steroids and pathways related to central metabolism whereas adapted pathway activity level method identified steroid hormone metabolism and fatty acid metabolism that were not identified by GSEA.

The pathway activity method is adapted to analyze multiple groups over time. With this adaption, we could address the pathway level analysis of transcriptional data with both control and treated profiles across different time points. This adaptation demonstrates biologically relevant changes after in utero DBP exposure. Testosterone concentration in the fetal testis was diminished by 50% within 1 hour and remained decreased for 24 hours (Thompson et al., 2005). Pathway activity analysis over time
reveals that pathway activity levels of certain biological functions accompany testosterone decrease such as steroid hormone metabolism and biosynthesis of steroids.
Chapter 6

Conclusions and Future Perspectives

6.1 Summary

Biochemical pathways “act in an orchestrated fashion to mediate the response of a cell toward internal and external signals”(Soyer and Bonhoeffer, 2006). The significance of the pathway based research is recognized to elucidate the mechanisms of the cell (Schaefer, 2004). One of the main challenges is to develop reliable representations of biological pathways that could help us to understand the dynamics and regulation of biochemical pathways. Our overall goal is to develop a pathway based approach that integrates different aspects of biological pathways, specifically the structure and the dynamics of a pathway in order to characterize cells’ behavior.

We propose an improvement over the reaction alignment method (Clemente et al., 2007) in Chapter 2, which is emerged as the most successful pathway comparison method, by including sequence similarity along with the reaction alignment. We test the reaction alignment algorithm with sequence similarity for two highly evolutionarily conserved pathways (glycolysis and citrate cycle). We demonstrate that our reaction alignment method with enzyme sequence similarity results in a better reconstruction of the known evolutionary relationships among these organisms, based on NCBI taxonomy, compared to the standard reaction alignment.

In Chapter 3, we show that pathway activity analysis of gene expression data enhances our ability to detect biologically relevant changes in pathway activity using two
sets of synthetic data. We demonstrate that pathway activity formulation can reveal the underlying dynamics within a pathway; however an additional significance analysis is required to represent the changes. The issue of the biological relevance of the pathway activity formulation is addressed in Chapter 4, where we show that the pathway activity formulation can identify the underlying circadian dynamics in the microarray data with an unsupervised approach and biologically relevant results were obtained. We believe that our analysis of circadian pathways based on transcriptional profiling can contribute to filling the gaps between circadian regulation and biochemical activity in liver. Finally, we show that pathway activity formulation can characterize the significant changes in cells after in utero DBP exposure in Chapter 5. Our results suggest that that the pathways that produce precursors to cholesterol synthesis exhibit more significant change compared to the rest of the significant pathways. Moreover, the pathway activity method enabled us to evaluate pathway changes over time based on gene expression data so that we could identify the difference between two treatment profiles.

The achievements of this thesis are summarized below:

- Addition of enzyme sequence similarity improved reaction alignment algorithm by resulting in a better reconstruction of the known evolutionary relationships among selected organisms, based on NCBI taxonomy. Further improvements can be made by integrating additional features of a metabolic pathway.

- We demonstrated that pathway activity formulation can reveal the underlying dynamics within a pathway; however an additional significance analysis is required to represent the changes. new formulation of pathway activity
method is a reliable representation of biological pathways. This formulation can be used to identify mode of action of drugs and xenobiotics as well as to understand the underlying biological phenomena.

- Pathway activity formulation enabled us to identify circadian signatures in pathways in rat liver. Characterizing the circadian regulation at the pathway level is an important piece of information that helps reveal the complex relationship between environment and circadian clock.

- Pathway activity formulation enabled us to address the pathway level analysis of transcriptional data with DBP treated and control profiles over time. Pathway activity analysis over time reveals that pathway activity levels of certain biological functions accompany testosterone decrease after in utero DBP exposure, which is one of the well-characterized MOAs for DBP, such as steroid hormone metabolism and biosynthesis of steroids.

6.2 Future directions

Due to the fact that transcriptomics provides a global view of cell’s state, it is not feasible to map the underlying biological phenomena onto every gene. However, pathway activity analysis translates gene expression levels into pathway activity levels, and consequently cell’s state is represented by pathway activity levels. Comparisons of pathway activity levels across species integrated with cross-species pathway comparison methods could yield great insights into underlying biological phenomena.

In addition, to incorporate individual genes to i.e. pharmacodynamics models, either empirical or mechanism based, is simply not feasible. Therefore, modeling cellular
processes was proposed by (Kremling et al., 2000). Producing the emergent phenotype of the cell via pathway activity levels and appropriate mathematical representation would enable us to predict i.e. different experimental conditions such as different dose and knock-out experiments. Ideally, a possible novel approach is to integrate every layer of information (gene function, gene expression, protein sequence, metabolic and signaling pathways) (Gatzidou et al., 2007).

It is important to realize that computational methods cannot to substitute biological experiments, however they can be used to rationalize the decision making process. One of the broader purposes of this research is to serve as a bridge between experimental observations of microarray data and cross-species approaches. Such an effort will enable researchers to compare cell’s response based on microarray data in different species. We believe that the advances in technologies and data analyzing techniques, which will improve to monitor the cell’s molecular response, can be used to extrapolate across species, so that the resources would be spent more efficiently; i.e. the cost of the experiments and the number of sacrificed animals will diminish greatly.
Bibliography


Barlow, N. J., Foster, P. M. 2003. Pathogenesis of male reproductive tract lesions from gestation through adulthood following in utero exposure to Di(n-butyl) phthalate. Toxicol Pathol 31, 397-410.

Barlow, N. J., McIntyre, B. S., Foster, P. M. 2004. Male reproductive tract lesions at 6, 12, and 18 months of age following in utero exposure to di(n-butyl) phthalate. Toxicol Pathol 32, 79-90.


Ema, M., Miyawaki, E. 2001a. Adverse effects on development of the reproductive system in male offspring of rats given monobutyl phthalate, a metabolite of dibutyl phthalate, during late pregnancy. Reprod Toxicol 15, 189-194.


Kleymenova, E., Swanson, C., Boekelheide, K., Gaido, K. W. 2005. Exposure in utero to di(n-butyl) phthalate alters the vimentin cytoskeleton of fetal rat Sertoli cells and disrupts Sertoli cell-gonocyte contact. Biol Reprod 73, 482-490.


Llaneras, F., Picó, J. 2007b. A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. BMC Bioinformatics 8, 421.


Mylchreest, E., Sar, M., Cattley, R. C., Foster, P. M. 1999. Disruption of androgen-regulated male reproductive development by di(n-butyl) phthalate during late gestation in rats is different from flutamide. Toxicol Appl Pharmacol 156, 81-95.


procedure and example from nucleotide metabolism. Journal of Mathematical biology 45, 153-181.


Thomas, C. E., Ganji, G. 2006. Integration of genomic and metabonomic data in systems biology - are we 'there' yet? Curr Opin Drug Disc 9, 92-100.


Appendix: Transcriptional and Metabolic Flux Profiling of Triadimefon Effects on Cultured Hepatocytes

Conazoles are triazole containing fungicides used in both agriculture to protect crops and pharmaceutics to prevent and treat fungal infections. Conazoles have the ability to block lanosterol 14-alpha-demethylase (CYP51), which catalyzes the synthesis of ergosterol (Georgopapadakou and Walsh, 1996). Blocking of the synthesis of ergosterol, an essential component of the fungal cell membrane, hinders the viability of the fungus. In addition, the main target of conazoles, CYP51, is the only enzyme in the cytochrome P450 family which is present across various biological kingdoms such as vertebrates and prokaryotes (Yoshida et al., 2000; Lepesheva et al., 2003). Because of their ability to inhibit the cytochrome P450 dependent enzymes involved in the biosynthesis of steroid hormones in mammalian cells, conazoles may produce hepatic, thyroid and developmental toxicity (Wolf et al., 2006; Goetz et al., 2007). The identified modes of action of conazole toxicity are conserved across rat and mice and are believed to be relevant to risk assessment for human health (Allen et al., 2006; Hester et al., 2006; Goetz and Dix, 2009c).

Myclobutanil, propiconazole and triadimefon are examples of conazoles possessing a 1,2,4-triazole ring. Despite all three conazoles being triazole-based, and all known to be CYP51 inhibitors, their toxic responses in rats are quite different. Triadimefon was found to be tumorigenic in rats and mice, propiconazole was a rat tumorigen while myclobutanil was neither a rat or mouse liver tumorigen (Hester et al., 2006). Triazole conazole detoxification in rodent liver has been predominantly studied by gene expression studies
(Hester et al., 2006; Ward et al., 2006; Hester and Nesnow, 2008; Goetz and Dix, 2009b; Goetz and Dix, 2009a) and more recently metabolite profiling (Ekman et al., 2006). Many of these studies were performed in vivo and over a long time frame (Hester et al., 2006; Tully et al., 2006; Goetz and Dix, 2009a). Recently, a microarray study over a 3 day period was performed using cultured hepatocytes with low concentrations of conazoles (≤ 0.1 mM), and exhibited very few changes in probe sets due to conazole exposure (Goetz and Dix, 2009b).

A greater understanding of xenobiotic toxicity at the cellular level is facilitated by the use of a cell culture model that approximates the in vivo functions of hepatocytes. In vitro primary rat hepatocyte cultures using collagen sandwich have been routinely used for studying drug metabolism (Kern et al., 1997; Lau et al., 2002; Plant, 2004), for inflammatory studies (Zupke et al., 1998; Jayaraman et al., 2005) and towards improving their application in bioartificial livers (Chan et al., 2002; Chan et al., 2003a). The collagen sandwich configuration mimics the organization of the liver sinusoid, allowing the cultured hepatocytes to attain polarity and maintain structural integrity, and stable differentiated functions for periods up to two months (Dunn et al., 1991). Such an in vitro culture system also facilitates quantification of intracellular fluxes in hepatic metabolism under various stimuli (Chan et al., 2003a; Chan et al., 2003b). Metabolic flux, defined as the rate of conversion of biochemical molecules in a metabolic network (Stephanopoulos et al., 1998) has been characterized as an important critical parameter, and along with the metabolite concentrations define a basic determinant of cellular physiology (Stephanopoulos, 1999).
Flux balance analysis (FBA) is a method that helps in quantification of intracellular fluxes and describing cellular states using an optimization framework (Varma and Palsson, 1994). This approach was first applied to unicellular organisms for which growth is the primary objective but more recently also to mammalian systems including the liver (Sharma et al., 2005; Nolan et al., 2006; Nagrath et al., 2007). A variation of FBA called metabolic network flexibility analysis (MNFA) uses thermodynamic constraints and can be applied to explore the effect of xenobiotics on the interaction between hepatic metabolism and detoxification pathways. Quantification of the intracellular fluxes can be performed in conjunction with genome wide transcriptional profiling, which represents the physiological state of the cell at the mRNA level. The interplay between metabolic flux and gene expression eventually affects homeostasis, and direct cell fate thus enabling the emergence of particular pathogenic phenotypes. The complete characterization of the two end points of cellular processes, metabolic flux and gene expression, can be used to identify key regulatory points of xenobiotic exposure.

Because triadimefon has been found to be tumorigenic in rats and mice, it was chosen in this study. Freshly isolated primary rat hepatocytes were exposed to two different concentrations (0.3 mM and 0.15 mM) of triadimefon for three days. Subsequently, gene expression and metabolic flux data were analyzed to demonstrate the effect of triadimefon on metabolism.
A.1 Materials and Methods

A.1.1 Hepatocyte isolation and culture

Hepatocytes were isolated from adult male Fisher F344 rat (150-200g) based on the two-step collagenase-perfusion method via the portal vein as described previously (Dunn et al., 1991). Cells were cultured on a collagen sandwich at a density of 1x10^6 cells/mL in six-well plates. Hepatocytes were cultured for 4 days in standard hepatocyte C+H medium, which consists of DMEM supplemented with 10% heat inactivated fetal bovine serum (FBS), 7.0 ng/mL glucagon, 7.5 g/mL hydrocortisone, 20 g/L epidermal growth factor, 200 U/mL penicillin, 200 g/mL streptomycin, and 500 mU/L insulin. Medium was changed daily. Triadimefon (Sigma, St. Louis, MO) was dissolved in sterile DMSO (Fisher Scientific, Pittsburgh, PA) and stored in aliquots in -20°C for further use. After 4 days, hepatocytes were treated to C+H medium containing one of the following media for another 3 days: 0.3 mM triadimefon, 0.15 mM triadimefon and DMSO control. Dilutions were performed in DMSO-containing media so that the total concentration of DMSO was constant (0.04 % v/v with respect to the C+H medium) for all three treatments.

Media was changed daily and supernate was subsequently collected and analyzed for metabolite measurements. There were three replicates at each time point. LDH toxicity assay (Roche Diagnostics, Indianapolis, IN) was carried out daily to account for cell death. The total cell count was performed using calcein AM cell viability assay (Invitrogen, Carlsbad, CA) on day 3 to cross-check the cell count. For microarray analysis, cells were harvested at three time points 6, 24 and 72 hours, respectively, after triadimefon treatment. The cells were immediately frozen in liquid nitrogen at these time
points and stored in the -80 deg freezer for further analysis. There were four replicates at each time point.

**A.1.2 RNA Extraction and Subsequent Microarray Experiments**

Tissues were homogenized in 1 mL of TRIzol reagent (Invitrogen, Carlsbad, CA) using the GenoGrinder 2000 homogenizer. Samples were centrifuged at 12,000g for 10 min at 4°C and supernate was incubated for 5 minutes at 30°C for complete dissociation of nucleoprotein complexes. Fresh 0.2 mL of chloroform per 1 mL of TRIzol reagent was added and centrifuged at 12,000g for 15 minutes at 4°C. The aqueous phase was transferred to a new tube with 70% ethanol and further purified with Qiagen RNAeasy columns (Qiagen, Valencia, CA) with DNase treatment according to the manufacturer’s recommendations. The RNA was eluted in 30 µL DNase/RNase-free water. RNA quality was assessed by capillary electrophoresis using the Agilent Bioanalyzer 2100 and spectrophotometric analysis. All samples passed the quality criteria with $A_{260}/A_{280}$ ratios greater than 1.8 and no signs of RNA degradation observed in the electropherograms. Fifty nanograms of total RNA from each sample was amplified to cDNA, fragmentated and biotinylated using the Nugen kits WT Ovation Pico and FL Ovation Biotin (NuGen, San Carlos, CA). The labeled samples were hybridized to Affymetrix GeneChip® Rat Genome 230 2.0 Array according to manufacturer’s recommendations for hybridization, washing and scanning.

**A.1.3 Microarray Data Analysis:**

To increase the confidence that a measurement is accurate at a certain experimental condition, the variation of the replicates were evaluated. First, the standard deviation of
the replicates was normalized by the mean \( \left( \frac{\sigma}{\mu} \right) \) and compared to randomly permutated microarray data at each experimental condition. If the normalized standard deviation \( \left( \frac{\sigma}{\mu} \right) \) of the randomly generated matrices exceeded the experimental normalized standard deviation more than 1 of the 10000 trials, then that measurement was attributed to a random variation in the microarray data (p-value < 0.0001). The probe sets that exhibited higher p-value than the significance level (p-value < 0.0001) at any experimental condition were eliminated.

A.1.4 Clustering Analysis

The aim of the clustering analysis is to identify the subset of probe sets that exhibit the same dynamic response. The dynamic characteristics of the probe sets were evaluated by using trajectory clustering method (Phang et al., 2003). The general idea of trajectory clustering algorithm is to look at the transitions between any pairs of temporal measurements. This algorithm not only considers adjacent time-points but also overall change in case a gene is gradually increasing in expression, such that no adjacent time points are significantly different. When such a change exists, the trajectory clustering algorithm assigns the change to the appropriate time interval. The details of the trajectory clustering algorithm can be found in (Phang et al., 2003). In the present study, a p-value-based test was used instead of a rank-based test to use information from all of the replicates, rather than just averaging them together. Intervals that did not contain significant differential expression were called constant (C), while those that were significantly different were labeled either up-regulated (U) or down-regulated (D). This
resulted in a symbolic representation of the probe sets with \( n-1 \) letters, where \( n \) is the number of time points in the experiment.

Gene expression data were collected for three different treatment conditions; DMSO control, 0.15 and 0.3 mM triadimefon at three time points after treatment 6\(^{th}\), 24\(^{th}\) and 72\(^{nd}\) hours respectively. The transcriptional profiles of a probe set were translated to their corresponding symbolic representation according to the treatment condition separately. So, each experimental condition is characterized by 2 symbolic representations. At the end, the transcriptional profile of a probe set was identified by 6 symbolic representations. Each two consecutive symbols in this representation correspond to the time intervals for DMSO control, 0.15 and 0.3 mM triadimefon treatments, respectively. For example, a symbolic representation of DC CC CC corresponds to the following; the transcriptional DMSO control profile was down regulated between 6\(^{th}\) and 24\(^{th}\) hour and exhibited no change between 24\(^{th}\) and 72\(^{nd}\) hours, whereas the transcriptional profiles of both 0.15 and 0.3 mM triadimefon treated profiles exhibited no change between 6\(^{th}\) and 24\(^{th}\) hour and between 24\(^{th}\) and 72\(^{nd}\) hours. The profiles of the probe sets that exhibited the same behavior for all treatment conditions (e.g. DC DC DC) were excluded from the analysis, and the probe sets that have the same symbolic representation were clustered together. The outline of the clustering algorithm is illustrated in Figure A.-1.

The identified clusters can be present in random data, i.e. randomly permutated microarray data. A significance value can be assigned to the clusters based on their population size to identify the clusters that characterize non-random data (Foteinou et al., 2009). Hence, microarray data were permutated 1000 times and a distribution for the cluster sizes that corresponded to random data was generated. The randomly generated
cluster size distribution was then compared to the original cluster size distribution and the clusters that were likely to be generated by a random model were filtered out. Specifically, the $p$-value of each cluster size for random data was modeled by Eq. (A.1)

$$p-value = e^{-\lambda x}$$  \hspace{1cm} (A.1)

where $x$ is the cluster size and $\lambda$ was estimated from the distribution of random dataset. The significance threshold used corresponds to $p$-value of 0.0001.

Figure A.-1: The outline of the clustering algorithm. First the transcriptional profiles of a probe set were translated to their corresponding symbolic representation according to the treatment condition separately. Subsequently, the probe sets that have the same symbolic representation were clustered together.
A.1.5 Functional analyses of microarray data

The functional analyses of the significant gene groups were performed through the use of Ingenuity Pathways Analysis (IPA, Ingenuity® Systems, www.ingenuity.com). To identify signaling and metabolic pathways those were overrepresented by the gene groups, the library of pathways provided by IPA was used. The significance ($p$-value $< 0.05$) of the association between the data set and the canonical pathway was measured by Fisher’s exact test. In addition to the library of pathways provided by IPA, IPA-TOX® database was used. IPA-TOX® provides functional gene groupings and toxicity lists based on critical biological processes and key toxicological responses such as adaptive, defensive, or reparative responses to xenobiotic insult. Moreover, IPA-TOX® provides toxicity functions that link genes to possible clinical pathology endpoints for identification of mechanism of toxicity. The significance ($p$-value $< 0.05$) of toxicity functions was calculated by Fisher’s exact test.

A.1.6 Computational Identification of Transcriptional Regulators

Given the assumption that co-expressed genes could be governed by some common regulatory mechanism (Roth et al., 1998; Tavazoie et al., 1999; Altman and Raychaudhuri, 2001), it is hypothesized that this regulation process is mainly controlled by the interplay between transcriptional factors (TFs) and their corresponding transcription factor binding sites (TFBSs) on the non-coding regions (promoters) of the target genes (Lemon and Tjian, 2000; Levine and Tjian, 2003; Maston et al., 2006).

The promoter regions of the genes in the analysis were extracted from the comprehensive database Genomatix (www.genomatix.de) that provides the promoters
and alternative transcripts of the corresponding gene. A promoter can be assigned to several alternative transcripts within a locus. Genomatix marks the alternative transcripts in the order of relevancy to the promoter of each gene. Thus, a better representation of the main promoters for subsequent analysis was retrieved. If there was no associated prior length information of a promoter, a default suggested by Genomatix (500bp upstream plus 100bp downstream the TSSs) was used. MatInspector (Cartharius et al., 2005) was then applied to scan for PWM matches on those promoter sequences using optimal parameters from MatBase (www.genomatix.de). The details of the computational analysis can be found in (Nguyen et al., 2010). The transcription factors that were found in more than 70% of the promoters of the co-expressed genes were identified as significant.

A.1.7 Extracellular metabolite measurements

The levels of various metabolites were measured both in the media and supernate to determine the extracellular fluxes (uptake or production). Glucose, Triglyceride (TG) and glycerol levels were measured using commercial available kits (Sigma, St. Louis, MO). Urea was determined using a kit available from Fisher Scientific (Pittsburgh, PA). Albumin was quantified using an enzyme-linked immunosorbent (ELISA) assay described previously (Dunn et al., 1991; Yang et al., 2009a). Lactate was assayed using the kit available from Trinity Biotech Berkeley Heights, NJ). β-hydroxybutyrate was analyzed using a commercial kit (Fisher Scientific, Pittsburgh, PA) and acetoacetate was analyzed using the protocol adapted by Chan et. al (Chan et al., 2003a). Cholesterol levels were measured using the commercially available EnzyChrom kit (Bioassay
Systems, Hayward, CA) and were found to be negligible. The non-esterified fatty acids (NEFA) in the supernate were measured using a colorimetric assay based commercial kit (Roche Diagnostics, Indianapolis, IN). The levels of 19 amino acids (Asp, Glu, Gly, Arg, Thr, Ala, Pro, Tyr, Val, Met, Lys, Ile, Leu, Phe, Ser, Cys, Orn, Asn, His) and NH$_4$ were measured using HPLC and the AccqTag method (Waters, Milford, MA), whereas glutamine was quantified using a kit from Sigma (St. Louis, MO).

Varian Saturn 2000 GC/ITMS was used to analyze triadimefon and triadimenol. Chromatographic separation was achieved on a Varian 3400CX gas chromatograph equipped with a 30 m capillary (DB-XLB) column, with a 0.32 mm ID and 0.25 µm film thickness. Helium was used as a carrier gas with flow rate 1 mL/min. Injector start temperature, 100 °C, was held for 0.5 min then increased to end temperature, 260 °C, at a rate 150 °C/min. The temperature was then held for 10 min at 260 °C. GC column temperature program was used: initial oven temperature was 45 °C, kept for 3 min then ramped to 320 °C in 3 steps. Detection was carried out in SIS mode (Selective Ion Storage) with ions m/z 57, 181 for triadimefon and m/z 168, 112 for triadimenol.

A.1.8 Metabolic network flexibility analysis (MNFA)

Metabolic flux is defined as the net rate of conversion of a precursor metabolite to a product in a metabolic pathway. Metabolic flux characterizes the end point of functional changes in a cell and can be used in comparing genetic and environmental variants. As such, it has been identified as one of the most critical parameters and a basic determinant of cellular physiology (Stephanopoulos, 1999). Due to the lack of detailed kinetic information (e.g.: enzyme kinetics, metabolite concentrations), an alternative approach
called flux balance analysis (FBA) has been extensively applied for the quantification of metabolic fluxes.

Flux balance analysis (FBA) involves determination of intracellular fluxes in a biological system, using constraints based on reaction stoichiometry, and extracellular measurements in an optimization framework (Varma and Palsson, 1994; Stephanopoulos et al., 1998). Assuming pseudo-steady state conditions, balances around each intracellular metabolite are written and the equations are solved for the intracellular fluxes. The fluxes in rat hepatocytes are considerably higher than the actual changes in the intracellular metabolite concentration, thereby indicating negligible accumulation (Chan et al., 2003a; Yang et al., 2009b). Hence, we believe that the steady-state assumption is justified over each 24 hour time period. In our study a modified approach called metabolic network flexibility analysis (MNFA) was used to quantify the intracellular fluxes. This procedure involves solving, separately, minimization and maximization linear programming problems for each of the unmeasured fluxes, thereby calculating a flux range for each unmeasured flux (Wiback et al., 2004; Llaneras and Picó, 2007b), subject to system constraints as explained below (Eq.(A.2)).

\[
\begin{align*}
\text{Max/Min} & \quad v_j \\
\text{s.t.} & \quad \sum_j S_{ij} v_j = 0 \quad i \in M \\
& \quad v^\min_j < v_j < v^\max_j \quad j \in K \\
& \quad \sum_j \Delta G^0_p \cdot v_j \leq 0 \quad p \in P
\end{align*}
\]

where \( v_j \) is the reaction rate of reaction \( j \); \( S_{ij} \) is the stoichiometric coefficient of metabolite \( i \) in reaction \( j \). The fluxes \( v^\min_j \) and \( v^\max_j \) are the lower bound and upper bound of
constrained reactions, respectively; $M$ is the set of metabolites; $N$ is the total number of reactions involved in the hepatic network; $K$ is the set of constrained reactions (based on measurements and/or irreversibility), $P$ is the number of pathways and $T$ is the set of unknown reactions. The matrix of elementary pathways weighted by Gibbs energy of reactions is denoted by $\Delta G_p$ ($P \times N$ dimension).

The main assumptions for the development of the MNFA model are as follows: (1) The internal metabolites are assumed to be maintained at pseudo-steady state, which means their rate of change is small compared to their turnover; (2) The constraints for irreversible reactions, $v_j \geq 0$, are imposed based on the information from KEGG; (3) The value of each measured flux was constrained by an interval $[v_{\min}, v_{\max}]$ corresponding to their average and standard derivation of triplicate measurements; (4) Pathway energy balance (PEB) constraints were added to reduce and more accurately describe the feasible range of intracellular fluxes. The standard Gibbs free energies of metabolites ($\Delta G_i^0$) were estimated using a group-contribution method (Mavrovouniotis, 1991) and used with the reaction stoichiometry to calculate the free energy of reaction, $\Delta G_{\text{RXN}}^0$.

We used the standard Gibbs free energy of reaction in our calculations. Previous calculations with a hepatocyte network have indicated that exclusion of the activity term in the Gibbs free energy has a very minor effect on metabolic thermodynamic constraints (Nolan et al., 2006). Elementary modes were computed by generating extreme vectors of pointed convex cones (Schuster et al., 2000; Schuster et al., 2002), implemented in Matlab software Fluxanalyzer (Klamt et al., 2003). The matrix of elementary modes $E$ ($P \times N$ dimension) is multiplied by the Gibbs energy of each reaction, element-by-
element, to form a matrix of pathway Gibbs free energies of reactions, written as $\Delta G_p$ ($P \times N$ dimension) in Eq. (A.3).

$$\Delta G_p^0 = E \cdot \Delta G_{\text{RXN}}^0$$  \hspace{1cm} (A.3)

By weighting the pathway energy ($\Delta G_p$) and flux ($v$), pathway energy constraints are written as in Eq. (A.4).

$$\sum_j \Delta G_p^0 \cdot v_j \leq 0 \hspace{1cm} p \in P$$  \hspace{1cm} (A.4)

Statistical analysis between treatments was performed using analysis of variance (ANOVA) followed by Tukey’s student range test, performed with SAS software (SAS Institute, Inc., Cary, NC).

The metabolic network constructed for the hepatic metabolism in conjunction with triadimefon detoxification pathways is illustrated in Figure 1-4. The metabolic network for hepatic metabolism consisted of important pathways of glycolysis/gluconeogenesis, pentose phosphate pathway, lactate production, malate to pyruvate anaplerotic reaction, tricarboxylic acid (TCA) cycle, urea cycle, albumin production, lipid metabolism, glycogen metabolism, electron transport chain reactions, and amino acid metabolism reactions (Chan et al., 2003a; Yang et al., 2009b). Additionally, the detoxification pathways of triadimefon included oxidation to triadimenol, 1-(4-chlorophenoxy)-4-hydroxy-3,3-dimethyl-1-(1H-1,2,4-triazol-1-yl)-2-butanone (kwg1323) and $\beta$-(4-chlorophenoxy)-$\alpha$-(1,1-dimethylethyl)-1H-1,2,4-triazole-1-ethanol (desmethyl kwg1342) respectively. Subsequently, desmethyl kwg1342 was conjugated to 1-(4-chlorophenoxy)-3,3-dimethyl-1-(1H-1,2,4-triazol-1-yl)-2-hydroxybutane-4-O-glucuronic acid (desmethyl kwg1342-ga) and 1-[(4-chlorophenoxy)(2,2-dihydro-2,2-dihydroxy-5-methyl-1,3,2-
dioxathian-4-yl)methyl]-1H-1,2,4-triazole (desmethyl kwg1342-sulf). kwg1323 was conjugated to 1-(4-chlorophenoxy)-3,3-dimethyl-1-(1H-1,2,4-triazol-1-yl)-2-oxo-butanol-4-O-glucuronic acid (kwg1323-ga) (FAO and WHO, 2007). In total, there were 90 reactions (Table A.1), 55 metabolites and 33 extracellular fluxes Using the aforementioned network, MNFA was performed for all three days at every experimental condition and the fluxes are calculated.

Table A.1 Reaction stoichiometry for central hepatic metabolism along with triadimefon detoxification pathways. Reactions 1 to 32 indicate reactions relating to extracellular measurements and 33 to 89 represent the intracellular reactions. Some pathways have different reactions in the forward and reverse directions (e.g; glycolysis and gluconeogenesis and fatty acid oxidation and synthesis). These are represented by the suffix a and b respectively.

<table>
<thead>
<tr>
<th>#</th>
<th>Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 G6P + H2O = 1 Glc + 1 Pi</td>
</tr>
<tr>
<td>2</td>
<td>1 Pyr + 1 NADH + H+ = 1 Lac + 1 NAD+</td>
</tr>
<tr>
<td>3</td>
<td>1 Arg + H2O = 1 Orn + 1 Urea</td>
</tr>
<tr>
<td>4</td>
<td>1 Asp =</td>
</tr>
<tr>
<td>5</td>
<td>1 Ser =</td>
</tr>
<tr>
<td>6</td>
<td>1 Glu =</td>
</tr>
<tr>
<td>7</td>
<td>= 1 Gly</td>
</tr>
<tr>
<td>8</td>
<td>1 NH4 =</td>
</tr>
<tr>
<td>9</td>
<td>= 1 Arg</td>
</tr>
<tr>
<td>10</td>
<td>1 Thr =</td>
</tr>
<tr>
<td>11</td>
<td>1 Ala =</td>
</tr>
<tr>
<td>12</td>
<td>1 Pro =</td>
</tr>
<tr>
<td>13</td>
<td>1 Cys =</td>
</tr>
<tr>
<td>14</td>
<td>1 Tyr =</td>
</tr>
<tr>
<td>15</td>
<td>1 Val =</td>
</tr>
<tr>
<td>16</td>
<td>= 1 Met</td>
</tr>
<tr>
<td>17</td>
<td>1 Orn =</td>
</tr>
<tr>
<td>18</td>
<td>1 Lys =</td>
</tr>
<tr>
<td>19</td>
<td>1 Ile =</td>
</tr>
<tr>
<td></td>
<td>Reaction</td>
</tr>
<tr>
<td>---</td>
<td>----------</td>
</tr>
<tr>
<td>20</td>
<td>1 Leu =</td>
</tr>
<tr>
<td>21</td>
<td>1 Gln =</td>
</tr>
<tr>
<td>22</td>
<td>1 Asn =</td>
</tr>
<tr>
<td>23</td>
<td>1 NADH + 1 Ace + 1 H+ = 1 β-hydroxybutyrate</td>
</tr>
<tr>
<td>24</td>
<td>1 Gol =</td>
</tr>
<tr>
<td>25</td>
<td>1 NADH + 1 Ace + 1 H+ = 1 β-hydroxybutyrate</td>
</tr>
<tr>
<td>26</td>
<td>1 Pal =</td>
</tr>
<tr>
<td>27</td>
<td>1 TG =</td>
</tr>
<tr>
<td>28</td>
<td>1 Tri =</td>
</tr>
<tr>
<td>29</td>
<td>1 Tri =</td>
</tr>
<tr>
<td>30</td>
<td>1 G6P = 1 F6P</td>
</tr>
<tr>
<td>31</td>
<td>1 F6P = 1 G6P</td>
</tr>
<tr>
<td>32</td>
<td>1 F6P + 1Pi = 1 F16BP + 1 H2O</td>
</tr>
<tr>
<td>33</td>
<td>1 F16BP + 1 H2O = 1 F6P + 1 Pi</td>
</tr>
<tr>
<td>34a</td>
<td>1 F16BP + = 2 G3P</td>
</tr>
<tr>
<td>35a</td>
<td>2 G3P = 1 F16BP</td>
</tr>
<tr>
<td>36a</td>
<td>1 G3P + 1 Pi + 1 NAD+ + 1 ADP = 1 PEP + 1 NADH + 1 H+ + 1 ATP + 1 H2O</td>
</tr>
<tr>
<td>37a</td>
<td>1 F16BP + 1 H2O = 1 G3P + 1 Pi + 1 NAD+ + 1 ADP</td>
</tr>
<tr>
<td>38a</td>
<td>1 PEP + 1 ADP = 1 Pyr + 1 ATP</td>
</tr>
<tr>
<td>39a</td>
<td>1 OAA + 1 GTP = 1 PEP + 1 GDP + 1 CO2</td>
</tr>
<tr>
<td>39b</td>
<td>1 Pyr + 1 CoA + 1 NAD+ = 1 ACoA + 1 CO2 + 1 NADH</td>
</tr>
<tr>
<td>40</td>
<td>1 ACoA + 1 H2O = 1 ACoA + 1 Cit + 1 H+</td>
</tr>
<tr>
<td>41</td>
<td>1 Cit + 1 NAD+ = 1 NADH + 1 αKG + 1 CO2</td>
</tr>
<tr>
<td>42</td>
<td>1 αKG + 1 CoA + 1 NAD+ = 1 NADH + 1 SCoA + 1 CO2 + 1 H+</td>
</tr>
<tr>
<td>43</td>
<td>1 SCoA + 1 GDH + 1 Pi + 1 FAD = 1 CoA + 1 FADH2 + 1 Fum + 1 GTP</td>
</tr>
<tr>
<td>44</td>
<td>1 Fum + 1 H2O = 1 Mal</td>
</tr>
<tr>
<td>45</td>
<td>1 Mal + 1 NAD+ = 1 NADH + 1 OAA + 1 H+</td>
</tr>
<tr>
<td>46</td>
<td>1 OAA + 1 NH4 + 1 CO2 + 2ATP + 1 H2O = 1 Citu + 2ADP + 2Pi + 3H+</td>
</tr>
<tr>
<td>47</td>
<td>1 Asp + 1 Citu + 1 ATP = 1 Arg + 1 Fum + 1 AMP + 1 Ppi</td>
</tr>
<tr>
<td>48</td>
<td>1 Pyr + 1 NH4 + 0.5 NADH + 0.5 NADPH + 1 H+ = 1 Ala + 0.5 NAD+ + 0.5 NADP+ + 1 H2O</td>
</tr>
<tr>
<td>49</td>
<td>1 Ser = 1 Pyr + 1 NH4</td>
</tr>
<tr>
<td>50</td>
<td>1 Pyr + 0.5 NADH + 0.5 NADPH + 1 NH4 + 1 thiosulfate + 1 H+ = 1 Cys + 0.5 NAD+ + 0.5 NADP+ + 1 H2O + 1 SO32-</td>
</tr>
<tr>
<td>51</td>
<td>1 NADH + 1 Gly + 1 ACoA = 1 Thr + 1 NAD+</td>
</tr>
<tr>
<td>52</td>
<td>1 Gly + 1 Methylene THF + 1 H2O = 1 THF + 1 Ser</td>
</tr>
</tbody>
</table>
1. SCoA + 1 AMP + 1 PPI = 1 PCoA + 1 ATP + 1 CO2
2. 2 NH4+ + 5 H+ + 5 NADH+ + 2 CO2 + 1 FADH2 + 1 AceCoA = 1 Lysine + 3 H2O + 3 O2 + 1 CoA
3. 1 Phe + 1 O2 + 1 H4biopterin = 1 Tyr + 1 H2O + 1 H2biopterin
4. 1 Tyr + 1 aKG + 2 O2 + 2H2O = 1 Glu + 1 Ace + 1 Fum + 1 CO2 + 2H+
5. 1 Glu + 1 H2O = 0.5 NADH + 1 aKG + 1 NH4 + 0.5 NADPH + 1 H+
# Reaction
6. 1 Gln + 1 H2O = 1 Glu + 1 NH4
7. 1 Orn + 1 H2O = 1 Glu + 1 NADH + 1 NH4 + 1 NADPH + 1 H+
8. 1 Glu + 0.5 NADH + 0.5 NADPH + 1 H+ = 1 Pro + 0.5 O2
9. 1 His + 1 H4folate + 2 H2O = 1 Glu + 1 NH4 + 1 N5formiminoH4folate
10. 1 Ser + 1 ATP + 1 Met + 1 CoA = 1 PPI + 1 Pi + 1 Adenosine + 1 Cys + 1 PCoA + 1 Co2 + 1 NADH + 1 NH4
11. 1 Asp + 1 H2O = 1 OAA + 1 NH4 + 0.5 NADH + 0.5 NADPH + 1 H+
12. 1 Glu + 1 H2O = 1 AceCoA + 1 H2O = 1 Ace + 1 CoA
13. 1 NADH + 0.5 O2 + 1 H+ + 3 ADP = 3 ATP + 1 H2O + 1 NAD+
14. 1 FADH2 + 0.5 O2 + 2 ADP = 2 ATP + 1 H2O + 1 FAD
15. = 1 O2
16. 12 NADPH + 6 CO2 + 12 H+ + 1 Pi = 1 G6P + 12 NADP+ + 7 H2O
17. 1 NH4 + 0.5 NADPH + 3.5 NADH + 1 FADH2 + 2 CO2 + 3 H+ + 1 PCoA = 1 Val + 1 CoA + 2 H2O
18. 1 NH4 + 0.5 NADPH + 2.5 NADH + 1 FADH2 + 1 CO2 + 3 H+ + 1 ACoA + 1 PCoA = 1 Ileu + 2 CoA + 2 H2O
19. 1 NH4 + 0.5 NADPH + 1.5 NADH + 1 FADH2 + 1 ADP + 1 Pi + 2 H+ + 1 Ace + 1 ACoA = 1 Leu + 1 CoA + 1 H2O + 1 ATP
20. 1 TG + 3 H2O = 1 Gol + 3 Palm + 3 H+
21. 1 Glycogen + 2 Pi + 1 UDP = 1 G6P + 1 UTP + 1 H2O
22. 1 Gol + 1 NAD+ = 1 NADH + 1 G3P + 1 H+
23. 1 TG =
24. 1 G6P = 1 G1P
25. 1 G1P + 1 UTP + 1 H+ = 1 UDPG + 1 Ppi
26. 1 UDPG + 1 H2O = 1 UDP-GA + 2 NADH + 3 H+
27. 1 SO3 + 1 ATP = 1 APS + 1 Ppi
28. 1 APS + 1 ATP = 1 PAPS + 1 ADP + 2 H+
29. 1 Mal + 1 NADP+ = 1 Pyr + 1 NADPH + 1 H+
30. 1 Tri + 1 NADPH + 1 O2 = 1 KWG1323 + 1 NADP+ + 1 H2O
31. 1 KWG1323 + 1 O2 + 1 NADH = 1 KWG1342 + 1 H2O + 1 CO2 + 1 NAD+
32. 1 KWG1323 + 1 UDPGA = 1 KWG1323GA + 1 UDP
33. 1 KWG1342 + 1 UDPGA = 1 KWG1342GA + 1 UDP
34. 1 KWG1342 + 1 PAPS = 1 KWG1342SULF + 1 PAP
35. 1 PAP + 1 H2O = 1 Pi + 1 AMP
A.2 Results

Primary rat hepatocytes were cultured in an in vitro collagen sandwich culture for 4 days in standard hepatocyte C+H medium in order to stabilize their phenotype (Dunn et al., 1991; Chan et al., 2003a; Yang et al., 2009b). Initially, we carried out an experiment with multiple concentrations (≤ 0.6 mM) and observed cell death for concentrations higher than 0.3 mM (FigureA.-3). We also observed no significant differences between concentrations lower than 0.15 mM on metabolism (FigureA.-4). Hence, the high concentration was chosen to be 0.3 mM, where significant changes in hepatic metabolism were expected without causing cell death. A lower concentration of 0.15 mM was used to determine the effects of concentration where the changes in metabolism may be less pronounced.
Figure A.-2: Cell count data using Calcein assay. Cells were treated to multiple concentrations of triadimefon ($\leq 0.6$ mM). Cell death was not observed for concentrations $\leq 0.3$ mM triadimefon treatment.

Subsequently, cells were treated to 2 different triadimefon concentrations, 0.3 mM (high concentration) and 0.15 mM (low concentration) and also a control containing DMSO for a period of an additional three days. No significant differences were observed in metabolism between the DMSO control and regular C+H hepatocyte medium (Supplementary Figures 3 and 4). In addition, the total cell count was performed using calcein AM cell viability assay on day 3 to check for cell death (FigureA.-5). Metabolites were measured in the media and supernate daily to analyze the changes in the extracellular fluxes at 24, 36 and 72 hours after triadimefon treatment. Using a metabolic network flexibility analysis (MNFA) framework (Iyer et al., 2010b), intracellular fluxes under different concentrations of triadimefon were evaluated. Cells were also harvested at 6, 24 and 72 hours after triadimefon treatment for microarray analysis. Affymetrix
GeneChip® Rat Genome 230 2.0 Array was employed to detect the changes in gene expression due to two different concentrations of triadimefon treatment. An increase in expression of CYP1A1 (ethoxyresorufin-O-dealkylase, EROD) and CYP1A2 (methoxyresorufin-O-dealkylase, MROD) was observed in triadimefon treated cells compared to DMSO control and is presented in Figure A.-6. respectively and is indicative of conazole activity (Wolfe et al., 2005; Hester et al., 2006).

![Urea production graph]

![Lactate flux graph]

**Figure A.- 3 Extracellular flux measurements for** (A) Urea production (B) Lactate consumption for cells treated to triadimefon. No significant differences were observed between 0.15 mM and 0.075 mM triadimefon treated cells.
Figure A.- 4 Cell count data measured on day 3 using Calcein assay measured on day 3. Cell death was not observed for concentrations $\leq 0.3$ mM triadimefon treatment and cell count was comparable between triadimefon treated cells, DMSO control (DMSO) and regular C+H medium (Control).

### A.2.1 Clustering and Functional Analysis of Microarray Data

In this study, we identified the effect of 0.15 mM (low concentration) and 0.3 mM (high concentration) triadimefon treatment on gene expression. For this purpose, first a subset of probe sets was identified to be a representative of significant dynamic responses in the microarray via clustering analysis. In addition, possible transcriptional regulators for the selected probe sets were determined. Ingenuity Pathway Analysis (IPA®) was then used for functional annotation of the selected subset of the probe sets.

Prior to the clustering analysis, the genes that have high variance and therefore attributed random variations at any experimental condition were filtered ($p$-value $<0.0001$). The clustering analysis focused on the dynamic responses of the filtered probe sets and distinguished probe sets that were most significant to the microarray data. The
clustering of the filtered probe sets yielded 138 unique clusters. Then, a *p*-value was estimated for each cluster based on their population size. The cluster sizes (the number of probe sets in a given cluster) and corresponding estimated *p*-values are depicted in **Figure A.-6**. Based on the significance threshold (*p*-value < 0.0001), there were 12 significant clusters that are characterized by a cluster size greater than 55.

![Figure A.-5 Dose response of CYP1A1 and CYP1A2 in triadimefon treated rat hepatocytes.](image)

**Figure A.- 5 Dose response of CYP1A1 and CYP1A2 in triadimefon treated rat hepatocytes.** Solid line(-) is the average intensity values for DMSO exposed replicates (circles). Dashed-dotted line (···) is the average intensity values for 0.15 mM Triadimefon exposed replicates (triangles). Dotted line (·-·) is the average intensity values for 0.30 Triadimefon exposure replicates (squares).

The 12 significant clusters were arranged in three main groups based on their responses over time between the three treatments (**Table A. 2**). The most populated group is group III that includes 62% of the selected probe sets. The probe sets in group III exhibited various dynamic responses when the cells were treated with 0.3 mM triadimefon and no change over time when cells were treated with DMSO control and 0.15 mM of triadimefon. Both group I and II included ~ 19% of the selected probe sets. The probe sets in group I showed change over time when treated with DMSO control but
no change, when the cells were treated with 0.15 mM and 0.3 mM triadimefon. The probe sets in group II indicated dynamic change when the cells were treated with 0.15 mM of triadimefon but no change, over time when cells were treated with DMSO control and 0.3 mM of triadimefon.

**Figure A.-** Estimated *p*-value vs. cluster size. Cluster size of 55 corresponds to a *p*-value < 0.0001

**Table A. 2** The groups of the dynamic responses and their size (number of probe sets). Each two consecutive symbols in this representation correspond to the time intervals (6th - 24th and 24th and 72nd) for DMSO control, 0.15 and 0.30 mM triadimefon, respectively. Intervals that did not contain significant differential expression were called constant (C), while those that were significantly different were labeled either up-regulated (U) or down-regulated (D).

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Response</td>
<td>Size</td>
<td>Dynamic Response</td>
</tr>
<tr>
<td>DMSO 0.15mM 0.30mM</td>
<td>DMSO 0.15mM 0.30mM</td>
<td>DMSO 0.15mM 0.30mM</td>
</tr>
<tr>
<td>DC CC CC</td>
<td>140</td>
<td>CC CU CC</td>
</tr>
<tr>
<td>CU CC CC</td>
<td>113</td>
<td>CC DD CC</td>
</tr>
<tr>
<td>CC DC CC</td>
<td>64</td>
<td>CC CC CU</td>
</tr>
<tr>
<td>CC UU CC</td>
<td>55</td>
<td>CC CC UU</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CC CC DU</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CC CC CD</td>
</tr>
</tbody>
</table>
After analyzing the significant clusters of the microarray data, the possible transcriptional factors (TFs) were determined via a computational approach (explained in materials and methods). Nine of the identified TFs were associated with liver in Genomatix® database (Table A.3). Though a significant overlap among identified TFs was observed across various clusters, AHHR was specific to group I and HNF1 and PERO were specific to group III. Furthermore, HOMF and PARF were associated with both group II and III, whereas AP1R, FKHD, NRF2 and RXRF were related to all three groups. The identified TFs are characteristic of xenobiotic metabolism (AHHR, (Schultz et al., 2003; Kawakami et al., 2009)), fatty acid metabolism and oxidation (PERO and NRF2 (Pawar and Jump, 2003; Hertz et al., 2005; Capobianco et al., 2008)), cell differentiation (HOMF, PARF and NRF2 (Tanaka et al., 1999; Kimata et al., 2006)) and cell proliferation (HOMF, PARF and NRF2 (Mueller et al., 1990; Chiba et al., 2005). The significant overlap across various clusters is reasonable due to the fact that transcription factors are characterized by pleiotropic effects (Rodriguez-Caso et al., 2005).
Table A. 3 Computationally identified Transcription Factors associated with liver in all the clusters.

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Transcription Factor Information</th>
<th>Cluster ID</th>
<th>Grouping ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHRR</td>
<td>AHR-arnt heterodimers and AHR-related factors</td>
<td>4</td>
<td>I</td>
</tr>
<tr>
<td>AP1R</td>
<td>MAF and AP1 related factors</td>
<td>1, 5, 6, 7, 8, 9, 12</td>
<td>I, II, III</td>
</tr>
<tr>
<td>FKHD</td>
<td>Fork head domain factors</td>
<td>1, 2, 4, 6, 10, 11</td>
<td>I, II, III</td>
</tr>
<tr>
<td>HNF1</td>
<td>Hepatic Nuclear Factor 1</td>
<td>12</td>
<td>II, III</td>
</tr>
<tr>
<td>HOMF</td>
<td>Homeodomain transcription factors</td>
<td>2, 6, 9, 10, 11, 12</td>
<td>II, III</td>
</tr>
<tr>
<td>NR2F</td>
<td>Nuclear receptor subfamily 2 factors</td>
<td>1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12</td>
<td>I, II, III</td>
</tr>
<tr>
<td>PARF</td>
<td>PAR/bZIP family</td>
<td>10, 12</td>
<td>II, III</td>
</tr>
<tr>
<td>PERO</td>
<td>Peroxisome proliferator-activated receptor</td>
<td>12</td>
<td>III</td>
</tr>
<tr>
<td>RXRF</td>
<td>RXR heterodimer binding sites</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12</td>
<td>I, II, III</td>
</tr>
</tbody>
</table>

The annotation of the biological functions of the significant clusters was determined from the IPA® library. The significant pathways ($p$-value $< 0.05$) and the representative profiles of the associated clusters are given in Table A. 4. The significant pathways of group I (Table A. 4) were mainly related to signaling pathways (e.g. ILK signaling, HIF1a signaling, MAPK signaling and NFKB activation), whereas the significant pathways of Group II were distributed in both signaling pathways (e.g. T-cell receptor signaling, regulation of IL-2 expression) and metabolic pathways (e.g. glutathione metabolism, metabolism of xenobiotics by cytochrome P450). Group III was characterized mostly by metabolic pathways that were related to carbohydrate metabolism (pyruvate, propanoate and butanoate metabolism), amino acid metabolism (D-glutamate metabolism, Valine, Leucine and Isoleucine Degradation and beta-alanine metabolism) and lipid metabolism (bile acid biosynthesis and fatty acid metabolism). In addition, the significant toxicity functions ($p$-value $< 0.05$) are illustrated in Figure A.7.
We observed that both significant pathways and significant toxicity functions are highly specific to the identified group, indicating that Groups I, II and III characterize distinct biological functions.

Table A. 4 The enriched pathways (*p*-value < 0.01) of each group. The representative profiles of the dynamic responses in each group are illustrated. The solid line represents DMSO exposure, the dashed line represents low dose triadimefon exposure, and the dotted line represents the high dose triadimefon exposure. (The dots represent the replicates). The *p*-values of the enriched pathways are listed in Supplementary Table 8.

<table>
<thead>
<tr>
<th>GROUP I</th>
<th>GROUP II</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILK Signaling, MIF Regulation of Innate Immunity, IL-9 Signaling, Thrombopoietin Signaling, HIF1α Signaling, HMGB1 Signaling, Erythropoietin Signaling, IL-17 Signaling, LPS-stimulated MAPK Signaling, NF-κB Activation by Viruses, BMP signaling pathway</td>
<td>Metabolism of Xenobiotics by Cytochrome P450, Glyoxylate and Dicarboxylate Metabolism, Coagulation System, T Cell Receptor Signaling, Role of Macrophages, Glutathione Metabolism</td>
</tr>
</tbody>
</table>
GROUP III

Valine, Leucine and Isoleucine Degradation, Propanoate Metabolism, Fatty Acid Metabolism, Bile Acid Biosynthesis, Butanoate Metabolism, Pyruvate Metabolism, β-alanine Metabolism, D-glutamine and D-glutamate Metabolism
Figure A.7: The significant toxicity lists (p-value < 0.05) provided by IPA-TOX® of each gene groups.

A.2.2 Metabolite measurements

Due to the centrality of metabolism to hepatic function, the concentrations of metabolites were measured in the media and supernate daily. Triadimefon uptake and triadimenol production were measured using GC-MS. As expected, we found that triadimefon uptake was significantly higher in 0.3 mM (high concentration) compared to 0.15 mM (low concentration) triadimefon treatment (Figure A.8). Triadimefon uptake, however, remained comparable from day 1 through day 3. However, triadimenol production showed a gradual increase from day 1 through day 3 for both high and 0.15 mM triadimefon treatment. Triadimenol production was also higher for the 0.3 mM compared to 0.15 mM triadimefon treatment for all three days. Urea and albumin production are
important markers of hepatic function. We observed that urea production was not affected in either high or 0.15 mM triadimefon treatment compared to DMSO control (Figure A.-8). Albumin production also did not show any statistically significant difference between 0.15 mM triadimefon treatment and DMSO control for all three days. For 0.3 mM triadimefon treatment, initially there was no difference but by the end of the three day period, albumin was significantly reduced (Figure A.-8).

Figure A.-8: Comparison of selected extracellular fluxes (µmol/million cells/day) between high and low concentration triadimefon exposure and DMSO control. Star indicates statistically significantly different relative to DMSO control (p < 0.05)

Lactate production is a relevant flux to glycolytic or gluconeogenic pathways of hepatic metabolism. Lactate flux was comparable between 0.15 mM triadimefon treatment and DMSO control for all three days. Lactate was initially produced during day 1 but then switched over to lactate consumption for day 2 and day 3 in case of both 0.15 mM triadimefon treatment and DMSO control. In case of 0.3 mM triadimefon treatment, lactate production was significantly higher relative to DMSO control but exhibited a
gradual decrease from day 1 to day 3 (Figure A.-8). β-hydroxybutyrate flux can give valuable information regarding lipid metabolism in hepatocytes. β-hydroxybutyrate flux was significantly higher for both low and 0.3 mM triadimefon treatment compared to DMSO control (Figure A.-8). There was also a gradual increase in β-hydroxybutyrate flux from day 1 through day 3 for both low and 0.3 mM triadimefon treatment compared to DMSO control. Thus, extracellular flux measurements in triadimefon treated cells showed significant changes compared to DMSO control. Further, this effect was more pronounced in 0.3 mM triadimefon treatment compared to 0.15 triadimefon treatment.

A.2.3 Metabolic network flexibility analysis (MNFA)

Fluxes were evaluated using a MNFA framework, which involves solving simultaneously, a set of minimization and maximization linear programming problems for the intracellular fluxes, thereby calculating a flux range for each unmeasured flux (Llaneras and Picó, 2007b; Llaneras and Picó, 2007a) using extracellular measurements, stoichiometric balances and thermodynamic constraints (See Materials and Methods). Calculated flux ranges were tight for all the treatments and are represented as a bar graph in Figure A.-9, 10 and 11.

Phase I of triadimefon metabolism involved oxidation to conjugate metabolite kwg1323 and subsequently to desmethyl kwg1342 (See Materials and Method for IUPAC nomenclature) (FAO and WHO, 2007). These conjugate metabolites were further detoxified to either glucuronide conjugates (kwg1323-ga and desmethyl kwg1342-ga) or sulfate conjugates (desmethyl kwg1342-sulf) (Figure A.-9) (FAO and WHO, 2007). For 0.3 mM and 0.15 mM triadimefon treatment, MNFA indicated that most of the
triadimefon was converted to the sulfate conjugate (Figure A-9). The detoxification fluxes were however higher for 0.3 mM compared to 0.15 mM triadimefon treatment, thereby reflecting higher triadimefon uptake in 0.3 mM triadimefon treatment (Figure A-9). In addition, these trends remained the same for day 1 and day 2 for triadimefon treated cells. Triadimefon detoxification steps are linked to hepatic metabolism by their requirement of O₂ and energy for cytochrome P450 enzymes (Nelson, 2005; Sun et al., 2007). There is also an involvement of glucose-6-phosphate and sulfate for glucuronide and sulfate conjugation reactions respectively (Figure 1-4).

**Figure A-9:** (A) Simplified metabolic network representing triadimefon detoxification. (B) MNFA demonstrated that triadimefon was oxidized to conjugate metabolites (desmethyl kwg1342 and kwg1323) and then further detoxified mostly to sulfate conjugate (desmethyl kwg1342-sulf) for high concentration (0.3 mM) and low concentration (0.15 mM) triadimefon exposure.

The intracellular fluxes in hepatic metabolism were then compared to detect pathways that vary the most between triadimefon treated cells and DMSO control. Cells treated to the DMSO control exhibited glucose and lactate consumption. MNFA indicated that glycolytic pathway was active in cells treated to DMSO control (Figure A-10). Further, fatty acids synthesis and intracellular triglyceride (TG) accumulation was observed. \( \beta \)-hydroxybutyrate production was low and so was the flux through the TCA cycle in cells
treated to DMSO control. Day 3 metabolic flux data are represented in Figure A.-10; however, the trends remain the same for day 1 and day 2 for DMSO control.

Cells exposed to 0.3 mM of triadimefon exhibited lactate production and glucose consumption with an active glycolytic pathway for day 1 and day 2 (Figure A.-10). However, fatty acids were oxidized in the 0.3 mM triadimefon treatment (Figure A.-10). The fatty acid oxidation was driven by intracellular TG depletion. Fatty acid oxidation also caused an increase in β-hydroxybutyrate production and TCA cycle flux relative to DMSO control. By day 3, 0.3 mM triadimefon treatment exhibited a major switch in metabolism compared to DMSO control. Gluconeogenic pathway was active by day 3, leading to glucose and lactate production (Figure A.-10). There was a substantial increase in fatty acid oxidation, β-hydroxybutyrate production and the TCA cycle flux by day 3. In the case of 0.15 mM triadimefon treatment, lactate and glucose were consumed leading to an active glycolytic pathway for all three days (Figure A.-10). Similar to 0.3 mM triadimefon treatment, fatty acid oxidation was active, and β-hydroxybutyrate production and TCA cycle flux was higher in 0.15 mM triadimefon treatment relative to DMSO control.
Figure A.- 10: (A) Simplified metabolic network representing hepatic metabolism involving glycolysis and lipid metabolism. Corresponding metabolic fluxes for (B) All three treatments for day 3, and (C) 0.3 mM triadimefon treatment for all three days.
Triadimefon detoxification requires CYP450 enzymes along with O$_2$ and energy (Nelson, 2005). Fatty acid oxidation is a substantial source of energy in the form of NADH and FADH$_2$, which in turn promote the electron transport chain reactions for the production of ATP. In the case of 0.3 mM triadimefon treatment, the large fatty acid oxidation flux generates substantial acetyl-CoA that drives a high flux through the TCA cycle and β-hydroxybutyrate production (Figure A.-10). This large flux through the TCA cycle might in turn cause gluconeogenesis in 0.3 mM triadimefon treatment by day 3. It is interesting to note that the low concentration (0.15 mM) of triadimefon though causes a switch from fatty acid synthesis to fatty acid oxidation, is not large enough to drive gluconeogenesis (Figure A.-10).

Figure A.-11 demonstrates the differences in reactions contributing to energy metabolism for all the treatments for day 3. The main reactions contributing to cofactor production and in turn electron transport chain (ETC) to produce ATP, include glycolysis, malate to pyruvate anaplerotic reaction, fatty acid oxidation and TCA cycle. The ETC reactions were highest for 0.3 mM, followed by 0.15 mM triadimefon treatment and finally lowest for DMSO control. The fatty acid oxidation and subsequent TCA cycle flux were mainly responsible for the large increase in the ETC flux.
Figure A.- 11 Reactions contributing to energy metabolism for (A) 0.3 mM triadimefon treatment for all three days, and (B) All three treatments for day 3.

A.3 Discussion

In this study, the effects of two concentrations of triadimefon (0.15 mM and 0.3 mM) on gene expression and metabolic fluxes in cultured rat hepatocytes were compared. Effects of triadimefon exposure have been evaluated in rats previously in both in vivo and in vitro models (Hester et al., 2006; Goetz and Dix, 2009b). A single previous study has reported gene expression changes in cultured rat hepatocytes exposed to much lower concentration of triadimefon (than those utilized in our study), in which relatively few genes exhibited differential expression (Goetz and Dix, 2009b). Therefore, we studied higher concentrations (0.3 mM and 0.15 mM) where more significant changes were observed without causing cell death. Furthermore, we quantified the changes in metabolic fluxes to assess the changes in primary metabolism of hepatocytes caused by triadimefon detoxification.

In this study, we observed significant differential gene expression when the cells were treated with triadimefon. Clustering analysis of differentially expressed probe sets across different treatment periods and concentrations revealed three main groups. The most
significant response in the micro array data was captured in what we term group III (62% of the filtered probe sets), representing the probe sets that showed dynamic responses to only 0.3 mM triadimefon treatment. Specifically, the genes in fatty acid metabolism were affected due to 0.3 mM triadimefon treatment. The changes in fatty acid metabolism due to conazole treatment have been previously observed in both in vivo and in vitro rat and human studies (Goetz and Dix, 2009b; Goetz and Dix, 2009a). Here, in addition to the changes in gene expression of fatty acid metabolism, we also observe corresponding changes in metabolic fluxes related to lipid metabolism, which have not been quantified previously. In 0.3 mM triadimefon treatment, intracellular flux quantification illustrated that fatty acid oxidation was highly active (Figure A.-10). The increased fatty acid oxidation flux in 0.3 mM triadimefon treatment compared to fatty acid synthesis in DMSO control is most probably a response for energy requirement of triadimefon detoxification. Detoxification of triadimefon involves CYP450 enzymes along with O$_2$ and energy (Nelson, 2005). The energy and O$_2$ requirements are provided from fatty acid oxidation, which is a substantial source of energy in the form of NADH and FADH$_2$.

The large fatty acid oxidation flux also generated substantial acetyl-CoA that increased flux through TCA cycle, β-hydroxybutyrate production and probably induced gluconeogenesis (Figure A.-10). We additionally observed that the genes in pyruvate metabolism were affected only in 0.3 mM triadimefon treatment. This is interesting because pyruvate is highly linked to TCA cycle, gluconeogenesis/glycolysis and amino acid metabolism. Correspondingly, we also detected changes in gene expression and metabolic fluxes related to valine, leucine and isoleucine (VLI) metabolism (Supplementary Tables 4-6 and 9). Thus, the concentration of 0.3 mM triadimefon
treatment produced substantial and comparable changes in gene expression and metabolic fluxes. This response was unique to 0.3 mM triadimefon treatment.

The remaining filtered probe sets were distributed equally between Group I (~19%) and Group II (~19%). The probe sets in group II showed dynamic responses only to 0.15 mM triadimefon treatment. The lack of similar response of these probe sets in the 0.3 mM triadimefon treatment was unexpected. The biological annotation of the probe sets belonging to Group II were related to glutathione metabolism, metabolism of xenobiotics by CYP450 and signaling pathways. Further, in case of 0.15 mM triadimefon treatment, the response of the fluxes related to hepatic metabolism was less pronounced compared to 0.3 mM triadimefon treatment. Specifically, the fluxes related to lipid metabolism were much lower in 0.15 mM triadimefon treatment compared to 0.3 mM triadimefon treatment (Figure A.-10). Subsequently, glycolytic pathway remained active through all three days for 0.15 mM triadimefon treatment (Figure A.-10). It appears that the cells may be able to cope with 0.15 mM triadimefon treatment with relatively small changes in hepatic metabolism compared to large changes in 0.3 mM triadimefon treatment. Meanwhile, it is likely that the genes related to xenobiotic metabolism and signaling pathways demonstrated dynamic response in 0.15 mM triadimefon treatment in order to maintain detoxification.

Finally, the probe sets in group I showed dynamic response over time when treated to DMSO control but no change when exposed to 0.15 mM and 0.3 mM triadimefon treatment. The probe sets in group I are related to three major areas; cellular growth, cellular stress, and cellular defense response (Figure A.-7 and Table A. 4), which have been previously identified as common responses to conazole exposure in both in vivo and
in vitro studies (Hester et al., 2006; Tully et al., 2006; Goetz and Dix, 2009b). Probe sets in Group I may thus characterize a xenobiotic effect because the gene expression profiles are affected in a similar manner in both triadimefon treatments, compared to DMSO control.

Transcriptional factors are important in elucidating the regulation of xenobiotic response. Among the computationally identified transcription factors, we observed RXRF to be associated with all the clusters of co-expressed genes. RXRF is a common heterodimerization partner for members of the subfamily 1 nuclear receptors including CAR and PXR (Laudet et al., 1992; Allenby et al., 1993). Interestingly, transcriptional factors CAR and PXR have been proposed to be responsible in regulating general triazole metabolism in rats (Goetz and Dix, 2009b). We also found a specific transcriptional regulator (PERO) consistent with the observed gene expression and metabolic flux changes. PERO is related to fatty acid metabolism (Pawar and Jump 2003, Capobianco 2008) and it was found to be associated only with the 0.3 mM triadimefon treatment. Hence, PERO may be responsible for the fatty acid related changes in the 0.3 mM triadimefon treatment.

In summary, the most significant changes in gene expression and metabolic fluxes were observed in 0.3 mM triadimefon treatment. These changes were related predominantly to primary hepatic metabolism, in response to xenobiotic effects. Specifically, fatty acid metabolism demonstrated a similar response on a gene expression and metabolic flux level, confirming its importance in triadimefon detoxification. On the other hand, genes related to xenobiotic metabolism were affected at the relatively lower concentration of 0.15 mM triadimefon treatment to sustain detoxification, in the absence
of pronounced changes in hepatic fluxes. Overall, we find that the two types of measurements, metabolic flux and gene expression, provide consistent but non-redundant views of the hepatic response to triadimefon treatment in vitro, indicating that their simultaneous analysis enables a richer understanding of the emergent phenotypes.
Curriculum Vitae

Ayse Meric Ovacik

Education

09/2005 – 1/2011  PhD in Chemical and Biochemical Engineering
Rutgers University, Piscataway NJ

Koc University, Rumelifeneri, ISTANBUL

9/2000 - 6/2003  B.S. in Chemical Engineering
Middle East Technical University, ANKARA

Middle East Technical University, ANKARA

Referred Journal Publications


- Ovacik M., Ierapetritou M.G., Euling S.Y., Gaido K., Androulakis I.P. Pathway modeling of microarray data: A case study of pathway changes in the testis following in utero exposure to dibutyl phthalate (DBP). Toxicol Appl Pharmacol. Special Issue: Toxicogenomics and Risk Assessment

- Ovacik M, Androulakis I.P. Cross-species Comparison Analysis of Metabolic Pathways: Informing inter-species differences. Toxicol Appl Pharmacol. Special Issue: Toxicogenomics and Risk Assessment


**Federal Documents**